UNIVERSITY OF SOUTHAMPTON

On Variance Estimation Under Complex Sampling Designs

by

Emilio López Escobar

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Social and Human Sciences
Division of Social Statistics and Demography

March, 2013

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL AND HUMAN SCIENCES
DIVISION OF SOCIAL STATISTICS AND DEMOGRAPHY

Doctor of Philosophy

by Emilio López Escobar

This thesis is formed of three manuscripts (chapters) about variance estimation. Each of the chapters focuses on developing new original variance estimators. The Chapter 1 proposes a novel jackknife variance estimator for self-weighted two-stage sampling. Customary jackknifes for these designs rely only on the first sampling stage. This omission may induce a bias in the variance estimation when cluster sizes vary, second stage sampling fractions are small or when there is low variability between clusters. The proposed jackknife accounts of all sampling stages via deletion of clusters and observations within clusters. It does not need join-inclusion probabilities and naturally includes finite population corrections. Its asymptotic design-consistency is shown. A simulation study show that it can be more accurate than the customary jackknife used for this kind of sampling designs (Rao, Wu and Yue, 1992). The Chapter 2 proposes a totally new replication variance estimator for any unequal-probability without-replacement sampling design. The proposed replication estimator is approximately equal to the linearisation variance estimators obtained by the Demnati and Rao (2004) approach. It is more general than the Campbell (1980); Berger and Skinner (2005) generalised jackknife. Its asymptotic design-consistency is shown. A simulation study shows it is more stable than standard jackknifes (Quenouille, 1956; Tukey, 1958) with *ad hoc* finite population corrections and than the generalised jackknife (Campbell, 1980; Berger and Skinner, 2005). The Chapter 3 proposes a new variance estimator which accounts the item non-response under unequal-probability without-replacement sampling when estimating a change from rotating (overlapping) repeated surveys. The proposed estimator combines the original approach by Berger and Priam (2010, 2012) and the non-response reverse approach for variance estimation (Fay, 1991; Shao and Steel, 1999). It gives design-consistent estimation of the variance of change when the sampling fraction is small. The proposed estimator uses random Hot-deck imputation, but it can be implemented with other imputation techniques. Further, there are two more complementary chapters. One introduces the R package called *samplingVarEst* which implements of some methods for variance estimation utilised for the simulations. Finally, there is a brief chapter which discusses future research work.

Contents

List of Tables

List of Figures

Declaration of Authorship

I, Emilio López Escobar, declare that this thesis entitled: *"On Variance Estimation Under Complex Sampling Designs"* and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Either none of this work has been published before submission, or parts of this work have been published as:

   · Berger, Y. G. and Escobar, E. L. (2013a) Variance estimation of hot-deck imputed estimators of change for rotating repeated surveys. (*Submitted*).

- Berger, Y. G. and Escobar, E. L. (2012b) Variance estimation of imputed estimators of change over time from repeated surveys. In *Proceeding of the XIèmes Journées de Méthodologie Statistique de l'Insee.* Paris: Institut National de la Statistique et des Études Économiques (National Institute of Statistics and Economic Studies).
- Escobar, E. L. and Barrios, E. (2012). SamplingVarEst: Sampling Variance Estimation. R package version 0.9-1.
- Escobar, E. L. and Berger, Y. G. (2010) A novel jackknife variance estimator for two-stage without replacement sampling designs. In *Abstracts of Communications of the 10th International Vilnius Conference on Probability Theory and Mathematical Statistics*, 144. Vilnius, Lithuania: International Statistical Institute.
- Escobar, E. L. and Berger, Y. G. (2011) Jackknife variance estimation for functions of Horvitz & Thompson estimators under unequal probability sampling without replacement. In *Proceeding of the 58th World Statistics Congress.* Dublin: International Statistical Institute.
- Escobar, E. L. and Berger, Y. G. (2013a) A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica*, 23, 595-613.
- Escobar, E. L. and Berger, Y. G. (2013b) A new replicate variance estimator for unequal probability sampling without replacement. *Canadian Journal of Statistics* (*to appear*).

Signed: ....

Date: ....

Acknowledgements

I owe my deepest gratitude to the following people involved in my supervision. Their example of professionalism was the most valuable and inspiring lesson:
· Yves G. Berger, University of Southampton, U.K.,
· Chris J. Skinner, London School of Economics and Political Sciences, U.K.,
· Ray L. Chambers, University of Wollongong, Australia.

I thank my examiners for their constructive comments and feedback:
· Camelia Goga, Université de Bourgogne, France,
· Nikolaos Tzavidis, University of Southampton, U.K.

I am grateful to the following people for their useful advise, constructive feedback and very helpful motivation:
· Pedro L. D. N. Silva, IBGE - Escola Nacional de Ciências Estatísticas, Brazil,
· Danny Pfeffermann, Univ. Southampton & Hebrew Univ. of Jerusalem, Israel,
· Juan F. Muñoz Rosas & Encarnación Álvarez Verdejo, Univ. Granada, Spain,
· Guillermina Eslava Gómez, National Autonomous University of Mexico,
· Rubén Hernández Cid, Mexico's Autonomous Institute of Technology.

Additionally, I would like to thank Yves Tillé, University of Neuchâtel Switzerland and Jay Breidt, Colorado State University U.S.A. for their questions and sharp comments at conferences which helped me to improve the presentation of these results. Further, I am thankful to a couple of journals' editors and five anonymous referees for their edifying comments and observant suggestions during the publication process of these manuscripts.

I thank my sponsor CONACyT - Mexico's National Council for Science and Technology; and to these institutions which eased expenses of conferences and seminars:
· Statistics Sweden, Statistiska Centralbyrån, Sweden,
· Australian Bureau of Statistics, Australia,
· School of Social Sciences, Univ. Southampton, U.K.

I would like to show my gratitude to these colleagues with whom I shared my days whilst completing my research duties: Juan Correa Allamand, Omar De la Riva, Dimitrios Gkountanis, Nicholas Lazarou, Mauro Testaverde, Thomas King, Ivonne Nava Ledezma, Clare Woodford, Ngianga II Kandala, Lana Chikhungu, Yahia El-Horbaty, Dayang Awang Marikan, Sharon Holder, Ronaldo Nazare, Norimah Rambeli, Riayati Ahmad, Derya Tas, Rebecca Vassallo, Lorraine Waller, Katie Bruce, Claire Bailey, David Clifford, Bernard Baffour-Awuah, Guy Abel, Marcos Gómez Mella, Damião Silva, Denize Barbosa.

Despite the geographical distance, I would also like to thank my parents and my sister who greatly supported me.

Finally, I am indebted to my wife Pilar. This thesis would not have been possible without her continuous and joyful support. I dedicate this thesis to her.

To Pilar

Introduction

In survey sampling the aim is to make inferences about a (finite) population from a sample. We are often interested in estimating characteristics of a population or in producing parameter-estimates of an assumed population model. In making inferences from survey data, we have ultimately to deal with the quality of sampling strategies and the accuracy of the obtained estimates. As it will be introduced below, the output of the variance estimation problem gives input matter for that quality and accuracy scrutiny. In practice, the variance estimation thus emerges as central and imperative for accomplishing the survey sampling aim correctly (e.g. Shao and Tu, 1995, ch. 1 & 6).

Two major conceptions

Inference from survey data has two major conceptions, either the *design-based approach* or the *model-based approach.* Names to these approaches may vary and mixtures are also available, e.g. the model-assisted approach (Särndal, Swensson and Wretman, 1992). An overview of approaches can be encountered in Kish (1995, sec. 7), Särndal *et al.* (1992, sec. 1.10) and in Smith (2001). Further, a very good and complete summary can be found in Brewer and Gregoire (2009).

This thesis is confined to the design-based approach, also know as the *classical* or as the *randomised* approach. The design-based conception is widely acknowledged and used in practice. It can be found in any standard sampling techniques textbook, e.g. Kish (1965); Särndal *et al.* (1992); Lohr (1999); Lehtonen and Pahkinen (2004). Literature comprising the model-based approach includes, for example, books by Valliant, Dorfman and Royall (2000) and, recently, by Chambers and Clark (2012).

The variance estimation problem

Under the design-based approach, the inference is subject to the variability that comes from the randomised draw of the sample. Accordingly, the population is assumed fixed and the variability emerges from the changeable output from sample to sample. This variability is measured by the variance of the used point estimators. In practice, this variance is unknown as it would be necessary to have available all the possible samples. Usually only one sample is available and thus the variance has to be estimated from it.

The estimated variance is important because it provides an input for the construction of confidence intervals, coefficients of variation, hypothesis testing and design effects. These serve as accuracy measures and ways to asses the suitability of certain sampling strategies.

Variance estimation under complex sampling designs

The variance estimation often becomes complicated when the involved point estimators are non-linear and when the used sampling designs are complex (e.g. Kish and Frankel, 1974). Under that complex framework, standard statistical theory is no longer suitable (e.g. Chambers and Skinner, 2003, sec. 1.1). That is, the usual *i.i.d.* statistical inference assumptions no longer hold and standard methods fail. Furthermore, estimating the variance of an estimator becomes even more complex in the presence of certain features encountered in practice, e.g. non-response, overlapping samples, multiple-frames (see Wolter, 2007, p. 2).

Solutions for the variance estimation problem include the traditional Taylor linearisation (or delta) methods and resampling methods such as the Jackknife, the Bootstrap and the Balanced half-sampling. An introductory overview about variance estimation can be found in Wolter (2007). Whereas a more theoretical compendium can be found in Shao and Tu (1995).

Organisation of this thesis

This thesis is formed of three manuscripts about variance estimation. The focus of each manuscript is on developing new original variance estimators for different type of point estimators, sampling designs and/or particular features encountered in practice. Each manuscript embodies a chapter of this thesis. The chapters are related, nevertheless, each one is intended to be a self-contained and a standalone piece of research.

Further, as pointed by the examiners, there are two more complementary chapters. One introduces R software implementations of some methods for variance estimation utilised for the simulations. Finally, there is a last brief chapter which discusses future research work.

The first chapter

The first chapter proposes a novel jackknife-type variance estimator for self-weighted two-stage sampling designs. This kind of two-stage sampling is the most utilised in practice as it simplifies field-work. Currently, the available jackknife variance estimators for this kind of designs rely only on the first sampling stage, i.e. deletion of clusters. However, the omission of the second stage may induce a bias in the variance estimation. Particularly when cluster sizes vary, second stage sampling fractions are small or when there is low variability between clusters. On the contrary, the proposed jackknife from Chapter 1 accounts of all sampling stages via deletion of clusters and observations within clusters. In Chapter 1, it is also shown that the proposed jackknife is asymptotically design-consistent, and thus, valid for inference. The proposed two-stage jackknife does not need join-inclusion probabilities and naturally includes finite population corrections. It is further shown that it can be more accurate than the customary jackknife utilised for these kind of sampling designs (Rao, Wu and Yue, 1992).

The second chapter

Following, the second chapter proposes a totally new replication variance estimator for any unequal-probability without-replacement sampling design. The proposed replication estimator has the advantage of being approximately equal to the linearisation variance estimators but without the need of deriving derivatives. It is shown that the proposed replication method is asymptotically design-consistent and more general than the generalised jackknife from Campbell (1980); Berger and Skinner (2005). Moreover, a simulation study shows that the proposed estimator is more stable than the standard jackknife (Quenouille, 1956; Tukey, 1958) with the *ad hoc* finite population correction, and than the (Campbell, 1980; Berger and Skinner, 2005) generalised jackknife .

The third chapter

The third chapter proposes a new variance estimator which accounts the item non-response under unequal-probability without-replacement sampling when estimating a change from rotating (overlapping) repeated surveys. The proposed method combines the original approach by Berger and Priam (2010, 2012) and the Fay (1991) non-response reverse approach for variance estimation (e.g. Rao and Shao, 1992; Shao and Steel, 1999). The proposed variance estimator gives design-consistent estimation of the variance of change when the sampling fraction is small, i.e. when the finite population corrections are negligible. The proposed approach is illustrated using random Hot-deck imputation, although the proposed estimator can be implemented with other imputation techniques.

The fourth chapter

The fourth chapter introduces an R package called *samplingVarEst*. It implements several well-known variance estimators and it also introduces some of the novel methods developed in earlier chapters. The package is under continuous updating and we invite readers of this thesis to check for the last version published at the *The Comprehensive R Archive Network*, (CRAN). To illustrate details, the appendix of this chapter includes the full user's manual of the version 0.9-1 is.

The fifth chapter

The fifth chapter briefly discusses some possible extensions and further research work which is planned in coming years.

Chapter 1

A jackknife variance estimator for self-weighted two-stage samples

Abstract

Self-weighted two-stage sampling designs are popular in practice as they simplify fieldwork. It is common in practice to compute variance estimates only from the first sampling stage, neglecting the second stage. This omission may induce a bias in variance estimation; especially in situations where there is low variability between clusters or when sampling fractions are non-negligible.

We propose a design-consistent jackknife variance estimator which takes account of all stages via deletion of clusters and observations within clusters. The proposed jackknife can be used for a wide class of point estimators. It does not need joint-inclusion probabilities and naturally includes finite population corrections. A simulation study shows that the proposed estimator can be more accurate than standard jackknifes (Rao, Wu and Yue, 1992) for self-weighted two-stage sampling designs.

Keywords and phrases: Linearisation; pseudovalues; Sen-Yates-Grundy form; smooth function of means; stratification.

1.1 Introduction

In survey sampling, the accuracy of point estimates are assessed using variance estimates. Variance estimation becomes difficult when we have non-linear point estimators and complex sampling designs. This is a well known problem which has been broadly covered in the survey sampling literature, e.g. Kish and Frankel (1974), Särndal *et al.* (1992) and Wolter (2007). Resampling techniques for variance estimation often overcome these difficulties. The Jackknife, was first introduced by Quenouille (1956) for bias reduction and later by Tukey (1958) for variance estimation. This resampling technique has been widely studied, e.g. Krewski and Rao (1981), Kovar *et al.* (1988), Rao *et al.* (1992), and Shao and Tu (1995) among others.

Campbell (1980) proposed a totally different generalised jackknife variance estimator based on the analogy between linearisation and jackknife techniques. Berger and Skinner (2005) showed its design consistency for single stage designs under a set of regularity conditions. They also compare the empirical performance of Campbell's jackknife (in a single stage context) with standard single stage jackknifes such as Tukey (1958), Kish and Frankel (1974), and Rao *et al.* (1992). Further, Berger and Rao (2006) extended Campbell's approach for imputation. Berger (2007) proposed a modified Campbell's estimator which incorporates the Hájek (1964) approximation for the joint inclusion probabilities.

The regularity conditions in Berger and Skinner (2005) for the design-consistency of the Campbell estimator are too restrictive for two-stage sampling. For example, in two-stage simple random sampling the total number of sampled units would need to be fixed as population size tends to infinity for the Berger and Skinner (2005) regularity conditions to hold. In section 1.3, we propose new less restrictive regularity conditions which accommodate two-stage sampling. We also propose a Sen (1953) and Yates and Grundy (1953) version of the Campbell's jackknife which overcomes the possibility of getting negative variance estimates. Further, the asymptotic design-consistency of these jackknife estimators is established under two-stage sampling.

In section 1.4, we propose a jackknife variance estimator for self-weighted two-stage (stratified) without replacement sampling. These sampling designs are very common in practice; examples include, the Youth Risk Behavior Survey in the U.S.A., the Labour Force Survey for São Paulo in Brazil, and the Living Standards

Survey for countries like South Africa, Ghana and Côte d'Ivoire. We focus on self-weighted two-stage designs. However, there are different self-weighted designs that are widely used in practice. Some utilise three or more stages, and some others use unequal probabilities at the final stage. Examples include the US National Health and Nutrition Examination Survey (NHANES) and the Australian and New Zealand Labour Force Surveys.

The proposed jackknife for self-weighted two-stage sampling involves deletion of both, clusters and observations. The proposed jackknife estimator does not have double sums and does not need joint inclusion probabilities. Further, we show that this novel estimator is asymptotically design-consistent. To ease computing efforts, a subsampling version is also proposed in subsection 1.4.2 for its most computer intensive part which involve deleting observations.

In section 1.5, Monte-Carlo simulations show that the proposed jackknife can be more accurate than customary jackknife estimators for more than one stage such as the Rao *et al.* (1992) stratified multi-stage delete-cluster jackknife.

1.2 The class of point estimators

Let $\mathcal{U}$ denote a finite population of size N whose elements are grouped into N_I clusters of size M_i, $i = 1, \dots, N_I$. Consider a without replacement sample s of elements drawn according to a self-weighted two-stage fixed sample size design. That is, n_I clusters are drawn using a without-replacement probability proportional to the size of the clusters, then a simple random sample without-replacement of m fixed elements is drawn within each sampled cluster. Therefore, the sample size is fixed and given by $n = n_I m$ elements grouped in n_I clusters.

Let $\pi_{Ii} > 0$ and π_{Iij} denote, respectively, the first and the second order inclusion probabilities for the clusters $i, j = 1, \dots, N_I$; and also let $\pi_k > 0$ and $\pi_{k\ell}$ denote the inclusion probabilities for the elements $k, \ell = 1, \dots, N$. For a self-weighted sampling design the clusters inclusion probabilities are

$$\pi_{Ii} \;=\; n_I \frac{M_i}{N},$$

and thus

$$\pi_k \;=\; f,$$

where $f = n/N$.

Let y_{qk} denote the value of the survey variable q $(q = 1, \dots, Q)$ for $k \in \mathcal{U}$. Suppose we are interested in the population parameter

$$\theta \;=\; g(\mu_1, \dots, \mu_q, \dots, \mu_Q),$$

which is a smooth and differentiable function of population means

$$\mu_q \;=\; \frac{1}{N} \sum_{k \in \mathcal{U}} y_{qk}, \quad q = 1, \dots, Q.$$

Further, assume θ is estimated by the substitution point estimator

$$\tilde{\theta} \;=\; g(\tilde{\mu}_1, \dots, \tilde{\mu}_q, \dots, \tilde{\mu}_Q),$$

where

$$\tilde{\mu}_q \;=\; \sum_{k \in s} \tilde{w}_k y_{qk}, \quad q = 1, \dots, Q,$$

is the Hájek (1971) mean estimator for μ_q, with normalised sampling weights

$$\tilde{w}_k \;=\; \frac{w_k}{\widehat{N}},$$

where

$$\widehat{N} \;=\; \sum_{k \in s} w_k,$$

and $w_k \;=\; 1/\pi_k$.

1.3 Generalised jackknife variance estimators

The Campbell (1980) generalised jackknife variance estimator of $\tilde{\theta}$ is defined by (see Berger and Skinner, 2005),

$$\widehat{\text{var}}(\tilde{\theta})_{HT} \;=\; \sum_{k \in s} \sum_{\ell \in s} \mathcal{D}_{k\ell} \; \varepsilon_{(k)} \; \varepsilon_{(\ell)}, \tag{1.1}$$

with

$$\begin{aligned} \mathcal{D}_{k\ell} \;&=\; \frac{\pi_{k\ell} - \pi_k \pi_\ell}{\pi_{k\ell}}, \qquad (1.2)\\ \varepsilon_{(k)} \;&=\; (1 - \tilde{w}_k)(\tilde{\theta} - \tilde{\theta}_{(k)}), \end{aligned}$$

where $\tilde{\theta}_{(k)}$ has same functional form as $\tilde{\theta}$ but after omitting the observation k. That is,

$$\tilde{\theta}_{(k)} \;=\; g(\tilde{\mu}_{1(k)}, \ldots, \tilde{\mu}_{q(k)}, \ldots, \tilde{\mu}_{Q(k)}),$$

where

$$\tilde{\mu}_{q(k)} \;=\; \sum_{\ell \in s-\{k\}} \tilde{w}_{\ell(k)}\, y_{q\ell},$$

with

$$\tilde{w}_{\ell(k)} \;=\; \frac{w_\ell}{\sum_{\ell \in s-\{k\}} w_\ell},$$

and with $s-\{k\}$ denoting s after deleting the k-th observation.

It can clearly be seen that the Equation (1.1) may take negative values. To overcome this issue, we propose the following alternative Sen (1953) and Yates and Grundy (1953) form,

$$\widehat{\text{var}}(\tilde{\theta})_{SYG} \;=\; \frac{-1}{2} \sum_{k \in s} \sum_{\ell \in s} \mathcal{D}_{k\ell}\; (\varepsilon_{(k)} - \varepsilon_{(\ell)})^2, \tag{1.3}$$

which is always positive if the Sen-Yates-Grundy condition, $\mathcal{D}_{k\ell} < 0$, holds. Note that the Equation (1.3) is suitable for unequal-probability fixed sample size designs (e.g. Chao, 1982).

For single-stage sampling, Berger and Skinner (2005) showed the asymptotic design-consistency of Equation (1.1) and also illustrated the better empirical performance of (Eq. 1.1) in comparison with standard jackknifes such as Tukey (1958), Kish and Frankel (1974), and Rao *et al.* (1992). Further, Berger and Rao (2006) extended (Eq. 1.1) for imputation and Berger (2007) proposed a modified version incorporating the Hájek (1964) approximation for the joint inclusion probabilities.

Note that under uni-stage simple random sampling, both Equations (1.1) and (1.3) reduce to the *standard jackknife* (e.g. Shao and Tu, 1995, p. 239),

$$\widehat{\text{var}}(\tilde{\theta})_{STD} \;=\; \Big(1 - \frac{n}{N}\Big)\, \frac{n-1}{n} \sum_{k \in s} (\tilde{\theta}_{(k)} - \tilde{\theta}_{(\cdot)})^2,$$

where

$$\tilde{\theta}_{(\cdot)} \;=\; \frac{1}{n} \sum_{k \in s} \tilde{\theta}_{(k)}.$$

1.3.1 Consistency of the generalised jackknifes for two-stage sampling

The consistency of $\widehat{\mathrm{var}}(\tilde{\theta})_{HT}$ and $\widehat{\mathrm{var}}(\tilde{\theta})_{SYG}$ is now set under new less restrictive regularity conditions than those specified by Berger and Skinner (2005). These new conditions will allow two-stage sampling.

We use the Isaki and Fuller (1982) asymptotic framework which considers a sequence of nested populations of size $N_{[t]}$ $(0 < N_{[t]} < N_{[t+1]})$, and a sequence of samples of size $n_{[t]}$ $(n_{[t]} < n_{[t+1]}, n_{[t]} < N_{[t]}$, for all $t)$. To simplify notation, we drop the index t in what follows. Thus, if $t \to \infty$, it implies: $N \to \infty$, $n \to \infty$ and $n_I \to \infty$. We consider that $f = n/N$, $f_I = n_I/N_I$ and m are constants free of the limiting process.

For the vector of means

$$\boldsymbol{\mu} \;=\; (\mu_1, \ldots, \mu_Q)^T,$$

and the vector of point estimators

$$\tilde{\boldsymbol{\mu}} \;=\; (\tilde{\mu}_1, \ldots, \tilde{\mu}_Q)^T,$$

the multivariate Horvitz-Thompson and Sen-Yates-Grundy design variances and variance estimators of $\tilde{\boldsymbol{\mu}}$ are defined by (see Särndal *et al.*, 1992, secs. 5.5, 5.7)

$$\begin{aligned}
\mathbf{var}(\tilde{\boldsymbol{\mu}})_{HT} &\doteq \sum_{k\in\mathcal{U}}\sum_{\ell\in\mathcal{U}} \mathcal{D}_{k\ell}\,\pi_{k\ell}\,\boldsymbol{z}_k\,\boldsymbol{z}_\ell^T,\\
\widehat{\mathbf{var}}(\tilde{\boldsymbol{\mu}})_{HT} &\doteq \sum_{k\in s}\sum_{\ell\in s} \mathcal{D}_{k\ell}\,\check{\boldsymbol{z}}_k\,\check{\boldsymbol{z}}_\ell^T,\\
\mathbf{var}(\tilde{\boldsymbol{\mu}})_{SYG} &\doteq \frac{-1}{2}\sum_{k\in\mathcal{U}}\sum_{\ell\in\mathcal{U}} \mathcal{D}_{k\ell}\,\pi_{k\ell}\,\{\boldsymbol{z}_k-\boldsymbol{z}_\ell\}\{\boldsymbol{z}_k-\boldsymbol{z}_\ell\}^T,\\
\widehat{\mathbf{var}}(\tilde{\boldsymbol{\mu}})_{SYG} &\doteq \frac{-1}{2}\sum_{k\in s}\sum_{\ell\in s} \mathcal{D}_{k\ell}\,\{\check{\boldsymbol{z}}_k-\check{\boldsymbol{z}}_\ell\}\{\check{\boldsymbol{z}}_k-\check{\boldsymbol{z}}_\ell\}^T,
\end{aligned}$$

with

$$\begin{aligned}
\boldsymbol{z}_k &= \frac{w_k}{N}(\boldsymbol{y}_k-\boldsymbol{\mu}),\\
\check{\boldsymbol{z}}_k &= \tilde{w}_k(\boldsymbol{y}_k-\tilde{\boldsymbol{\mu}}),\\
\boldsymbol{y}_k &= (y_{1k}, \ldots, y_{Qk})^T.
\end{aligned}$$

Now, assume the following regularity conditions:

C1. $\widehat{\text{var}}(\tilde{\theta})_L/\text{var}(\tilde{\theta})_L \to_p 1$, $\text{var}(\tilde{\theta})_L \neq 0$ where

$$\begin{aligned} \text{var}(\tilde{\theta})_L &= \boldsymbol{\nabla}(\boldsymbol{\mu})^T \, \mathbf{var}(\tilde{\boldsymbol{\mu}})_{HT} \, \boldsymbol{\nabla}(\boldsymbol{\mu}), \\ \widehat{\text{var}}(\tilde{\theta})_L &= \boldsymbol{\nabla}(\tilde{\boldsymbol{\mu}})^T \, \widehat{\mathbf{var}}(\tilde{\boldsymbol{\mu}})_{HT} \, \boldsymbol{\nabla}(\tilde{\boldsymbol{\mu}}). \end{aligned} \tag{1.4}$$

Alternatively for fixed sample-size designs,

$$\begin{aligned} \text{var}(\tilde{\theta})_L &= \boldsymbol{\nabla}(\boldsymbol{\mu})^T \, \mathbf{var}(\tilde{\boldsymbol{\mu}})_{SYG} \, \boldsymbol{\nabla}(\boldsymbol{\mu}), \\ \widehat{\text{var}}(\tilde{\theta})_L &= \boldsymbol{\nabla}(\tilde{\boldsymbol{\mu}})^T \, \widehat{\mathbf{var}}(\tilde{\boldsymbol{\mu}})_{SYG} \, , \boldsymbol{\nabla}(\tilde{\boldsymbol{\mu}}), \end{aligned} \tag{1.5}$$

where $\boldsymbol{\nabla}(\boldsymbol{x}) = (\partial g(\boldsymbol{\mu})/\partial\mu_1, \ldots, \partial g(\boldsymbol{\mu})/\partial\mu_Q)^T_{\boldsymbol{\mu}=\boldsymbol{x}}$ is the gradient of $g(\cdot)$ at $\boldsymbol{x} \in \Re^Q$ with $g(\cdot)$ continuous and differentiable at $\boldsymbol{\mu}$.

C2. $|1 - \tilde{w}_k| \geq \alpha > 0$, for all $k \in \mathcal{U}$, α is a constant.

C3. $\lim\inf \{n \; \text{var}(\tilde{\theta})_L\} > 0$.

C4. $n^{-1} \sum_{k\in s} \tilde{w}_k^\tau \, \|\boldsymbol{y}_k - \tilde{\boldsymbol{\mu}}\|^\tau = \mathcal{O}_p(n^{-\tau})$ for all $\tau \geq 2$, where $\|\boldsymbol{A}\| = \text{tr}(\boldsymbol{A}^T\boldsymbol{A})^{1/2}$ denotes the Euclidean norm.

C5. $G_s = n^{-\beta} \sum\sum_{(k\neq\ell)\in s} (\mathcal{D}^-_{k\ell})^2 = \mathcal{O}_p(1)$, with $0 \leq \beta < 1$, where $\mathcal{D}^-_{k\ell} = -\mathcal{D}_{k\ell}$ if $\mathcal{D}_{k\ell} < 0$ and 0 otherwise.

C6. $H_s = n^{-\beta} \sum\sum_{(k\neq\ell)\in s} (\mathcal{D}^+_{k\ell})^2 = \mathcal{O}_p(1)$, with $0 \leq \beta < 1$, where $\mathcal{D}^+_{k\ell} = \mathcal{D}_{k\ell}$ if $\mathcal{D}_{k\ell} \geq 0$ and 0 otherwise.

C7. $\boldsymbol{\nabla}(\boldsymbol{x})$ is Lipschitz continuous, $\|\boldsymbol{\nabla}(\boldsymbol{x}_1) - \boldsymbol{\nabla}(\boldsymbol{x}_2)\| \leq \lambda \, \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^\delta$, $\lambda > 0$ and $\delta > 0$ constants, $0 \leq \beta/2 < \delta$, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ in the neighbourhood of $\boldsymbol{\mu}$.

C8. $\|\boldsymbol{\nabla}(\tilde{\boldsymbol{\mu}})\| = \mathcal{O}_p(1)$.

The regularity conditions C1 and C5 to C7 are similar but different from the ones proposed by Berger and Skinner (2005). These conditions now allow two stage sampling. The condition C1 sets the consistency of the linearisation variance estimator recalling the Robinson and Särndal (1983) approach (see Särndal *et al.*, 1992, secs. 5.5, 5.7). The conditions C2 to C4 are typical in asymptotics (e.g. Shao and Tu, 1995, p. 258): C2 sets that none of the normalised weights reach 1, C3 implies $\text{var}(\tilde{\theta})_L$ decreases with rate n^{-1}, and C4 is a Lyapunov-type condition for the existence of moments. The regularity conditions C5 and C6 are mild conditions on the design, similar to ones in Isaki and Fuller (1982); the conditions C7 and C8 are usual smoothness requirements for the jackknife: C7 sets the smoothness of

$g(\cdot)$ (regardless of the sampling design). Note that for two stage-sampling, $\beta < 1$ means that there are more observations than clusters in s.

Theorem 1.1. *For sampling designs of fixed size, if the regularity conditions C1 to C8 hold, then the proposed generalised jackknife variance estimator $\widehat{var}(\tilde{\theta})_{SYG}$ from Equation (1.3) is asymptotically design-consistent for the approximate linearised variance $var(\tilde{\theta})_L \neq 0$ in (Eq. 1.5). That is,*

$$\frac{\widehat{var}(\tilde{\theta})_{SYG}}{var(\tilde{\theta})_L} \rightarrow_p 1.$$

A proof of Theorem 1.1 is given in the Appendix 1.A of this chapter.

Corollary 1.2. *If the regularity conditions C1 to C8 hold, then $\widehat{var}(\tilde{\theta})_{HT}$ from Equation (1.1) is also asymptotically design-consistent for the approximate linearised variance $var(\tilde{\theta})_L \neq 0$ in (Eq. 1.4), i.e.*

$$\frac{\widehat{var}(\tilde{\theta})_{HT}}{var(\tilde{\theta})_L} \rightarrow_p 1.$$

The Corollary 1.2 can be shown from Berger and Skinner (2005) proof taking into account the changes in the conditions C5 to C7.

From the Theorem 1.1, Corollary 1.2 and the Slutsky's theorem (e.g. Valliant, Dorfman and Royall, 2000, p. 414), when $\tilde{\theta}$ is asymptotically normal, it follows that

$$\frac{\tilde{\theta} - \theta}{\sqrt{\widehat{\mathrm{var}}(\tilde{\theta})_{SYG}}} \rightarrow_d \mathrm{N}(0, 1), \tag{1.6}$$

and

$$\frac{\tilde{\theta} - \theta}{\sqrt{\widehat{\mathrm{var}}(\tilde{\theta})_{HT}}} \rightarrow_d \mathrm{N}(0, 1).$$

Thus, allowing valid confidence intervals of $\tilde{\theta}$ for θ.

1.4 The proposed jackknife for self-weighted two-stage sampling

For self-weighted two-stage sampling we have that

$$\pi_{Ii} = n_I \frac{M_i}{N},$$

and

$$\pi_k = \frac{n}{N} = f.$$

Now, by using the Hájek's approximation (Hájek, 1964, eq. 5.27, p. 1511), the clusters' joint inclusion probabilities π_{Iij} are approximated by

$$\begin{aligned} \pi_{Iij} &\doteq \pi_{Ii}\, \pi_{Ij} \left\{ 1 - \frac{(1-\pi_{Ii})(1-\pi_{Ij})}{d} \right\}, \\ d &= \sum_{Ii \in \mathcal{U}} \pi_{Ii}\, (1-\pi_{Ii}). \end{aligned}$$

This approximation was originally developed for $d \to \infty$, i.e. in our case $N_I \to \infty$, under the maximum entropy sampling design (see Hájek, 1981, Theorem 3.3, ch. 3 & 6); namely the Rejective Sampling design, a. k. a. the Conditional Poisson Sampling design. It requires that the utilised sampling design (of clusters) is of large entropy. An overview can be found in Berger and Tillé (2009). An account of different sampling designs, π_{Iij}'s approximations and approximate variances under large-entropy designs can be found in Tillé (2006); Brewer and Donadio (2003); Haziza *et al.* (2004, 2008). Recently, Berger (2011) gives sufficient conditions under which Hájek's results still hold for large entropy sampling designs that are not the maximum entropy one.

Low entropy sampling designs, such as the systematic probability proportional-to-size design, are not suitable for the above approximation. However, the randomized systematic sampling is suitable as it is of large entropy (e.g. Brewer and Gregoire, 2009; Berger and Tillé, 2009).

Following, given the conditional inclusion probabilities

$$\pi_{k|Ii} = \frac{m}{M_i},$$

and

$$\pi_{k\ell|Ii} = \frac{m\,(m-1)}{M_i\,(M_i-1)},$$

the elements' joint inclusion probabilities $\pi_{k\ell}$ are,

$$\pi_{k\ell} \;\doteqdot\; \begin{cases} \pi_{Ii}\,\pi_{k|Ii} \;=\; f & \text{if } (k=\ell)\in s_i, \\ \pi_{Ii}\,\pi_{k\ell|Ii} \;=\; f(m-1)/(M_i-1) & \text{if } (k\neq\ell)\in s_i, \\ \pi_{Iij}\,\pi_{k|Ii}\,\pi_{\ell|Ij} \;\doteqdot\; f^2\{1-d^{-1}(1-\pi_{Ii})(1-\pi_{Ij})\} & \text{if } k\in s_i, \ell\in s_j, i\neq j, \end{cases}$$

where s_i denotes the observations from the i-th cluster. Therefore, by substituting for $\mathcal{D}_{k\ell}$ from Equation (1.2) we obtain

$$\mathcal{D}_{k\ell} \;\doteqdot\; \begin{cases} 1-f & \text{if } (k=\ell)\in s_i, \\ 1-\pi^*_{Ii} & \text{if } (k\neq\ell)\in s_i, \\ (1-\pi_{Ii})(1-\pi_{Ij})\{(1-\pi_{Ii})(1-\pi_{Ij})-d\}^{-1} & \text{if } k\in s_i, \ell\in s_j, i\neq j, \end{cases}$$

where

$$\pi^*_{Ii} \;=\; \pi_{Ii}\left(\frac{m}{m-1}\right)\left(\frac{M_i-1}{M_i}\right).$$

Thus, it can be shown (see the Appendix 1.C of this chapter) that by substituting these values of $\mathcal{D}_{k\ell}$ into (Eq. 1.3), the Equation (1.3) reduces to the following jackknife variance estimator suitable for self-weighted two-stage sampling designs,

$$\widehat{\mathrm{var}}(\tilde{\theta})_{prop} \;=\; v_{clu} \;+\; v_{obs}, \tag{1.7}$$

where

$$v_{clu} \;=\; \sum_{i\in s}(1-\pi^*_{Ii})\,\varsigma^2_{(Ii)} \;-\; \frac{1}{d}\left(\sum_{i\in s}(1-\pi_{Ii})\,\varsigma_{(Ii)}\right)^2, \tag{1.8}$$

$$v_{obs} \;=\; \sum_{k\in s}\phi_k\,\varepsilon^2_{(k)}, \tag{1.9}$$

with

$$\phi_k \;=\; \pi^*_{Ii}\,\frac{M_i-m}{M_i-1},$$

for $k\in s_i$, where the *delete cluster pseudo-values* $\varsigma_{(Ii)}$ are given by

$$\varsigma_{(Ii)} \;=\; \frac{n_I-1}{n_I}(\tilde{\theta}-\tilde{\theta}_{(Ii)}), \tag{1.10}$$

with $\tilde{\theta}_{(Ii)}$ of same functional form as $\tilde{\theta}$ but excluding the observations from the i-th cluster, i.e.

$$\tilde{\theta}_{(Ii)} \;=\; g(\tilde{\mu}_{1(Ii)},\ldots,\tilde{\mu}_{q(Ii)},\ldots,\tilde{\mu}_{Q(Ii)}),$$

where

$$\tilde{\mu}_{q(Ii)} = \sum_{k \in s(Ii)} \tilde{w}_{k(Ii)} \, y_{qk},$$

and

$$\tilde{w}_{k(Ii)} = \frac{w_k}{\sum_{k \in s(Ii)} w_k},$$

with $s(Ii) = s - s_i$ denoting the sample without observations from the i-th cluster; and where the *delete observation pseudo-values* $\varepsilon_{(k)}$, previously defined at section 1.3, are given by

$$\varepsilon_{(k)} = \frac{n-1}{n}(\tilde{\theta} - \tilde{\theta}_{(k)}).$$

The proposed variance estimator (Eq. 1.7) has two terms. One which deletes observations within clusters and another which deletes clusters. The term v_{clu} from Equation (1.8) computes variability between clusters, and v_{obs} from Equation (1.9) computes variability of observations within clusters.

If d is unknown we can replace d by

$$\hat{d} = \sum\nolimits_{i \in s} (1 - \pi_{Ii}).$$

As $\phi_k \propto f(M_i - m)(m-1)^{-1}$ the term v_{obs} is zero if $m = M_i$ and diminishes for small f. Conversely, it may become large if f is large, if the sampling fractions within clusters are small, or if the M_i vary.

To simplify notation, we derive the proposed estimator (Eq. 1.7) for non-stratified designs. However, it can be generalised by treating the strata separately. The number of strata has to be bounded and large sample regularity conditions must hold within each stratum. Therefore, the applicability of the proposed jackknife variance estimator excludes highly-stratified sampling designs with very few sampling units per stratum.

1.4.1 Consistency of the proposed jackknife

Let $\operatorname{var}(\tilde{\theta})_{HL}$ denote the Hájek approximation to the approximate linearised variance $\operatorname{var}(\tilde{\theta})_L$.

Theorem 1.3. *If the regularity conditions C1, C3, C4, C7 and C8 hold, and if $M_i \geq m \geq 2$, then the proposed jackknife variance estimator for self-weighted* *two-stage sampling designs $\widehat{var}(\tilde{\theta})_{prop}$ from Equation (1.7) is asymptotically design-consistent for the Hájek approximate linearised variance $var(\tilde{\theta})_{HL} \doteq var(\tilde{\theta})_L \neq 0$. That is,*

$$\frac{\widehat{var}(\tilde{\theta})_{prop}}{var(\tilde{\theta})_L} \rightarrow_p 1.$$

A proof of Theorem 1.3 is given in the Appendix 1.B of this chapter. Furthermore, the Equation (1.6) also holds for (Eq. 1.7) when $\tilde{\theta}$ is asymptotically normal.

1.4.2 A less computationally intensive version of the proposed jackknife

The delete-observation term v_{obs} in Equation (1.9) may become laborious with large datasets. To ease computing, we propose to treat v_{obs} as a total which can therefore be estimated from a subsample via the Horvitz and Thompson (1952) estimator. Hence, we subsample $\tilde{n}$ elements from the sample s.

Let $\tilde{s}$ denote this subsample and let $\tilde{\pi}_k$ be the first order inclusion probabilities of $\tilde{s}$. We propose to estimate v_{obs} using the unbiased Horvitz-Thompson point estimator

$$\tilde{v}_{obs} = \sum_{k \in \tilde{s}} \frac{\phi_k \; \varepsilon^2_{(k)}}{\tilde{\pi}_k}. \tag{1.11}$$

Thus, a less computationally intensive estimator than Equation (1.7) is given by

$$\widetilde{\text{var}}(\tilde{\theta})_{prop} = v_{clu} + \tilde{v}_{obs}. \tag{1.12}$$

It is recommended to use inclusion probabilities proportional to ϕ_k; that is,

$$\tilde{\pi}_k = \tilde{n}\frac{\phi_k}{\Phi},$$

where

$$\Phi = \sum_{k \in s} \phi_k,$$

implying

$$\tilde{v}^{\pi ps}_{obs} = \sum_{k=1}^{\tilde{n}} \frac{\phi_k \; \varepsilon^2_{(k)}}{\tilde{\pi}_k} = \frac{\Phi}{\tilde{n}} \sum_{k=1}^{\tilde{n}} \varepsilon^2_{(k)}.$$

Note that $\tilde{\pi}_k$ should be approximately proportional to $\phi_k \varepsilon^2_{(k)}$. Hence, this will give an efficient Horvitz-Thompson estimator.

In the context of two-phase sampling, Kim and Sitter (2003) proposed also a less computationally intensive approach. In further research, it would be good to explore the applicability of (Eq. 1.11) for two-phase sampling designs.

1.4.3 Customary jackknife variance estimator

A customary jackknife variance estimator for sampling designs of more than one stage is the *stratified multi-stage delete cluster* jackknife estimator by Rao, Wu and Yue (1992), which is originally purposed for functions of totals and for with-replacement sampling designs, that is, for negligible sampling fractions. The Rao *et al.* (1992) estimator is defined as

$$\widehat{\text{var}}(\tilde{\theta})_{RWY} \;=\; \sum_{i \in s} \varsigma_{(Ii)}^{2}, \tag{1.13}$$

where $\varsigma_{(Ii)}$ is defined as Equation (1.10). When the sampling fraction is large, the estimator (Eq. 1.13) is usually adjusted by an overall clusters' finite population correction (FPC),

$$\widehat{\text{var}}(\tilde{\theta})_{RWY}^{FPC} \;=\; \sum_{i \in s} \left(1 - \frac{n_I}{N_I}\right) \varsigma_{(Ii)}^{2}. \tag{1.14}$$

Comparing the proposed jackknife $\widehat{\text{var}}(\tilde{\theta})_{prop}$ in (Eq. 1.7) and the above FPC-adjusted customary $\widehat{\text{var}}(\tilde{\theta})_{RWY}^{FPC}$ in (Eq. 1.14) we note that they differ in two main aspects:

(i) $\widehat{\text{var}}(\tilde{\theta})_{prop}$ adds the term v_{obs} which computes variability of observations within clusters,

(ii) $\widehat{\text{var}}(\tilde{\theta})_{prop}$ uses a different FPC $(1-\pi_{Ii}^{*})$ for each cluster i, whereas $\widehat{\text{var}}(\tilde{\theta})_{RWY}^{FPC}$ uses the fixed FPC $(1 - n_I/N_I)$.

1.5 Simulation study

We illustrate two simulation examples from two datasets: the Labour Force Population from Valliant *et al.* (2000, Appendix B.5) and the MU284 Swedish Municipalities Population from Särndal *et al.* (1992, Appendix B). For both datasets, we duplicated 3 times the number of clusters and 3 times the number of observations within each cluster. We therefore use two population frames of $N = 4\,302$ and $2\,556$ observations which are grouped into $N_I = 345$ and 150 clusters, respectively. The minimum/maximum cluster sizes are: 6/39 and 15/27, respectively.

We use four variables of interest, two from each population frame: the weekly wages (y_1) and number of hours worked per week (y_2) from the first population frame; and, the number of Social-Democratic seats in municipal council (y_3) and the number of Conservative seats in municipal council (y_4) from the second population frame.

The homogeneity measures $ICC(\cdot)$, i.e. the intra-class correlation coefficient (as defined in Särndal *et al.*, 1992, secs. 3.4.3 & 4.2.2), for each of the variables of interest are: $ICC(y_1) = 0.2965$, $ICC(y_2) = 0.1951$, $ICC(y_3) = 0.3181$ and $ICC(y_4) = 0.4958$.

The parameters of interest are the ratios:

$$R_{12} = \frac{\mu_1}{\mu_2} = 7.697$$

and,

$$R_{34} = \frac{\mu_3}{\mu_4} = 2.439$$

which are estimated by

$$\hat{R}_{12} = \frac{\tilde{\mu}_1}{\tilde{\mu}_2},$$

and

$$\hat{R}_{34} = \frac{\tilde{\mu}_3}{\tilde{\mu}_4},$$

where $\tilde{\mu}_1, \ldots, \tilde{\mu}_4$ are Horvitz and Thompson (1952) point estimators.

Clusters were selected using Brewer (1975) unequal probability sampling design with clusters' inclusion probabilities proportional to the cluster size; then, a simple random without replacement sample of individuals is selected within clusters using sample size $m = 2,\ 4,\ 6$. For the labour force population frame, it is important to note that 20.34% of the clusters of the minimum cluster size, meaning that with $m = 6$ we are doing selection of all the elements within many clusters.

For the estimator (Eq. 1.12), we use Brewer (1975) unequal probability design with subsampling rate 0.25 and with subsampling inclusion probabilities proportional to ϕ_k as defined in subsection 1.4.2.

For each simulation and for each simulation example, $N_{Sim1} = 100\,000$ and $N_{Sim2} = 1\,000\,000$ samples were selected to compute:

- The empirical relative bias

  $$\mathrm{RB} = \frac{\mathrm{B}(\widehat{\mathrm{var}}(\hat{R}_{ab}))}{\mathrm{var}(\hat{R}_{ab})},$$

  where

  $$\mathrm{B}(\widehat{\mathrm{var}}(\hat{R}_{ab})) = \mathrm{E}(\widehat{\mathrm{var}}(\hat{R}_{ab})) - \mathrm{var}(\hat{R}_{ab});$$

- The empirical relative root mean square error defined by

  $$\mathrm{RRMSE} = \frac{\sqrt{\mathrm{MSE}(\widehat{\mathrm{var}}(\hat{R}_{ab}))}}{\mathrm{var}(\hat{R}_{ab})};$$

- The coverage at a 95% confidence level.

The $\mathrm{var}(\hat{R}_{ab})$ is the empirical variance computed from the N_{Sim1} (and N_{Sim2}) observed values of $\hat{R}_{ab}$ (ab =12 and 34). These quantities were computed for the estimators (Eq. 1.7), (Eq. 1.12), (Eq. 1.13) and (Eq. 1.14).

1.5.1 Example 1: Point estimator $\hat{R}_{12}$

Results for this example are summarised in Tables 1.1, 1.2 and 1.3. Additional graphical representations for these results are provided in the Appendix 1.D of this chapter.

The Table 1.1 illustrates in terms of RB that the customary estimators (Eq. 1.13) and (Eq. 1.14), respectively, over-estimates and under-estimates the variance for increasing values of f (f_I). In general, this effect is more pronounced for small second-stage sampling sizes $m = 2, 4$. That undesirable effect decreases with

Table 1.1: Relative Bias (%) of variance estimators for the point estimator $\hat{R}_{12}$ where $ICC(y_1) = 0.2965$ and $ICC(y_2) = 0.1951$.

					Proposed		Customary	
m	n_I	n	f_I	f	(Eq. 1.7)	(Eq. 1.12)	(Eq. 1.13)	(Eq. 1.14)
2	20	40	0.058	0.009	-4.40	-4.41	4.52	-1.54
	40	80	0.116	0.019	-1.73	-1.69	9.07	-3.58
	60	120	0.174	0.028	-1.76	-1.80	12.87	-6.76
	80	160	0.232	0.037	-1.33	-1.35	18.07	-9.31
	100	200	0.290	0.046	-0.62	-0.71	24.25	-11.76
	120	240	0.348	0.056	-0.78	-0.82	30.12	-15.14
	140	280	0.406	0.065	-0.59	-0.62	37.15	-18.50
	160	320	0.464	0.074	-0.54	-0.51	44.77	-22.37
4	20	80	0.058	0.019	-4.77	-4.78	5.52	-0.60
	40	160	0.116	0.037	-3.06	-3.07	10.56	-2.26
	60	240	0.174	0.056	-2.73	-2.73	16.60	-3.68
	80	320	0.232	0.074	-1.81	-1.82	24.70	-4.21
	100	400	0.290	0.093	-0.90	-0.91	34.15	-4.73
	120	480	0.348	0.112	-0.01	-0.03	45.05	-5.40
	140	560	0.406	0.130	-0.01	0.03	56.37	-7.09
	160	640	0.464	0.149	-1.23	-1.23	67.54	-10.16
6	20	120	0.058	0.028	-4.93	-4.93	6.05	-0.10
	40	240	0.116	0.056	-2.99	-2.99	12.17	-0.83
	60	360	0.174	0.084	-1.28	-1.29	20.93	-0.10
	80	480	0.232	0.112	-1.27	-1.27	29.32	-0.66
	100	600	0.290	0.139	-0.82	-0.83	39.90	-0.65
	120	720	0.348	0.167	-1.38	-1.38	50.89	-1.59
	140	840	0.406	0.195	-0.22	-0.22	66.91	-0.82
	160	960	0.464	0.223	-1.06	-1.07	82.67	-2.04

$m = 6$ (census within several clusters) for the FPC-adjusted customary estimator (Eq. 1.14).

On the other hand, the RB for the proposed variance estimator (Eq. 1.7) and its subsampling version (Eq. 1.12) remains tightly around zero as the sampling fractions increases regardless of the second-stage sample sizes. The reason for this is that the proposed estimators correctly incorporate the finite population corrections at both stages. Note that there is a particular FPC for each cluster in the expression of the proposed variance estimators (Eq. 1.7) and (Eq. 1.12).

In terms of RRMSE, it can also be seen in the Table 1.2 that the proposed estimator (Eq. 1.7) has always the smallest RRMSE followed by the FPC-adjusted Rao *et al.*

Table 1.2: Relative Root Mean-Square-Error (%) of variance estimators for the point estimator $\hat{R}_{12}$ where $ICC(y_1) = 0.2965$ and $ICC(y_2) = 0.1951$.

					Proposed		Customary	
m	n_I	n	f_I	f	(Eq. 1.7)	(Eq. 1.12)	(Eq. 1.13)	(Eq. 1.14)
2	20	40	0.058	0.009	46.12	46.41	51.52	48.37
	40	80	0.116	0.019	31.90	32.91	38.04	32.86
	60	120	0.174	0.028	25.25	27.22	33.49	26.42
	80	160	0.232	0.037	21.05	24.38	32.89	23.07
	100	200	0.290	0.046	18.32	23.36	35.25	21.64
	120	240	0.348	0.056	16.03	23.09	38.50	21.77
	140	280	0.406	0.065	14.38	23.87	43.74	23.04
	160	320	0.464	0.074	13.12	25.49	50.09	25.40
4	20	80	0.058	0.019	42.13	42.15	47.22	44.18
	40	160	0.116	0.037	29.24	29.30	35.49	30.04
	60	240	0.174	0.056	23.21	23.39	33.01	23.86
	80	320	0.232	0.074	19.59	19.88	35.77	20.31
	100	400	0.290	0.093	16.97	17.44	41.86	17.84
	120	480	0.348	0.112	15.05	15.79	50.75	16.16
	140	560	0.406	0.130	13.36	14.50	60.74	15.20
	160	640	0.464	0.149	11.94	13.63	71.07	15.63
6	20	120	0.058	0.028	40.06	40.07	45.03	42.04
	40	240	0.116	0.056	27.70	27.72	34.44	28.50
	60	360	0.174	0.084	22.10	22.13	34.57	22.72
	80	480	0.232	0.112	18.44	18.50	38.30	18.93
	100	600	0.290	0.139	16.16	16.26	46.25	16.61
	120	720	0.348	0.167	14.16	14.33	55.53	14.59
	140	840	0.406	0.195	12.65	12.91	70.39	13.01
	160	960	0.464	0.223	11.28	11.71	85.43	11.73

(1992) from (Eq. 1.14) and by the subsampling version of the proposed estimator (Eq. 1.12).

In terms of the coverage of the 95% confidence intervals, it can also be seen in Table 1.3 that the original Rao *et al.* (1992) (Eq. 1.13) has the correct coverage for small sampling fractions, although this variance estimator had the worst performance in terms of RB and RRMSE. Hence, discarding (Eq. 1.13), the proposed estimator (Eq. 1.7) has presumably the best coverage for increasing sampling fractions. In general, this also happens for the subsampling version (Eq. 1.12) which has similar RB, RRMSE and coverage as (Eq. 1.7).

Finally, both Tables 1.1, 1.2 and 1.3 suggest that, although the FPC corrections

TABLE 1.3: Coverage at 95% confidence level of variance estimators for the point estimator $\hat{R}_{12}$ where $ICC(y_1) = 0.2965$ and $ICC(y_2) = 0.1951$.

					Proposed		Customary	
m	n_I	n	f_I	f	(Eq. 1.7)	(Eq. 1.12)	(Eq. 1.13)	(Eq. 1.14)
2	20	40	0.058	0.009	91.86	91.84	92.89	92.13
	40	80	0.116	0.019	93.39	93.35	94.45	93.12
	60	120	0.174	0.028	93.83	93.65	95.25	93.10
	80	160	0.232	0.037	94.24	94.02	95.96	93.06
	100	200	0.290	0.046	94.37	94.05	96.53	92.70
	120	240	0.348	0.056	94.43	94.07	97.02	92.40
	140	280	0.406	0.065	94.73	94.22	97.54	91.95
	160	320	0.464	0.074	94.68	94.00	97.90	91.29
4	20	80	0.058	0.019	92.13	92.12	93.32	92.62
	40	160	0.116	0.037	93.47	93.44	94.93	93.52
	60	240	0.174	0.056	93.90	93.87	95.78	93.72
	80	320	0.232	0.074	94.21	94.20	96.54	93.89
	100	400	0.290	0.093	94.45	94.42	97.27	93.88
	120	480	0.348	0.112	94.66	94.63	97.84	93.94
	140	560	0.406	0.130	94.74	94.73	98.31	93.83
	160	640	0.464	0.149	94.65	94.62	98.69	93.42
6	20	120	0.058	0.028	92.07	92.07	93.41	92.71
	40	240	0.116	0.056	93.52	93.50	95.12	93.75
	60	360	0.174	0.084	94.12	94.11	96.16	94.23
	80	480	0.232	0.112	94.30	94.28	96.93	94.37
	100	600	0.290	0.139	94.50	94.48	97.56	94.51
	120	720	0.348	0.167	94.46	94.44	98.09	94.42
	140	840	0.406	0.195	94.59	94.59	98.63	94.52
	160	960	0.464	0.223	94.50	94.50	99.03	94.40

improves the Rao *et al.* (1992) estimator in terms of bias and stability, these artificial corrections are not always the best way to proceed; particularly, for situations where the second stage sampling may use small sampling fractions within certain clusters.

1.5.2 Example 2: Point estimator $\hat{R}_{34}$

The results for this example were summarised in Tables 1.4, 1.5 and 1.6. Additional graphical representations for these results are provided in the Appendix 1.E of this chapter.

TABLE 1.4: Relative Bias (%) of variance estimators for the point estimator $\hat{R}_{34}$ where $ICC(y_3) = 0.3181$ and $ICC(y_4) = 0.4958$.

					Proposed		Customary	
m	n_I	n	f_I	f	(Eq. 1.7)	(Eq. 1.12)	(Eq. 1.13)	(Eq. 1.14)
2	18	36	0.120	0.014	-3.5	-3.5	12.8	-0.8
	26	52	0.173	0.020	-2.4	-2.4	17.3	-3.0
	35	70	0.233	0.027	-1.7	-1.7	23.5	-5.3
	44	88	0.293	0.034	-1.4	-1.4	30.5	-7.8
	53	106	0.353	0.041	-1.1	-1.1	38.8	-10.3
	62	124	0.413	0.049	-1.1	-1.1	47.8	-13.3
	69	138	0.460	0.054	-0.9	-0.9	56.2	-15.7
4	18	72	0.120	0.028	-3.8	-3.8	14.1	0.4
	26	104	0.173	0.041	-2.6	-2.6	19.9	-0.9
	35	140	0.233	0.055	-2.0	-2.0	27.4	-2.3
	44	176	0.293	0.069	-1.7	-1.6	36.3	-3.7
	53	212	0.353	0.083	-1.4	-1.5	46.7	-5.1
	62	248	0.413	0.097	-1.2	-1.2	59.1	-6.7
	69	276	0.460	0.108	-1.0	-1.0	70.4	-8.0
6	18	108	0.120	0.042	-4.1	-4.1	14.6	0.8
	26	156	0.173	0.061	-2.9	-2.9	20.7	-0.2
	35	210	0.233	0.082	-1.9	-1.9	29.3	-0.8
	44	264	0.293	0.103	-1.6	-1.6	39.0	-1.7
	53	318	0.353	0.124	-1.7	-1.7	49.9	-3.0
	62	372	0.413	0.146	-1.5	-1.5	63.7	-4.0
	69	414	0.460	0.162	-1.0	-1.0	77.0	-4.4

In terms of RB, it can be seen that the FPC-adjusted Rao *et al.* (1992) (1.14) estimator has the best performance when the sampling fractions are very small. This might be useful for highly stratified sampling designs. However, for increasing sampling fractions, both versions of the Rao *et al.* (1992), the estimators (Eq. 1.13) and (Eq. 1.14), tend respectively to increasingly over and under estimate the variance. Again, this is more noticeable with small sample sizes at the second stage (small values of m).

On the other hand, in terms of RB, both proposed variance estimators from Equations (1.7) and (1.12) tend consistently to zero for increasing sampling fractions and regardless of the utilised sample size at the second stage. This is something desirable in business surveys for example, where sampling fractions are large or in situations where stratification is moderate. Note that the RB for the proposed estimators always showed a slight negative bias. This is something expected and well documented when using the Hájek (1964) approximations (see Haziza *et al.*, 2004, 2008; Brewer and Donadio, 2003).

TABLE 1.5: Relative Root Mean-Square-Error (%) of variance estimators for the point estimator $\hat{R}_{34}$ where $ICC(y_3) = 0.3181$ and $ICC(y_4) = 0.4958$.

					Proposed		Customary	
m	n_I	n	f_I	f	(Eq. 1.7)	(Eq. 1.12)	(Eq. 1.13)	(Eq. 1.14)
2	18	36	0.120	0.014	42.2	43.0	51.7	44.1
	26	52	0.173	0.020	33.6	35.1	44.9	34.4
	35	70	0.233	0.027	27.8	30.5	43.2	28.3
	44	88	0.293	0.034	23.8	27.8	44.9	24.6
	53	106	0.353	0.041	20.9	27.0	49.6	22.5
	62	124	0.413	0.049	18.5	26.8	56.2	21.8
	69	138	0.460	0.054	17.0	27.7	63.1	22.1
4	18	72	0.120	0.028	39.9	39.9	49.5	41.7
	26	104	0.173	0.041	31.9	32.0	44.2	32.6
	35	140	0.233	0.055	26.4	26.6	44.2	26.6
	44	176	0.293	0.069	22.7	22.9	48.2	22.7
	53	212	0.353	0.083	19.8	20.3	55.4	20.0
	62	248	0.413	0.097	17.6	18.3	65.7	18.2
	69	276	0.460	0.108	16.2	17.1	75.9	17.2
6	18	108	0.120	0.042	38.9	38.9	48.5	40.8
	26	156	0.173	0.061	31.1	31.1	43.8	31.9
	35	210	0.233	0.082	25.8	25.8	44.9	26.1
	44	264	0.293	0.103	22.1	22.2	50.0	22.1
	53	318	0.353	0.124	19.3	19.4	57.9	19.2
	62	372	0.413	0.146	17.1	17.2	69.7	17.0
	69	414	0.460	0.162	15.7	15.9	81.8	15.7

In terms of RRMSE, that is in terms of stability of the studied variance estimators, the Table 1.5 shows that the proposed estimator (Eq. 1.7) has the smallest RRMSE in all considered situations. It can also be seen that the subsampling version (Eq. 1.12) have small but slightly higher RRMSE.

In terms of the coverage of the 95% confidence intervals, it can be seen that the Rao *et al.* (1992) estimator (Eq. 1.13) has better coverage than the FPC-adjusted (Eq. 1.14). The coverage of the proposed estimators (Eq. 1.7) and (Eq. 1.12) become closer to 95% for increasing sampling fractions. Overall, the worst coverage was

TABLE 1.6: Coverage at 95% confidence level of variance estimators for the point estimator $\hat{R}_{34}$ where $ICC(y_3) = 0.3181$ and $ICC(y_4) = 0.4958$.

					Proposed		Customary	
m	n_I	n	f_I	f	(Eq. 1.7)	(Eq. 1.12)	(Eq. 1.13)	(Eq. 1.14)
2	18	36	0.120	0.014	93.1	93.0	94.8	93.3
	26	52	0.173	0.020	93.7	93.6	95.7	93.6
	35	70	0.233	0.027	94.1	93.9	96.4	93.5
	44	88	0.293	0.034	94.3	94.0	97.0	93.4
	53	106	0.353	0.041	94.4	94.1	97.5	93.2
	62	124	0.413	0.049	94.5	94.0	98.0	92.8
	69	138	0.460	0.054	94.6	94.0	98.3	92.4
4	18	72	0.120	0.028	92.9	92.9	94.8	93.4
	26	104	0.173	0.041	93.6	93.6	95.8	93.8
	35	140	0.233	0.055	94.0	94.0	96.6	94.0
	44	176	0.293	0.069	94.2	94.2	97.3	93.9
	53	212	0.353	0.083	94.4	94.3	97.9	93.9
	62	248	0.413	0.097	94.5	94.5	98.4	93.8
	69	276	0.460	0.108	94.5	94.5	98.7	93.6
6	18	108	0.120	0.042	92.9	92.9	94.9	93.5
	26	156	0.173	0.061	93.6	93.6	95.9	93.9
	35	210	0.233	0.082	94.0	94.0	96.8	94.1
	44	264	0.293	0.103	94.2	94.2	97.4	94.2
	53	318	0.353	0.124	94.3	94.3	98.0	94.2
	62	372	0.413	0.146	94.4	94.4	98.5	94.1
	69	414	0.460	0.162	94.5	94.5	98.9	94.1

showed by the FPC-adjusted Rao *et al.* (1992) estimator (Eq. 1.14). This suggest again the fixed *ad hoc* FPC correction might not be always suitable.

1.6 Conclusion

Self-weighted two-stage sampling designs are very common in practice. Besides the popularity of such designs, it is also common in practice to compute variance estimates relying only on the first sampling stage (e.g. Särndal *et al.*, 1992, ch. 4).

A customary jackknife variance estimator for sampling designs of more than one stage is the Rao *et al.* (1992) estimator which is originally designed for functions of totals and for negligible sampling fractions. This customary jackknife would work well when most of the variability is between clusters and with very small sampling fractions (highly stratified samples) but this may not necessarily be the case.

First, we propose an alternative Sen-Yates-Grundy form of the generalised unequal-probability without-replacement jackknife variance estimator (Campbell, 1980). This estimator is extended to two-stage sampling by proposing new less restrictive regularity conditions than those from Berger and Skinner (2005), and thus allowing two-stage sampling for the Horvitz-Thompson (original) form of the Campbell (1980) generalised jackknife as well.

Secondly, we propose a novel design-consistent jackknife variance estimator for self-weighted two-stage without-replacement sampling. The proposed estimator does not need joint-inclusion probabilities, allows stratification, naturally includes FPC and comprises a wide class of point estimators (functions of means).

Monte-Carlo simulations show that the proposed estimator can be more accurate than customary jackknife estimators, specially in situations where the first stage sampling fraction is large or in cases where the second stage sampling fractions are small.

The proposed estimator incorporates not only clustering effects but also the underlying unequal-probabilities of both, clusters and observations.

Appendices to Chapter 1

1.A. Proof of Theorem 1.1

The proof uses the standard arguments in proving jackknife variance estimators design consistency (see Miller, 1964; Shao and Tu, 1995, sub-sec. 2.1.1). Hence, from the mean value theorem we have that

$$\begin{aligned}\tilde{\theta} - \tilde{\theta}_{(k)} &= g(\tilde{\boldsymbol{\mu}}) - g(\tilde{\boldsymbol{\mu}}_{(k)}) \\ &= \boldsymbol{\nabla}(\boldsymbol{\xi}_k)^T\,(\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}_{(k)}) \\ &= \boldsymbol{\nabla}(\tilde{\boldsymbol{\mu}})^T\,(\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}_{(k)}) \ + \ r_k^*, \end{aligned} \tag{1.15}$$

where $\boldsymbol{\xi}_k$ denotes a point between $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\mu}}_{(k)}$, and where

$$r_k^* \ = \ (\boldsymbol{\nabla}(\boldsymbol{\xi}_k) - \boldsymbol{\nabla}(\tilde{\boldsymbol{\mu}}))^T\,(\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}_{(k)})$$

is the remainder term. Thus,

$$\varepsilon_{(k)} \ = \ \boldsymbol{\nabla}(\tilde{\boldsymbol{\mu}})^T\,(1 - \tilde{w}_k)\,(\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}_{(k)}) \ + \ r_k,$$

where

$$r_k \ = \ (1 - \tilde{w}_k)\ r_k^*. \tag{1.16}$$

It can be shown that

$$(1 - \tilde{w}_k)\,(\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}_{(k)}) \ = \ \tilde{w}_k\,(\boldsymbol{y}_k - \tilde{\boldsymbol{\mu}}), \tag{1.17}$$

implying that

$$\varepsilon_{(k)} \ = \ \boldsymbol{\nabla}(\tilde{\boldsymbol{\mu}})^T\,\tilde{w}_k\,(\boldsymbol{y}_k - \tilde{\boldsymbol{\mu}}) \ + \ r_k. \tag{1.18}$$

Furthermore, the Cauchy inequality together with Equations (1.16) and (1.17) imply

$$|r_k| \ \leq \ ||\boldsymbol{\nabla}(\boldsymbol{\xi}_k) - \boldsymbol{\nabla}(\tilde{\boldsymbol{\mu}})||\,\tilde{w}_k\,||\boldsymbol{y}_k - \tilde{\boldsymbol{\mu}}||. \tag{1.19}$$

Besides, the regularity condition C7 implies that there are constants $\lambda > 0$, δ and $0 \leq \beta < 1$ where $\beta/2 < \delta$ such that,

$$||\boldsymbol{\nabla}(\boldsymbol{\xi}_k) - \boldsymbol{\nabla}(\tilde{\boldsymbol{\mu}})|| \ \leq \ \lambda\,||\boldsymbol{\xi}_k - \tilde{\boldsymbol{\mu}}||^{\delta}. \tag{1.20}$$

As $\boldsymbol{\xi}_k$ is between $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\mu}}_{(k)}$, we have that

$$||\boldsymbol{\xi}_k - \tilde{\boldsymbol{\mu}}|| \;\leq\; ||\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}_{(k)}||.$$

Combining this with Equation (1.17), we obtain

$$||\boldsymbol{\xi}_k - \tilde{\boldsymbol{\mu}}|| \;\leq\; ||(1-\tilde{w}_k)^{-1}\,\tilde{w}_k\,(\boldsymbol{y}_k - \tilde{\boldsymbol{\mu}})||,$$

which by Equation (1.20) gives

$$||\boldsymbol{\nabla}(\boldsymbol{\xi}_k) - \boldsymbol{\nabla}(\tilde{\boldsymbol{\mu}})|| \;\leq\; \lambda\,|1-\tilde{w}_k|^{-\delta}\,\tilde{w}_k^{\delta}\,||\boldsymbol{y}_k - \tilde{\boldsymbol{\mu}}||^{\delta}.$$

Then, by condition C2 this becomes

$$||\boldsymbol{\nabla}(\boldsymbol{\xi}_k) - \boldsymbol{\nabla}(\tilde{\boldsymbol{\mu}})|| \;\leq\; \lambda\,\alpha^{-\delta}\,\tilde{w}_k^{\delta}\,||\boldsymbol{y}_k - \tilde{\boldsymbol{\mu}}||^{\delta},$$

which combined with the Equation (1.19) imply

$$|r_k| \;\leq\; \lambda\,\alpha^{-\delta}\,\tilde{w}_k^{1+\delta}\,||\boldsymbol{y}_k - \tilde{\boldsymbol{\mu}}||^{1+\delta}. \tag{1.21}$$

Moreover, the regularity condition C3 implies that

$$\{n\ \mathrm{var}(\tilde{\theta})_L\}^{-2} \;=\; \mathcal{O}(1). \tag{1.22}$$

By substituting (Eq. 1.18) in Equation (1.3), we obtain

$$\widehat{\mathrm{var}}(\tilde{\theta})_{SYG} \;=\; A \;+\; 2\,(E-C) \;+\; D-B,$$

where

$$\begin{aligned}
A &= \boldsymbol{\nabla}(\tilde{\boldsymbol{\mu}})^T\ \widehat{\mathbf{var}}(\tilde{\boldsymbol{\mu}})_{SYG}\ \boldsymbol{\nabla}(\tilde{\boldsymbol{\mu}}),\\
B &= \sum_{k\in s}\sum_{\ell\in s}\mathcal{D}_{k\ell}\ r_k\ r_\ell,\\
C &= \sum_{k\in s}\sum_{\ell\in s}\mathcal{D}_{k\ell}\ r_k\ \tilde{w}_\ell\,(\boldsymbol{y}_\ell - \tilde{\boldsymbol{\mu}})^T\ \boldsymbol{\nabla}(\tilde{\boldsymbol{\mu}}),\\
D &= \sum_{k\in s}\sum_{\ell\in s}\mathcal{D}_{k\ell}\ r_k^2, && (1.23)\\
E &= \sum_{k\in s}\sum_{\ell\in s}\mathcal{D}_{k\ell}\ r_k\ \tilde{w}_k\,(\boldsymbol{y}_k - \tilde{\boldsymbol{\mu}})^T\ \boldsymbol{\nabla}(\tilde{\boldsymbol{\mu}}). && (1.24)
\end{aligned}$$

Hence, Theorem 1.1 follows if we may show

$$\begin{aligned} A/\mathrm{var}(\tilde{\theta})_L &\rightarrow_p 1, & (1.25)\\ B/\mathrm{var}(\tilde{\theta})_L &\rightarrow_p 0, & (1.26)\\ C/\mathrm{var}(\tilde{\theta})_L &\rightarrow_p 0, & (1.27)\\ D/\mathrm{var}(\tilde{\theta})_L &\rightarrow_p 0, & (1.28)\\ E/\mathrm{var}(\tilde{\theta})_L &\rightarrow_p 0, & (1.29)\end{aligned}$$

The condition C1 implies Equation (1.25), whereas Equations (1.26) and (1.27) can be shown following the Berger and Skinner (2005) proof taking into account the changes in regularity conditions C5 to C7. Hence, it remains to show Equations (1.28) and (1.29). We start with Equation (1.28). By the triangle and by the Cauchy inequalities, the Equation (1.23) implies

$$\begin{aligned} |D| &\leq \sum_{k\in s}\sum_{\ell\in s} |\mathcal{D}_{k\ell}|\,|r_k|^2\\ &= D_1 + D_2\\ &\leq (G_s^{1/2} + H_s^{1/2})\,D_3^{1/2},\end{aligned}$$

where

$$\begin{aligned} D_1 &= \sum_{k\in s}\sum_{\ell\in s} \mathcal{D}^-_{k\ell}\,|r_k|^2\\ &\leq G_s^{1/2}D_3^{1/2},\\ D_2 &= \sum_{k\in s}\sum_{\ell\in s} \mathcal{D}^+_{k\ell}\,|r_k|^2\\ &\leq H_s^{1/2}D_3^{1/2},\end{aligned}$$

and

$$\begin{aligned} D_3 &= n^{\beta}\sum_{k\in s}\sum_{\ell\in s}|r_k|^4\\ &= n^{1+\beta}\sum_{k\in s}|r_k|^4. \qquad (1.30)\end{aligned}$$

Thus, (Eq. 1.28) follows from conditions C5 and C6, if we show $D_3\{\mathrm{var}(\tilde{\theta})_L\}^{-2} \rightarrow_p 0$. Hence, using the Equation (1.21) in (Eq. 1.30), we have that

$$\frac{D_3}{\mathrm{var}(\tilde{\theta})^2_L} \leq \frac{\lambda^4}{\alpha^{4\delta}}\frac{n^{4+\beta}}{\{n\ \mathrm{var}(\tilde{\theta})_L\}^2}\left(\frac{1}{n}\sum_{k\in s}\tilde{w}_k^{4(1+\delta)}\,||\boldsymbol{y}_k - \tilde{\boldsymbol{\mu}}||^{4(1+\delta)}\right). \qquad (1.31)$$

Conditions C3, C4 and Equations (1.22) and (1.31) imply

$$D_3\{\mathrm{var}(\tilde{\theta})_L\}^{-2} \;=\; n^{\beta}\mathcal{O}_p(n^{-4\delta}).$$

From condition C7, $\beta < 4\delta$. Thus $D_3\{\mathrm{var}(\tilde{\theta})_L\}^{-2} \rightarrow_p 0$, implying (1.28). We now show Equation (1.29). Using the triangle and the Cauchy inequalities in (Eq. 1.24) gives

$$\begin{aligned}
|E| &\leq \sum_{k\in s}\sum_{\ell\in s} |\mathcal{D}_{k\ell}|\;|r_k|\;|\tilde{y}_k| \\
&= E_1 + E_2 \\
&\leq (G_s^{1/2} + H_s^{1/2})\,E_3^{1/2}, \qquad (1.32)
\end{aligned}$$

where

$$\begin{aligned}
E_1 &= \sum_{k\in s}\sum_{\ell\in s} \mathcal{D}_{k\ell}^{-}\;|r_k|\;|\tilde{y}_k|, \\
E_2 &= \sum_{k\in s}\sum_{\ell\in s} \mathcal{D}_{k\ell}^{+}\;|r_k|\;|\tilde{y}_k|,
\end{aligned}$$

and

$$\begin{aligned}
E_3 &= n^{\beta}\sum_{k\in s}\sum_{\ell\in s} |r_k|^2\;|\tilde{y}_k|^2 \\
&= n^{1+\beta}\sum_{k\in s} |r_k|^2\;|\tilde{y}_k|^2, \qquad (1.33) \\
\tilde{y}_k &= \tilde{w}_k\;(\boldsymbol{y}_k - \tilde{\boldsymbol{\mu}})^T\;\boldsymbol{\nabla}(\tilde{\boldsymbol{\mu}}). \qquad (1.34)
\end{aligned}$$

Thus, Equation (1.29) follows from conditions C5 and C6, if we show $E_3\{\mathrm{var}(\tilde{\theta})_L\}^{-2} \rightarrow_p 0$. Hence, by using the Cauchy inequality in (Eq. 1.34) we have that

$$|\tilde{y}_k| \;\leq\; \tilde{w}_k\;||\boldsymbol{y}_k - \tilde{\boldsymbol{\mu}}||\;||\boldsymbol{\nabla}(\tilde{\boldsymbol{\mu}})||.$$

This inequality together with Equations (1.21) and (1.33) imply that

$$\frac{E_3}{\mathrm{var}(\tilde{\theta})_L^2} \;\leq\; ||\boldsymbol{\nabla}(\tilde{\boldsymbol{\mu}})||^2\frac{\lambda^2}{\alpha^{2\delta}}\frac{n^{4+\beta}}{\{n\;\mathrm{var}(\tilde{\theta})_L\}^2}\left(\frac{1}{n}\sum_{k\in s}\tilde{w}_k^{4+2\delta}||\boldsymbol{y}_k - \tilde{\boldsymbol{\mu}}||^{4+2\delta}\right). \qquad (1.35)$$

From condition C7, $\beta < 2\delta$. This, together with (Eq. 1.35), (Eq. 1.22), conditions C4 and C8 imply $E_3\{\mathrm{var}(\tilde{\theta})_L\}^{-2} = n^{\beta}\mathcal{O}_p(n^{-2\delta})$, i.e. $E_3\{\mathrm{var}(\tilde{\theta})_L\}^{-2} \rightarrow_p 0$. ■

1.B. Proof of Theorem 1.3

We use the Theorem 1.1 that sets the consistency of the variance estimator $\widehat{\text{var}}(\tilde{\theta})_{SYG}$ from Equation (1.3), which was utilised to develop the proposed variance estimator $\widehat{\text{var}}(\tilde{\theta})_{prop}$ in Equation (1.7). Hence, given conditions C1, C3, C4, C7 and C8, it remains to show that $M_i \geq m \geq 2$, for all $i = 1, \ldots, N_I$ implies that C2, C5 and C6 hold.

From self-weighting, it can easily be shown that the condition C2 holds. We now show the conditions C5 and C6 hold.

Let $q_{Ii} = 1 - \pi_{Ii}$ and $q^*_{Ii} = 1 - \pi^*_{Ii}$, and also let

$$\beta \;=\; \frac{\text{Log}(n_I)}{\text{Log}(n)} \;<\; 1,$$

be such that $n^\beta = n_I$. It can be shown that $|q^*_{Ii}| = \mathcal{O}(1)$ for $M_i \geq m \geq 2$, for all $i = 1, \ldots, N_I$, and also that $d = \mathcal{O}(N_I)$ as $q_{Ii} = \mathcal{O}(1)$.

If $q^*_{Ii} > 0$, we have from the conditions C5 and C6 that:

$$\mathcal{D}^-_{k\ell} \;=\; q_{Ii}\, q_{Ij}/d \;=\; \mathcal{O}_p({N_I}^{-1}) \;\text{ for }\; k \in s_i, \ell \in s_j, i \neq j,$$

and that

$$\mathcal{D}^+_{k\ell} \;=\; q^*_{Ii} \;\text{ for }\; (k \neq \ell) \in s_i.$$

Thus, we obtain

$$\begin{aligned}
G_s &= \frac{1}{n_I} \sum_{i=1}^{n_I} \sum_{j=1, i\neq j}^{n_I} \sum_{k=1}^{m} \sum_{\ell=1}^{m} \Big(\frac{q_{Ii}\, q_{Ij}}{d}\Big)^2 \\
&= \frac{m^2 {n_I}^2}{n_I {N_I}^2} \mathcal{O}_p(1) \\
&= f_I^2 m^2 \mathcal{O}_p(n_I^{-1}), \\
H_s &= \frac{1}{n_I} \sum_{i=1}^{n_I} \sum_{j=1, i=j}^{n_I} \sum_{k=1}^{m} \sum_{\ell=1, k\neq\ell}^{m} (q^*_{Ii})^2 \\
&= m(m-1)\mathcal{O}_p(1),
\end{aligned}$$

where $f_I = n_I/N_I$ and m are constants. Moreover, if $\hat{d}$ was used instead of d in Equation (1.8), then:

$$G_s \;=\; m^2 \mathcal{O}_p(n_I^{-1}).$$

Following, if $q^*_{Ii} < 0$, we have that:

$$\mathcal{D}^-_{k\ell} \;=\; q_{Ii}\, q_{Ij}/d \;=\; \mathcal{O}_p({N_I}^{-1}) \;\text{ for }\; k \in s_i, \ell \in s_j, i \neq j,$$

and

$$\mathcal{D}^-_{k\ell} \;=\; q^*_{Ii} \;\text{ for }\; (k \neq \ell) \in s_i,$$

and also that

$$\mathcal{D}^+_{k\ell} \;=\; 0.$$

Hence,

$$G_s \;=\; f_I^2 m^2 \mathcal{O}_p(n_I^{-1}) \;+\; m(m-1)\mathcal{O}_p(1),$$

and

$$H_s \;=\; 0.$$

Again, if $\hat{d}$ was used instead of d, then

$$G_s \;=\; m^2 \mathcal{O}_p(n_I^{-1}) \;+\; m(m-1)\mathcal{O}_p(1).$$

Thus, G_s and H_s are $\mathcal{O}_p(1)$ completing the proof. ■

1.C. Proof of Equation (1.7)

We first introduce some results and approximations which will be used. Hence, for the designs and estimators defined in sections 1.2 and 1.4, we have that

$$\varsigma_{(Ii)} = \sum_{k \in s_i} \varepsilon_{(k)}. \tag{1.36}$$

We show Equation (1.36) by recalling that the sampling design is self weighted. Hence,

$$\begin{aligned} n &= n_I \; m, \\ \tilde{w}_k &= \frac{1}{n}, \\ \tilde{w}_{\ell(k)} &= \frac{1}{n-1}, \\ \tilde{w}_{k(Ii)} &= \frac{1}{n-m}, \end{aligned}$$

implying

$$\begin{aligned} \frac{n_I - 1}{n_I}(\tilde{\mu}_q - \tilde{\mu}_{q(Ii)}) &= \frac{1}{n_I}\left[(n_I - 1)\tilde{\mu}_q - \frac{n_I - 1}{m(n_I - 1)}\left(n_I m \tilde{\mu}_q - \sum_{k \in s_i} y_{qk}\right)\right] \\ &= \sum_{k \in s_i} \frac{(n-1)\tilde{\mu}_q - n\tilde{\mu}_q + y_{qk}}{n} \\ &= \frac{n-1}{n} \sum_{k \in s_i} (\tilde{\mu}_q - \tilde{\mu}_{q(k)}). \end{aligned}$$

Then, as $g(\cdot)$ is linearisable, we obtain Equation (1.36). Now, under the asymptotic framework from subsection 3.1 (Isaki and Fuller, 1982) we have that $N \to \infty$. Thus, implying $N_I \to \infty$ as m is assumed fixed. Then we have that $d \to \infty$, i.e. the Hájek (1964) asymptotic framework for the clusters' selection stage. Further, we have that $q_{Ii} = \mathcal{O}(1)$. Letting $q_{Ii} = 1 - \pi_{Ii}$ and $q^*_{Ii} = 1 - \pi^*_{Ii}$, we also introduce the following approximations which are suitable for large values of d:

$$q_{Ii} \; q_{Ij} \; \{q_{Ii} \; q_{Ij} - d\}^{-1} \doteq -q_{Ii} q_{Ii}/d, \tag{1.37}$$

$$(\hat{d} - q_{Ii})/d \doteq 1, \tag{1.38}$$

$$q^*_{Ii} + q^2_{Ii}/d \doteq q^*_{Ii}, \tag{1.39}$$

Hence, from Equation (1.3), we have that

$$\begin{aligned}
\widehat{\mathrm{var}}(\tilde{\theta})_{SYG} &= \sum_{k\in s}\ \sum_{\ell\in s,\ell\neq k} \mathcal{D}_{kl}\ \varepsilon_{(k)}\,\varepsilon_{(\ell)} - \sum_{k\in s}\ \sum_{\ell\in s,\ell\neq k} \mathcal{D}_{kl}\ \varepsilon^2_{(k)} \\
&= \sum_{i\in s}\ \sum_{j\in s,j=i}\ \sum_{k\in s_i}\ \sum_{\ell\in s_j,\ell\neq k} \mathcal{D}_{kl}\ \varepsilon_{(k)}\,\varepsilon_{(\ell)} \\
&\quad + \sum_{i\in s}\ \sum_{j\in s,j\neq i}\ \sum_{k\in s_i}\ \sum_{\ell\in s_j} \mathcal{D}_{kl}\ \varepsilon_{(k)}\,\varepsilon_{(\ell)} \\
&\quad - \sum_{i\in s}\ \sum_{j\in s,j=i}\ \sum_{k\in s_i}\ \sum_{\ell\in s_j,\ell\neq k} \mathcal{D}_{kl}\ \varepsilon^2_{(k)} \\
&\quad - \sum_{i\in s}\ \sum_{j\in s,j\neq i}\ \sum_{k\in s_i}\ \sum_{\ell\in s_j} \mathcal{D}_{kl}\ \varepsilon^2_{(k)}.
\end{aligned}$$

Then, substituting $\mathcal{D}_{kl}$ (see section 1.4) and using Equations (1.36) and (1.37) we obtain,

$$\begin{aligned}
\widehat{\mathrm{var}}(\tilde{\theta})_{SYG} &\bumpeq \sum_{i\in s} q^*_{Ii}\left[\varsigma^2_{(Ii)} - \sum_{k\in s_i}\ \sum_{\ell\in s_i \ell=k} \varepsilon_{(k)}\ \varepsilon_{(\ell)}\right] \\
&\quad - \sum_{i\in s}\ \sum_{j\in s,j\neq i} \frac{q_{Ii}\ q_{Ij}}{d}\ \varsigma_{(Ii)}\ \varsigma_{(Ij)} \\
&\quad - (m-1)\sum_{i\in s} q^*_{Ii}\sum_{k\in s_i}\varepsilon^2_{(k)} \\
&\quad + m\sum_{i\in s}\ \sum_{j\in s,j\neq i} \frac{q_{Ii}\ q_{Ij}}{d}\ \sum_{k\in s_i}\varepsilon^2_{(k)} \\
&= \sum_{i\in s}\left[q^*_{Ii} + \frac{q^2_{Ii}}{d}\right]\varsigma^2_{(Ii)}\ -\ \frac{1}{d}\left[\sum_{i\in s} q_{Ii}\ \varsigma_{(Ii)}\right]^2 \\
&\quad + m\sum_{i\in s}\left[q_{Ii}\ \frac{\widehat{d} - q_{Ii}}{d}\ -\ q^*_{Ii}\right]\sum_{k\in s_i}\varepsilon^2_{(k)}.
\end{aligned}$$

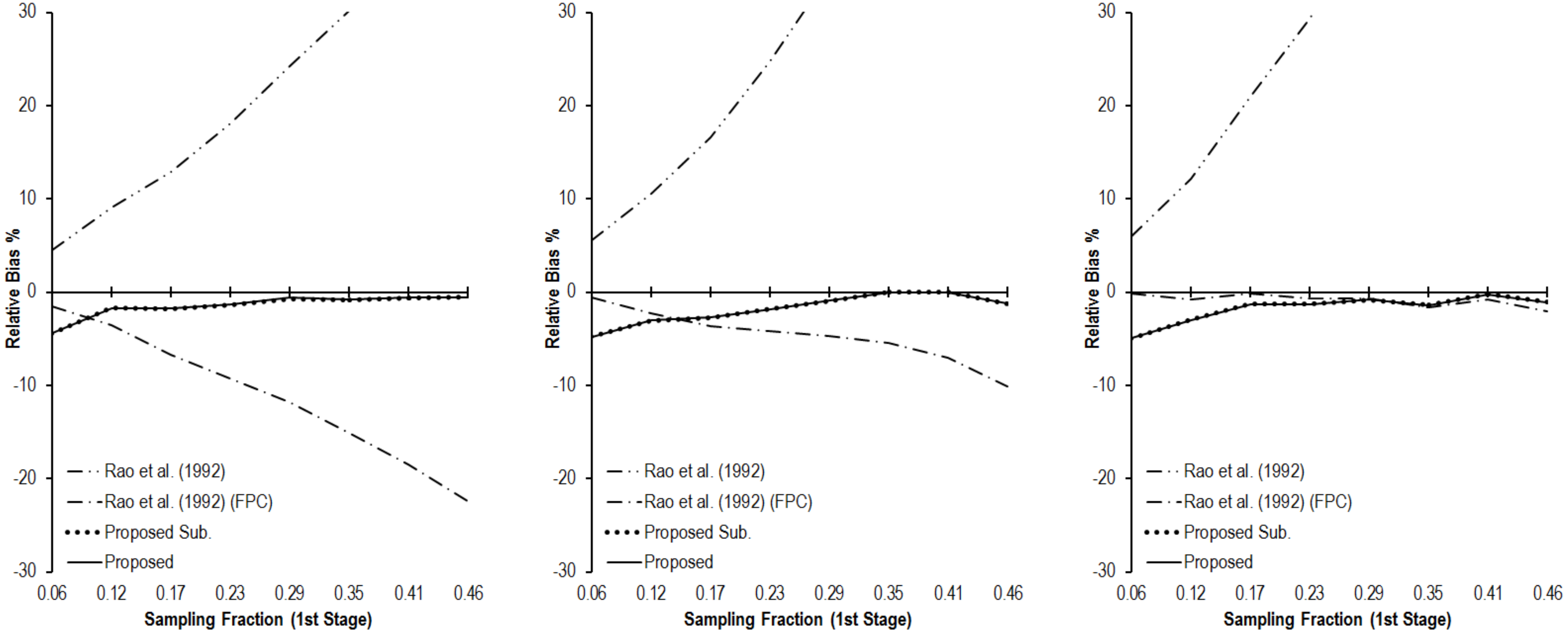
Now, by using Equations (1.38) and (1.39), combined with

$$\begin{aligned}
q_{Ii} - q^*_{Ii} &= \frac{\pi_{Ii} - f}{m-1} \\
&= \frac{\pi_{Ii}(M_i - m)}{M_i(m-1)},
\end{aligned}$$

we obtain the proposed estimator $\widehat{\mathrm{var}}(\tilde{\theta})_{prop}$ from Equation (1.7). ■

1.D. Figures of the simulation study, Example 1

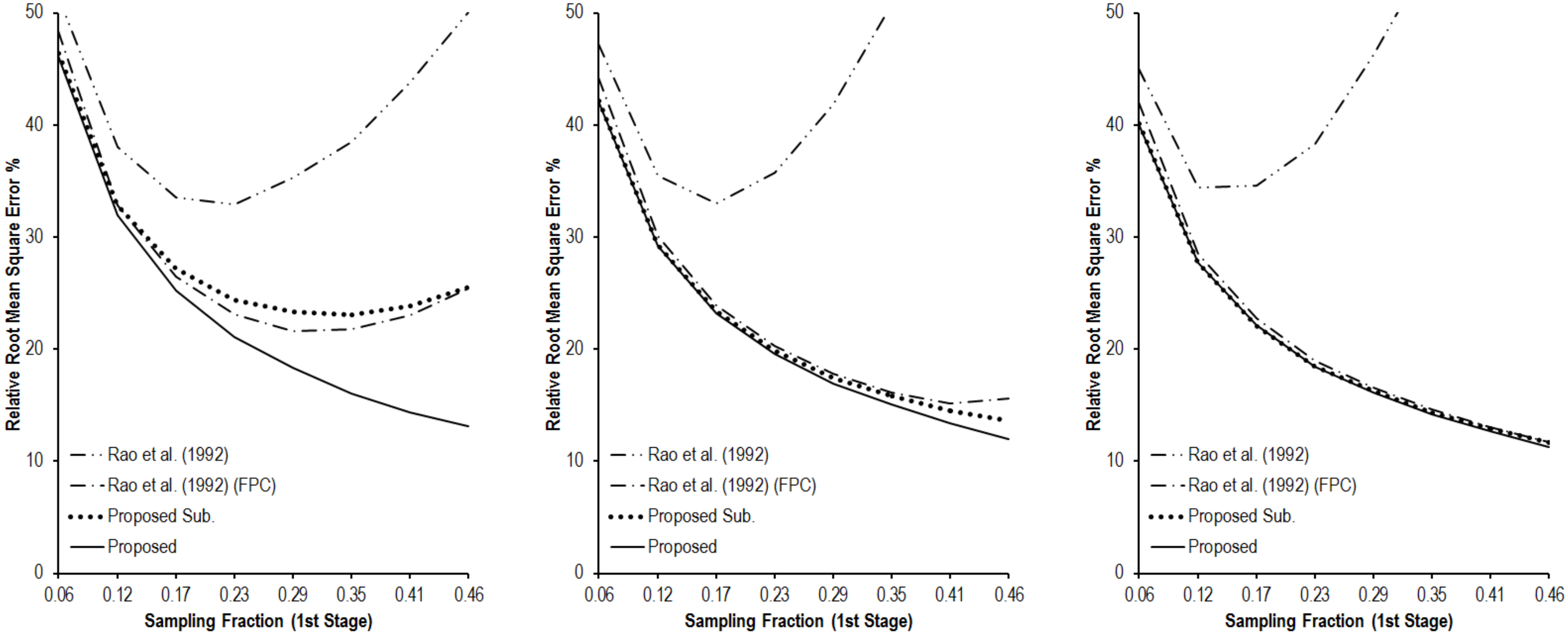
Relative Bias for the variance of $\hat{R}_{12}$



$m = 2$ $m = 4$ $m = 6$

FIGURE 1.1: Relative Bias (%) of variance estimators for the point estimator $\hat{R}_{12}$ where $ICC(y_1) = 0.2965$ and $ICC(y_2) = 0.1951$.

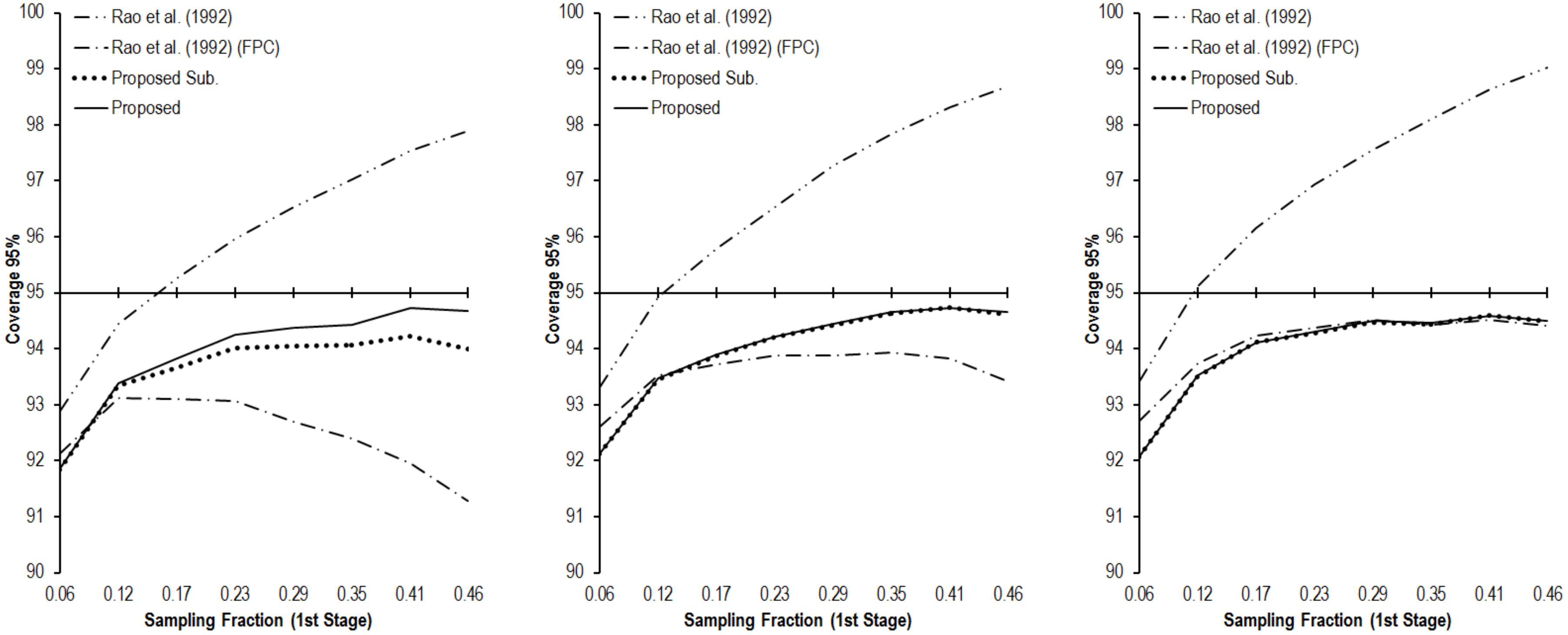
Relative Root Mean-Square-Error for the variance of $\hat{R}_{12}$



$m = 2$ $m = 4$ $m = 6$

FIGURE 1.2: Relative Root Mean-Square-Error (%) of variance estimators for the point estimator $\hat{R}_{12}$ where $ICC(y_1) = 0.2965$ and $ICC(y_2) = 0.1951$.

Coverage at 95% confidence level for the variance of $\hat{R}_{12}$

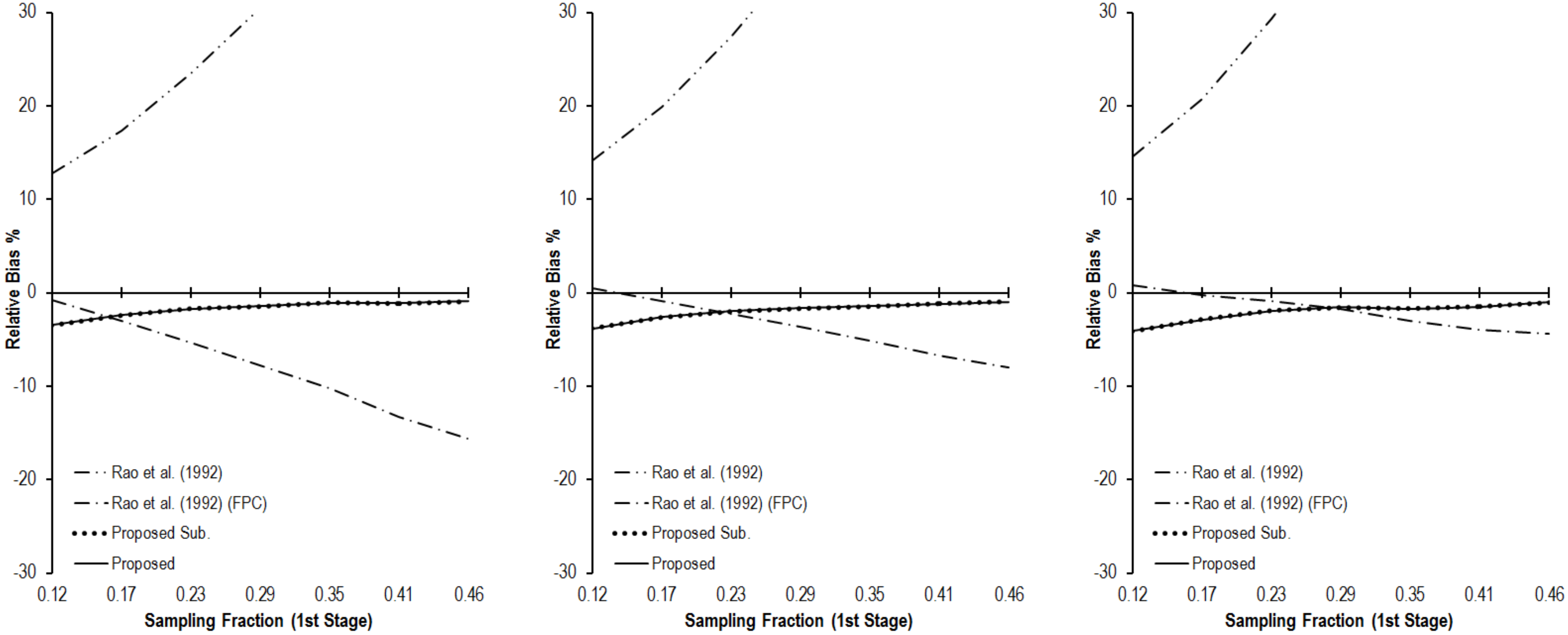


$m = 2$ $m = 4$ $m = 6$

FIGURE 1.3: Coverage at 95% confidence level of variance estimators for the point estimator $\hat{R}_{12}$ where $ICC(y_1) = 0.2965$ and $ICC(y_2) = 0.1951$.

1.E. Figures of the simulation study, Example 2

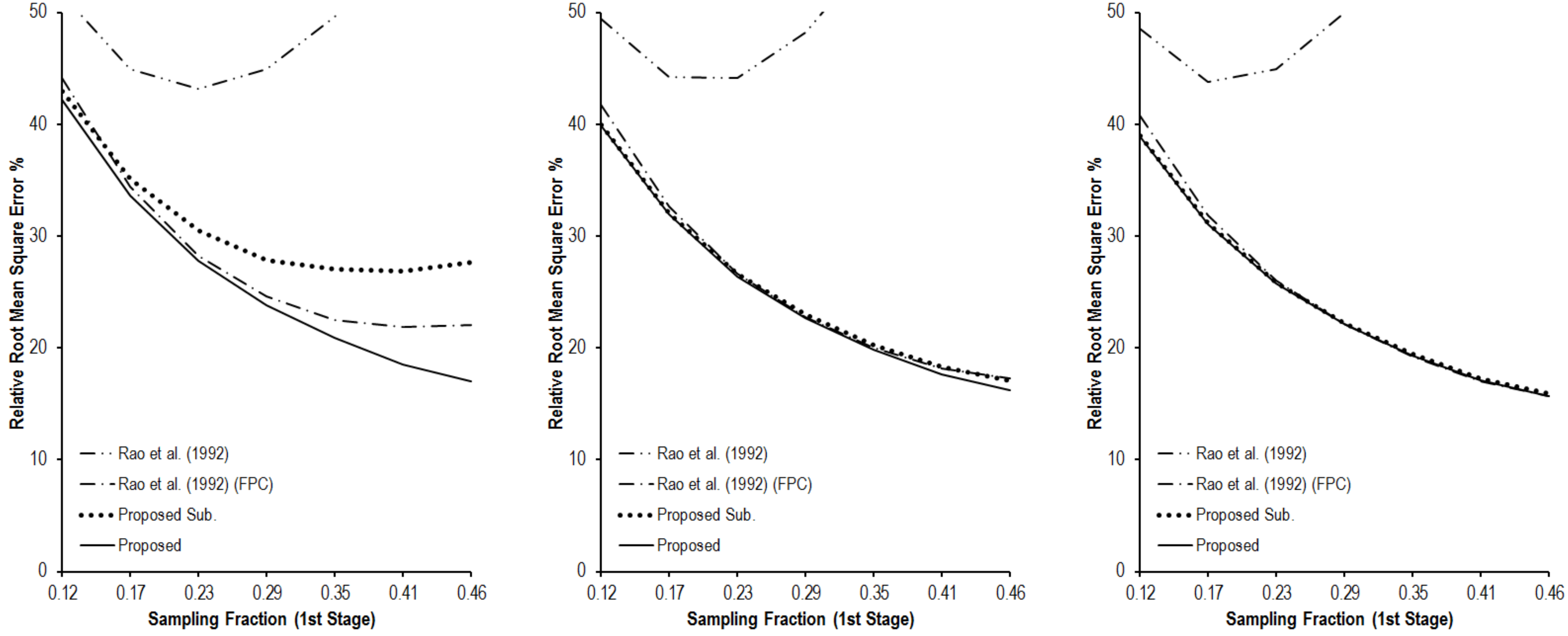
Relative Bias for the variance of $\hat{R}_{34}$



$m = 2$ $m = 4$ $m = 6$

FIGURE 1.4: Relative Bias (%) of variance estimators for the point estimator $\hat{R}_{34}$ where $ICC(y_3) = 0.3181$ and $ICC(y_4) = 0.4958$.

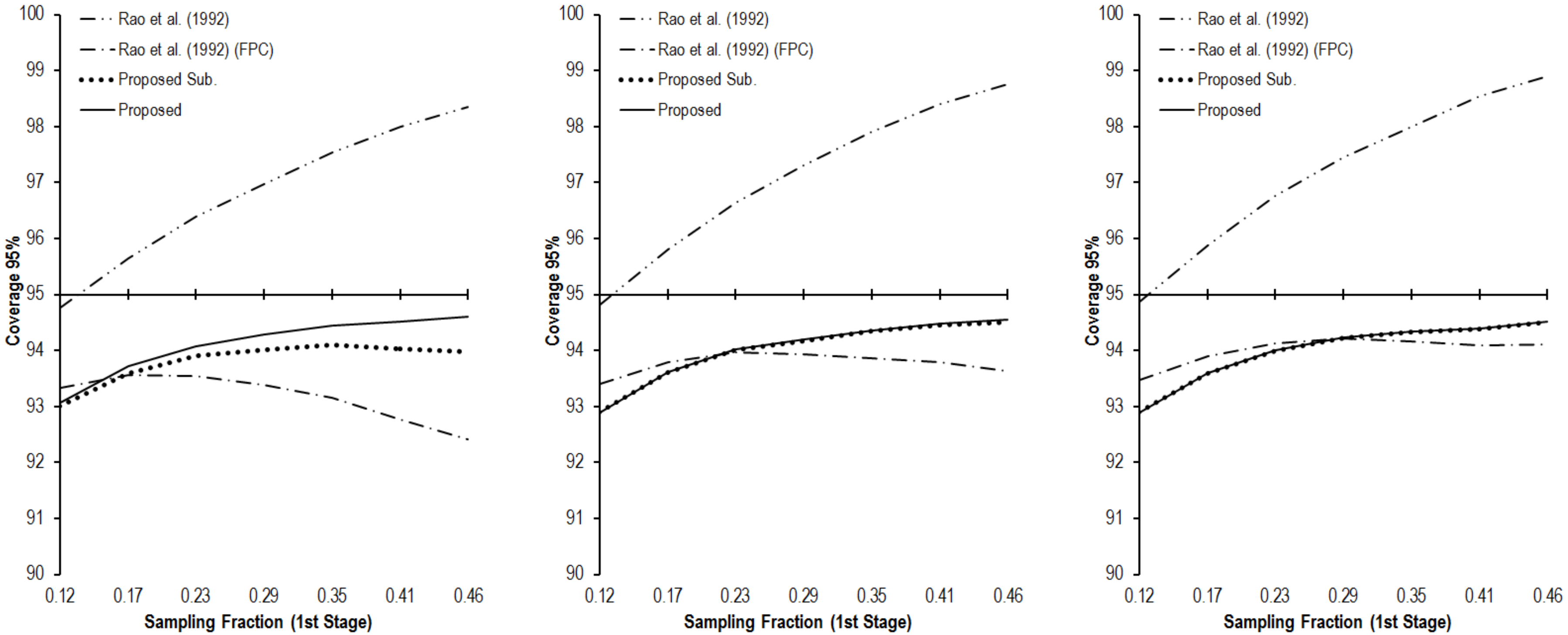
Relative Root Mean-Square-Error for the variance of $\hat{R}_{34}$



$m = 2$ $m = 4$ $m = 6$

FIGURE 1.5: Relative Root Mean-Square-Error (%) of variance estimators for the point estimator $\hat{R}_{34}$ where $ICC(y_3) = 0.3181$ and $ICC(y_4) = 0.4958$.

Coverage at 95% confidence level for the variance of $\hat{R}_{34}$



$m = 2$ $m = 4$ $m = 6$

FIGURE 1.6: Coverage at 95% confidence level of variance estimators for the point estimator $\hat{R}_{34}$ where $ICC(y_3) = 0.3181$ and $ICC(y_4) = 0.4958$.

Chapter 2

A new replicate variance estimator for unequal probability sampling without replacement

Abstract

We propose a new replicate variance estimator suitable for differentiable functions of estimated totals. The proposed variance estimator is defined for any unequal-probability without-replacement sampling design, it naturally includes finite population corrections and it allows two-stage sampling. We show its design-consistency and its close relationship with linearisation variance estimators.

When estimating a total, the proposed estimator reduces to the Horvitz and Thompson (1952) variance estimator. Monte-Carlo simulations suggest that the proposed variance estimator is more stable than its replicate competitors.

Keywords and phrases: Derivative; jackknife; pseudo-value; self-weighted; stratification; Taylor linearisation.

2.1 Introduction

Replication methods for variance estimation such as the Jackknife, the Bootstrap and the Balanced Half-Samples are very popular in practice (e.g. Shao and Tu, 1995; Davison and Hinkley, 1997; Lehtonen and Pahkinen, 2004; Wolter, 2007). However, there are only limited applications under unequal-probability without-replacement sampling designs (e.g. Berger and Skinner, 2005; Berger and Rao, 2006; Berger, 2007).

Linearisation is an alternative to replication methods (e.g. Robinson and Särndal, 1983; Binder, 1996; Deville, 1999; Demnati and Rao, 2004, 2010). Although slightly design-biased, the linearisation variance estimators are more stable than its replication counterparts (e.g. Kish and Frankel, 1974; Kovar *et al.*, 1988; Shao and Tu, 1995, pp. 32, 69). There are different approaches for deriving linearisation variance estimators which may give asymptotically equivalent but different estimators (Binder, 1996, pp. 17, 18). Deville (1999) proposes an approach based upon derivatives of the population parameter of interest. Demnati and Rao (2004) propose an approach which is based on derivatives of the estimator of the parameter of interest. Nevertheless, linearisation involves deriving analytic derivatives; a well documented practical drawback (e.g. Shao and Tu, 1995, pp. 69, 281). Skinner (2004) and Demnati and Rao (2004, p. 21) raised the need of a replication estimator which overcomes that practical drawback.

We propose a new replicate variance estimator suitable for differentiable functions of estimated totals. The estimator is defined for any unequal-probability without-replacement sampling design and it naturally includes finite population corrections. The proposed approach consists in repeatedly perturbing the sampling weights. In Subsections 2.3.1 and 2.3.2, we show that this novel approach can be interpreted in several ways depending on its configuration. Further, we show that the proposed replicate variance estimator is approximately equal to linearisation variance estimators. Moreover, it can be seen as an approximation to the linearisation estimators obtained by the Demnati and Rao (2004) approach.

We also show that it is asymptotically design-consistent and that it can handle two-stage sampling. For the Horvitz and Thompson (1952) point estimator, the proposed variance estimator reduces to the Horvitz and Thompson (1952) and the Sen (1953); Yates and Grundy (1953) variance estimator.

2.2 The class of point estimators

Let $\mathcal{U} = \{1, \ldots, k, \ell, \ldots, N\}$ denote a finite population and let $s = \{1, \ldots, n\} \subseteq \mathcal{U}$ denote a sample whose elements are randomly selected with an unequal probability sampling design without replacement. We assume full response.

Consider the population parameter

$$\theta = h(t_1, \ldots, t_q, \ldots, t_Q),$$

where $h(\cdot)$ is a smooth and differentiable function (e.g. Shao and Tu, 1995, ch. 2) of population totals t_q, $(q = 1, \ldots, Q)$ of Q survey variables,

$$t_q = \sum_{k \in \mathcal{U}} y_{qk},$$

with y_{qk} denoting the value of the variable q for the unit $k \in \mathcal{U}$. Consider that θ is estimated by its substitution point estimator

$$\hat{\theta} = h(\hat{t}_1, \ldots, \hat{t}_q, \ldots, \hat{t}_Q),$$

where $\hat{t}_q$, is the Horvitz and Thompson (1952) point estimator

$$\hat{t}_q = \sum_{k \in s} w_k \, y_{qk},$$

with survey weights $w_k = 1/\pi_k$; where $\pi_k > 0$ is the inclusion probability of the unit k.

2.3 The proposed replicate variance estimator

We propose to estimate the variance of $\hat{\theta}$ by

$$\widehat{\text{var}}(\hat{\theta})^{HT}_{prop} \;=\; \sum_{k \in s} \sum_{\ell \in s} \mathcal{D}_{k\ell} \; w_k \nu_k \; w_\ell \nu_\ell, \tag{2.1}$$

where

$$\mathcal{D}_{k\ell} \;=\; \frac{\pi_{k\ell} - \pi_k \pi_\ell}{\pi_{k\ell}},$$

with $\pi_{k\ell} > 0$ denoting the joint inclusion probability of the units k and ℓ, and where

$$\nu_k = \frac{\hat{\theta} - \hat{\theta}^*_k}{\varrho_k}, \tag{2.2}$$

with

$$\varrho_k = w_k^{1-\alpha_k},$$

for some $\alpha_k \geq 0$ (see Subsection 2.3.1), where $\hat{\theta}^*_k$ has the same functional form as $\hat{\theta}$ but using $\hat{t}^*_{qk}$ instead of $\hat{t}_q$, i.e.

$$\hat{\theta}^*_k = h(\hat{t}^*_{1k}, \ldots, \hat{t}^*_{qk}, \ldots, \hat{t}^*_{Qk}),$$

with

$$\hat{t}^*_{qk} = \sum_{\ell \in s} w^*_{\ell(k)}\, y_{q\ell}, \tag{2.3}$$

where

$$w^*_{\ell(k)} = \begin{cases} w_\ell & \text{if } \ell \neq k, \\ w_k - \varrho_k & \text{if } \ell = k. \end{cases}$$

Alternatively, with fixed sample size, we propose to use the estimator,

$$\widehat{\text{var}}(\hat{\theta})^{SYG}_{prop} = \frac{-1}{2} \sum_{k \in s} \sum_{\ell \in s} \mathcal{D}_{k\ell}\, (w_k \nu_k - w_\ell \nu_\ell)^2, \tag{2.4}$$

which is positive provided that the Sen (1953); Yates and Grundy (1953) condition, $\mathcal{D}_{k\ell} < 0$, holds.

2.3.1 The value of α_k

We originally developed the proposed replication variance estimators $\widehat{\text{var}}(\hat{\theta})^{HT}_{prop}$ and $\widehat{\text{var}}(\hat{\theta})^{SYG}_{prop}$ for $\alpha_k = 1$ (Escobar and Berger, 2011). However, to avoid restricting α_k, we explore its range of values. In Section 2.5, we show that Equations (2.1) and (2.4) are valid for any $\alpha_k \geq 0$. We recommend to use $\alpha_k = 1$ or $\alpha_k > 1$ (see below comments and Subsections 2.3.3, 2.4.2, 2.4.3 and Section 2.7).

Using $\alpha_k = 0$ gives $\varrho_k = w_k$ which corresponds to a naïve jackknife which deletes the unit k, i.e. using $w^*_{\ell(k)} = 0$ if $\ell = k$. In Section 2.7 we show that this case produces biased and unstable estimates.

Using $\alpha_k = 1$ gives $\varrho_k = 1$. This implies that Equations (2.2) and (2.3) reduce respectively to $\nu_k = \hat{\theta} - \hat{\theta}^*_k$ and $\hat{t}^*_{qk} = \hat{t}_q - y_{qk}$, obtaining the Escobar and Berger

(2011) jackknife. In this case, note that α_k (and therefore ϱ_k) is a constant free of k.

Using $\alpha_k > 1$ results in approximating the linearisation variance estimators obtained by the Demnati and Rao (2004) approach. The larger the value of α_k, the closer the approximation (Subsections 2.3.3, 2.4.3 and Section 2.7). This feature can be used when the involved derivatives in linearisation become extremely cumbersome, e.g. when $h(\cdot)$ is an implicit function (Shao and Tu, 1995, p. 29).

2.3.2 Example of a total (underlying idea)

Consider the case of estimating a total,

$$\hat{\theta} = \hat{t} = \sum_{k \in s} w_k y_k.$$

The Equations (2.2) and (2.3) imply that,

$$\begin{aligned}\nu_k &= \frac{\hat{\theta} - \hat{\theta}^*_k}{\varrho_k} \\ &= \frac{\hat{t} - (\hat{t} - \varrho_k\, y_k\,)}{\varrho_k} \\ &= \frac{\varrho_k\, y_k}{\varrho_k} \\ &= y_k.\end{aligned}$$

Hence, (Eq. 2.1) and (Eq. 2.4) reduce respectively to the Horvitz and Thompson (1952), and the Sen (1953); Yates and Grundy (1953) unbiased estimators of $\mathrm{var}(\hat{t})$. Note that this is true for any value of α_k.

2.3.2.1 Intuitive underlying idea

As suggested by a referee, we give an intuitive explanation. Consider the case $\alpha_k = 1$. Let $\mho$ be the artificial population obtained from expanding the $y_k \in s$ by their w_k's, i.e. the expanded sample. Accordingly, we are omitting y_k from $\mho$ and estimating θ via $\hat{\theta}^*_k = \hat{t} - y_k$ with a $\mathrm{Bias}(\hat{\theta}^*_k) = y_k$. Hence, $\nu_k = \hat{\theta} - \hat{\theta}^*_k = y_k$.

Comparing with the customary Jackknife (Quenouille, 1956; Tukey, 1958) we have that: (i) The proposed replication removes units from $\mho$ instead of from s; (ii) The customary assumes the pseudo-values are unbiased and approx. i.i.d., whereas in the proposed the ν_k's are neither unbiased nor i.i.d. If $\alpha_k > 1$, the proposed deletes bits of units, i.e. it perturbs the w_k's by ϱ_k. The proposed variance estimator can thus be seen as a *post-expansion* or as a *delete-weight* Jackknife.

From a Bootstrap (Efron, 1979) perspective, the proposed estimator can also be seen as a Bootstrap which deterministically subsamples n different subsets of size $\#(\mho) - \varrho_k$ from $\mho$, instead of randomly subsampling (say) L resamples of size n from $\mho$. Note that there are at most n different pairs $w_k y_k$ in $\mho$.

As it has been mentioned, and as it will be shown, the proposed replication estimator can also be seen as an approximation to linearisation estimators.

2.3.3 Example of a ratio

Let $R = t_y/t_x$, be the parameter of interest, where $t_y = \sum_{k\in\mathcal{U}} y_k$ and $t_x = \sum_{k\in\mathcal{U}} x_k$ are the population totals of the variables y and x. Assume that R is estimated with the point estimator

$$\hat{R} \;=\; \frac{\hat{t}_y}{\hat{t}_x}.$$

A linearisation variance estimator of $\hat{R}$ is given by (e.g. Demnati and Rao, 2004, example 2.1),

$$\widehat{\mathrm{var}}(\hat{R})_{Lin} \;=\; \sum_{k\in s}\sum_{\ell\in s} \mathcal{D}_{k\ell}\; w_k u_k\; w_\ell u_\ell, \tag{2.5}$$

with

$$u_k \;=\; \frac{y_k \;-\; \hat{R}\, x_k}{\hat{t}_x}. \tag{2.6}$$

In this case, the ν_k values in Equation (2.2) are given by

$$\nu_k \;=\; u_k \left(\frac{\hat{t}_x}{\hat{t}_x \;-\; \varrho_k\, x_k} \right). \tag{2.7}$$

From (Eq. 2.6) and (Eq. 2.7) we see that

$$\nu_k \;\doteq\; u_k,$$

if,

$$\frac{\hat{t}_x}{\hat{t}_x \;-\; \varrho_k x_k} \;\doteq\; 1.$$

Thus, for large α_k, i.e. small ϱ_k, we have that (Eq. 2.1) is a suitable approximation of (Eq. 2.5) and its sensitivity to highly-skewed weights should be low (see Section 2.7).

2.4 Alternative estimators for the variance

2.4.1 Generalised jackknife for functions of means

The Campbell (1980); Berger and Skinner (2005) generalised jackknife is defined by,

$$\widehat{\mathrm{var}}(\hat{\theta})^{CBS}_{Jack} \;=\; \sum_{k\in s}\sum_{\ell\in s} \mathcal{D}_{k\ell}\; w_k\varepsilon_k\; w_\ell\varepsilon_\ell, \tag{2.8}$$

with

$$\varepsilon_k \;=\; \left(\frac{1}{w_k} - \frac{1}{\hat{N}}\right)(\tilde{\theta} - \tilde{\theta}^{(k)}), \tag{2.9}$$

where

$$\tilde{\theta} \;=\; g(\tilde{\mu}_1,\, \ldots,\, \tilde{\mu}_q,\, \ldots,\, \tilde{\mu}_Q),$$

is a function of Hájek (1971) mean estimators of Q variables with

$$\begin{aligned}\tilde{\mu}_q &= \frac{\hat{t}_q}{\hat{N}},\\ \hat{N} &= \sum_{k\in s} w_k,\end{aligned}$$

and where

$$\tilde{\theta}^{(k)} \;=\; g(\tilde{\mu}^{(k)}_1, \ldots, \tilde{\mu}^{(k)}_q, \ldots, \tilde{\mu}^{(k)}_Q)$$

has the same functional form as $\tilde{\theta}$ but replacing $\tilde{\mu}_q$ by

$$\tilde{\mu}^{(k)}_q \;=\; \frac{\hat{t}_q \;-\; w_k y_k}{\hat{N} \;-\; w_k}.$$

Note that the generalised jackknife (Eq. 2.8) is designed for functions of means whereas the proposed estimator (Eq. 2.1) is designed for functions of totals. Thus, the proposed estimator is more general.

2.4.1.1 Example of a ratio (revisited)

When $\hat{\theta} = \hat{R}$, we have that $\tilde{\theta} = \hat{\theta}$ and that the ε_k's from (Eq. 2.9) are given by

$$\varepsilon_k \; = \; u_k \left(\frac{\hat{t}_x}{\hat{t}_x \, - \, w_k x_k} \right) \left(\frac{\hat{N} \, - \, w_k}{\hat{N}} \right) . \tag{2.10}$$

From Equations (2.6), (2.7) and (2.10) we see that

$$\varepsilon_k \; \doteq \; u_k,$$

if,

$$\frac{\hat{N} \, - \, w_k}{\hat{N}} \; \doteq \; 1,$$

and if,

$$\frac{\hat{t}_x}{\hat{t}_x \, - \, w_k \, x_k} \; \doteq \; 1.$$

Note that (Eq. 2.7) is a better approximation of (Eq. 2.6) than (Eq. 2.10). Hence, the proposed estimator (Eq. 2.1) should be as precise as (Eq. 2.5).

As suggested by a referee, we show the above example for Poisson sampling. It can be shown that the Equations (2.5), (2.8) and (2.1) reduce respectively to

$$\widehat{\text{var}}(\hat{R})_{PoiLin} \;=\; \sum_{k \in s} \frac{w_k - 1}{w_k} \, (w_k \, u_k)^2, \tag{2.11}$$

$$\widehat{\text{var}}(\hat{R})^{CBS}_{PoiJack} \;=\; \sum_{k \in s} \frac{w_k - 1}{w_k} \left(\frac{\hat{t}_x}{\hat{t}_x \, - \, w_k \, x_k} \right)^2 \left(\frac{\hat{N} - w_k}{\hat{N}} \right)^2 (w_k \, u_k)^2, \tag{2.12}$$

$$\widehat{\text{var}}(\hat{R})^{HT}_{PoiProp} \;=\; \sum_{k \in s} \frac{w_k - 1}{w_k} \left(\frac{\hat{t}_x}{\hat{t}_x \, - \, \varrho_k \, x_k} \right)^2 (w_k \, u_k)^2, \tag{2.13}$$

with u_k as defined in (Eq. 2.6). Thus, we can see that (Eq. 2.13) is a better approximation to (Eq. 2.11) than (Eq. 2.12) for uniformly negligible ϱ_k (large α_k).

2.4.2 Linearisation based on the Gâteaux derivative

Let $\theta = T(M)$ be a functional where M is measure which allocates the unit mass to $k \in \mathcal{U}$ and let $\hat{\theta}$ be a functional $T(\hat{M})$, where $\hat{M}$ denotes a sample-based measure that allocates the mass w_k to the element $k \in s$, and let δ_{y_k} be the degenerate point mass at y_k. The *empirical* influence function of $T(\cdot)$ (e.g. Davison and Hinkley, 1997, sec. 2.7) in $\hat{M}$ (if it exists) is defined as

$$\mathrm{EIT}(\hat{M}, y_k) = \lim_{\zeta \to 0} \frac{T[\hat{M} + \zeta \delta_{y_k}] - T(\hat{M})}{\zeta},$$

i.e. the Gâteaux derivative of $T(\cdot)$ over the random measure $\hat{M}$. A variance estimator is given by the Equation (2.1) after substituting ν_k by $\mathrm{EIT}(\hat{M}, y_k)$.

The approach proposed by Deville (1999, p. 197) estimates the population influence function $\mathrm{IT}(M, y_k)$ by $\mathrm{IT}(\hat{M}, y_k)$, i.e. the influence function of $T(\cdot)$ in M evaluated at $M = \hat{M}$. Note that $\mathrm{IT}(\hat{M}, y_k)$ can be different from $\mathrm{EIT}(\hat{M}, y_k)$ which is the *empirical* influence function of $T(\cdot)$ in $\hat{M}$. Also note that, for the ratio (see Deville, 1999, p. 198), that approach does not always give the linearisation estimator (Eq. 2.5) with u_k as in Equation (2.6).

Using Equation (2.2), we have that

$$\nu_k = \frac{T[\hat{M} + \zeta_k \delta_{y_k}] - T(\hat{M})}{\zeta_k},$$

with

$$\zeta_k = -\varrho_k.$$

Thus, ν_k can be interpreted as an approximation of $\mathrm{EIT}(\hat{M}, y_k)$. Again, large α_k (small ϱ_k), ensures this approximation is better as ζ_k is close to zero.

Note that, for example, if $\alpha_k = 1$ then ϱ_k is a constant free of k.

An overview of the linearisation using Gâteaux differentiation can be found in Goga *et al.* (2009).

2.4.3 Demnati and Rao (2004) linearisation approach

Let us redefine the point estimator as $\hat{\theta} = g(d_1, d_2, \ldots, d_N)$ where $d_k = w_k$ if $k \in s$ and $d_k = 0$ otherwise. Demnati and Rao (2004) propose to estimate the variance using Equation (2.1) after substituting ν_k by

$$\begin{aligned} z_k &= \left. \frac{\partial g(a_1, \ldots, a_N)}{\partial a_k} \right|_{(a_k = d_k)} \\ &= \lim_{\zeta \to 0} \frac{g(d_1, \ldots, d_k + \zeta, \ldots, d_n) - g(d_1, \ldots, d_n)}{\zeta}, \end{aligned} \tag{2.14}$$

for constants $a_1, \ldots, a_N$. Now, by using Equation (2.2), we have that

$$\nu_k \;=\; \frac{g(d_1, \ldots, d_k + \zeta_k, \ldots, d_n) - g(d_1, \ldots, d_n)}{\zeta_k},$$

with

$$\zeta_k \;=\; -\,\varrho_k.$$

Thus, ν_k can be interpreted as an approximation of z_k. Again, large α_k's (small ϱ_k's) ensure that ζ_k is small.

Note again that the α_k can be defined such that ϱ_k is a constant free of k.

2.5 Design-consistency

We now set the validity of the proposed variance estimators (Eq. 2.1) and (Eq. 2.4) under the Isaki and Fuller (1982) asymptotic framework. Consider a sequence of nested populations of increasing sizes $\{N_t : 0 < N_t < N_{t+1}, \forall t\}$. Consider also a sequence of non-necessarily nested samples of increasing sizes $\{n_t : n_t < n_{t+1}; n_t < N_t, \forall t\}$. Thus, if $t \to \infty$, it implies that $N_t \to \infty$ and $n_t \to \infty$, with $f = n_t/N_t$ a constant free of the limiting process. In what follows, we drop the index t to simplify the notation.

Consider the Horvitz and Thompson (1952) estimator

$$\hat{\mu}_q \;=\; \sum_{k \in s} \bar{w}_k y_{qk},$$

of the population mean

$$\mu_q \;=\; \frac{1}{N} t_q,$$

for $q = 1, \ldots, Q$ where

$$\bar{w}_k \;=\; \frac{w_k}{N}.$$

Thus, for the vector of means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_Q)^T$ and the vector of estimators $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_Q)^T$, the multivariate Horvitz and Thompson (1952) and Sen (1953);

Yates and Grundy (1953) design variances and variance estimators of $\hat{\boldsymbol{\mu}}$ are

$$\begin{aligned}
\mathbf{var}(\hat{\boldsymbol{\mu}})_{HT} &= \sum_{k\in\mathcal{U}}\sum_{\ell\in\mathcal{U}} \mathcal{D}_{k\ell}\ \pi_{k\ell}\ \bar{w}_k\bar{w}_\ell\ \boldsymbol{y}_k\boldsymbol{y}_\ell^T, \\
\widehat{\mathbf{var}}(\hat{\boldsymbol{\mu}})_{HT} &= \sum_{k\in s}\sum_{\ell\in s} \mathcal{D}_{k\ell}\ \bar{w}_k\bar{w}_\ell\ \boldsymbol{y}_k\boldsymbol{y}_\ell^T, \qquad (2.15) \\
\mathbf{var}(\hat{\boldsymbol{\mu}})_{SYG} &= \frac{-1}{2}\sum_{k\in\mathcal{U}}\sum_{\ell\in\mathcal{U}} \mathcal{D}_{k\ell}\ \pi_{k\ell}\ \{\bar{w}_k\boldsymbol{y}_k - \bar{w}_\ell\boldsymbol{y}_\ell\}\{\bar{w}_k\boldsymbol{y}_k - \bar{w}_\ell\boldsymbol{y}_\ell\}^T, \\
\widehat{\mathbf{var}}(\hat{\boldsymbol{\mu}})_{SYG} &= \frac{-1}{2}\sum_{k\in s}\sum_{\ell\in s} \mathcal{D}_{k\ell}\ \{\bar{w}_k\boldsymbol{y}_k - \bar{w}_\ell\boldsymbol{y}_\ell\}\{\bar{w}_k\boldsymbol{y}_k - \bar{w}_\ell\boldsymbol{y}_\ell\}^T, \qquad (2.16)
\end{aligned}$$

with $\boldsymbol{y}_k = (y_{1k}, \ldots, y_{Qk})^T$.

Now, assume the following regularity conditions:

(a) $\widehat{\text{var}}(\hat{\theta})_L/\text{var}(\hat{\theta})_L \to_p 1$, $\text{var}(\hat{\theta})_L \neq 0$ where,

$$\begin{aligned}
\text{var}(\hat{\theta})_L &= \boldsymbol{\nabla}(\boldsymbol{\mu})^T\mathbf{var}(\hat{\boldsymbol{\mu}})_{HT}\boldsymbol{\nabla}(\boldsymbol{\mu}), \\
\widehat{\text{var}}(\hat{\theta})_L &= \boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})^T\widehat{\mathbf{var}}(\hat{\boldsymbol{\mu}})_{HT}\boldsymbol{\nabla}(\hat{\boldsymbol{\mu}}). \qquad (2.17)
\end{aligned}$$

Alternatively for fixed sample-size designs,

$$\begin{aligned}
\text{var}(\hat{\theta})_L &= \boldsymbol{\nabla}(\boldsymbol{\mu})^T\mathbf{var}(\hat{\boldsymbol{\mu}})_{SYG}\boldsymbol{\nabla}(\boldsymbol{\mu}), \\
\widehat{\text{var}}(\hat{\theta})_L &= \boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})^T\widehat{\mathbf{var}}(\hat{\boldsymbol{\mu}})_{SYG}\boldsymbol{\nabla}(\hat{\boldsymbol{\mu}}), \qquad (2.18)
\end{aligned}$$

where $\boldsymbol{\nabla}(\boldsymbol{x}) = (\partial h(\boldsymbol{\mu})/\partial\mu_1, \ldots, \partial h(\boldsymbol{\mu})/\partial\mu_Q)^T_{\boldsymbol{\mu}=\boldsymbol{x}}$ is the gradient of $h(\cdot)$ at $\boldsymbol{x} \in \Re^Q$ with $h(\cdot)$ continuous and differentiable at $\boldsymbol{\mu}$.

(b) $\lim\inf \{n\ \text{var}(\hat{\theta})_L\} > 0$.

(c) $n^{-1}\sum_{k\in s} \bar{w}_k{}^\tau \bar{\varrho}_k{}^\gamma ||\boldsymbol{y}_k||^{\tau+\gamma} = \mathcal{O}_p(n^{-(\tau+\gamma)})$, $\forall\tau \geq 2$, $\forall\gamma \geq 0$, where $\bar{\varrho}_k = \bar{w}_k^{1-\alpha_k}$, $\alpha_k \geq 0$, with $||\boldsymbol{A}|| = \text{tr}(\boldsymbol{A}^T\boldsymbol{A})^{1/2}$ the Euclidean norm.

(d) $G_s = n^{-\beta}\sum\sum_{(k\neq\ell)\in s}(\mathcal{D}_{k\ell}^-)^2 = \mathcal{O}_p(1)$, with $0 \leq \beta < 1$, where $\mathcal{D}_{k\ell}^- = -\mathcal{D}_{k\ell}$ if $\mathcal{D}_{k\ell} < 0$ and 0 otherwise.

(e) $H_s = n^{-\beta}\sum\sum_{(k\neq\ell)\in s}(\mathcal{D}_{k\ell}^+)^2 = \mathcal{O}_p(1)$, with $0 \leq \beta < 1$, where $\mathcal{D}_{k\ell}^+ = \mathcal{D}_{k\ell}$ if $\mathcal{D}_{k\ell} \geq 0$ and 0 otherwise.

(f) $\boldsymbol{\nabla}(\boldsymbol{x})$ is Lipschitz (Hölder) continuous of order δ, i.e. $\|\boldsymbol{\nabla}(\boldsymbol{x}_1) - \boldsymbol{\nabla}(\boldsymbol{x}_2)\| \leq \lambda\,\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^\delta$, $\lambda > 0$ and $\delta > 0$ constants, $0 \leq \beta/2 < \delta$, for $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ in the neighbourhood of $\boldsymbol{\mu}$.

(g) $||\boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})|| = \mathcal{O}_p(1)$.

Conditions (a), (c), (d), (e) and (f) are similar but different from those proposed in Berger and Skinner (2005). Condition (a) sets the consistency of the linearisation variance estimator (e.g. Robinson and Särndal, 1983; Särndal *et al.*, 1992, secs. 5.5, 5.7). Conditions (b) and (c) are usual conditions in asymptotic studies (e.g. Shao and Tu, 1995, subsec. 6.4.1): Condition (b) implies $\text{var}(\hat{\theta})_L$ decreases with rate n^{-1}, and the Condition (c) is a Lyapunov-type condition for the existence of moments; Conditions (d) and (e) are mild conditions on the design similar to ones in Isaki and Fuller (1982). These conditions allow two-stage sampling (e.g. Escobar and Berger, 2013a); Conditions (f) and (g) are smoothness and differentiability requirements.

Theorem 2.1. *Theorem 1. Under unequal probability sampling with fixed sample size, the regularity conditions (a)–(g) imply that*

$$\frac{\widehat{var}(\hat{\theta})^{SYG}_{prop}}{var(\hat{\theta})_L} \rightarrow_p 1.$$

A proof of Theorem 1 is given in the Appendix 2.A for this chapter.

Corollary 2.2. *Corollary 1. Under unequal probability sampling, if the regularity conditions (a)–(g) hold, then*

$$\frac{\widehat{var}(\hat{\theta})^{HT}_{prop}}{var(\hat{\theta})_L} \rightarrow_p 1.$$

The proof of Corollary 1 is also given in the Appendix 2.A for this chapter. Hence, from the Theorem 1, Corollary 1 and the Slutsky's theorem (e.g. Valliant *et al.*, 2000, p. 414), it follows that:

$$\frac{\hat{\theta} - \theta}{\sqrt{\widehat{\text{var}}(\hat{\theta})^{SYG}_{prop}}} \rightarrow_d \text{N}(0,1),$$

and

$$\frac{\hat{\theta} - \theta}{\sqrt{\widehat{\text{var}}(\hat{\theta})^{HT}_{prop}}} \rightarrow_d \text{N}(0,1),$$

when $\hat{\theta}$ is asymptotically Normal. Thus, allowing valid confidence intervals of $\hat{\theta}$ for θ.

2.6 Two-stage sampling

The proposed variance estimator can be used for two-stage sampling. For example, consider a self-weighted two-stage sampling design where n_I clusters are selected with inclusion probabilities π_{Ii} proportional to their sizes M_i, $(i = 1, \ldots, N_I)$. Within each selected cluster a simple random sample of m elements is drawn. Hence, the clusters' inclusion probabilities are

$$\pi_{Ii} \;=\; n_I \, \frac{M_i}{N},$$

and the elements' inclusion probabilities are

$$\pi_k \;=\; \frac{n}{N} \;=\; f.$$

By using the Hájek (1964, Eq. 5.27, p. 1511) approximation and by denoting $q_{Ii} = 1 - \pi_{Ii}$, the clusters' joint inclusion probabilities π_{Iij} are approximated by

$$\pi_{Iij} \;\doteq\; \pi_{Ii}\, \pi_{Ij} \left\{ 1 - \frac{q_{Ii} q_{Ij}}{d} \right\}, \quad (i \neq j = 1, \ldots, N_I), \tag{2.19}$$

where

$$d \;=\; \sum_{i \in \mathcal{U}} \pi_{Ii}\, q_{Ii}.$$

This approximation was originally developed for $d \to \infty$, i.e. for $N_I \to \infty$ with a fixed m, under the maximum entropy sampling design (see Hájek, 1981, ch. 3 & 6); namely the Rejective Sampling. It requires that the used sampling design (for clusters) is of large entropy. Low entropy sampling designs (e.g. systematic probability proportional-to-size design) are not suitable for the above approximation. However, the randomised systematic sampling is suitable as it is of large entropy. See Berger and Tillé (2009) for an overview. Berger (2011) gives sufficient conditions under which Hájek's results still hold for large entropy designs that are not the maximum entropy one.

Hence, given that the i-th cluster is selected, the elements' conditional inclusion probabilities are

$$\pi_{k|Ii} \;=\; \frac{m}{M_i},$$

and

$$\pi_{k\ell|Ii} \;=\; \frac{m\,(m-1)}{M_i\,(M_i - 1)}.$$

Using (Eq. 2.19), the elements' joint inclusion probabilities $\pi_{k\ell}$ are given by

$$\pi_{k\ell} \doteq \begin{cases} \pi_{Ii}\,\pi_{k|Ii} \;=\; f & \text{if } (k=\ell) \in s_i, \\ \pi_{Ii}\,\pi_{k\ell|Ii} \;=\; f(m-1)/(M_i-1) & \text{if } (k \neq \ell) \in s_i, \\ \pi_{Iij}\,\pi_{k|Ii}\,\pi_{\ell|Ij} \;\doteq\; f^2\{1-d^{-1}q_{Ii}q_{Ij}\} & \text{if } k \in s_i, \ell \in s_j, i \neq j, \end{cases}$$

where s_i denotes the sample of the i-th cluster. Substituting $\pi_{k\ell}$ in $\mathcal{D}_{k\ell}$, we obtain

$$\mathcal{D}_{k\ell} \;\doteq\; \begin{cases} 1-f & \text{if } (k=\ell) \in s_i, \\ 1-\pi^*_{Ii} & \text{if } (k \neq \ell) \in s_i, \\ q_{Ii}\,q_{Ij}/(q_{Ii}\,q_{Ij}-d) & \text{if } k \in s_i, \ell \in s_j, i \neq j. \end{cases} \tag{2.20}$$

where

$$\pi^*_{Ii} \;=\; \pi_{Ii}\,\frac{m}{m-1}\,\frac{M_i-1}{M_i}.$$

Thus, the proposed estimator is given by the Equations (2.1) or (2.4) with $\mathcal{D}_{k\ell}$ substituted by (Eq. 2.20).

2.6.1 Design-consistency for self-weighted two-stage sampling

The proposed variance estimators (Eq. 2.1) and (Eq. 2.4) are consistent under self-weighted two-stage sampling when the regularity conditions of Section 2.5 hold. Hence, assuming that the customary Conditions (a), (b), (c), (f) and (g) hold, it is only necessary to show that the Conditions (d) and (e) hold.

Let

$$\beta \;=\; \frac{\log(n_I)}{\log(n)} < 1,$$

such that $n^\beta = n_I$ and let $f_I = n_I/N_I$. It can be shown that $|1-\pi^*_{Ii}| = \mathcal{O}(1)$ for $M_i \geq m \geq 2, \forall i = 1,\ldots,N_I$ and that $d = \mathcal{O}(N_I)$ as $q_{Ii} = \mathcal{O}(1)$.

If $1-\pi^*_{Ii} > 0$, we have from Conditions (d) and (e), that

$$G_s \;=\; f_I^2 m^2 \mathcal{O}_p(n_I^{-1}),$$

and

$$H_s \;=\; m(m-1)\mathcal{O}_p(1).$$

If $1 - \pi^*_{Ii} < 0$, we have,

$$G_s \;=\; f_I^2 m^2 \mathcal{O}_p(n_I^{-1}) \;+\; m(m-1)\mathcal{O}_p(1),$$

and

$$H_s \;=\; 0.$$

Thus, G_s and H_s are $\mathcal{O}_p(1)$ when f_I and m are $\mathcal{O}(1)$. Note that this is also true if

$$\hat{d} \;=\; \sum_{i \in s} (1 - \pi_{Ii}),$$

is used instead of d, since $\hat{d} = \mathcal{O}_p(n_I)$. Thus, the Conditions (d) and (e) hold.

2.7 Simulation study

Consider the sugar cane farms dataset (Chambers and Dunstan, 1986) is a population frame of size $N = 338$. The variables of interest are:

- *Gross value of cane* (y_{1k}),
- *Total farm expenditure* (y_{2k}).

The parameter of interest is the ratio

$$R \;=\; \frac{t_1}{t_2},$$

with true value $R = 1.58$, which is estimated by the point estimator

$$\hat{R} \;=\; \frac{\hat{t}_1}{\hat{t}_2}.$$

For selecting the samples and for computing the joint inclusion probabilities we use the Midzuno (1951) method.

We consider four scenarios where the inclusion probabilities π_k are proportional to the variables:

- *Total cane harvested* (x_k), to obtain π_k's correlated to y_{1k} and y_{2k},

- *Square root of the total cane harvested* ($\sqrt{x_k}$), to obtain π_k's mildly correlated to y_{1k} and y_{2k} through a non-linear relationship,
- *Variable with ones.* The π_k's are equal in this case, $\pi_k = n/N$,
- *Generated variable* (ψ_k^{-1}) where $\psi_k \sim$ Log-Normal$(0, 1/2)$, to use independent and randomly highly-skewed sampling weights w_k's.

For each simulation, 1 000 000 samples were selected to compute: the empirical relative bias

$$\mathrm{RB} \;=\; \frac{\mathrm{B}(\widehat{\mathrm{var}}(\hat{R}))}{\mathrm{var}(\hat{R})},$$

where

$$\mathrm{B}(\widehat{\mathrm{var}}(\hat{R})) \;=\; \mathrm{E}(\widehat{\mathrm{var}}(\hat{R})) - \mathrm{var}(\hat{R}),$$

and the empirical relative root mean square error

$$\mathrm{RRMSE} \;=\; \frac{\sqrt{\mathrm{MSE}(\widehat{\mathrm{var}}(\hat{R}))}}{\mathrm{var}(\hat{R})}.$$

The term $\mathrm{var}(\hat{R})$ is the empirical variance computed from the 1 000 000 observed values of $\hat{R}$. The used variance estimators are:

- The Quenouille (1956); Tukey (1958) standard jackknife,

$$\widehat{\mathrm{var}}(\hat{\theta})_{STD} \;=\; \left(1 - \frac{n}{N}\right)\frac{n-1}{n}\sum_{k\in s}(\hat{\theta}_{(k)} - \hat{\theta}_{(\cdot)})^2, \tag{2.21}$$

  with an *ad hoc* finite population correction, where $\hat{\theta}_{(k)}$ has the same functional form as $\hat{\theta}$ but using $\hat{t}_{q(k)}$ instead of $\hat{t}_q$, i.e.

$$\hat{\theta}_{(k)} \;=\; h(\hat{t}_{1(k)}, \ldots, \hat{t}_{Q(k)}),$$

  with

$$\hat{t}_{q(k)} \;=\; \sum_{\ell \neq k \in s} w_\ell \, y_{q\ell},$$

  and

$$\hat{\theta}_{(\cdot)} \;=\; \frac{1}{n}\sum_{k\in s}\hat{\theta}_{(k)},$$

- The Campbell (1980); Berger and Skinner (2005) generalised jackknife from the Equation (2.8),

- The proposed variance estimator (Eq. 2.1) with $\alpha_k = b_k$ and $\alpha_k = 0,\ 1,\ 2$ where
$$b_k \ = \ 1 + \frac{\log(n)}{\log(w_k \ + \ 1/n)},$$
i.e. $\varrho_k \doteqdot n^{-1}$ and $\varrho_k = w_k, 1, w_k^{-1}$, respectively.
- The linearisation variance estimator from Equation (2.5).

In Table 2.1, we see that the standard jackknife (Eq. 2.21) has increasing RB, in absolute value, for increasing n under the unequal probability scenarios $\pi_k \propto x_k, \sqrt{x_k}$ and ψ_k^{-1}, and it has a decreasing RB for increasing n with $\pi_k = n/N$. The proposed estimator (Eq. 2.1) with $\alpha_k = 0$, has the largest but always positive RB which decreases with increasing n.

Further, in all four scenarios, we can observe that for increasing values of α_k, the RB of the proposed (Eq. 2.1) tends to replicate the RB of the linearisation estimator (Eq. 2.5), confirming that (Eq. 2.1) is approximately equal to (Eq. 2.5) when $\alpha_k > 1$. The CBS generalised jackknife (Eq. 2.8) has a decreasing RB for increasing n. Note that the RB of the proposed estimator can be smaller than the RB of (Eq. 2.8) in absolute value. This is more noticeable with $\alpha_k \neq 0$ for the scenarios $\pi_k \propto x_k$ and $\sqrt{x_k}$. The estimator (Eq. 2.8) tends to be less biased than (Eq. 2.5) and (Eq. 2.1) with independent or highly skewed sampling weights. However, the RB of (Eq. 2.8) tends to be greater than (Eq. 2.5) and (Eq. 2.1) with $\alpha_k > 0$ or $\alpha_k = b_k$ in the case where there is a non-linear relationship between the inclusion probabilities and the variables of interest.

Table 2.2 shows that the linearisation estimator (Eq. 2.5) has the smallest RRMSE among all other variance estimators, except for $\pi_k \propto \psi_k^{-1}$ and $f = 0.201, 0.302$ where the standard estimator (Eq. 2.21) has the smallest RRMSE. Again, in all four scenarios, we can observe, for increasing values of α_k, that the RRMSE of the estimator (Eq. 2.1) tends to replicate the RRMSE of the linearisation estimator (Eq. 2.5).

In almost all cases, note that the proposed estimator (Eq. 2.1) with $\alpha_k \neq 0$ has smaller RRMSE than the estimator (Eq. 2.8). Thus, the estimator (Eq. 2.1) is more stable than the estimator (Eq. 2.8). This is more noticeable for small n, when the inclusion probabilities are poorly correlated with the variables of interest, or with highly skewed sampling weights. However, as previously suggested, the estimator (Eq. 2.1) may become unstable if the extreme value $\alpha_k = 0$ is used.

TABLE 2.1: Relative Bias (%) of the ratio point estimator.

		Std.	CBS	Proposed replication				Taylor
		Jack.	Jack.	(Eq. 2.1) with				Lin.
n	$f(\%)$	(Eq. 2.21)	(Eq. 2.8)	$\alpha_k = b_k$	$\alpha_k = 0$	$\alpha_k = 1$	$\alpha_k = 2$	(Eq. 2.5)
$\pi_k \propto x_k$								
2	0.6	18.8	10.9	-2.5	343.3	-2.1	-2.9	-2.9
4	1.2	4.8	4.2	-1.6	84.4	-1.1	-1.8	-1.8
7	2.1	2.7	1.8	-1.0	38.7	-0.3	-1.1	-1.1
10	3.0	2.1	1.1	-0.8	25.0	-0.1	-0.8	-0.9
17	5.0	2.1	0.5	-0.5	13.7	0.3	-0.5	-0.5
34	10.1	3.1	-0.2	-0.6	6.2	0.1	-0.5	-0.6
68	20.1	7.8	0.1	0.0	3.3	0.7	0.2	0.0
102	30.2	14.2	-0.1	0.0	2.2	0.7	0.2	0.0
$\pi_k \propto \sqrt{x_k}$								
2	0.6	15.0	12.3	-8.2	326.3	-7.9	-8.5	-8.5
4	1.2	5.9	11.7	-5.0	86.6	-4.5	-5.2	-5.2
7	2.1	3.2	6.7	-3.0	40.0	-2.3	-3.1	-3.2
10	3.0	2.5	4.6	-2.1	25.9	-1.4	-2.2	-2.2
17	5.0	2.0	2.5	-1.3	14.1	-0.5	-1.3	-1.4
34	10.1	2.6	1.1	-0.7	6.6	0.1	-0.7	-0.8
68	20.1	5.5	0.4	-0.5	3.1	0.4	-0.3	-0.5
102	30.2	10.0	0.3	-0.2	2.1	0.6	0.1	-0.2
$\pi_k = n/N$								
2	0.6	8.5	-3.6	-24.1	285.6	-23.8	-24.3	-24.3
4	1.2	6.8	5.1	-16.8	86.8	-16.4	-16.9	-17.0
7	2.1	5.1	4.7	-11.1	42.6	-10.5	-11.2	-11.2
10	3.0	4.0	3.9	-8.1	28.3	-7.4	-8.2	-8.2
17	5.0	2.5	2.4	-5.1	15.7	-4.3	-5.1	-5.2
34	10.1	1.0	1.0	-2.9	7.2	-2.0	-2.9	-3.0
68	20.1	0.9	0.9	-1.1	4.0	-0.1	-0.9	-1.1
102	30.2	0.2	0.2	-1.1	2.2	-0.2	-0.8	-1.1
$\pi_k \propto 1/\psi_k$								
2	0.6	7.0	-28.2	-41.3	252.9	-41.1	-41.5	-41.5
4	1.2	10.4	-12.8	-29.1	94.4	-28.6	-29.2	-29.2
7	2.1	6.8	-8.9	-21.6	47.7	-21.0	-21.6	-21.6
10	3.0	5.0	-7.0	-17.2	32.2	-16.6	-17.3	-17.3
17	5.0	3.1	-4.2	-11.5	19.1	-10.8	-11.5	-11.6
34	10.1	-0.3	-2.2	-6.9	9.5	-6.1	-6.8	-6.9
68	20.1	-4.9	-1.4	-4.3	4.6	-3.4	-4.1	-4.3
102	30.2	-9.2	-1.1	-3.2	3.1	-2.3	-3.0	-3.2

Additional graphical representations for all these results are provided in the Appendix 2.B of this chapter.

Table 2.2: Relative Root Mean-Square Error (%) of the ratio point estimator.

		Std.	CBS	Proposed replication				Taylor
		Jack.	Jack.	(Eq. 2.1) with				Lin.
n	$f(\%)$	(Eq. 2.21)	(Eq. 2.8)	$\alpha_k = b_k$	$\alpha_k = 0$	$\alpha_k = 1$	$\alpha_k = 2$	(Eq. 2.5)
$\pi_k \propto x_k$								
2	0.6	202.1	189.5	150.4	814.9	151.0	149.9	149.9
4	1.2	101.7	102.9	88.3	199.2	88.8	88.1	88.1
7	2.1	68.7	68.5	63.2	101.3	63.6	63.1	63.1
10	3.0	55.0	54.7	51.7	72.4	52.1	51.7	51.7
17	5.0	40.4	40.0	38.8	47.5	39.1	38.8	38.8
34	10.1	27.4	26.8	26.4	29.3	26.6	26.4	26.4
68	20.1	20.0	18.1	18.0	19.0	18.1	18.0	18.0
102	30.2	20.3	14.3	14.3	14.8	14.4	14.3	14.3
$\pi_k \propto \sqrt{x_k}$								
2	0.6	197.4	184.8	144.7	774.9	145.2	144.3	144.3
4	1.2	108.9	118.3	89.1	209.7	89.7	89.0	89.0
7	2.1	74.0	78.3	65.9	108.3	66.4	65.8	65.8
10	3.0	59.1	61.3	54.5	77.3	54.9	54.5	54.4
17	5.0	42.9	43.5	40.7	50.0	41.1	40.7	40.7
34	10.1	28.8	28.2	27.4	30.4	27.7	27.4	27.4
68	20.1	19.7	17.9	17.7	18.7	17.9	17.7	17.7
102	30.2	17.6	13.3	13.2	13.7	13.3	13.2	13.2
$\pi_k = n/N$								
2	0.6	186.3	161.3	132.3	705.4	132.6	132.0	131.9
4	1.2	127.4	124.2	90.7	237.2	91.1	90.6	90.5
7	2.1	94.0	93.5	73.3	134.0	73.8	73.2	73.2
10	3.0	76.4	76.2	63.4	98.2	63.9	63.4	63.4
17	5.0	55.4	55.4	49.6	64.4	50.0	49.6	49.5
34	10.1	36.4	36.4	34.6	39.3	34.9	34.6	34.6
68	20.1	23.8	23.8	23.2	24.8	23.4	23.2	23.2
102	30.2	17.9	17.9	17.6	18.3	17.8	17.6	17.6
$\pi_k \propto 1/\psi_k$								
2	0.6	191.5	145.0	126.1	681.5	126.5	125.8	125.8
4	1.2	153.9	112.2	91.5	289.9	91.9	91.4	91.4
7	2.1	129.2	95.0	80.5	187.7	80.9	80.5	80.5
10	3.0	114.1	88.6	77.2	150.6	77.6	77.2	77.2
17	5.0	91.7	78.7	71.4	110.9	71.8	71.4	71.4
34	10.1	63.6	61.7	58.1	73.9	58.5	58.1	58.1
68	20.1	42.2	46.2	44.6	50.7	44.8	44.6	44.6
102	30.2	33.1	39.3	38.3	41.8	38.5	38.3	38.3

2.8 Discussion

We propose a new design-consistent replication variance estimator for any unequal-probability without-replacement sampling design. The proposed replication estimator is approximately equal to the linearisation variance estimators proposed by Demnati and Rao (2004).

As it is suitable for functions of Horvitz and Thompson (1952) totals, the proposed estimator enjoys of broad applicability, being more general than the generalised jackknife (Campbell, 1980; Berger and Skinner, 2005) which is designed for functions of Hájek (1971) means. Our empirical results suggest that, the proposed estimator is more stable than its competitors.

The proposed replicate estimator can be extended in a number of ways. For example, by embedding the Hájek (1964) approximation for the joint inclusion probabilities as in Berger (2007) and as in Escobar and Berger (2013a) for self-weighted two-stage sampling; or by addressing non-response adjustments as in Berger and Rao (2006). Further, another possibility is to address the variance estimation of model parameters (e.g. Demnati and Rao, 2010).

Appendices to Chapter 2

2.A. Proof of Theorem 2.1 (and Corollary 2.2)

We use standard arguments in proving design-consistency (e.g. Miller, 1964; Shao and Tu, 1995, subsecs. 2.1.1, 3.1.5). Hence, from the mean value theorem we have that,

$$\begin{aligned}\hat{\theta} - \hat{\theta}_k^* &= h(\hat{\boldsymbol{\mu}}) - h(\hat{\boldsymbol{\mu}}_k^*) \\ &= \boldsymbol{\nabla}(\boldsymbol{\xi}_k)^T (\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_k^*) \\ &= \boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})^T (\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_k^*) \ + \ r_k^*,\end{aligned}$$

where $\boldsymbol{\xi}_k$ is a point between $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\mu}}_k^*$, and

$$r_k^* \ = \ \{\boldsymbol{\nabla}(\boldsymbol{\xi}_k) - \boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})\}^T (\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_k^*),$$

is the remainder. Now, from the Equation (2.3) it can be shown that

$$\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_k^* \ = \ \bar{\varrho}_k \, \boldsymbol{y}_k, \tag{2.22}$$

where $\bar{\varrho}_k = \bar{w}_k^{1-\alpha_k}$ and $\bar{w}_k = w_k/N$, $\alpha_k \geq 0$. Combining with (Eq. 2.2) implies,

$$\nu_k \ = \ \boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})^T \, \boldsymbol{y}_k \ + \ r_k, \tag{2.23}$$

where

$$\begin{aligned}r_k &= \bar{\varrho}_k^{-1} \, r_k^* \\ &= \{\boldsymbol{\nabla}(\boldsymbol{\xi}_k) - \boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})\}^T \, \boldsymbol{y}_k.\end{aligned}$$

For r_k, the Cauchy inequality implies

$$|r_k| \ \leq \ ||\boldsymbol{\nabla}(\boldsymbol{\xi}_k) - \boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})|| \ ||\boldsymbol{y}_k||. \tag{2.24}$$

As $\boldsymbol{\xi}_k$ is between $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\mu}}_k^*$ we have that

$$||\boldsymbol{\xi}_k - \hat{\boldsymbol{\mu}}|| \ \leq \ ||\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_k^*||.$$

This combined with Condition (f) and (Eq. 2.22) imply, for constants $\lambda > 0$ and $0 \leq \beta/2 < \delta$, that

$$\begin{aligned} ||\boldsymbol{\nabla}(\boldsymbol{\xi}_k) - \boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})|| &\leq \lambda\, ||\boldsymbol{\xi}_k - \hat{\boldsymbol{\mu}}||^{\delta} \\ &\leq \lambda\, ||\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_k^{*}||^{\delta} \\ &\leq \lambda\, \bar{\varrho}_k^{\,\delta}\, ||\boldsymbol{y}_k||^{\delta}. \qquad (2.25) \end{aligned}$$

Now, let denote

$$\begin{aligned} \tilde{r}_k &= \bar{w}_k\, r_k, & (2.26) \\ \tilde{y}_k &= \bar{w}_k\, \boldsymbol{y}_k^T \boldsymbol{\nabla}(\hat{\boldsymbol{\mu}}), & (2.27) \\ \Psi &= n\, \mathrm{var}(\hat{\theta})_L. & (2.28) \end{aligned}$$

By combining Equations (2.24), (2.25) and (2.26) and multiplying both sides by $\bar{w}_k$, we obtain

$$|\tilde{r}_k| \;\leq\; \lambda\; \bar{w}_k\; \bar{\varrho}_k^{\,\delta} ||\boldsymbol{y}_k||^{1+\delta}. \qquad (2.29)$$

Using the Cauchy inequality and Condition (b) on Equations (2.27) and (2.28) give,

$$\begin{aligned} |\tilde{y}_k| &\leq \bar{w}_k\, ||\boldsymbol{y}_k||\; ||\boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})||, & (2.30) \\ \Psi^{-2} &= \mathcal{O}(1). & (2.31) \end{aligned}$$

By substituting (Eq. 2.23) in (Eq. 2.4) we obtain:

$$\widehat{\mathrm{var}}(\hat{\theta})^{SYG}_{prop} \;=\; A \;+\; 2(E \;-\; C) \;+\; D \;-\; B,$$

where

$$\begin{aligned} A &= \boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})^T\; \widehat{\mathbf{var}}(\hat{\boldsymbol{\mu}})_{SYG}\; \boldsymbol{\nabla}(\hat{\boldsymbol{\mu}}), & (2.32) \\ B &= \sum\sum\nolimits_{k,\ell \in s} \mathcal{D}_{k\ell}\; \tilde{r}_k\; \tilde{r}_\ell, & (2.33) \\ C &= \sum\sum\nolimits_{k,\ell \in s} \mathcal{D}_{k\ell}\; \tilde{r}_k\; \tilde{y}_\ell, & (2.34) \\ D &= \sum\sum\nolimits_{k,\ell \in s} \mathcal{D}_{k\ell}\; \tilde{r}_k^2, & (2.35) \\ E &= \sum\sum\nolimits_{k,\ell \in s} \mathcal{D}_{k\ell}\; \tilde{r}_k\; \tilde{y}_k, & (2.36) \end{aligned}$$

with $\widehat{\mathbf{var}}(\hat{\boldsymbol{\mu}})_{SYG}$, $\tilde{r}_k$ and $\tilde{y}_k$ as Equations (2.16), (2.26) and (2.27). We have to show:

$$
\begin{aligned}
A/\mathrm{var}(\hat{\theta})_L &\to_p 1, && (2.37)\\
B/\mathrm{var}(\hat{\theta})_L &\to_p 0, && (2.38)\\
C/\mathrm{var}(\hat{\theta})_L &\to_p 0, && (2.39)\\
D/\mathrm{var}(\hat{\theta})_L &\to_p 0, && (2.40)\\
E/\mathrm{var}(\hat{\theta})_L &\to_p 0. && (2.41)
\end{aligned}
$$

If proving for Corollary 1, note that substituting (Eq. 2.23) in (Eq. 2.1) gives:

$$\widehat{\mathrm{var}}(\hat{\theta})^{HT}_{prop} = A + B + 2C,$$

with A as (Eq. 2.32) but replacing $\widehat{\mathbf{var}}(\hat{\boldsymbol{\mu}})_{SYG}$ by $\widehat{\mathbf{var}}(\hat{\boldsymbol{\mu}})_{HT}$ defined in (Eq. 2.15), and setting Equations (2.35) and (2.36) equal to zero. Thus, it would suffice to show Equations (2.37), (2.38) and (2.39).

Condition (a) implies the Equation (2.37). We now show (Eq. 2.38). From Equation (2.33) and Conditions (d) and (e) we have,

$$
\begin{aligned}
B &= \frac{-1}{2}\sum_{k\in s}\sum_{\ell\in s}\mathcal{D}_{k\ell}\,(\tilde{r}_k-\tilde{r}_\ell)^2 + \frac{1}{2}\sum_{k\in s}\sum_{\ell\in s}\mathcal{D}_{k\ell}\,(\tilde{r}_k^2+\tilde{r}_\ell^2)\\
&\leq \frac{B_1+B_2}{2}, \qquad (2.42)
\end{aligned}
$$

where

$$B_1 = \sum_{k\in s}\sum_{\ell\in s}\mathcal{D}^{-}_{k\ell}\,(\tilde{r}_k-\tilde{r}_\ell)^2,$$

and

$$B_2 = \sum_{k\in s}\sum_{\ell\in s}\mathcal{D}^{+}_{k\ell}\,(\tilde{r}_k^2+\tilde{r}_\ell^2).$$

Now, using the Cauchy inequality on B_1, we have that

$$B_1^2 \leq G_s n^{\beta}\sum\sum_{k,\ell\in s}(\tilde{r}_k-\tilde{r}_\ell)^4,$$

but as

$$\sum\sum_{k,\ell\in s}(\tilde{r}_k-\tilde{r}_\ell)^4 = 2n\sum_{k\in s}(\tilde{r}_k-\bar{r})^4 + 6\left\{\sum_{k\in s}(\tilde{r}_k-\bar{r})^2\right\}^2,$$

with

$$\bar{r} \;=\; \frac{1}{n}\sum_{k\in s}\tilde{r}_k,$$

we have that

$$\begin{aligned} B_1^2 \;&\leq\; G_s\left[2n^{1+\beta}\sum_{k\in s}(\tilde{r}_k-\bar{r})^4 \;+\; 6n^{\beta}\left\{\sum_{k\in s}(\tilde{r}_k-\bar{r})^2\right\}^2\right] \\ &\leq\; G_s(B_3 \;+\; B_4), \end{aligned} \tag{2.43}$$

with

$$B_3 \;=\; 2n^{1+\beta}\sum_{k\in s}\tilde{r}_k^4,$$

and

$$B_4 \;=\; 6n^{\beta}\left(\sum_{k\in s}\tilde{r}_k^2\right)^2.$$

Hence, the Equation (2.43) and the Condition (d) imply that $B_1/\text{var}(\hat{\theta})_L \rightarrow_p 0$, if we show $(B_3+B_4)/\text{var}(\hat{\theta})_L^2 \rightarrow_p 0$. Thus, by using the Equation (2.29), we obtain

$$\frac{B_3+B_4}{\text{var}(\hat{\theta})_L^2} \;\leq\; \frac{\lambda^4 n^{4+\beta}}{\Psi^2}\left[\frac{2}{n}\sum_{k\in s}(\bar{w}_k\bar{\varrho}_k^{\delta}||\boldsymbol{y}_k||^{1+\delta})^4 \;+\; 6\left\{\frac{1}{n}\sum_{k\in s}(\bar{w}_k\bar{\varrho}_k^{\delta}||\boldsymbol{y}_k||^{1+\delta})^2\right\}^2\right].$$

From Condition (f), we have $\beta < 4\delta$. This combined with Condition (c) and (Eq. 2.31) imply

$$\frac{B_3+B_4}{\text{var}(\hat{\theta})_L^2} \;=\; n^{4+\beta}\,\mathcal{O}_p(n^{-4(1+\delta)}) \;+\; n^{4+\beta}\,\mathcal{O}_p(n^{-2(1+\delta)})^2,$$

that is,

$$\frac{B_3+B_4}{\text{var}(\hat{\theta})_L^2} \;\rightarrow_p\; 0. \tag{2.44}$$

Thus, the Condition (d) and Equations (2.43) and (2.44) all together imply that

$$\frac{B_1}{\text{var}(\hat{\theta})_L} \;\rightarrow_p\; 0. \tag{2.45}$$

We now show that $B_2/\text{var}(\hat{\theta})_L \rightarrow_p 0$. As

$$\sum_{k\in s}\sum_{\ell\in s}(\tilde{r}_k^2+\tilde{r}_\ell^2)^2 \;=\; 2n\sum_{k\in s}\tilde{r}_k^4 \;+\; 2\left(\sum_{k\in s}\tilde{r}_k^2\right)^2,$$

we have by the Cauchy inequality that

$$\begin{aligned} B_2^2 &\leq H_s\, n^\beta \sum_{k\in s}\sum_{\ell\in s}(\tilde r_k^2 \,+\, \tilde r_\ell^2)^2 \\ &= H_s\left(B_3 \,+\, \frac{B_4}{3}\right). \end{aligned}$$

Thus, the Condition (e) and the Equation (2.44) imply that

$$\frac{B_2}{\mathrm{var}(\hat\theta)_L} \;\to_p\; 0,$$

which together with Equations (2.45) and (2.42) imply (Eq. 2.38).

We now show (Eq. 2.39). By the triangle and Cauchy inequalities, (Eq. 2.34) implies

$$\begin{aligned} |C| &\leq \sum_{k\in s}\sum_{\ell\in s}|\mathcal{D}_{k\ell}||\tilde r_k||\tilde y_\ell| \\ &= \sum_{k\in s}\sum_{\ell\in s}\mathcal{D}_{k\ell}^-|\tilde r_k||\tilde y_\ell| \,+\, \sum_{k\in s}\sum_{\ell\in s}\mathcal{D}_{k\ell}^+|\tilde r_k||\tilde y_\ell| \\ &\leq (G_s^{1/2} \,+\, H_s^{1/2})\,\tilde C^{1/2}, \end{aligned}$$

where

$$\tilde C \;=\; n^\beta \sum_{k\in s}\tilde r_k^2 \sum_{\ell\in s}|\tilde y_\ell|^2. \tag{2.46}$$

From Conditions (d) and (e), Equation (2.39) follows if we show $\tilde C/\mathrm{var}(\hat\theta)_L^2 \to_p 0$. Substituting Equations (2.29) and (2.30) in the Equation (2.46) implies

$$\frac{\tilde C}{\mathrm{var}(\hat\theta)_L^2} \;\leq\; ||\boldsymbol\nabla(\hat{\boldsymbol\mu})||^2\frac{\lambda^2 n^{4+\beta}}{\Psi^2}\left[\frac{1}{n}\sum_{k\in s}\left(\bar w_k\,\bar\varrho_k^\delta||\boldsymbol y_k||^{1+\delta}\right)^2\right]\left[\frac{1}{n}\sum_{\ell\in s}\bar w_\ell^2||\boldsymbol y_\ell||^2\right]. \tag{2.47}$$

From Condition (f), $\beta < 2\delta$, which by Condition (c) and Equations (2.31) and (2.47) imply

$$\frac{\tilde C}{\mathrm{var}(\hat\theta)_L^2} \;=\; n^\beta\,\mathcal{O}_p(n^{-2\delta}),$$

which implies Equation (2.39).

We now show the (Eq. 2.40). Using the triangle and Cauchy inequalities on the Equation (2.35),

$$\begin{aligned} |D| &\leq \sum_{k\in s}\sum_{\ell\in s}\mathcal{D}^{-}_{k\ell}|\tilde{r}_k|^2 \;+\; \sum_{k\in s}\sum_{\ell\in s}\mathcal{D}^{+}_{k\ell}|\tilde{r}_k|^2 \\ &\leq (G_s^{1/2} \;+\; H_s^{1/2})\,\tilde{D}^{1/2}, \end{aligned} \tag{2.48}$$

where

$$\begin{aligned} \tilde{D} &= n^{\beta}\sum_{k\in s}\sum_{\ell\in s}|\tilde{r}_k|^4 \\ &= n^{1+\beta}\sum_{k\in s}|\tilde{r}_k|^4. \end{aligned} \tag{2.49}$$

From Conditions (d) and (e), the Equation (2.40) follows if we show that $\tilde{D}/\text{var}(\hat{\theta})^2_L \rightarrow_p 0$. By substituting the Equation (2.29) in (Eq. 2.49) we obtain,

$$\frac{\tilde{D}}{\text{var}(\hat{\theta})^2_L} \;\leq\; \frac{\lambda^4\, n^{4+\beta}}{\Psi^2}\;\left[\frac{1}{n}\sum_{k\in s}\left(\bar{w}_k\,\bar{\varrho}_k^{\,\delta}||\boldsymbol{y}_k||^{1+\delta}\right)^4\right]. \tag{2.50}$$

which by Condition (c), (Eq. 2.31) and as $\beta < 4\delta$, we have that

$$\frac{\tilde{D}}{\text{var}(\hat{\theta})^2_L} \;=\; n^{\beta}\,\mathcal{O}_p(n^{-4\delta}),$$

implying the Equation (2.40). We now show the Equation (2.41). Using the triangle and the Cauchy inequalities on the Equation (2.36) gives

$$\begin{aligned} |E| &\leq \sum_{k\in s}\sum_{\ell\in s}\mathcal{D}^{-}_{k\ell}|\tilde{r}_k||\tilde{y}_k| \;+\; \sum_{k\in s}\sum_{\ell\in s}\mathcal{D}^{+}_{k\ell}|\tilde{r}_k||\tilde{y}_k| \\ &\leq (G_s^{1/2} \;+\; H_s^{1/2})\,\tilde{E}^{1/2}, \end{aligned}$$

where

$$\begin{aligned} \tilde{E} &= n^{\beta}\sum_{k\in s}\sum_{\ell\in s}|\tilde{r}_k|^2|\tilde{y}_k|^2 \\ &= n^{1+\beta}\sum_{k\in s}|\tilde{r}_k|^2|\tilde{y}_k|^2. \end{aligned} \tag{2.51}$$

From Conditions (d) and (e), the Equation (2.41) follows if $\tilde{E}/\mathrm{var}(\hat{\theta})^2_L \to_p 0$. By substituting Equations (2.29) and (2.30) in the Equation (2.51),

$$\frac{\tilde{E}}{\mathrm{var}(\hat{\theta})^2_L} \;\leq\; ||\boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})||^2 \, \frac{\lambda^2 \, n^{4+\beta}}{\Psi^2} \left[\frac{1}{n} \sum_{k \in s} \bar{w}_k{}^4 \bar{\varrho}_k{}^{2\delta} ||\boldsymbol{y}_k||^{4+2\delta} \right]. \tag{2.52}$$

From Condition (f), we have $\beta < 2\delta$, which by Condition (c) and Equations (2.31) and (2.52) imply

$$\frac{\tilde{E}}{\mathrm{var}(\hat{\theta})^2_L} \;=\; n^{\beta} \mathcal{O}_p(n^{-2\delta}),$$

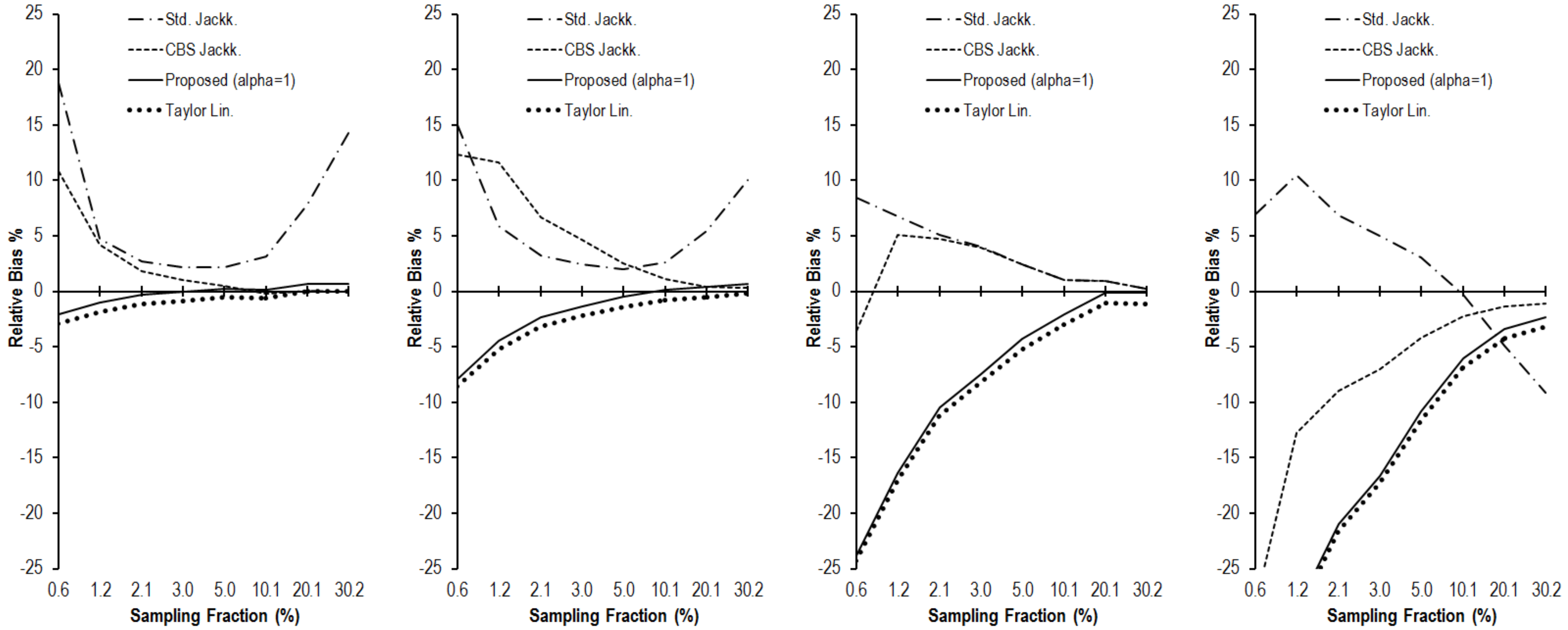
that is

$$\frac{\tilde{E}}{\mathrm{var}(\hat{\theta})^2_L} \;\to_p\; 0.$$

■

2.B. Figures of the simulation study

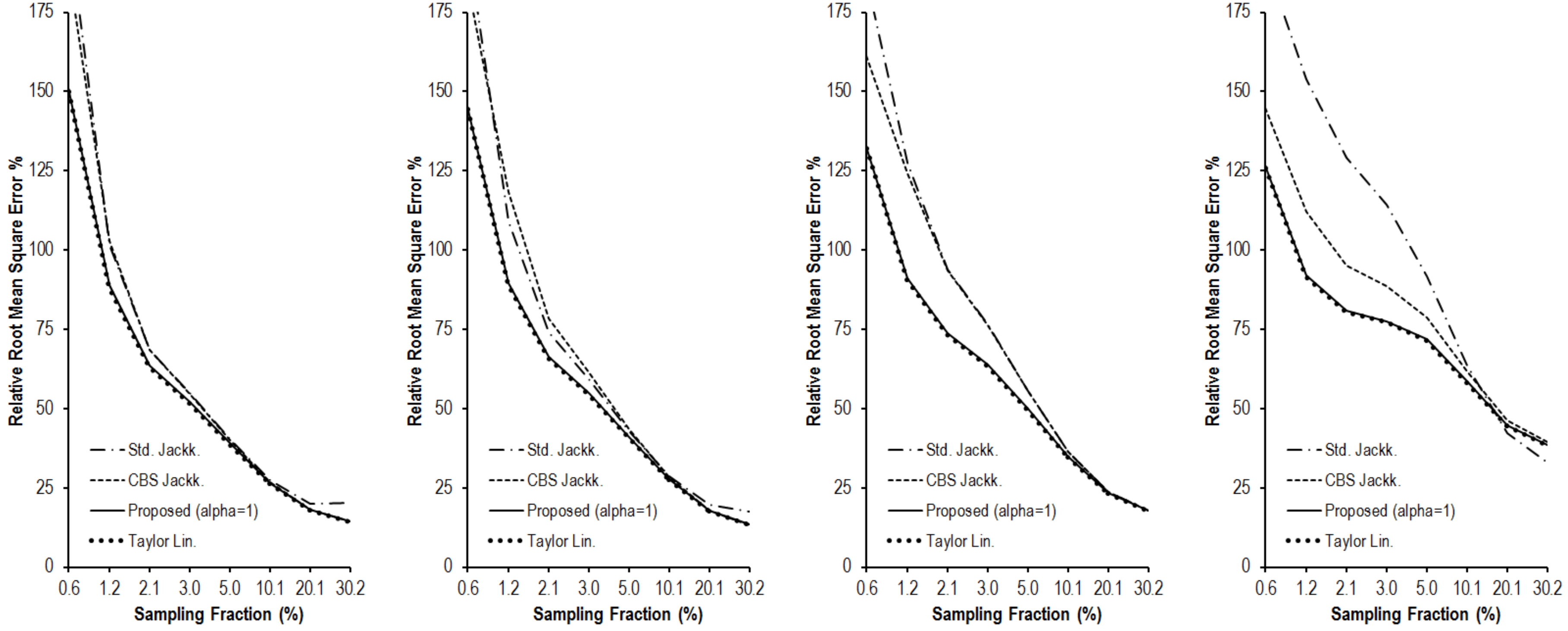
Relative Bias for the variance of $\hat{R}$



$\pi_k \propto x_k$ $\pi_k \propto \sqrt{x_k}$ $\pi_k = n/N$ $\pi_k \propto 1/\psi_k$

FIGURE 2.1: Relative Bias (%) of variance estimators for the ratio point estimator.

Relative Root Mean-Square Error for the variance of $\hat{R}$



$\pi_k \propto x_k$ $\pi_k \propto \sqrt{x_k}$ $\pi_k = n/N$ $\pi_k \propto 1/\psi_k$

FIGURE 2.2: Relative Root Mean-Square Error (%) of variance estimators for the ratio point estimator.

Chapter 3

Variance estimation of Hot-deck imputed estimators of change for rotating repeated surveys

Abstract

A primary interest of many users is often in changes or trends from one time period to another. It is common to compare two cross-sectional estimates for the same study variable taken on two different waves or occasions. These cross-sectional estimates often include imputed values to compensate for item non-response. The estimation of the sampling variance of an estimator of change is useful to judge whether the observed change is statistically significant. Covariances play an important role in estimating that variance and they are not straightforward to estimate due to rotation in repeated surveys. We propose to use a multivariate linear regression approach to estimate these covariances. The proposed estimator is not a model-based estimator, as it is valid even if the underlying model does not fit the data (Berger and Priam, 2010, 2012). We show how this approach can be used to accommodate the effect of imputation. This approach gives design-consistent estimation of the variance of change when the sampling fraction is small and the finite population corrections are negligible. We illustrate the proposed approach using random Hot-deck imputation, although the proposed estimator can be implemented with other imputation techniques.

Keywords and phrases: Deterministic imputation; longitudinal; missing data; nonresponse; overlapping; unequal inclusion probabilities.

3.1 Introduction

Measuring change over time is a central problem for many users of social, economic and demographic data. Smith *et al.* (2003) recognised that assessing change is one of the most important challenges in survey statistics. A primary interest of many users is often in changes or trends from one time period to another. A common problem is to compare two cross-sectional estimates for the same study variable taken on two different waves or occasions. These cross-sectional estimates often include imputed values to compensate for item non-response. That is, missing information for certain variables from an observation contained in the sample (e.g. Lohr, 1999, ch. 8). The estimation of the sampling variance of an estimator of change is useful to judge whether the observed change is statistically significant. Covariances play an important role in estimating the variance of an estimated change and they are not straightforward to estimate due to rotation in repeated surveys.

We propose to use a multivariate linear regression approach to estimate these covariances. The proposed estimator is not a model-based estimator, as it is valid even if the underlying model does not fit the data (Berger and Priam, 2010, 2012). We show how this approach can be used to accommodate the effect of imputation. The regression approach gives design-consistent estimation of the variance of change when the sampling fraction is small and the finite population corrections are negligible. We illustrate the proposed approach using random Hot-deck imputation, although the proposed estimator can be implemented with other imputation techniques.

3.2 Rotating surveys

The estimation of variance of change would be relatively straightforward if cross-sectional estimates were based upon the same sample. Furthermore, because of rotations used in repeated surveys, cross-sectional estimates are not independent. Let s_1 and s_2 denote respectively the first and the second wave samples. The samples s_1 and s_2 are usually not completely overlapping sets of units, because repeated surveys use rotation designs which consist in selecting new units ($k \in s_2 \setminus s_1$) to replace old units ($k \in s_1 \setminus s_2$) that have been in the survey for a specified number of waves. We assume that s_1 and s_2 have the same sample size n. Let n_{12}

denote the sample size of the common sample,

$$s_{12} = s_1 \cap s_2.$$

The units sampled on s_{12} represent usually a large fraction of the sample s_1; that is, n_{12}/n is usually large. We denote the overall sample by $\tilde{s}$ where,

$$\tilde{s} = s_1 \cup s_2.$$

The size of the overall sample is denoted by $\tilde{n} = \#(s)$. We assume that the rotation sampling design is such that n and n_{12} are fixed quantities.

This class contains standard rotating sampling designs such as the rotating systematic sampling design (Holmes and Skinner, 2000), the rotation groups sampling design (e.g. Kalton, 2009; Gambino and Silva, 2009), the rotating design proposed by Tam (1984) and the permanent random numbers rotating design (e.g. Ohlsson, 1995; Nordberg, 2000).

Let $y_{\ell;k}$ denote the value of the variable of interest y for the wave ($\ell = 1, 2$). Suppose, we wish to estimate the absolute change

$$\Delta = \tau_2 - \tau_1, \tag{3.1}$$

between the two population totals

$$\tau_\ell = \sum_{k \in \mathcal{U}} y_{\ell;k},$$

from waves $\ell = 1, 2$; where $\mathcal{U}$ denotes the population of size N, assumed to be the same at both waves. In Section 3.3 we introduce the utilised non-response setting for rotating sampling designs and we show how random Hot-deck imputed values can be used to compensate for item non-response. In Section 3.4, we propose to use a reverse approach (Fay, 1991) to estimate the variance of the imputed estimator of change.

The proposed variance estimator depends on a covariance matrix which will be estimated using a multivariate (general) linear regression approach described in Section 3.6. Further, in Section 3.7 we treat the proposed variance estimator using multiple imputation-classes. A simulation study in Section 3.8 illustrate our findings.

3.3 The non-response

The main objective of this article is to address the problem of variance estimation rather than the non-response problem itself. Literature on variance estimation in surveys with imputed data is vast. For example, when addressing the non-response under a design-based approach we can mention Fay (1991); Rao and Shao (1992); Rao and Sitter (1995); Shao and Steel (1999) among others. On the other hand using models to address the non-response is also popular, a model-assisted approach can be found in Deville and Särndal (1994); Fay (1994); Steel and Fay (1995); Särndal and Lundström (2005) and a Bayesian treatment of imputation can be found for example in the book of Rubin (1987). See Brick and Montaquila (2009) for a wide discussion on non-response. A discussion on which inference-approach to use for non-response in surveys can be found in Haziza (2009). Here we propose to use a design-based approach particularly Hot Deck imputation. A recent review on Hot Deck imputation can be found in Andridge and Little (2010).

Due to non-response, some of the values $y_{\ell;k}$ can be missing in each wave sample s_ℓ, $(\ell = 1, 2)$. We propose to impute this item non-response. Let $z_{\ell;k} = 1\{k \in s_\ell\}$ be the wave sample indicator variables, defined by

$$z_{\ell;k} = \begin{cases} 1 & \text{if } k \in s_\ell, \\ 0 & \text{otherwise,} \end{cases} \qquad (\ell = 1, 2),$$

and let $a_{\ell;k} = 1\{y_{\ell;k} \text{ observed}\}$ be the random variables representing the response mechanism.

$$a_{\ell;k} = \begin{cases} 1 & \text{if } y_{\ell;k} \text{ observed,} \\ 0 & \text{if } y_{\ell;k} \text{ missing,} \end{cases} \qquad (\ell = 1, 2).$$

The observed values of $z_{\ell;k}$ and $a_{\ell;k}$ are known. To simplify, we use the same notation for the random variables and their observed values. For the response mechanism, we consider the usual cross-sectional design-based assumption (e.g. Fay, 1991; Rao and Shao, 1992; Rao and Sitter, 1995; Shao and Steel, 1999), but adapted for rotating sampling designs:

Assumption 1 (single imputation-class). *The response probability for the variable of interest in each wave is uniform on $\mathcal{U}$ and it is strictly positive (i.e. $P\{a_{\ell;k} = 1\} > 0$). The units' responses within waves are independent; and the responses between waves can be dependent. The imputation is conducted independently between waves.*

The setting of one imputation class is the simplest case when handling non-response. In practice, however, it may be considered unrealistic (e.g. Rao and Shao, 1992, p. 818). We therefore also consider, in Section 3.7, multiple imputation-classes where different values are imputed according to a categorical variable. The following assumption is then used:

> **Assumption 2 (multiple imputation-classes)**. *The population $\mathcal{U}$ can be divided into C imputation classes according to a categorical variable x_k, which is observed for all units and remains unchanged at waves. The response probability for the variable of interest is uniform within wave-class combinations and it is strictly positive. The units' responses within and across classes are independent; and responses between waves can be dependent. The imputation is conducted independently within and across wave-class combinations.*

To simplify, we proceed using a single imputation-class setting. In Section 3.7 we address an extension under a multiple imputation-classes.

3.3.1 The imputed estimator of change

Suppose that the change Δ from the Equation (3.1) is estimated by

$$\hat{\Delta}^* \;=\; \hat{\tau}_2^* \;-\; \hat{\tau}_1^*, \tag{3.2}$$

where

$$\hat{\tau}_\ell^* \;=\; \sum_{k \in \tilde{s}} \frac{y_{\ell;k}^*}{\pi_{\ell;k}}, \quad (\ell = 1, 2), \tag{3.3}$$

are two cross-sectional imputed Horvitz and Thompson (1952) estimators where $y_{\ell;k}^*$, are the imputed values given by

$$y_{\ell;k}^* \;=\; z_{\ell;k}\,\{(1 - a_{\ell;k})\,\Phi_{\ell;k} \;+\; a_{\ell;k}\,y_{\ell;k}\}, \tag{3.4}$$

where $\Phi_{\ell;k}$ depends on the imputation technique. For example, in what follows $\Phi_{\ell;k}$ is defined for random Hot-deck imputation, although the proposed approach can be generalised for other imputation techniques. The deterministic mean imputation is also considered in what follows as a particular case of the Hot-deck imputation. Note that the imputation will only be used for missing data due to non-response, and not to impute the values $y_{2;k}$ of $k \in s_2 \setminus s_1$ and the values $y_{1;k}$ of $k \in s_1 \setminus s_2$ which rotate in and out.

3.3.2 Random Hot-deck imputation

The random Hot-deck imputation has the advantage of guaranteeing unbiased estimation of population distributions. In this case, the values $\Phi_{\ell;k}$ used in Equation (3.4) are,

$$\begin{aligned} \Phi_{\ell;k} &= \hat{\mu}_\ell^r + e_{\ell;k}, \\ e_{\ell;k} &= y_{\ell;j} - \hat{\mu}_\ell^r, \end{aligned} \tag{3.5}$$

where j is a donor selected with replacement with probabilities

$$p_{\ell;k} = \frac{\check{a}_{\ell;k}}{\hat{N}_\ell^r},$$

from the wave sample of respondents

$$s_\ell^r = \{k : z_{\ell;k} = 1 \text{ and } a_{\ell;k} = 1\},$$

and where

$$\hat{\mu}_\ell^r = \frac{\hat{\tau}_\ell^r}{\hat{N}_\ell^r},$$

is the estimator of the respondents' mean,

$$\hat{\tau}_\ell^r = \sum_{k \in \tilde{s}} \check{y}_{\ell;k},$$

is the estimator of the respondents' totals, and

$$\hat{N}_\ell^r = \sum_{k \in \tilde{s}} \check{a}_{\ell;k},$$

is the estimator of the number of respondents for waves $\ell = 1, 2$; with $\tilde{s} = s_1 \cup s_2$, and where

$$\check{y}_{\ell;k} = \pi_{\ell;k}^{-1}\, z_{\ell;k}\, a_{\ell;k}\, y_{\ell;k}, \tag{3.6}$$

$$\check{a}_{\ell;k} = \pi_{\ell;k}^{-1}\, z_{\ell;k}\, a_{\ell;k}, \tag{3.7}$$

$$\check{z}_{\ell;k} = \pi_{\ell;k}^{-1}\, z_{\ell;k}. \tag{3.8}$$

The $\pi_{\ell;k}$ denote the first-order inclusion probability of the unit k at wave sample s_ℓ.

If we set $e_{\ell;k} = 0$ in Equation (3.5), then the $y^*_{\ell;k}$ from Equation (3.4) are the deterministic mean imputed values.

3.4 Population variance of the Hot-deck imputed estimator of change

We propose to estimate the variance of $\hat{\Delta}^*$ from (Eq. 3.2) using a reverse approach for non-response (Fay, 1991; Shao and Steel, 1999). Let $\mathcal{U}_1^r$ and $\mathcal{U}_2^r$ be respectively the population of respondents at wave 1 and 2. In other words, at both waves, the population is randomly split into a population of respondents and a population of non-respondents according to an unknown response mechanism. Let $E_r\{\cdot\}$, $V_r(\cdot)$ and $Corr_r(\cdot)$ denote respectively the expectation, variance and the correlation operators with respect to the response mechanism. Rotation samples s_1 and s_2 are selected from the population $\mathcal{U}$ according to a rotation sampling design (see Section 4.1). The wave samples of respondents are given by

$$s_\ell^r \;=\; \mathcal{U}_\ell^r \cap s_\ell, \qquad (\ell = 1, 2).$$

Let $E_d\{\cdot\}$ and $V_d(\cdot)$ denote the expectation and the variance operators with respect to the sampling design. Furthermore, we suppose that the random Hot-deck imputation from Section 3.3.1 is used. Let $E_I\{\cdot\}$ and $V_I(\cdot)$ denote the expectation and the variance operators with respect to the random imputation.

The overall variance of the imputed estimator of change $\hat{\Delta}^*$ from (Eq. 3.2) is given by

$$V(\hat{\Delta}^*) \;=\; A \;+\; B \;+\; C, \tag{3.9}$$

which is an overall three stage variance, where

$$\begin{aligned} A &= E_r\{V_d(E_I\{\hat{\Delta}^*|S,R\}|R)\}, && (3.10)\\ B &= E_r\{E_d\{V_I(\hat{\Delta}^*|S,R)|R\}\}, && (3.11)\\ C &= V_r(E_d\{E_I\{\hat{\Delta}^*|S,R\}|R\}), && (3.12) \end{aligned}$$

with $S = \{s_1, s_2\}$, $R = \{s_1^r, s_2^r\}$.

The overall variance from the Equation (3.9) includes the effect of the response mechanism, the sampling design and the imputation. We now focus on each of its terms.

The term A

Turning to the term A in Equation (3.10). As $E_I\{e_{\ell;k}|S,R\}=0$, from (Eq. 3.5) we have that

$$E_I\{\Phi_{\ell;k}|S,R\} \;=\; \hat{\mu}^r_\ell.$$

Hence, from the Equations (3.3) and (3.4), it can be shown

$$E_I\{\hat{\tau}^*_\ell|S,R\} \;=\; \hat{N}_\ell\,\frac{\hat{\tau}^r_\ell}{\hat{N}^r_\ell}.$$

We thus have,

$$E_I\{\hat{\Delta}^*|S,R\} \;=\; \hat{N}_2\,\frac{\hat{\tau}^r_2}{\hat{N}^r_2} \;-\; \hat{N}_1\,\frac{\hat{\tau}^r_1}{\hat{N}^r_1}, \tag{3.13}$$

where

$$\hat{N}_\ell \;=\; \sum_{k\in\tilde{s}} \check{z}_{\ell;k}, \quad (\ell=1,2),$$

is an estimator of N. The $E_I\{\hat{\Delta}^*|S,R\}=f(\hat{\boldsymbol{\tau}})$ is a function $f(\cdot)$ of estimated totals $\hat{\boldsymbol{\tau}}=(\hat{\boldsymbol{\tau}}_1^T,\hat{\boldsymbol{\tau}}_2^T)^T$, where

$$\hat{\boldsymbol{\tau}}_\ell \;=\; \left(\hat{N}_\ell,\, \hat{N}^r_\ell,\, \hat{\tau}^r_\ell\right)^T, \tag{3.14}$$

is a vector of Horvitz and Thompson (1952) totals. Using the Taylor approximation (e.g. Särndal *et al.*, 1992, secs. 5.5, 5.7), we have that

$$E_I\{\hat{\Delta}^*|S,R\} - \Delta \;\doteq\; \boldsymbol{\nabla}(\boldsymbol{\tau})^T\,(\hat{\boldsymbol{\tau}}-\boldsymbol{\tau}),$$

where

$$\boldsymbol{\nabla}(\boldsymbol{\tau}) \;=\; \left(\frac{-\tau^r_1}{N^r_1},\, \frac{N\tau^r_1}{(N^r_1)^2},\, \frac{-N}{N^r_1},\, \frac{\tau^r_2}{N^r_2},\, \frac{-N\tau^r_2}{(N^r_2)^2},\, \frac{N}{N^r_2}\right)^T, \tag{3.15}$$

is the gradient of $f(\boldsymbol{\tau})$ at $\boldsymbol{\tau}=(\boldsymbol{\tau}_1^T,\boldsymbol{\tau}_2^T)^T$, with

$$\boldsymbol{\tau}_\ell \;=\; (N,\, N^r_\ell,\, \tau^r_\ell)^T, \tag{3.16}$$

where τ^r_ℓ is the population total of the variable of interest over the respondents at wave ℓ; and N^r_ℓ is the total number of respondents at wave ℓ, $(\ell=1,2)$.

The Taylor approximation of $V_d(E_I\{\hat{\Delta}^*|S,R\}|R)$ is given by

$$V_d(E_I\{\hat{\Delta}^*|S,R\}|R) \;\doteq\; \boldsymbol{\nabla}(\boldsymbol{\tau})^T\,\boldsymbol{V_d}(\hat{\boldsymbol{\tau}}|R)\,\boldsymbol{\nabla}(\boldsymbol{\tau}), \tag{3.17}$$

where $\boldsymbol{V_d}(\hat{\boldsymbol{\tau}}|R)$ is the design covariance matrix of the vector $\hat{\boldsymbol{\tau}}$. Thus, an approximately design-based unbiased estimator for $V_d(E_I\{\hat{\Delta}^*|S,R\}|R)$ in the Equation (3.17) is given by

$$\hat{V}_d(E_I\{\hat{\Delta}^*|S,R\}|R) \;=\; \boldsymbol{\nabla}(\hat{\boldsymbol{\tau}})^T\, \hat{\boldsymbol{V}}_{\boldsymbol{d}}(\hat{\boldsymbol{\tau}}|R)\, \boldsymbol{\nabla}(\hat{\boldsymbol{\tau}}), \tag{3.18}$$

where $\hat{\boldsymbol{V}}_{\boldsymbol{d}}(\hat{\boldsymbol{\tau}}|R)$ is the approximately design unbiased estimator, defined below in (Eq. 3.25), of $\boldsymbol{V_d}(\hat{\boldsymbol{\tau}}|R)$. Note that in the Equation (3.18), the $a_{\ell;k}$'s are treated as fixed quantities, as $\hat{V}_d(E_I\{\hat{\Delta}^*|S,R\}|R)$ is a conditional variance given R.

The term B

We now turn to the term B in Equation (3.11). From *Assumption 1* we have,

$$\begin{aligned} V_I(\hat{\Delta}^*|S,R) \;&=\; \sum_{\ell=1}^{2} V_I(\hat{\tau}_\ell^*|S,R) \\ &=\; \sum_{\ell=1}^{2} V_I(\Phi_{\ell;k}|S,R) \sum_{k\in\tilde{s}} \frac{z_{\ell;k}}{\pi_{\ell;k}^2}(1-a_{\ell;k}), \end{aligned} \tag{3.19}$$

with

$$\begin{aligned} V_I(\Phi_{\ell;k}|S,R) \;&=\; V_I(e_{\ell;k}|S,R) \\ &=\; \sum_{k\in\tilde{s}} a_{\ell;k}\, p_{\ell;k}\, e_{\ell;k}^2, \end{aligned}$$

as

$$E_I\{e_{\ell;k}|S,R\} \;=\; 0.$$

Note that we use the same notation for the random variables $e_{\ell;k}$'s and their observed values. Also note that, under deterministic mean imputation, we have

$$V_I(\hat{\Delta}^*|S,R) = 0.$$

The term C

We now turn to the term C in Equation (3.12). By denoting

$$\Upsilon_\ell \;=\; E_d\{E_I\{\hat{t}_\ell^*|S,R\}|R\} \;=\; N_\ell\,\frac{\tau_\ell^r}{N_\ell^r},$$

we have from (3.2) that,

$$E_d\{E_I\{\hat{\Delta}^*|S,R\}|R\} \;=\; \Upsilon_2 \;-\; \Upsilon_1.$$

Hence, from Equation (3.12),

$$C \;=\; V_r\,(\Upsilon_1) \;+\; V_r\,(\Upsilon_2) \;-\; 2\,Corr_r\,(\Upsilon_1,\Upsilon_2)\,\sqrt{V_r\,(\Upsilon_1)\;V_r\,(\Upsilon_2)}, \qquad (3.20)$$

where

$$V_r\,(\Upsilon_\ell) \;=\; V_r(E_d\{E_I\{\hat{t}_\ell^*|S,R\}|R\}),$$

is the cross-sectional variance for the wave ℓ under the response mechanism given the random imputation and the sampling design.

As the correlation $Corr_r\,(\Upsilon_1,\Upsilon_2) = \mathcal{O}(1)$ in Equation (3.20), we recall from Shao and Steel (1999, pp. 256, 257) that the cross-sectional variances $V_r\,(\Upsilon_\ell)$ are of order $\mathcal{O}(N_\ell)$ implying

$$C \;=\; \mathcal{O}(N_\ell).$$

Given standard assumptions for linearised variances of functions of totals (e.g. Robinson and Särndal, 1983; Särndal *et al.*, 1992, secs. 5.5, 5.7), the linearised version of the term A from (Eq. 3.17) is of order $\mathcal{O}(N_\ell^2/n)$, being the dominant term of the overall variance $V(\hat{\Delta}^*)$ from Equation (3.9). Furthermore,

$$\frac{C}{A} \;=\; \mathcal{O}\left(\frac{n}{N_\ell}\right).$$

Thus, for negligible n/N_ℓ the contribution of C to (Eq. 3.9) should be negligible (e.g. Haziza, 2009, pp. 238-240). We thus have that

$$V(\hat{\Delta}^*) \;\simeq\; A \,+\, B. \qquad (3.21)$$

Also note from Equation (3.20) that the response mechanism can be correlated between waves.

3.5 The proposed variance estimator

We proposed to estimate the variance of the imputed estimator of change $\hat{\Delta}^*$ from Equation (3.2) by

$$\hat{V}(\hat{\Delta}^*) \;=\; \hat{V}_d(E_I\{\hat{\Delta}^*|S,R\}|R) \;+\; V_I(\hat{\Delta}^*|S,R), \tag{3.22}$$

where $\hat{V}_d(E_I\{\hat{\Delta}^*|S,R\}|R)$ and $V_I(\hat{\Delta}^*|S,R)$ are as defined in Equations (3.18) and (3.19). In the following section 3.6 we propose a multivariate (or general) linear regression model to estimate the covariance matrix $\boldsymbol{V_d}(\hat{\boldsymbol{\tau}}|R)$ involved in the computation of (Eq. 3.18).

Note that (Eq. 3.22) can be generalised for other types of imputation, as long as $E_I\{\hat{\Delta}^*|S,R\}$ is a function of Horvitz and Thompson (1952) totals. In that situation $\boldsymbol{\nabla}(\boldsymbol{\tau})^T$ would have a different expression which depends on the used imputation.

The proposed estimator (Eq. 3.22) is an approximately unbiased estimator of the variance $V(\hat{\Delta}^*)$ from Equation (3.9), as the overall expectation of (Eq. 3.22) is given by

$$\begin{aligned} E_r\{E_d\{E_I\{\hat{V}(\hat{\Delta}^*)|S,R\}|R\}\} &= E_r\{E_d\{E_I\{\hat{V}_d(E_I\{\hat{\Delta}^*|S,R\}|R)|S,R\}|R\}\} \\ &\quad + E_r\{E_d\{E_I\{V_I(\hat{\Delta}^*|S,R)\}|R\}\} \\ &\doteq E_r\{V_d(E_I\{\hat{\Delta}^*|S,R\}|R)\} \\ &\quad + E_r\{E_d\{V_I(\hat{\Delta}^*|S,R)|R\}\} \\ &\doteq V(\hat{\Delta}^*), \end{aligned}$$

by using the Equation (3.21) and the fact that (Eq. 3.18) does not depends on the $e_{\ell;k}$'s for $\ell = 1, 2$.

An advantage of the proposed variance estimator (Eq. 3.22) is that it is approximately unbiased under the unknown response mechanism without making strong assumptions about it.

3.6 Variance estimation using the multivariate regression approach

We derive here an expression to estimate the covariance matrix $\boldsymbol{V_d}(\hat{\boldsymbol{\tau}}|R)$ in Equation (3.17) under the rotation sampling design. Note that this covariance is not straightforward to estimate because it involves covariance between components of $\hat{\boldsymbol{\tau}}$ defined from different samples, s_1 and s_2 that are partially overlapped. Several methods can be used to estimate it (e.g. Kish, 1965; Tam, 1984; Holmes and Skinner, 2000; Nordberg, 2000; Berger, 2004; Qualité and Tillé, 2008; Wood, 2008; Goga *et al.*, 2009). We propose to use a multivariate (or general) linear regression model to estimate this covariance matrix.

Consider the following $\tilde{n} \times 6$ matrix

$$\check{\boldsymbol{Y}}_{(\tilde{n}\times 6)} = (\check{\boldsymbol{y}}_1, \ldots, \check{\boldsymbol{y}}_k, \ldots, \check{\boldsymbol{y}}_{\tilde{n}})^T,$$

where $\tilde{n} = \#(s_1 \cup s_2)$, $\check{\boldsymbol{y}}_k = (\check{\boldsymbol{y}}_{1k}, \check{\boldsymbol{y}}_{2k})$ and

$$\check{\boldsymbol{y}}_{\ell k} = (\check{z}_{\ell;k}, \check{a}_{\ell;k}, \check{y}_{\ell;k}), \tag{3.23}$$

with $\check{z}_{\ell;k}$, $\check{a}_{\ell;k}$ and $\check{y}_{\ell;k}$ as in Equations (3.8), (3.7) and (3.6), ($\ell = 1, 2$). Consider the following multivariate (general) regression model

$$\check{\boldsymbol{Y}} = \boldsymbol{Z}_s\,\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \tag{3.24}$$

where $\boldsymbol{\alpha}$ is a 3×6 matrix of regression parameters, the residuals $\boldsymbol{\varepsilon}$ have a 6×6 covariance matrix $\boldsymbol{\Sigma}$, and $\boldsymbol{Z}_s$ is a $\tilde{n} \times 3$ design matrix which specifies the fixed-size constraints of the rotation design. The matrix $\boldsymbol{Z}_s$ is defined by

$$\boldsymbol{Z}_s = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k, \ldots, \boldsymbol{z}_{\tilde{n}})^T,$$

with

$$\boldsymbol{z}_k = (z_{1;k}, z_{2;k}, z_{1;k} \times z_{2;k}).$$

The model (Eq. 3.24) belongs to the class of general linear model. In fact, (Eq.

3.24) is also a multivariate analysis of variance (MANOVA) model, as the covariates are all dummy variables. Note that we have the fixed size constraints

$$\begin{aligned}\sum_{k\in\tilde{s}} z_{\ell;k} &= n \\ \sum_{k\in\tilde{s}} z_{1;k}\, z_{2;k} &= n_{12},\end{aligned}$$

which restrict us to samples with only these sample sizes. Thus, by using the design variables as covariates in the model, we are conditioning on them. This takes into account the fixed size constraints in the estimation of the covariance (see Berger and Priam, 2010). Note that the model from Equation (3.24) includes interactions between the variable $z_{1;k}$ and $z_{2;k}$. These interactions capture the rotation of the sampling design which is represented by the constraint $\sum_{k\in\tilde{s}} z_{1;k}z_{2;k} = n_{12}$.

To estimate $\boldsymbol{V_d}(\hat{\boldsymbol{\tau}}|R)$, Berger and Priam (2010) proposed the estimator

$$\hat{\boldsymbol{V}}_{\boldsymbol{d}}(\hat{\boldsymbol{\tau}}|R) \;=\; \hat{\boldsymbol{D}}^T\, \hat{\boldsymbol{\Sigma}}\, \hat{\boldsymbol{D}}, \tag{3.25}$$

where the matrix $\hat{\boldsymbol{\Sigma}}$ is the *Ordinary Least Squares* residual covariance matrix estimate of the model from Equation (3.24) and $\hat{\boldsymbol{D}}$ is a diagonal matrix with the diagonal elements:

$$\sqrt{\frac{\hat{V}(\hat{\tau}_q|R)}{\hat{\Sigma}_{qq}}},$$

where $\hat{V}(\hat{\tau}_q|R)$ is a design-based variance estimator of the q-th component of $\hat{\boldsymbol{\tau}}$ and $\hat{\Sigma}_{qq}$ is the q-th diagonal component of $\hat{\boldsymbol{\Sigma}}$. Any unbiased standard variance estimator can be used to calculate $\hat{V}(\hat{\tau}_q|R)$. Note that (Eq. 3.25) is positive definite, as $\hat{\boldsymbol{\Sigma}}$ is always positive definite. Hence, the proposed variance estimator (Eq. 3.22) is always positive.

Berger and Priam (2010) showed that the estimator (Eq. 3.25) is an approximately design unbiased estimator for $\boldsymbol{V_d}(\hat{\boldsymbol{\tau}}|R)$ when the finite population corrections are negligible. It is a design-based consistent estimator for $\boldsymbol{V_d}(\hat{\boldsymbol{\tau}}|R)$ even when model from Equation (3.24) does not fit the data (Berger and Priam, 2010). Note that (Eq. 3.25) takes into account the unequal probabilities.

In a series of simulations based on the Swedish Labour Force Survey, Andersson *et al.* (2011a,b) showed that (Eq. 3.25) gives more accurate estimates than standard variance estimators (e.g. Tam, 1984; Qualité and Tillé, 2008) when we are interested in change between strata domains.

3.7 Multiple imputation-classes

We now consider the situation of multiple imputation-classes for the imputation. Hence, the Hot-deck imputation setting uses the *Assumption 2* in Section 3.3 instead of the *Assumption 1.*

Let $b_k^{(c)} = 1\{x_k = c\}$, $(c = 1, 2, \ldots, C)$. That is,

$$b_k^{(c)} = \begin{cases} 1 & \text{if } x_k = c, \\ 0 & \text{otherwise,} \end{cases}$$

where x_k is the categorical variable from *Assumption 2.* The random Hot-deck imputed values $y^*_{\ell;k}$ from Equation (3.4) now use

$$\Phi_{\ell;k} = \sum_{c=1}^{C} b_k^{(c)} \left(\hat{\mu}_\ell^{r(c)} + e_{\ell;k}^{(c)} \right),$$

where

$$e_{\ell;k}^{(c)} = y_{\ell;j} - \hat{\mu}_\ell^{r(c)},$$

instead of (Eq. 3.5), where j is a donor selected with replacement with probabilities,

$$p_{\ell;k} = \frac{b_k^{(c)} \breve{a}_{\ell;k}}{\hat{N}_\ell^{r(c)}},$$

from the wave-class combination sample of respondents,

$$s_\ell^{r(c)} = \{k : z_{\ell;k} = 1,\ a_{\ell;k} = 1,\ b_k^{(c)} = 1\},$$

and where

$$\begin{aligned} \hat{\mu}_\ell^{r(c)} &= \frac{\hat{\tau}_\ell^{r(c)}}{\hat{N}_\ell^{r(c)}}, \\ \hat{\tau}_\ell^{r(c)} &= \sum_{k \in \tilde{s}} b_k^{(c)} \, \breve{y}_{\ell;k} \\ \hat{N}_\ell^{r(c)} &= \sum_{k \in \tilde{s}} b_k^{(c)} \, \breve{a}_{\ell;k}, \end{aligned}$$

estimate, respectively, the respondents' mean, totals and the number of respondents for each wave-class combination $(\ell = 1, 2; c = 1, 2, \ldots, C)$.

Again, as it was the case for a single-imputation class, if we set

$$e_{\ell;k}^{(c)} \;=\; 0,$$

then we are using deterministic mean imputation.

With multiple imputation-classes the population variance $V(\hat{\Delta}^*)$ from Equation (3.9) changes. The term A in (Eq. 3.10) now uses the Equations (3.13), (3.14), (3.15) and (3.16) replaced by,

$$\begin{aligned}
E_I\{\hat{\Delta}^*|S,R\} &= \hat{N}_2\,\frac{\sum_{c=1}^{C}\hat{\tau}_2^{r(c)}}{\sum_{c=1}^{C}\hat{N}_2^{r(c)}} \;-\; \hat{N}_1\,\frac{\sum_{c=1}^{C}\hat{\tau}_1^{r(c)}}{\sum_{c=1}^{C}\hat{N}_1^{r(c)}}, && (3.26)\\
\hat{\boldsymbol{\tau}}_\ell &= \left(\hat{N}_\ell,\, \hat{N}_\ell^{r(1)}, \ldots, \hat{N}_\ell^{r(C)},\, \hat{\tau}_\ell^{r(1)}, \ldots,\, \hat{\tau}_\ell^{r(C)}\right)^T, && (3.27)\\
\boldsymbol{\nabla}(\boldsymbol{\tau}) &= \left(\frac{-\tau_1^r}{N_1^r},\, \underbrace{\frac{N\tau_1^r}{(N_1^r)^2}, \ldots,}_{\text{C times}}\, \underbrace{\frac{-N}{N_1^r}, \ldots,}_{\text{C times}}\, \frac{\tau_2^r}{N_2^r},\, \underbrace{\frac{-N\tau_2^r}{(N_2^r)^2}, \ldots,}_{\text{C times}}\, \underbrace{\frac{N}{N_2^r}, \ldots}_{\text{C times}}\right)^T && (3.28)\\
\boldsymbol{\tau}_\ell &= \left(N,\, N_\ell^{r(1)}, \ldots,\, N_\ell^{r(C)},\, \tau_\ell^{r(1)}, \ldots,\, \tau_\ell^{r(C)}\right)^T, && (3.29)
\end{aligned}$$

with

$$\tau_\ell^r \;=\; \sum_{c=1}^{C}\tau_\ell^{r(c)},$$

and

$$N_\ell^r \;=\; \sum_{c=1}^{C}N_\ell^{r(c)},$$

where $\tau_\ell^{r(c)}$ and $N_\ell^{r(c)}$ are the respondents population total of the variable y_k and the number of respondents at each wave-class combination, $(\ell = 1, 2; c = 1, 2, \ldots, C)$. Now, revisiting the term B in (Eq. 3.11). From *Assumption 2* we have that

$$V_I(\hat{\Delta}^*|S,R) \;=\; \sum_{\ell=1}^{2} V_I(\Phi_{\ell;k}|S,R)\sum_{c=1}^{C}\sum_{k\in\tilde{s}}\frac{b_k^{(c)}\,z_{\ell;k}}{\pi_{\ell;k}^2}(1-a_{\ell;k}), \qquad (3.30)$$

with

$$V_I(\Phi_{\ell;k}|S,R) \;=\; \sum_{c=1}^{C}\sum_{k\in\tilde{s}} a_{\ell;k}\,p_{\ell;k}\,\{e_{\ell;k}^{(c)}\}^2.$$

Hence, the Equation (3.30) replaces (Eq. 3.19). About the term C in (Eq. 3.12). The *Assumption 2* implies an stratified setting which does not affect the obtained results for a single imputation class. Thus, the proposed estimator is given by

Equation (3.22) with $\hat{V}_d(E_I\{\hat{\Delta}^*|S,R\}|R)$ and $V_I(\hat{\Delta}^*|S,R)$ now given by Equations (3.18) and (3.30) but using the Equations (3.26), (3.27), (3.28) and (3.29).

As in section 3.6, the covariance matrix $\boldsymbol{V_d}(\hat{\boldsymbol{\tau}}|R)$ in the Equation (3.17) can be estimated using a multivariate (or general) linear regression model. With multiple imputation-classes, the model (Eq. 3.24) now uses a $\tilde{n} \times (2+4C)$ matrix

$$\check{\boldsymbol{Y}}_{(\tilde{n}\times(2+4C))} \;=\; (\check{\boldsymbol{y}}_1, \ldots, \check{\boldsymbol{y}}_k, \ldots, \check{\boldsymbol{y}}_{\tilde{n}})^T,$$

where $\check{\boldsymbol{y}}_k = (\check{\boldsymbol{y}}_{1k}, \check{\boldsymbol{y}}_{2k})$ with

$$\check{\boldsymbol{y}}_{\ell k} = (\check{z}_{\ell;k},\, b_k^{(1)}\,\check{a}_{\ell;k},\, \ldots,\, b_k^{(C)}\,\check{a}_{\ell;k},\, b_k^{(1)}\,\check{y}_{\ell;k},\, \ldots,\, b_k^{(C)}\,\check{y}_{\ell;k}),$$

replacing the Equation (3.23), and now $\boldsymbol{\alpha}$ is a $3 \times (2+4C)$ matrix and $\boldsymbol{\Sigma}$ is a $(2+4C) \times (2+4C)$ matrix.

3.8 Simulation study

We use the *Labor Force Population* dataset from Valliant *et al.* (2000, Appendix B.5) available at the John Wiley worldwide website. The dataset is duplicated 50 times to obtain a large population suitable for different levels of rotation and small sampling fractions in the sampling design.

We consider using two variables to build the overtime wave variables and for the inclusion probabilities (explained below):

- *Weekly wages*,
- *Hours worked per week* (*HW*).

Their respective maximum values 99 and 999 are removed. We thus obtain a population frame of size $N = 23\ 550$.

Further, we build the overtime waves $(y_{1;k})$ and $(y_{2;k})$ from the variable *Weekly wages*, with

$$\begin{aligned} y_{1;k} &= \textit{Weekly wages}, \\ y_{2;k} &= y_{1;k} \;+\; \sqrt{y_{1;k}} \;+\; \psi_k, \end{aligned}$$

where ψ_k denotes randomly generated values according to a Normal distribution $N(0, 5^2)$.

The population parameter of interest Δ is the absolute change between the two wave totals $\tau_\ell = \sum_{k\in\mathcal{U}} y_{\ell;k}$ $(\ell = 1, 2)$ as defined in section 4.1; and with true value $\Delta \ = 377\ 960.66$. We estimate Δ by the Hot-deck imputed point estimator $\hat{\Delta}^*$ defined in subsection 3.3.1.

In selecting the corresponding wave sample s_1, we use the sampling method by Rao (1965); Sampford (1967) considering two scenarios:

- The $\pi_{1;k} = n/N$,
- The $\pi_{1;k} \propto$ *Hours worked per week.*

Note that the Rao-Sampford sampling is of large entropy as shown in Berger (2011). Then, for selecting the wave sample s_2 we first select a simple random sample of n_{12} units taken from s_1 where $g = n_{12}/n = \{0.40, 0.60, 0.80, 0.95\}$. Following, we sample $n - n_{12}$ units from $\mathcal{U} \setminus s_1$ with probabilities proportional to

$$\pi_{2;k} \ = \ \frac{\pi_{1;k}}{1 - \pi_{1;k}}.$$

Each wave respondents are selected randomly using a Poisson sampling from $\mathcal{U}$. At wave 1, they are selected with probabilities

$$q_{1;k} \ = \ P\{a_{1;k} = 1\} \ = \ \{0.70,\ 0.90\,\}.$$

Then wave 2 respondents are selected with conditional probabilities given their response at wave 1,

$$P\{a_{2;k} = 1|a_{1;k}\} \ = \ (0.95)\, a_{1;k} \ + \ (0.65)\, (1 \ - \ a_{1;k}).$$

These imply the wave 2 response probabilities

$$q_{2;k} \ = \ P\{a_{2;k} = 1\} \ \simeq \ (0.95)\,(0.90) \ + \ (0.65)\,(0.10) \ = \ \{0.86,\ 0.92\}.$$

Note that at each wave the number of respondents is not fixed. Also note that it could happen an observation has missing information at both waves. The item non-response is imputed using random Hot-deck according to the procedure described in subsections 3.3.1 and 3.3.2.

For each simulation, 10 000 set of respondents and set of samples were selected to compute: the empirical relative bias

$$\mathrm{RB} \;=\; \frac{\mathrm{B}(\widehat{\mathrm{var}}(\hat{\Delta}^{*}))}{\mathrm{var}(\hat{\Delta}^{*})},$$

where

$$\mathrm{B}(\widehat{\mathrm{var}}(\hat{\Delta}^{*})) \;=\; \mathrm{E}(\widehat{\mathrm{var}}(\hat{\Delta}^{*})) - \mathrm{var}(\hat{\Delta}^{*}),$$

the empirical relative root mean square error

$$\mathrm{RRMSE} \;=\; \frac{\sqrt{\mathrm{MSE}(\widehat{\mathrm{var}}(\hat{\Delta}^{*}))}}{\mathrm{var}(\hat{\Delta}^{*})},$$

and the coverage at a 95% confidence level. The term $\mathrm{var}(\hat{\Delta}^{*})$ denotes the empirical variance computed from the 10 000 observed values of $\hat{\Delta}^{*}$.

We compare the proposed variance estimator $\hat{V}(\hat{\Delta}^{*})$ from Equation (3.22) versus a naïve approach which consists in computing variance estimates directly from imputed values and neglecting the fact that imputation is used.

Both Tables 1 and 2, in terms of RB show that for different values of response probabilities $q_{1;k}$ and for different values of the overlapping fraction g between waves, the proposed approach tends to consistently give values which are closer to zero than the naïve approach. As expected, the naïve approach which neglects imputation, tends to severely under estimate the variance; in particular, when non-response is large; that is, if $q_{1;k}$ is smaller. Further, by comparing Table 1 and 2, we can observe that using unequal inclusion probabilities improves both approaches' bias.

In terms of RRMSE it can also be seen in both tables that for different values of g and $q_{1;k}$, the proposed approach has smaller values than the naïve approach. Now, comparing tables 1 and 2, it can be seen that both approaches are more unstable with unequal probabilities than when using simple random sampling. We recall that this can be generally improved by using a more complex imputation. Regarding the coverage, the tables show that the proposed approach provides closer values to 95% than the naïve approach.

Additional graphical representations for these simulation results are illustrated in the Appendix 3.A of this chapter.

Table 3.1: RB, RRMSE and Coverage at 95% confidence level of the variance estimators for the Hot-deck imputed point estimator $\hat{\Delta}^*$ using $\pi_{1;k} = n/N$.

$q_{1;k}$	$q_{2;k}$	g	f	RB		RRMSE		Coverage	
				Prop.	Naïve	Prop.	Naïve	Prop.	Naïve
	($\doteq$)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
0.70	0.86	40	0.5	-2.8	-33.8	15.5	35.3	95.0	88.7
			1.0	-0.7	-32.3	11.2	33.2	94.8	89.3
			1.5	-0.4	-32.1	9.1	32.7	94.7	89.5
			2.0	-2.7	-33.7	8.2	34.1	94.6	88.8
		60	0.5	-1.8	-31.3	17.6	33.7	94.7	89.1
			1.0	-1.2	-30.9	12.5	32.2	94.8	89.7
			1.5	-1.1	-30.8	10.2	31.7	94.7	89.2
			2.0	0.0	-30.1	8.7	30.8	94.8	89.9
		80	0.5	-1.8	-28.8	20.1	32.7	94.7	89.8
			1.0	-0.4	-27.5	14.4	29.8	95.0	90.4
			1.5	-0.4	-27.5	11.6	29.0	95.0	90.5
			2.0	-2.2	-29.0	10.0	30.0	94.6	90.0
		95	0.5	-1.8	-25.3	22.8	31.9	94.8	90.8
			1.0	-1.9	-25.5	16.0	29.0	94.5	90.8
			1.5	-0.9	-24.8	13.1	27.2	94.8	90.7
			2.0	-1.6	-25.3	11.2	27.0	94.7	90.9
0.90	0.92	40	0.5	-0.7	-15.9	14.5	20.2	94.8	92.9
			1.0	0.2	-15.2	10.2	17.6	95.3	93.2
			1.5	-2.1	-17.2	8.5	18.5	94.7	92.6
			2.0	-0.7	-15.9	7.2	17.1	95.1	92.8
		60	0.5	0.4	-14.4	17.2	21.0	94.9	93.2
			1.0	0.2	-14.6	12.1	18.2	94.9	92.9
			1.5	0.6	-14.2	9.9	16.7	95.1	93.1
			2.0	0.0	-14.8	8.5	16.7	94.8	92.7
		80	0.5	-2.2	-15.3	21.4	25.5	94.7	93.0
			1.0	-2.0	-15.0	15.0	20.8	95.0	93.1
			1.5	-0.2	-13.7	12.3	18.2	94.8	92.9
			2.0	-1.0	-14.4	10.7	17.7	94.6	92.9
		95	0.5	-2.9	-13.4	27.7	33.0	94.5	92.9
			1.0	-2.4	-13.2	19.4	24.8	94.5	93.2
			1.5	-1.1	-12.0	15.9	21.3	95.1	93.6
			2.0	-0.9	-12.0	13.8	19.3	94.9	93.5

Table 3.2: RB, RRMSE and Coverage at 95% confidence level of the variance estimators for the Hot-deck imputed point estimator $\hat{\Delta}^*$ using $\pi_{1;k} \propto HW$.

$q_{1;k}$	$q_{2;k}$	g	f	RB		RRMSE		Coverage	
				Prop.	Naïve	Prop.	Naïve	Prop.	Naïve
	($\hat{=}$)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
0.70	0.86	40	0.5	-1.6	-29.1	32.3	52.7	94.3	88.7
			1.0	-2.5	-29.9	23.3	42.8	93.5	87.3
			1.5	-3.4	-30.5	19.0	39.7	93.0	86.8
			2.0	-1.5	-29.4	16.4	36.7	92.4	86.5
		60	0.5	-2.0	-27.7	36.2	57.1	94.3	89.3
			1.0	-1.2	-27.5	26.1	44.0	94.2	88.9
			1.5	-0.8	-27.1	21.4	39.1	94.1	88.9
			2.0	-0.7	-27.5	18.2	36.4	93.5	87.8
		80	0.5	-0.1	-25.6	40.7	59.2	94.8	90.4
			1.0	0.0	-25.2	29.3	45.5	94.9	89.7
			1.5	-0.4	-25.1	23.6	40.4	94.8	90.2
			2.0	-0.6	-25.8	20.4	37.1	94.5	89.7
		95	0.5	-1.5	-24.3	43.9	63.6	94.5	90.8
			1.0	0.4	-22.9	31.7	48.5	95.2	91.3
			1.5	0.5	-23.4	26.0	41.4	94.9	91.2
			2.0	-0.8	-24.3	22.3	38.1	94.9	90.6
0.90	0.92	40	0.5	-0.5	-15.5	34.3	51.2	94.1	91.7
			1.0	-1.5	-15.6	23.5	37.7	93.1	90.4
			1.5	-0.5	-14.8	19.9	33.1	92.9	90.2
			2.0	-1.7	-16.0	16.9	29.7	91.8	88.9
		60	0.5	-0.1	-14.2	41.2	61.1	94.3	92.3
			1.0	-2.4	-15.3	29.2	48.1	93.8	91.6
			1.5	0.2	-13.4	23.6	38.1	93.9	91.3
			2.0	-1.0	-14.4	20.5	34.3	93.0	90.6
		80	0.5	-0.3	-12.8	51.9	78.6	94.5	93.1
			1.0	-0.3	-11.7	36.3	59.9	94.7	93.1
			1.5	-1.3	-12.8	29.3	49.3	94.2	92.0
			2.0	-0.4	-12.4	25.6	42.1	94.2	92.1
		95	0.5	-0.8	-11.5	64.3	99.0	94.7	94.4
			1.0	-1.9	-11.5	44.0	71.4	94.3	93.5
			1.5	-1.5	-11.9	35.5	58.8	94.6	93.2
			2.0	-0.8	-11.3	30.7	49.7	94.6	93.4

3.9 Discussion

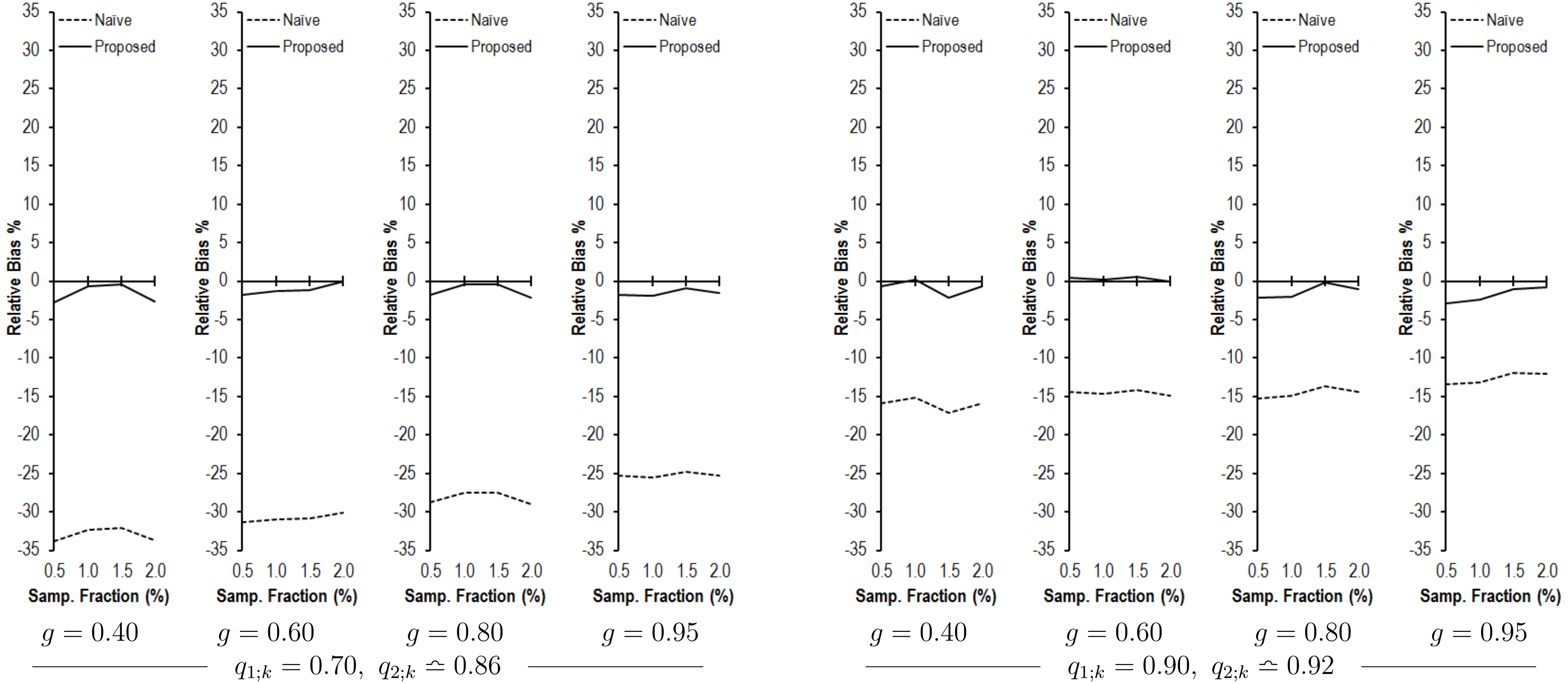
The proposed variance estimator is applicable for unequal rotation sampling designs when random Hot-deck imputation is used at both waves and the sampling fractions are negligible. The proposed variance estimator may be extended in various ways. Point estimators, such as calibration estimators (Huang and Fuller, 1978; Deville and Särndal, 1992) which employ auxiliary population information may often be expressible as functions of totals. The proposed variance estimator from Equation (3.18) can be modified to accommodate this situation.

The proposed approach is not limited to Hot-deck imputation, as it can be extended to other method of imputation, as long as the expectation of the imputed estimator of change under the random imputation method can be expressed as a function of totals. We have explored the use of single and multiple classes for imputation. At the second wave, it is a common practice to impute using observations of the first wave. It would be useful to generalise the proposed estimator for this method of imputation.

Appendix to Chapter 3

3.A. Figures of the simulation study

Relative Bias using $\pi_{1;k} = n/N$



$g = 0.40$ $g = 0.60$ $g = 0.80$ $g = 0.95$

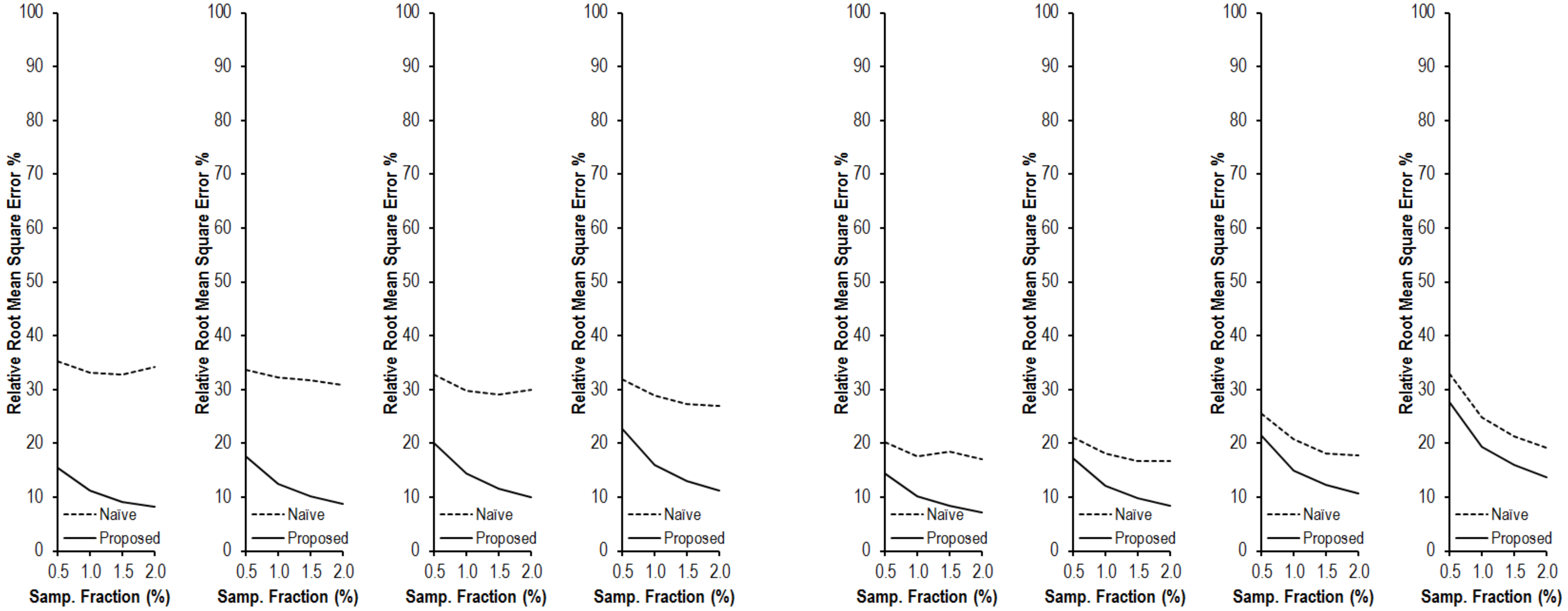
$q_{1;k} = 0.70,\ q_{2;k} \simeq 0.86$

$g = 0.40$ $g = 0.60$ $g = 0.80$ $g = 0.95$

$q_{1;k} = 0.90,\ q_{2;k} \simeq 0.92$

FIGURE 3.1: Relative Bias (%) of the variance estimators for the Hot-deck imputed point estimator $\hat{\Delta}^*$ using $\pi_{1;k} = n/N$.

Relative Root Mean-Square Error using $\pi_{1;k} = n/N$



$g = 0.40$ $g = 0.60$ $g = 0.80$ $g = 0.95$

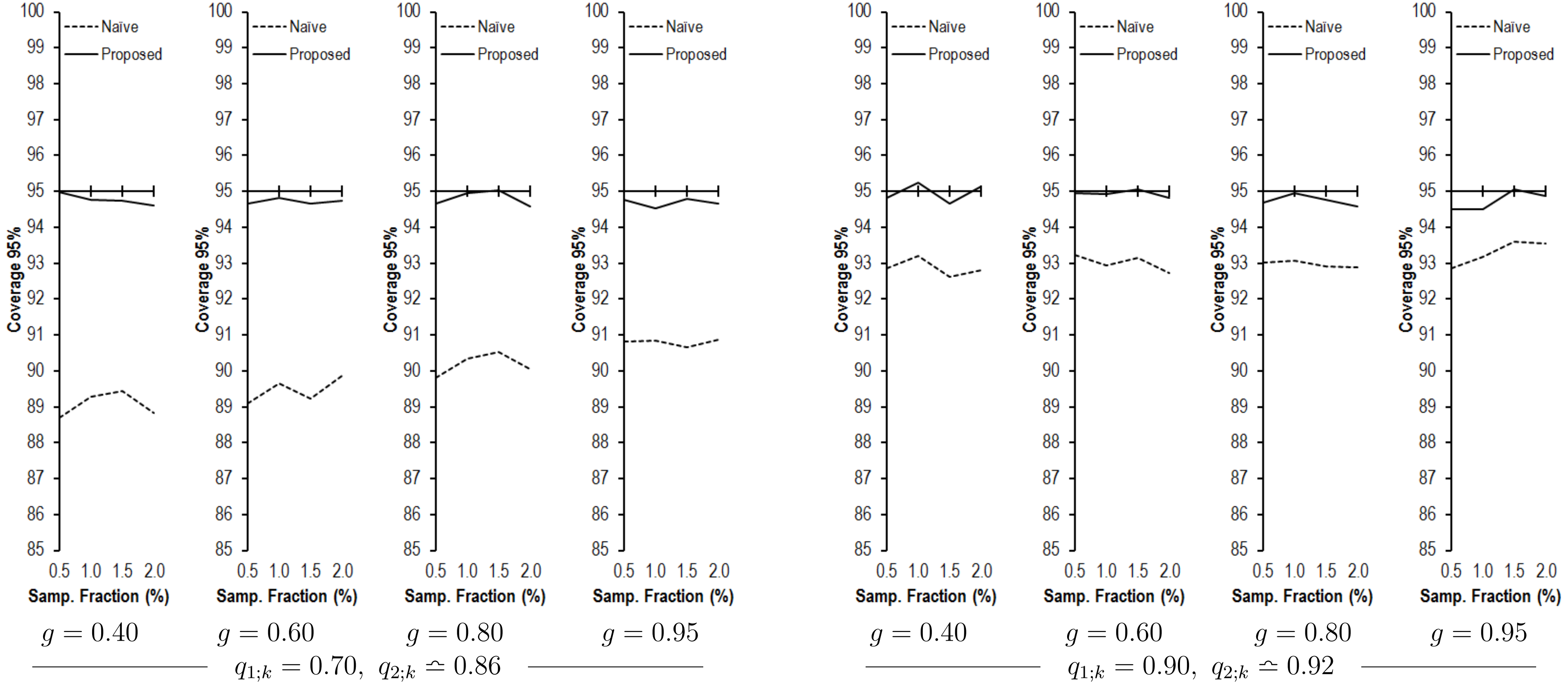
$q_{1;k} = 0.70, \; q_{2;k} \simeq 0.86$

$g = 0.40$ $g = 0.60$ $g = 0.80$ $g = 0.95$

$q_{1;k} = 0.90, \; q_{2;k} \simeq 0.92$

FIGURE 3.2: Relative Root Mean-Square Error (%) of the variance estimators for the Hot-deck imputed point estimator $\hat{\Delta}^*$ using $\pi_{1;k} = n/N$.

Coverage at 95% confidence level using $\pi_{1;k} = n/N$



$g = 0.40$ $g = 0.60$ $g = 0.80$ $g = 0.95$

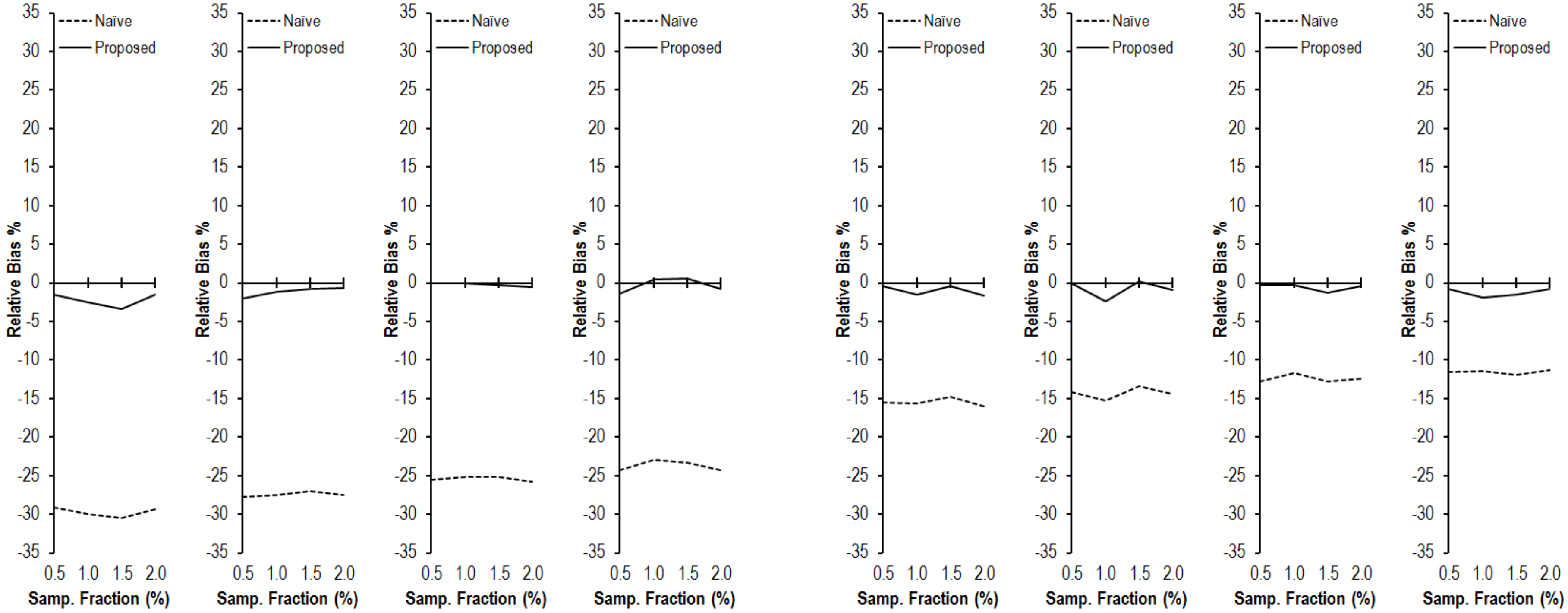
$q_{1;k} = 0.70,\ q_{2;k} \doteq 0.86$

$g = 0.40$ $g = 0.60$ $g = 0.80$ $g = 0.95$

$q_{1;k} = 0.90,\ q_{2;k} \doteq 0.92$

FIGURE 3.3: Coverage at 95% confidence level of the variance estimators for the Hot-deck imputed point estimator $\hat{\Delta}^*$ using $\pi_{1;k} = n/N$.

Relative Bias using $\pi_{1;k} \propto HW$



$g = 0.40$ $g = 0.60$ $g = 0.80$ $g = 0.95$

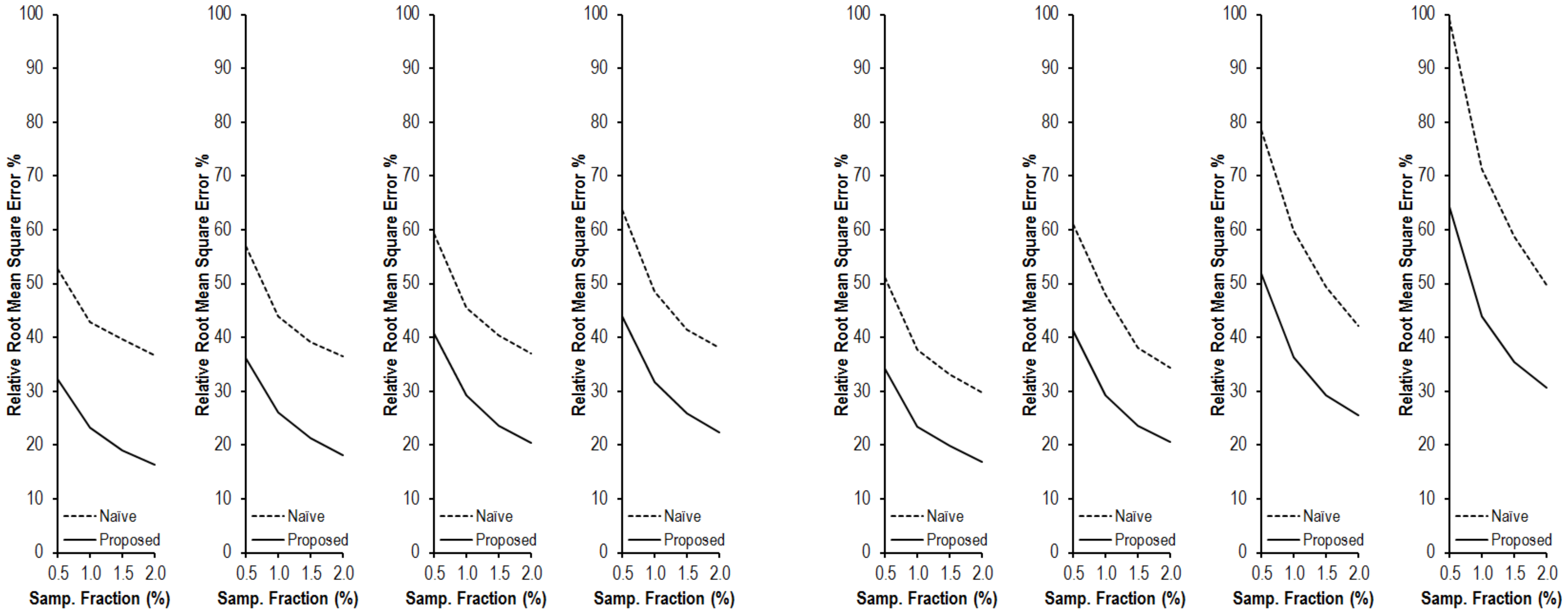
$q_{1;k} = 0.70,\ q_{2;k} \doteq 0.86$

$g = 0.40$ $g = 0.60$ $g = 0.80$ $g = 0.95$

$q_{1;k} = 0.90,\ q_{2;k} \doteq 0.92$

Figure 3.4: Relative Bias (%) of the variance estimators for the Hot-deck imputed point estimator $\hat{\Delta}^*$ using $\pi_{1;k} \propto HW$.

Relative Root Mean-Square Error using $\pi_{1;k} \propto HW$



$g = 0.40$ $g = 0.60$ $g = 0.80$ $g = 0.95$

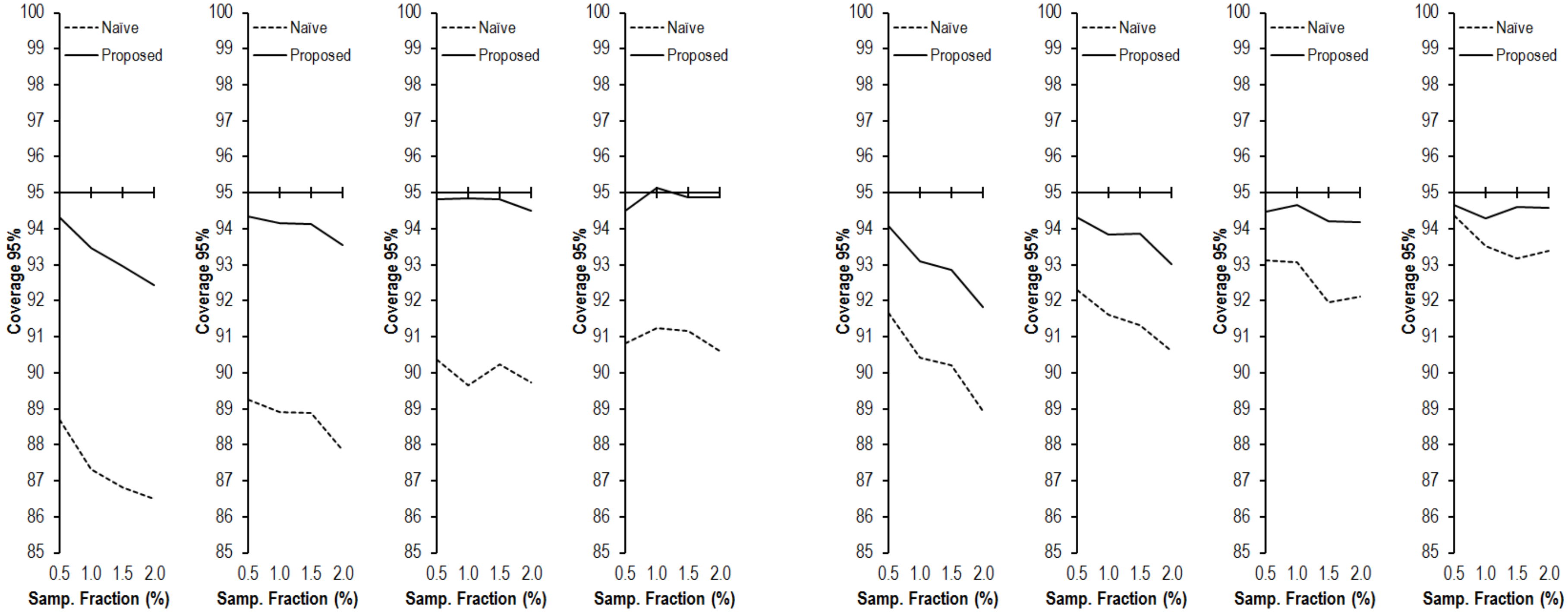
$q_{1;k} = 0.70,\ q_{2;k} \doteq 0.86$

$g = 0.40$ $g = 0.60$ $g = 0.80$ $g = 0.95$

$q_{1;k} = 0.90,\ q_{2;k} \doteq 0.92$

FIGURE 3.5: Relative Root Mean-Square Error (%) of the variance estimators for the Hot-deck imputed point estimator $\hat{\Delta}^*$ using $\pi_{1;k} \propto HW$.

Coverage at 95% confidence level using $\pi_{1;k} \propto HW$



$g = 0.40$ $g = 0.60$ $g = 0.80$ $g = 0.95$

$q_{1;k} = 0.70, \ q_{2;k} \simeq 0.86$

$g = 0.40$ $g = 0.60$ $g = 0.80$ $g = 0.95$

$q_{1;k} = 0.90, \ q_{2;k} \simeq 0.92$

FIGURE 3.6: Coverage at 95% confidence level of the variance estimators for the Hot-deck imputed point estimator $\hat{\Delta}^*$ using $\pi_{1;k} \propto HW$.

Chapter 4

Some R software implementations

Abstract

There is an increasing use of R software in many branches of Statistics. Moreover, there are R packages that comprise several other areas of knowledge and implementations beyond Statistics, e.g. Actuarial Science, Ecology and Finance. Here we introduce the *samplingVarEst* R package which incorporates some of the techniques utilised in earlier chapters of this thesis. The main purpose of creating an R package is to support other researchers interested in variance estimation. In our times, there is no doubt that open-source and freely-distributed software contributes towards dissemination and higher impact of research results.

Keywords and phrases: High-speed computing; variance estimation; R package.

4.1 The samplingVarEst R package

4.1.1 About the coding of the package

Routines for variance estimation are high-consuming in time and computing resources. To address this issue the package is mostly written in C compilable codes. These are later interfaced using R. All codes are now public at *The Comprehensive R Archive Network*, visit:

`http://cran.r-project.org/web/packages/samplingVarEst/`

4.1.2 About a description and a user's manual

To avoid repeating information here we do not describe the R package. A complete description is included in following pages as an appendix of this chapter. It is the *User's manual* of the samplingVarEst R package (version 0.9-1). It can be seen that the manual is self-contained and care has been taken in explaining with high detail each function.

The package is likely to suffer changes by the time this thesis is read. In fact, it is planned to be changing and updated continuously. We invite readers to check for the last version of the manual and, of course, the last version of the software.

Appendices to Chapter 4

4.A. User's manual of the samplingVarEst package

Package ‘samplingVarEst’

November 6, 2012

Version 0.9-1

Date 2012-09-20

Title Sampling Variance Estimation

Author Emilio Lopez Escobar, Ernesto Barrios Zamudio <ebarrios@itam.mx>

Maintainer Emilio Lopez Escobar <emilio.lopez@itam.mx>

Description Functions for estimating the sampling variance of some point estimators.

Classification/MSC 62D05, 62F40, 62G09, 62H12

Classification/JEL C130, C150, C420, C830

Classification/ACM G.3

Depends R (>= 2.10)

License GPL (>= 2)

Repository CRAN

Date/Publication 2012-11-06 07:01:21

R topics documented:

`samplingVarEst-package`

Sampling Variance Estimation package

Description

The package contains functions for estimating the variance of some point estimators under unequal-probability sampling. Emphasis has been put on the speed of routines. The package mostly uses C compilable code. The available functions are listed below matching: population parameters, point estimators and variance estimators.

parameters	**point estimators**
total:	`Est.Total.NHT`
	`Est.Total.Hajek`
mean:	`Est.Mean.NHT`
	`Est.Mean.Hajek`
ratio:	`Est.Ratio`
correlation coefficient:	`Est.Corr.NHT`
	`Est.Corr.Hajek`
regression coefficient:	`Est.RegCo.Hajek`

point estimators	**variance estimators** *(uni-stage samples)*
`Est.Total.NHT`:	`VE.HT.Total.NHT`
	`VE.SYG.Total.NHT`
	`VE.Hajek.Total.NHT`
`Est.Total.Hajek`:	`VE.Jk.Tukey.Total.Hajek`
	`VE.Jk.CBS.HT.Total.Hajek`
	`VE.Jk.CBS.SYG.Total.Hajek`
	`VE.Jk.B.Total.Hajek`
`Est.Mean.NHT`:	`VE.HT.Mean.NHT`
	`VE.SYG.Mean.NHT`
	`VE.Hajek.Mean.NHT`
`Est.Mean.Hajek`:	`VE.Jk.Tukey.Mean.Hajek`
	`VE.Jk.CBS.HT.Mean.Hajek`
	`VE.Jk.CBS.SYG.Mean.Hajek`
	`VE.Jk.B.Mean.Hajek`
`Est.Ratio`:	`VE.Jk.Tukey.Ratio`
	`VE.Jk.CBS.HT.Ratio`
	`VE.Jk.CBS.SYG.Ratio`
	`VE.Jk.B.Ratio`
`Est.Corr.NHT`:	`VE.Jk.Tukey.Corr.NHT`
`Est.Corr.Hajek`:	`VE.Jk.Tukey.Corr.Hajek`
	`VE.Jk.CBS.HT.Corr.Hajek`
	`VE.Jk.CBS.SYG.Corr.Hajek`
	`VE.Jk.B.Corr.Hajek`
`Est.RegCo.Hajek`:	`VE.Jk.Tukey.RegCo.Hajek`
	`VE.Jk.CBS.HT.RegCo.Hajek`
	`VE.Jk.CBS.SYG.RegCo.Hajek`
	`VE.Jk.B.RegCo.Hajek`

point estimators	**variance estimators** *(self-weighted two-stage samples)*
`Est.Total.Hajek`:	`VE.Jk.EB.SW2.Total.Hajek`
`Est.Mean.Hajek`:	`VE.Jk.EB.SW2.Mean.Hajek`
`Est.Ratio`:	`VE.Jk.EB.SW2.Ratio`
`Est.Corr.Hajek`:	`VE.Jk.EB.SW2.Corr.Hajek`
`Est.RegCo.Hajek`:	`VE.Jk.EB.SW2.RegCo.Hajek`

for the inclusion probabilities

1st order incl. probabilities:	Pk.PropNorm.U
2nd order (joint) incl. probs.:	Pkl.Hajek.s Pkl.Hajek.U

datasets

oaxaca

Details

To return to this description, type any time:
`help(samplingVarEst)`
To cite, use the given references or use:
`citation("samplingVarEst")`

Est.Corr.Hajek *Estimator of a correlation coefficient using the Hajek point estimator*

Description

Estimates a population correlation coefficient of two variables using the Hajek (1971) point estimator.

Usage

```
Est.Corr.Hajek(VecY.s, VecX.s, VecPk.s)
```

Arguments

`VecY.s`	vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be any missing value.
`VecX.s`	vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be any missing value.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.

Details

For the population correlation coefficient of two variables y and x:

$$C = \frac{\sum_{k\in U}(y_k - \bar{y})(x_k - \bar{x})}{\sqrt{\sum_{k\in U}(y_k - \bar{y})^2}\sqrt{\sum_{k\in U}(x_k - \bar{x})^2}}$$

the point estimator of C, assuming that N is unknown (see Sarndal et al., 1992, Sec. 5.9) (implemented by the current function), is:

$$\hat{C}_{Hajek} = \frac{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sqrt{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{Hajek})^2}\sqrt{\sum_{k\in s} w_k(x_k - \hat{\bar{x}}_{Hajek})^2}}$$

where $\hat{\bar{y}}_{Hajek}$ is the Hajek (1971) point estimator of the population mean $\bar{y} = N^{-1}\sum_{k\in U} y_k$,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s.

Value

The function returns a value for the correlation coefficient point estimator.

References

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

See Also

Est.Corr.NHT
VE.Jk.Tukey.Corr.Hajek
VE.Jk.CBS.HT.Corr.Hajek
VE.Jk.CBS.SYG.Corr.Hajek
VE.Jk.B.Corr.Hajek
VE.Jk.EB.SW2.Corr.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s     <- oaxaca$sHOMES00 #Defines the sample to be used for the example
y1    <- oaxaca$POP10    #Defines the variable of interest y1
y2    <- oaxaca$POPMAL10 #Defines the variable of interest y2
x     <- oaxaca$HOMES10  #Defines the variable of interest x
```

```
#Computes the correlation coefficient estimator for y1 and x
Est.Corr.Hajek(y1[s==1], x[s==1], pik.U[s==1])
#Computes the correlation coefficient estimator for y2 and x
Est.Corr.Hajek(y2[s==1], x[s==1], pik.U[s==1])
```

`Est.Corr.NHT` *Estimator of a correlation coefficient using the Narain-Horvitz-Thompson point estimator*

Description

Estimates a population correlation coefficient of two variables using the Narain (1951); Horvitz-Thompson (1952) point estimator.

Usage

```
Est.Corr.NHT(VecY.s, VecX.s, VecPk.s, N)
```

Arguments

`VecY.s` vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be any missing value.

`VecX.s` vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be any missing value.

`VecPk.s` vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.

`N` the population size.

Details

For the population correlation coefficient of two variables y and x:

$$C = \frac{\sum_{k\in U}(y_k - \bar{y})(x_k - \bar{x})}{\sqrt{\sum_{k\in U}(y_k - \bar{y})^2}\sqrt{\sum_{k\in U}(x_k - \bar{x})^2}}$$

the point estimator of C (implemented by the current function) is given by:

$$\hat{C} = \frac{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{NHT})(x_k - \hat{\bar{x}}_{NHT})}{\sqrt{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{NHT})^2}\sqrt{\sum_{k\in s} w_k(x_k - \hat{\bar{x}}_{NHT})^2}}$$

where $\hat{\bar{y}}_{NHT}$ is the Narain (1951); Horvitz-Thompson (1952) estimator for the population mean $\bar{y} = N^{-1}\sum_{k\in U} y_k$,

$$\hat{\bar{y}}_{NHT} = \frac{1}{N}\sum_{k\in s} w_k y_k$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s.

Value

The function returns a value for the correlation coefficient point estimator.

References

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

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

See Also

`Est.Corr.Hajek`
`VE.Jk.Tukey.Corr.NHT`

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s     <- oaxaca$sHOMES00 #Defines the sample to be used for the example
N     <- dim(oaxaca)[1]  #Defines the population size
y1    <- oaxaca$POP10    #Defines the variable of interest y1
y2    <- oaxaca$POPMAL10 #Defines the variable of interest y2
x     <- oaxaca$HOMES10  #Defines the variable of interest x
#Computes the correlation coefficient estimator for y1 and x
Est.Corr.NHT(y1[s==1], x[s==1], pik.U[s==1], N)
#Computes the correlation coefficient estimator for y2 and x
Est.Corr.NHT(y2[s==1], x[s==1], pik.U[s==1], N)
```

`Est.Mean.Hajek` *The Hajek estimator for a mean*

Description

Computes the Hajek (1971) estimator for a population mean.

Usage

```
Est.Mean.Hajek(VecY.s, VecPk.s)
```

Arguments

`VecY.s` vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s. There must not be any missing value.

`VecPk.s` vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.

Details

For the population mean of the variable y:

$$\bar{y} = \frac{1}{N}\sum_{k\in U} y_k$$

the approximately unbiased Hajek (1971) estimator of $\bar{y}$ (implemented by the current function) is given by:

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s.

Value

The function returns a value for the mean point estimator.

References

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

See Also

Est.Mean.NHT
VE.Jk.Tukey.Mean.Hajek
VE.Jk.CBS.HT.Mean.Hajek
VE.Jk.CBS.SYG.Mean.Hajek
VE.Jk.B.Mean.Hajek
VE.Jk.EB.SW2.Mean.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s     <- oaxaca$sHOMES00 #Defines the sample to be used for the example
y1    <- oaxaca$POP10    #Defines the variable of interest y1
y2    <- oaxaca$HOMES10  #Defines the variable of interest y2
Est.Mean.Hajek(y1[s==1], pik.U[s==1]) #Computes the Hajek est. for y1
Est.Mean.Hajek(y2[s==1], pik.U[s==1]) #Computes the Hajek est. for y2
```

`Est.Mean.NHT` *The Narain-Horvitz-Thompson estimator for a mean*

Description

Computes the Narain (1951); Horvitz-Thompson (1952) estimator for a population mean.

Usage

```
Est.Mean.NHT(VecY.s, VecPk.s, N)
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s. There must not be any missing value.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.
`N`	the population size.

Details

For the population mean of the variable y:

$$\bar{y} = \frac{1}{N}\sum_{k\in U} y_k$$

the unbiased Narain (1951); Horvitz-Thompson (1952) estimator of $\bar{y}$ (implemented by the current function) is given by:

$$\hat{\bar{y}}_{NHT} = \frac{1}{N}\sum_{k\in s}\frac{y_k}{\pi_k}$$

where π_k denotes the inclusion probability of the k-th element in the sample s.

Value

The function returns a value for the mean point estimator.

References

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

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

See Also

Est.Mean.Hajek
VE.HT.Mean.NHT
VE.SYG.Mean.NHT
VE.Hajek.Mean.NHT

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s     <- oaxaca$sHOMES00 #Defines the sample to be used for the example
N     <- dim(oaxaca)[1]  #Defines the population size
y1    <- oaxaca$POP10    #Defines the variable of interest y1
y2    <- oaxaca$HOMES10  #Defines the variable of interest y2
Est.Mean.NHT(y1[s==1], pik.U[s==1], N) #The NHT estimator for y1
Est.Mean.NHT(y2[s==1], pik.U[s==1], N) #The NHT estimator for y2
```

Est.Ratio *Estimator of a ratio*

Description

Estimates a population ratio of two totals/means.

Usage

```
Est.Ratio(VecY.s, VecX.s, VecPk.s)
```

Arguments

VecY.s	vector of the numerator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be any missing value.
VecX.s	vector of the denominator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be any missing value. All values of VecX.s must be greater than zero.
VecPk.s	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.

Details

For the population ratio of two totals/means of the variables y and x:

$$R = \frac{\sum_{k\in U} y_k/N}{\sum_{k\in U} x_k/N} = \frac{\sum_{k\in U} y_k}{\sum_{k\in U} x_k}$$

the ratio estimator of R (implemented by the current function) is given by:

$$\hat{R} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k x_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s.

Value

The function returns a value for the ratio point estimator.

References

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

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

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

See Also

VE.Jk.Tukey.Ratio
VE.Jk.CBS.HT.Ratio
VE.Jk.CBS.SYG.Ratio
VE.Jk.B.Ratio
VE.Jk.EB.SW2.Ratio

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s     <- oaxaca$sHOMES00 #Defines the sample to be used for the example
y1    <- oaxaca$POP10    #Defines the numerator variable of interest y1
y2    <- oaxaca$POPMAL10 #Defines the numerator variable of interest y2
x     <- oaxaca$HOMES10  #Defines the denominator variable of interest x
Est.Ratio(y1[s==1], x[s==1], pik.U[s==1]) #Ratio estimator for y1 and x
Est.Ratio(y2[s==1], x[s==1], pik.U[s==1]) #Ratio estimator for y2 and x
```

Est.RegCo.Hajek — *Estimator of the regression coefficient using the Hajek point estimator*

Description

Estimates the population regression coefficient using the Hajek (1971) point estimator.

Usage

```
Est.RegCo.Hajek(VecY.s, VecX.s, VecPk.s)
```

Arguments

`VecY.s` vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be any missing value.

`VecX.s` vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be any missing value.

`VecPk.s` vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.

Details

From Linear Regression Analysis, for an imposed population model

$$y = \alpha + \beta x$$

the population regression coefficient β, assuming that the population size N is unknown (see Sarndal et al., 1992, Sec. 5.9), can be estimated by:

$$\hat{\beta}_{Hajek} = \frac{\sum_{k\in s} w_k (y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sum_{k\in s} w_k (x_k - \hat{\bar{x}}_{Hajek})^2}$$

where $\hat{\bar{y}}_{Hajek}$ is the Hajek (1971) point estimator of the population mean $\bar{y} = N^{-1}\sum_{k\in U} y_k$,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s.

Value

The function returns a value for the regression coefficient point estimator.

References

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

See Also

VE.Jk.Tukey.RegCo.Hajek
VE.Jk.CBS.HT.RegCo.Hajek
VE.Jk.CBS.SYG.RegCo.Hajek
VE.Jk.B.RegCo.Hajek
VE.Jk.EB.SW2.RegCo.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s     <- oaxaca$sHOMES00 #Defines the sample to be used for the example
y1    <- oaxaca$POP10    #Defines the variable of interest y1
y2    <- oaxaca$POPMAL10 #Defines the variable of interest y2
x     <- oaxaca$HOMES10  #Defines the variable of interest x
#Computes the regression coefficient estimator for y1 and x
Est.RegCo.Hajek(y1[s==1], x[s==1], pik.U[s==1])
#Computes the regression coefficient estimator for y2 and x
Est.RegCo.Hajek(y2[s==1], x[s==1], pik.U[s==1])
```

Est.Total.Hajek *The Hajek estimator for a total*

Description

Computes the Hajek (1971) estimator for a population total.

Usage

```
Est.Total.Hajek(VecY.s, VecPk.s, N)
```

Arguments

VecY.s	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s. There must not be any missing value.
VecPk.s	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.
N	the population size.

Details

For the population total of the variable y:

$$t = \sum_{k \in U} y_k$$

the approximately unbiased Hajek (1971) estimator of t (implemented by the current function) is given by:

$$\hat{t}_{Hajek} = N \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s.

Value

The function returns a value for the total point estimator.

References

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

See Also

Est.Total.NHT
VE.Jk.Tukey.Total.Hajek
VE.Jk.CBS.HT.Total.Hajek
VE.Jk.CBS.SYG.Total.Hajek
VE.Jk.B.Total.Hajek
VE.Jk.EB.SW2.Total.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s     <- oaxaca$sHOMES00 #Defines the sample to be used for the example
N     <- dim(oaxaca)[1]  #Defines the population size
y1    <- oaxaca$POP10    #Defines the variable of interest y1
y2    <- oaxaca$HOMES10  #Defines the variable of interest y2
Est.Total.Hajek(y1[s==1], pik.U[s==1], N) #The Hajek estimator for y1
Est.Total.Hajek(y2[s==1], pik.U[s==1], N) #The Hajek estimator for y2
```

`Est.Total.NHT` *The Narain-Horvitz-Thompson estimator for a total*

Description

Computes the Narain (1951); Horvitz-Thompson (1952) estimator for a population total.

Usage

```
Est.Total.NHT(VecY.s, VecPk.s)
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s. There must not be any missing value.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.

Details

For the population total of the variable y:

$$t = \sum_{k \in U} y_k$$

the unbiased Narain (1951); Horvitz-Thompson (1952) estimator of t (implemented by the current function) is given by:

$$\hat{t}_{NHT} = \sum_{k \in s} \frac{y_k}{\pi_k}$$

where π_k denotes the inclusion probability of the k-th element in the sample s.

Value

The function returns a value for the total point estimator.

References

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

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

See Also

Est.Total.Hajek
VE.HT.Total.NHT
VE.SYG.Total.NHT
VE.Hajek.Total.NHT

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s     <- oaxaca$sHOMES00 #Defines the sample to be used for the example
y1    <- oaxaca$POP10    #Defines the variable of interest y1
y2    <- oaxaca$HOMES10  #Defines the variable of interest y2
Est.Total.NHT(y1[s==1], pik.U[s==1]) #Computes the NHT estimator for y1
Est.Total.NHT(y2[s==1], pik.U[s==1]) #Computes the NHT estimator for y2
```

oaxaca *Municipalities of the state of Oaxaca in Mexico*

Description

Dataset with information about the free and sovereign state of Oaxaca which is located in the south part of Mexico. The dataset contains information of population, surface, indigenous language, agriculture and income from years ranging from 2000 to 2010. The information was originally collected and processed by the Mexico's National Institute of Statistics and Geography (INEGI by its name in Spanish, 'Instituto Nacional de Estadistica y Geografia', http://www.inegi.org.mx/).

Usage

```
data(oaxaca)
```

Format

A data frame with 570 observations on the following 41 variables:

IDREGION region INEGI code.

LBREGION region name (without accents and Spanish language characters).

IDDISTRI district INEGI code.

LBDISTRI district name (without accents and Spanish language characters).

IDMUNICI municipality INEGI code.

LBMUNICI municipality name (without accents and Spanish language characters).

SURFAC05 surface in squared kilometres 2005.

POP00 population 2000.

POP10 population 2010.

HOMES00 number of homes 2000.

HOMES10 number of homes 2010.

POPMAL00 male population 2000.

POPMAL10 male population 2010.

POPFEM00 female population 2000.

POPFEM10 female population 2010.

INLANG00 5 or more years old population which speaks indigenous language 2000.

INLANG10 5 or more years old population which speaks indigenous language 2010.

INCOME00 gross income in thousands of Mexican pesos 2000.

INCOME01 gross income in thousands of Mexican pesos 2001.

INCOME02 gross income in thousands of Mexican pesos 2002.

INCOME03 gross income in thousands of Mexican pesos 2003.

PTREES00 planted trees 2000.

PTREES01 planted trees 2001.

PTREES02 planted trees 2002.

PTREES03 planted trees 2003.

MARRIA07 marriages 2007.

MARRIA08 marriages 2008.

MARRIA09 marriages 2009.

HARVBE07 harvested bean surface in hectares 2007.

HARVBE08 harvested bean surface in hectares 2008.

HARVBE09 harvested bean surface in hectares 2009.

VALUBE07 value of bean production in thousands of Mexican pesos 2007.

VALUBE08 value of bean production in thousands of Mexican pesos 2008.

VALUBE09 value of bean production in thousands of Mexican pesos 2009.

VOLUBE07 volume of bean production in tons 2007.

VOLUBE08 volume of bean production in tons 2008.

VOLUBE09 volume of bean production in tons 2009.

sHOMES00 a sample (column vector of ones and zeros; 1 = selected, 0 = otherwise) of 373 municipalities drawn using the Hajek (1964) maximum-entropy sampling design with inclusion probabilities proportional to the variable HOMES00.

sSURFAC a sample (column vector of ones and zeros; 1 = selected, 0 = otherwise) of 373 municipalities drawn using the Hajek (1964) maximum-entropy sampling design with inclusion probabilities proportional to the variable SURFAC05.

SIZEDIST the size of the district, i.e. the number of municipalities in each district.

sSW_10_3 a sample (column vector of ones and zeros; 1 = selected, 0 = otherwise) of 30 municipalities drawn using a self-weighted two-stage sampling design. The first stage draws 10 districts using the Hajek (1964) maximum-entropy sampling design with clusters' inclusion probabilities proportional to the size of the clusters (variable SIZEDIST). The second stage draws 3 municipalities within the selected districts at the first stage, using equal-probability without-replacement sampling.

Source

Mexico's National Institute of Statistics and Geography (INEGI), 'Instituto Nacional de Estadistica y Geografia' http://www3.inegi.org.mx/sistemas/descarga/

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Computes the INCOME00 mean (note that INCOME00 has NA's)
mean(oaxaca$INCOME00, na.rm= TRUE)
#Computes the INCOME00 median (note that INCOME00 has NA's)
median(oaxaca$INCOME00, na.rm= TRUE)
```

`Pk.PropNorm.U` *Inclusion probabilities proportional to a specified variable.*

Description

Creates and normalises the 1st order inclusion probabilities proportional to a specified variable. In the current context, normalisation means that the inclusion probabilities are less than or equal to 1. Ideally, they should sum up to n, the sample size.

Usage

```
Pk.PropNorm.U(n, VecMOS.U)
```

Arguments

`n`	the sample size.
`VecMOS.U`	vector of the variable called measure of size (MOS) to which the first-order inclusion probabilities are to be proportional; its length is equal to the population size. Values in VecMOS.U should be greater than zero (a warning message appears if this does not hold). There must not be any missing value.

Details

Although the normalisation procedure is well-known in the survey sampling literature, we follow the procedure described in Chao (1982, p. 654). Hence, we obtain a unique set of inclusion probabilities that are proportional to the MOS variable.

Value

The function returns a vector of length n with the inclusion probabilities.

References

Chao, M. T. (1982) A general purpose unequal probability sampling plan. *Biometrika* **69**, 653–656.

See Also

Pkl.Hajek.s
Pkl.Hajek.U

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Creates the normalised 1st order incl. probs. proportional
#to the variable oaxaca$HOMES00 and with sample size 373
pik.U       <- Pk.PropNorm.U(373, oaxaca$HOMES00)
sum(pik.U)   #Shows the sum is equal to the sample size 373
any(pik.U>1) #Shows there isn't any probability greater than 1
any(pik.U<0) #Shows there isn't any probability less than 0
```

Pkl.Hajek.s *The Hajek approximation for the 2nd order (joint) inclusion probabilities (sample based)*

Description

Computes the Hajek (1964) approximation for the 2nd order (joint) inclusion probabilities utilising only sample-based quantities.

Usage

```
Pkl.Hajek.s(VecPk.s)
```

Arguments

VecPk.s	vector of the first-order inclusion probabilities; its length is equal to the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.

Details

Let π_k denote the inclusion probability of the k-th element in the sample s, and let π_{kl} denote the joint-inclusion probabilities of the k-th and l-th elements in the sample s. If the joint-inclusion probabilities π_{kl} are not available, the Hajek (1964) approximation can be used. Note that this approximation is designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

The sample based version of the Hajek (1964) approximation for the joint-inclusion probabilities π_{kl} (implemented by the current function) is:

$$\pi_{kl} \doteq \pi_k\pi_l\{1 - \hat{d}^{-1}(1-\pi_k)(1-\pi_l)\}$$

where $\hat{d} = \sum_{k\in s}(1-\pi_k)$.

The approximation was originally developed for $d \to \infty$, under the maximum-entropy sampling design (see Hajek 1981, Theorem 3.3, Ch. 3 and 6), the Rejective Sampling design. It requires that the utilised sampling design be of large entropy. An overview can be found in Berger and Tille (2009). An account of different sampling designs, π_{kl} approximations, and approximate variances under large-entropy designs can be found in Tille (2006), Brewer and Donadio (2003), and Haziza, Mecatti, and Rao (2008). Recently, Berger (2011) gave sufficient conditions under which Hajek's results still hold for large-entropy sampling designs that are not the maximum-entropy one.

Value

The function returns a (n by n) square matrix with the estimated joint inclusion probabilities, where n is the sample size.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Berger, Y. G. (2011) Asymptotic consistency under large entropy sampling designs with unequal probabilities. *Pakistan Journal of Statististics*, **27**, 407–426.

Berger, Y. G. and Tille, Y. (2009) Sampling with unequal probabilities. In *Sample Surveys: Design, Methods and Applications* (eds. D. Pfeffermann and C. R. Rao), 39–54. Elsevier, Amsterdam.

Brewer, K. R. W. and Donadio, M. E. (2003) The large entropy variance of the Horvitz-Thompson estimator. *Survey Methodology* **29**, 189–196.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

Hajek, J. (1981) *Sampling From a Finite Population.* Dekker, New York.

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

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

See Also

Pkl.Hajek.U
Pk.PropNorm.U

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s      <- oaxaca$sHOMES00 #Defines the sample to be used for the example
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1]) #Approx. 2nd order incl. probs. from s
#First 5 rows/cols of (sample based) 2nd order incl. probs. matrix
pikl.s[1:5,1:5]
```

Pkl.Hajek.U *The Hajek approximation for the 2nd order (joint) inclusion probabilities (population based)*

Description

Computes the Hajek (1964) approximation for the 2nd order (joint) inclusion probabilities utilising population-based quantities.

Usage

```
Pkl.Hajek.U(VecPk.U)
```

Arguments

`VecPk.U`	vector of the first-order inclusion probabilities; its length is equal to the population size. Values in VecPk.U must be greater than zero and less than or equal to one. There must not be any missing value.

Details

Let π_k denote the inclusion probability of the k-th element in the sample s, and let π_{kl} denote the joint-inclusion probabilities of the k-th and l-th elements in the sample s. If the joint-inclusion probabilities π_{kl} are not available, the Hajek (1964) approximation can be used. Note that this approximation is designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

The population based version of the Hajek (1964) approximation for the joint-inclusion probabilities π_{kl} (implemented by the current function) is:

$$\pi_{kl} \doteq \pi_k \pi_l \{1 - d^{-1}(1-\pi_k)(1-\pi_l)\}$$

where $d = \sum_{k \in U} \pi_k(1-\pi_k)$.

The approximation was originally developed for $d \to \infty$, under the maximum-entropy sampling design (see Hajek 1981, Theorem 3.3, Ch. 3 and 6), the Rejective Sampling design. It requires that the utilised sampling design be of large entropy. An overview can be found in Berger and Tille (2009). An account of different sampling designs, π_{kl} approximations, and approximate variances under large-entropy designs can be found in Tille (2006), Brewer and Donadio (2003), and Haziza, Mecatti, and Rao (2008). Recently, Berger (2011) gave sufficient conditions under which Hajek's results still hold for large-entropy sampling designs that are not the maximum-entropy one.

Value

The function returns a (N by N) square matrix with the estimated joint inclusion probabilities, where N is the population size.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Berger, Y. G. (2011) Asymptotic consistency under large entropy sampling designs with unequal probabilities. *Pakistan Journal of Statististics*, **27**, 407–426.

Berger, Y. G. and Tille, Y. (2009) Sampling with unequal probabilities. In *Sample Surveys: Design, Methods and Applications* (eds. D. Pfeffermann and C. R. Rao), 39–54. Elsevier, Amsterdam.

Brewer, K. R. W. and Donadio, M. E. (2003) The large entropy variance of the Horvitz-Thompson estimator. *Survey Methodology* **29**, 189–196.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

Hajek, J. (1981) *Sampling From a Finite Population.* Dekker, New York.

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

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

See Also

Pkl.Hajek.s
Pk.PropNorm.U

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00)
#(This approximation is only suitable for large-entropy sampling designs)
pikl.U <- Pkl.Hajek.U(pik.U) #Approximates 2nd order incl. probs. from U
#First 5 rows/cols of (population based) 2nd order incl. probs. matrix
pikl.U[1:5,1:5]
```

VE.Hajek.Mean.NHT *The Hajek variance estimator for the Narain-Horvitz-Thompson point estimator for a mean*

Description

Computes the Hajek (1964) variance estimator for the Narain (1951); Horvitz-Thompson (1952) point estimator for a population mean.

Usage

```
VE.Hajek.Mean.NHT(VecY.s, VecPk.s, N)
```

Arguments

`VecY.s` vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s. There must not be any missing value.

`VecPk.s` vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.

`N` the population size.

Details

For the population mean of the variable y:

$$\bar{y} = \frac{1}{N} \sum_{k \in U} y_k$$

the unbiased Narain (1951); Horvitz-Thompson (1952) estimator of $\bar{y}$ is given by:

$$\hat{\bar{y}}_{NHT} = \frac{1}{N} \sum_{k \in s} \frac{y_k}{\pi_k}$$

where π_k denotes the inclusion probability of the k-th element in the sample s. For large-entropy sampling designs, the variance of $\hat{\bar{y}}_{NHT}$ is approximated by the Hajek (1964) variance:

$$V(\hat{\bar{y}}_{NHT}) = \frac{1}{N(N-1)} \left[\sum_{k \in U} \frac{y_k^2}{\pi_k}(1-\pi_k) - dG^2 \right]$$

with $d = \sum_{k \in U} \pi_k(1-\pi_k)$ and $G = d^{-1} \sum_{k \in U}(1-\pi_k)y_k$.

The variance $V(\hat{t}_{NHT})$ can be estimated by the variance estimator (implemented by the current function):

$$\hat{V}(\hat{\bar{y}}_{NHT}) = \frac{n}{N^2(n-1)} \left[\sum_{k \in s} \left(\frac{y_k}{\pi_k} \right)^2 (1-\pi_k) - \hat{d}\hat{G}^2 \right]$$

where $\hat{d} = \sum_{k \in s}(1-\pi_k)$ and $\hat{G} = \hat{d}^{-1} \sum_{k \in s}(1-\pi) y_k/\pi_k$.

Note that the Hajek (1964) variance approximation is designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

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

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

See Also

`VE.HT.Mean.NHT`
`VE.SYG.Mean.NHT`

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s     <- oaxaca$sHOMES00 #Defines the sample to be used for the example
N     <- dim(oaxaca)[1]  #Defines the population size
y1    <- oaxaca$POP10    #Defines the variable of interest y1
y2    <- oaxaca$HOMES10  #Defines the variable of interest y2
#Computes the (approximate) var. est. of the NHT point est. for y1
VE.Hajek.Mean.NHT(y1[s==1], pik.U[s==1], N)
#Computes the (approximate) var. est. of the NHT point est. for y2
VE.Hajek.Mean.NHT(y2[s==1], pik.U[s==1], N)
```

`VE.Hajek.Total.NHT` *The Hajek variance estimator for the Narain-Horvitz-Thompson point estimator for a total*

Description

Computes the Hajek (1964) variance estimator for the Narain (1951); Horvitz-Thompson (1952) point estimator for a population total.

Usage

```
VE.Hajek.Total.NHT(VecY.s, VecPk.s)
```

Arguments

`VecY.s` vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s. There must not be any missing value.

`VecPk.s` vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.

Details

For the population total of the variable y:

$$t = \sum_{k \in U} y_k$$

the unbiased Narain (1951); Horvitz-Thompson (1952) estimator of t is given by:

$$\hat{t}_{NHT} = \sum_{k \in s} \frac{y_k}{\pi_k}$$

where π_k denotes the inclusion probability of the k-th element in the sample s. For large-entropy sampling designs, the variance of $\hat{t}_{NHT}$ is approximated by the Hajek (1964) variance:

$$V(\hat{t}_{NHT}) = \frac{N}{N-1} \left[\sum_{k \in U} \frac{y_k^2}{\pi_k} (1 - \pi_k) - dG^2 \right]$$

with $d = \sum_{k \in U} \pi_k (1 - \pi_k)$ and $G = d^{-1} \sum_{k \in U} (1 - \pi_k) y_k$.

The variance $V(\hat{t}_{NHT})$ can be estimated by the variance estimator (implemented by the current function):

$$\hat{V}(\hat{t}_{NHT}) = \frac{n}{n-1} \left[\sum_{k \in s} \left(\frac{y_k}{\pi_k} \right)^2 (1 - \pi_k) - \hat{d}\hat{G}^2 \right]$$

where $\hat{d} = \sum_{k \in s} (1 - \pi_k)$ and $\hat{G} = \hat{d}^{-1} \sum_{k \in s} (1 - \pi) y_k / \pi_k$.

Note that the Hajek (1964) variance approximation is designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

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

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

See Also

`VE.HT.Total.NHT`
`VE.SYG.Total.NHT`

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U <- Pk.PropNorm.U(373, oaxaca$SURFAC05)
s     <- oaxaca$sSURFAC #Defines the sample to be used for the example
y1    <- oaxaca$POP10   #Defines the variable of interest y1
y2    <- oaxaca$HOMES10 #Defines the variable of interest y2
#Computes the (approximate) var. est. of the NHT point est. from y1
VE.Hajek.Total.NHT(y1[s==1], pik.U[s==1])
#Computes the (approximate) var. est. of the NHT point est. from y2
VE.Hajek.Total.NHT(y2[s==1], pik.U[s==1])
```

`VE.HT.Mean.NHT` *The Horvitz-Thompson variance estimator for the Narain-Horvitz-Thompson point estimator for a mean*

Description

Computes the Horvitz-Thompson (1952) variance estimator for the Narain (1951); Horvitz-Thompson (1952) point estimator for a population mean.

Usage

```
VE.HT.Mean.NHT(VecY.s, VecPk.s, MatPkl.s, N)
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s. There must not be any missing value.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.
`MatPkl.s`	matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in MatPkl.s must be greater than zero and less than or equal to one. There must not be any missing value.
`N`	the population size.

Details

For the population mean of the variable y:

$$\bar{y} = \frac{1}{N}\sum_{k\in U} y_k$$

the unbiased Narain (1951); Horvitz-Thompson (1952) estimator of $\bar{y}$ is given by:

$$\hat{\bar{y}}_{NHT} = \frac{1}{N}\sum_{k\in s}\frac{y_k}{\pi_k}$$

where π_k denotes the inclusion probability of the k-th element in the sample s. Let π_{kl} denotes the joint-inclusion probabilities of the k-th and l-th elements in the sample s. The variance of $\hat{\bar{y}}_{NHT}$ is given by:

$$V(\hat{\bar{y}}_{NHT}) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}(\pi_{kl} - \pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

which can therefore be estimated by the Horvitz-Thompson variance estimator (implemented by the current function):

$$\hat{V}(\hat{\bar{y}}_{NHT}) = \frac{1}{N^2}\sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

Value

The function returns a value for the estimated variance.

References

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

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

See Also

VE.SYG.Mean.NHT
VE.Hajek.Mean.NHT

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U  <- Pk.PropNorm.U(373, oaxaca$SURFAC05)
s      <- oaxaca$sSURFAC #Defines the sample to be used for the example
N      <- dim(oaxaca)[1] #Defines the population size
y1     <- oaxaca$POP10   #Defines the variable of interest y1
y2     <- oaxaca$HOMES10 #Defines the variable of interest y2
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1]) #Approx. 2nd order incl. probs. from s
#Computes the variance estimation of the NHT point estimator for y1
VE.HT.Mean.NHT(y1[s==1], pik.U[s==1], pikl.s, N)
#Computes the variance estimation of the NHT point estimator for y2
VE.HT.Mean.NHT(y2[s==1], pik.U[s==1], pikl.s, N)
```

`VE.HT.Total.NHT` *The Horvitz-Thompson variance estimator for the Narain-Horvitz-Thompson point estimator for a total*

Description

Computes the Horvitz-Thompson (1952) variance estimator for the Narain (1951); Horvitz-Thompson (1952) point estimator for a population total.

Usage

```
VE.HT.Total.NHT(VecY.s, VecPk.s, MatPkl.s)
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s. There must not be any missing value.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.
`MatPkl.s`	matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in MatPkl.s must be greater than zero and less than or equal to one. There must not be any missing value.

Details

For the population total of the variable y:

$$t = \sum_{k \in U} y_k$$

the unbiased Narain (1951); Horvitz-Thompson (1952) estimator of t is given by:

$$\hat{t}_{NHT} = \sum_{k \in s} \frac{y_k}{\pi_k}$$

where π_k denotes the inclusion probability of the k-th element in the sample s. Let π_{kl} denotes the joint-inclusion probabilities of the k-th and l-th elements in the sample s. The variance of $\hat{t}_{NHT}$ is given by:

$$V(\hat{t}_{NHT}) = \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

which can therefore be estimated by the Horvitz-Thompson variance estimator (implemented by the current function):

$$\hat{V}(\hat{t}_{NHT}) = \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

Value

The function returns a value for the estimated variance.

References

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

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

See Also

`VE.SYG.Total.NHT`
`VE.Hajek.Total.NHT`

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s      <- oaxaca$sHOMES00 #Defines the sample to be used for the example
y1     <- oaxaca$POP10    #Defines the variable of interest y1
y2     <- oaxaca$HOMES10  #Defines the variable of interest y2
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1]) #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the NHT point estimator for y1
VE.HT.Total.NHT(y1[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the NHT point estimator for y2
VE.HT.Total.NHT(y2[s==1], pik.U[s==1], pikl.s)
```

`VE.Jk.B.Corr.Hajek` *The Berger (2007) unequal probability jackknife variance estimator for the estimator of a correlation coefficient using the Hajek point estimator*

Description

Computes the Berger (2007) unequal probability jackknife variance estimator for the estimator of a correlation coefficient of two variables using the Hajek (1971) point estimator.

Usage

```
VE.Jk.B.Corr.Hajek(VecY.s, VecX.s, VecPk.s)
```

Arguments

`VecY.s` vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be any missing value.

`VecX.s` vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be any missing value.

`VecPk.s` vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.

Details

For the population correlation coefficient of two variables y and x:

$$C = \frac{\sum_{k \in U}(y_k - \bar{y})(x_k - \bar{x})}{\sqrt{\sum_{k \in U}(y_k - \bar{y})^2}\sqrt{\sum_{k \in U}(x_k - \bar{x})^2}}$$

the point estimator of C, assuming that N is unknown (see Sarndal et al., 1992, Sec. 5.9), is:

$$\hat{C}_{Hajek} = \frac{\sum_{k \in s} w_k(y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sqrt{\sum_{k \in s} w_k(y_k - \hat{\bar{y}}_{Hajek})^2}\sqrt{\sum_{k \in s} w_k(x_k - \hat{\bar{x}}_{Hajek})^2}}$$

where $\hat{\bar{y}}_{Hajek}$ is the Hajek (1971) point estimator of the population mean $\bar{y} = N^{-1}\sum_{k \in U} y_k$,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. The variance of $\hat{C}_{Hajek}$ can be estimated by the Berger (2007) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{C}_{Hajek}) = \sum_{k \in s} \frac{n}{n-1}(1 - \pi_k)\left(\varepsilon_k - \hat{B}\right)^2$$

where

$$\hat{B} = \frac{\sum_{k \in s}(1 - \pi_k)\varepsilon_k}{\sum_{k \in s}(1 - \pi_k)}$$

and

$$\varepsilon_k = (1 - \tilde{w}_k)\left(\hat{C}_{Hajek} - \hat{C}_{Hajek(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l \in s} w_l}$$

and where $\hat{C}_{Hajek(k)}$ has the same functional form as $\hat{C}_{Hajek}$ but omitting the k-th element from the sample s. Note that this variance estimator utilises implicitly the Hajek (1964) approximations that are designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Berger, Y. G. (2007) A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika* **94**, 953–964.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

See Also

VE.Jk.Tukey.Corr.Hajek
VE.Jk.CBS.HT.Corr.Hajek
VE.Jk.CBS.SYG.Corr.Hajek
VE.Jk.EB.SW2.Corr.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s      <- oaxaca$sHOMES00 #Defines the sample to be used for the example
y1     <- oaxaca$POP10    #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10 #Defines the variable of interest y2
x      <- oaxaca$HOMES10  #Defines the variable of interest x
#Computes the var. est. of the corr. coeff. point estimator using y1
VE.Jk.B.Corr.Hajek(y1[s==1], x[s==1], pik.U[s==1])
#Computes the var. est. of the corr. coeff. point estimator using y2
VE.Jk.B.Corr.Hajek(y2[s==1], x[s==1], pik.U[s==1])
```

VE.Jk.B.Mean.Hajek *The Berger (2007) unequal probability jackknife variance estimator for the Hajek estimator of a mean*

Description

Computes the Berger (2007) unequal probability jackknife variance estimator for the Hajek (1971) estimator of a mean.

Usage

```
VE.Jk.B.Mean.Hajek(VecY.s, VecPk.s)
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s. There must not be any missing value.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.

Details

For the population mean of the variable y:

$$\bar{y} = \frac{1}{N} \sum_{k \in U} y_k$$

the approximately unbiased Hajek (1971) estimator of $\bar{y}$ is given by:

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{\bar{y}}_{Hajek}$ can be estimated by the Berger (2007) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\bar{y}}_{Hajek}) = \sum_{k \in s} \frac{n}{n-1}(1 - \pi_k)\left(\varepsilon_k - \hat{B}\right)^2$$

where

$$\hat{B} = \frac{\sum_{k \in s}(1 - \pi_k)\varepsilon_k}{\sum_{k \in s}(1 - \pi_k)}$$

and

$$\varepsilon_k = (1 - \tilde{w}_k)\left(\hat{\bar{y}}_{Hajek} - \hat{\bar{y}}_{Hajek(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l \in s} w_l}$$

and

$$\hat{\bar{y}}_{Hajek(k)} = \frac{\sum_{l \in s, l \neq k} w_l y_l}{\sum_{l \in s, l \neq k} w_l}$$

Note that this variance estimator utilises implicitly the Hajek (1964) approximations that are designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Berger, Y. G. (2007) A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika* **94**, 953–964.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

See Also

VE.Jk.Tukey.Mean.Hajek
VE.Jk.CBS.HT.Mean.Hajek
VE.Jk.CBS.SYG.Mean.Hajek
VE.Jk.EB.SW2.Mean.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s      <- oaxaca$sHOMES00 #Defines the sample to be used for the example
y1     <- oaxaca$POP10    #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10 #Defines the variable of interest y2
#Computes the var. est. of the Hajek mean point estimator using y1
VE.Jk.B.Mean.Hajek(y1[s==1], pik.U[s==1])
#Computes the var. est. of the Hajek mean point estimator using y2
VE.Jk.B.Mean.Hajek(y2[s==1], pik.U[s==1])
```

VE.Jk.B.Ratio — *The Berger (2007) unequal probability jackknife variance estimator for the estimator of a ratio*

Description

Computes the Berger (2007) unequal probability jackknife variance estimator for the estimator of a ratio of two totals/means.

Usage

```
VE.Jk.B.Ratio(VecY.s, VecX.s, VecPk.s)
```

Arguments

`VecY.s`	vector of the numerator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be any missing value.
`VecX.s`	vector of the denominator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be any missing value. All values of VecX.s must be greater than zero.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.

Details

For the population ratio of two totals/means of the variables y and x:

$$R = \frac{\sum_{k \in U} y_k / N}{\sum_{k \in U} x_k / N} = \frac{\sum_{k \in U} y_k}{\sum_{k \in U} x_k}$$

the ratio estimator of R is given by:

$$\hat{R} = \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k x_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{R}$ can be estimated by the Berger (2007) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{R}) = \sum_{k \in s} \frac{n}{n-1} (1 - \pi_k) \left(\varepsilon_k - \hat{B} \right)^2$$

where

$$\hat{B} = \frac{\sum_{k \in s} (1 - \pi_k) \varepsilon_k}{\sum_{k \in s} (1 - \pi_k)}$$

and

$$\varepsilon_k = (1 - \tilde{w}_k) \left(\hat{R} - \hat{R}_{(k)} \right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l \in s} w_l}$$

and

$$\hat{R}_{(k)} = \frac{\sum_{l \in s, l \neq k} w_l y_l / \sum_{l \in s, l \neq k} w_l}{\sum_{l \in s, l \neq k} w_l x_l / \sum_{l \in s, l \neq k} w_l} = \frac{\sum_{l \in s, l \neq k} w_l y_l}{\sum_{l \in s, l \neq k} w_l x_l}$$

Note that this variance estimator utilises implicitly the Hajek (1964) approximations that are designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Berger, Y. G. (2007) A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika* **94**, 953–964.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

See Also

VE.Jk.Tukey.Ratio
VE.Jk.CBS.HT.Ratio
VE.Jk.CBS.SYG.Ratio
VE.Jk.EB.SW2.Ratio

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s      <- oaxaca$sHOMES00 #Defines the sample to be used for the example
y1     <- oaxaca$POP10    #Defines the numerator variable of interest y1
y2     <- oaxaca$POPMAL10 #Defines the numerator variable of interest y2
x      <- oaxaca$HOMES10  #Defines the denominator variable of interest x
#Computes the var. est. of the ratio point estimator using y1
VE.Jk.B.Ratio(y1[s==1], x[s==1], pik.U[s==1])
#Computes the var. est. of the ratio point estimator using y2
VE.Jk.B.Ratio(y2[s==1], x[s==1], pik.U[s==1])
```

VE.Jk.B.RegCo.Hajek *The Berger (2007) unequal probability jackknife variance estimator for the estimator of the regression coefficient using the Hajek point estimator*

Description

Computes the Berger (2007) unequal probability jackknife variance estimator for the estimator of the regression coefficient using the Hajek (1971) point estimator.

Usage

```
VE.Jk.B.RegCo.Hajek(VecY.s, VecX.s, VecPk.s)
```

Arguments

`VecY.s` vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be any missing value.

`VecX.s` vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be any missing value.

`VecPk.s` vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.

Details

From Linear Regression Analysis, for an imposed population model

$$y = \alpha + \beta x$$

the population regression coefficient β, assuming that the population size N is unknown (see Sarndal et al., 1992, Sec. 5.9), can be estimated by:

$$\hat{\beta}_{Hajek} = \frac{\sum_{k\in s} w_k (y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sum_{k\in s} w_k (x_k - \hat{\bar{x}}_{Hajek})^2}$$

where $\hat{\bar{y}}_{Hajek}$ is the Hajek (1971) point estimator of the population mean $\bar{y} = N^{-1}\sum_{k\in U} y_k$,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. The variance of $\hat{\beta}_{Hajek}$ can be estimated by the Berger (2007) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\beta}_{Hajek}) = \sum_{k\in s} \frac{n}{n-1}(1-\pi_k)\left(\varepsilon_k - \hat{B}\right)^2$$

where

$$\hat{B} = \frac{\sum_{k\in s}(1-\pi_k)\varepsilon_k}{\sum_{k\in s}(1-\pi_k)}$$

and

$$\varepsilon_k = (1-\tilde{w}_k)\left(\hat{\beta}_{Hajek} - \hat{\beta}_{Hajek(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l\in s} w_l}$$

and where $\hat{\beta}_{Hajek(k)}$ has the same functional form as $\hat{\beta}_{Hajek}$ but omitting the k-th element from the sample s. Note that this variance estimator utilises implicitly the Hajek (1964) approximations that are designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Berger, Y. G. (2007) A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika* **94**, 953–964.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

See Also

VE.Jk.Tukey.RegCo.Hajek
VE.Jk.CBS.HT.RegCo.Hajek
VE.Jk.CBS.SYG.RegCo.Hajek
VE.Jk.EB.SW2.RegCo.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s      <- oaxaca$sHOMES00 #Defines the sample to be used for the example
y1     <- oaxaca$POP10    #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10 #Defines the variable of interest y2
x      <- oaxaca$HOMES10  #Defines the variable of interest x
#Computes the var. est. of the regression coeff. point estimator using y1
VE.Jk.B.RegCo.Hajek(y1[s==1], x[s==1], pik.U[s==1])
#Computes the var. est. of the regression coeff. point estimator using y2
VE.Jk.B.RegCo.Hajek(y2[s==1], x[s==1], pik.U[s==1])
```

`VE.Jk.B.Total.Hajek` *The Berger (2007) unequal probability jackknife variance estimator for the Hajek estimator of a total*

Description

Computes the Berger (2007) unequal probability jackknife variance estimator for the Hajek (1971) estimator of a total.

Usage

```
VE.Jk.B.Total.Hajek(VecY.s, VecPk.s, N)
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s. There must not be any missing value.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.
`N`	the population size.

Details

For the population total of the variable y:

$$t = \sum_{k\in U} y_k$$

the approximately unbiased Hajek (1971) estimator of t (implemented by the current function) is given by:

$$\hat{t}_{Hajek} = N\frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{t}_{Hajek}$ can be estimated by the Berger (2007) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{t}_{Hajek}) = \sum_{k\in s}\frac{n}{n-1}(1-\pi_k)\left(\varepsilon_k - \hat{B}\right)^2$$

where

$$\hat{B} = \frac{\sum_{k\in s}(1-\pi_k)\varepsilon_k}{\sum_{k\in s}(1-\pi_k)}$$

and

$$\varepsilon_k = (1-\tilde{w}_k)\left(\hat{t}_{Hajek} - \hat{t}_{Hajek(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l\in s} w_l}$$

and

$$\hat{t}_{Hajek(k)} = N\frac{\sum_{l\in s,l\neq k} w_l y_l}{\sum_{l\in s,l\neq k} w_l}$$

Note that this variance estimator utilises implicitly the Hajek (1964) approximations that are designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Berger, Y. G. (2007) A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika* **94**, 953–964.

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

See Also

VE.Jk.Tukey.Total.Hajek
VE.Jk.CBS.HT.Total.Hajek
VE.Jk.CBS.SYG.Total.Hajek
VE.Jk.EB.SW2.Total.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s      <- oaxaca$sHOMES00 #Defines the sample to be used for the example
N      <- dim(oaxaca)[1]  #Defines the population size
y1     <- oaxaca$POP10    #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10 #Defines the variable of interest y2
#Computes the var. est. of the Hajek total point estimator using y1
VE.Jk.B.Total.Hajek(y1[s==1], pik.U[s==1], N)
#Computes the var. est. of the Hajek total point estimator using y2
VE.Jk.B.Total.Hajek(y2[s==1], pik.U[s==1], N)
```

VE.Jk.CBS.HT.Corr.Hajek

The Campbell-Berger-Skinner unequal probability jackknife variance estimator for the estimator of a correlation coefficient using the Hajek point estimator (Horvitz-Thompson form)

Description

Computes the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator for the estimator of a correlation coefficient of two variables using the Hajek (1971) point estimator. It uses the Horvitz-Thompson (1952) variance form.

Usage

```
VE.Jk.CBS.HT.Corr.Hajek(VecY.s, VecX.s, VecPk.s, MatPkl.s)
```

Arguments

`VecY.s` vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be any missing value.

`VecX.s` vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be any missing value.

`VecPk.s` vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.

`MatPkl.s` matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in MatPkl.s must be greater than zero and less than or equal to one. There must not be any missing value.

Details

For the population correlation coefficient of two variables y and x:

$$C = \frac{\sum_{k\in U}(y_k-\bar{y})(x_k-\bar{x})}{\sqrt{\sum_{k\in U}(y_k-\bar{y})^2}\sqrt{\sum_{k\in U}(x_k-\bar{x})^2}}$$

the point estimator of C, assuming that N is unknown (see Sarndal et al., 1992, Sec. 5.9), is:

$$\hat{C}_{Hajek} = \frac{\sum_{k\in s} w_k(y_k-\hat{\bar{y}}_{Hajek})(x_k-\hat{\bar{x}}_{Hajek})}{\sqrt{\sum_{k\in s} w_k(y_k-\hat{\bar{y}}_{Hajek})^2}\sqrt{\sum_{k\in s} w_k(x_k-\hat{\bar{x}}_{Hajek})^2}}$$

where $\hat{\bar{y}}_{Hajek}$ is the Hajek (1971) point estimator of the population mean $\bar{y} = N^{-1}\sum_{k\in U} y_k$,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. The variance of $\hat{C}_{Hajek}$ can be estimated by the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{C}_{Hajek}) = \sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}\varepsilon_k\varepsilon_l$$

where

$$\varepsilon_k = (1-\tilde{w}_k)\left(\hat{C}_{Hajek}-\hat{C}_{Hajek(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l\in s} w_l}$$

and where $\hat{C}_{Hajek(k)}$ has the same functional form as $\hat{C}_{Hajek}$ but omitting the k-th element from the sample s.

Value

The function returns a value for the estimated variance.

References

Campbell, C. (1980) A different view of finite population estimation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 319–324.

Berger, Y. G. and Skinner, C. J. (2005) A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, **67**, 79–89.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

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

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

See Also

VE.Jk.Tukey.Corr.Hajek
VE.Jk.CBS.SYG.Corr.Hajek
VE.Jk.B.Corr.Hajek
VE.Jk.EB.SW2.Corr.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s      <- oaxaca$sHOMES00 #Defines the sample to be used for the example
y1     <- oaxaca$POP10    #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10 #Defines the variable of interest y2
x      <- oaxaca$HOMES10  #Defines the variable of interest x
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1]) #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the corr. coeff. point estimator using y1
VE.Jk.CBS.HT.Corr.Hajek(y1[s==1], x[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the corr. coeff. point estimator using y2
VE.Jk.CBS.HT.Corr.Hajek(y2[s==1], x[s==1], pik.U[s==1], pikl.s)
```

VE.Jk.CBS.HT.Mean.Hajek

The Campbell-Berger-Skinner unequal probability jackknife variance estimator for the Hajek (1971) estimator of a mean (Horvitz-Thompson form)

Description

Computes the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator for the Hajek estimator of a mean. It uses the Horvitz-Thompson (1952) variance form.

Usage

```
VE.Jk.CBS.HT.Mean.Hajek(VecY.s, VecPk.s, MatPkl.s)
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s. There must not be any missing value.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.
`MatPkl.s`	matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in MatPkl.s must be greater than zero and less than or equal to one. There must not be any missing value.

Details

For the population mean of the variable y:

$$\bar{y} = \frac{1}{N}\sum_{k\in U} y_k$$

the approximately unbiased Hajek (1971) estimator of $\bar{y}$ is given by:

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{\bar{y}}_{Hajek}$ can be estimated by the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\bar{y}}_{Hajek}) = \sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}\varepsilon_k\varepsilon_l$$

where

$$\varepsilon_k = (1-\tilde{w}_k)\left(\hat{\bar{y}}_{Hajek} - \hat{\bar{y}}_{Hajek(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l\in s} w_l}$$

and

$$\hat{\bar{y}}_{Hajek(k)} = \frac{\sum_{l\in s, l\neq k} w_l y_l}{\sum_{l\in s, l\neq k} w_l}$$

Value

The function returns a value for the estimated variance.

References

Campbell, C. (1980) A different view of finite population estimation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 319–324.

Berger, Y. G. and Skinner, C. J. (2005) A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, **67**, 79–89.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

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

See Also

VE.Jk.Tukey.Mean.Hajek
VE.Jk.CBS.SYG.Mean.Hajek
VE.Jk.B.Mean.Hajek
VE.Jk.EB.SW2.Mean.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s      <- oaxaca$sHOMES00 #Defines the sample to be used for the example
y1     <- oaxaca$POP10    #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10 #Defines the variable of interest y2
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1]) #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the Hajek mean point estimator using y1
VE.Jk.CBS.HT.Mean.Hajek(y1[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the Hajek mean point estimator using y2
VE.Jk.CBS.HT.Mean.Hajek(y2[s==1], pik.U[s==1], pikl.s)
```

VE.Jk.CBS.HT.Ratio — *The Campbell-Berger-Skinner unequal probability jackknife variance estimator for the estimator of a ratio (Horvitz-Thompson form)*

Description

Computes the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator for the estimator of a ratio of two totals/means. It uses the Horvitz-Thompson (1952) variance form.

Usage

```
VE.Jk.CBS.HT.Ratio(VecY.s, VecX.s, VecPk.s, MatPkl.s)
```

Arguments

`VecY.s`	vector of the numerator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be any missing value.
`VecX.s`	vector of the denominator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be any missing value. All values of VecX.s must be greater than zero.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.
`MatPkl.s`	matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in MatPkl.s must be greater than zero and less than or equal to one. There must not be any missing value.

Details

For the population ratio of two totals/means of the variables y and x:

$$R = \frac{\sum_{k\in U} y_k/N}{\sum_{k\in U} x_k/N} = \frac{\sum_{k\in U} y_k}{\sum_{k\in U} x_k}$$

the ratio estimator of R is given by:

$$\hat{R} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k x_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{R}$ can be estimated by the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{R}) = \sum_{k\in s}\sum_{l\in s} \frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\varepsilon_k\varepsilon_l$$

where

$$\varepsilon_k = (1 - \tilde{w}_k)\left(\hat{R} - \hat{R}_{(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l\in s} w_l}$$

and

$$\hat{R}_{(k)} = \frac{\sum_{l\in s,l\neq k} w_l y_l / \sum_{l\in s,l\neq k} w_l}{\sum_{l\in s,l\neq k} w_l x_l / \sum_{l\in s,l\neq k} w_l} = \frac{\sum_{l\in s,l\neq k} w_l y_l}{\sum_{l\in s,l\neq k} w_l x_l}$$

Value

The function returns a value for the estimated variance.

References

Campbell, C. (1980) A different view of finite population estimation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 319–324.

Berger, Y. G. and Skinner, C. J. (2005) A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, **67**, 79–89.

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

See Also

VE.Jk.Tukey.Ratio
VE.Jk.CBS.SYG.Ratio
VE.Jk.B.Ratio
VE.Jk.EB.SW2.Ratio

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s      <- oaxaca$sHOMES00 #Defines the sample to be used for the example
y1     <- oaxaca$POP10    #Defines the numerator variable of interest y1
y2     <- oaxaca$POPMAL10 #Defines the numerator variable of interest y2
x      <- oaxaca$HOMES10  #Defines the denominator variable of interest x
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1]) #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the ratio point estimator using y1
VE.Jk.CBS.HT.Ratio(y1[s==1], x[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the ratio point estimator using y2
VE.Jk.CBS.HT.Ratio(y2[s==1], x[s==1], pik.U[s==1], pikl.s)
```

VE.Jk.CBS.HT.RegCo.Hajek

The Campbell-Berger-Skinner unequal probability jackknife variance estimator for the estimator of the regression coefficient using the Hajek point estimator (Horvitz-Thompson form)

Description

Computes the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator for the estimator of the regression coefficient using the Hajek (1971) point estimator. It uses the Horvitz-Thompson (1952) variance form.

Usage

```
VE.Jk.CBS.HT.RegCo.Hajek(VecY.s, VecX.s, VecPk.s, MatPkl.s)
```

Arguments

`VecY.s` vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be any missing value.

`VecX.s` vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be any missing value.

`VecPk.s` vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.

`MatPkl.s` matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in MatPkl.s must be greater than zero and less than or equal to one. There must not be any missing value.

Details

From Linear Regression Analysis, for an imposed population model

$$y = \alpha + \beta x$$

the population regression coefficient β, assuming that the population size N is unknown (see Sarndal et al., 1992, Sec. 5.9), can be estimated by:

$$\hat{\beta}_{Hajek} = \frac{\sum_{k\in s} w_k (y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sum_{k\in s} w_k (x_k - \hat{\bar{x}}_{Hajek})^2}$$

where $\hat{\bar{y}}_{Hajek}$ is the Hajek (1971) point estimator of the population mean $\bar{y} = N^{-1}\sum_{k\in U} y_k$,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. The variance of $\hat{\beta}_{Hajek}$ can be estimated by the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\beta}_{Hajek}) = \sum_{k\in s}\sum_{l\in s} \frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} \varepsilon_k \varepsilon_l$$

where

$$\varepsilon_k = (1 - \tilde{w}_k)\left(\hat{\beta}_{Hajek} - \hat{\beta}_{Hajek(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l\in s} w_l}$$

and where $\hat{\beta}_{Hajek(k)}$ has the same functional form as $\hat{\beta}_{Hajek}$ but omitting the k-th element from the sample s.

Value

The function returns a value for the estimated variance.

References

Campbell, C. (1980) A different view of finite population estimation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 319–324.

Berger, Y. G. and Skinner, C. J. (2005) A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, **67**, 79–89.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

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

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

See Also

VE.Jk.Tukey.RegCo.Hajek
VE.Jk.CBS.SYG.RegCo.Hajek
VE.Jk.B.RegCo.Hajek
VE.Jk.EB.SW2.RegCo.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s      <- oaxaca$sHOMES00 #Defines the sample to be used for the example
y1     <- oaxaca$POP10    #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10 #Defines the variable of interest y2
x      <- oaxaca$HOMES10  #Defines the variable of interest x
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1]) #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the regression coeff. point estimator using y1
VE.Jk.CBS.HT.RegCo.Hajek(y1[s==1], x[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the regression coeff. point estimator using y2
VE.Jk.CBS.HT.RegCo.Hajek(y2[s==1], x[s==1], pik.U[s==1], pikl.s)
```

VE.Jk.CBS.HT.Total.Hajek

The Campbell-Berger-Skinner unequal probability jackknife variance estimator for the Hajek (1971) estimator of a total (Horvitz-Thompson form)

Description

Computes the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator for the Hajek estimator of a total. It uses the Horvitz-Thompson (1952) variance form.

Usage

```
VE.Jk.CBS.HT.Total.Hajek(VecY.s, VecPk.s, MatPkl.s, N)
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s. There must not be any missing value.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.
`MatPkl.s`	matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in MatPkl.s must be greater than zero and less than or equal to one. There must not be any missing value.
`N`	the population size.

Details

For the population total of the variable y:

$$t = \sum_{k\in U} y_k$$

the approximately unbiased Hajek (1971) estimator of t (implemented by the current function) is given by:

$$\hat{t}_{Hajek} = N\frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{t}_{Hajek}$ can be estimated by the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{t}_{Hajek}) = \sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\varepsilon_k\varepsilon_l$$

where

$$\varepsilon_k = (1 - \tilde{w}_k)\left(\hat{t}_{Hajek} - \hat{t}_{Hajek(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l\in s} w_l}$$

and

$$\hat{t}_{Hajek(k)} = N\frac{\sum_{l\in s, l\neq k} w_l y_l}{\sum_{l\in s, l\neq k} w_l}$$

Value

The function returns a value for the estimated variance.

References

Campbell, C. (1980) A different view of finite population estimation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 319–324.

Berger, Y. G. and Skinner, C. J. (2005) A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, **67**, 79–89.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

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

See Also

VE.Jk.Tukey.Total.Hajek
VE.Jk.CBS.SYG.Total.Hajek
VE.Jk.B.Total.Hajek
VE.Jk.EB.SW2.Total.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s      <- oaxaca$sHOMES00 #Defines the sample to be used for the example
N      <- dim(oaxaca)[1]  #Defines the population size
y1     <- oaxaca$POP10    #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10 #Defines the variable of interest y2
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1]) #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the Hajek total point estimator using y1
VE.Jk.CBS.HT.Total.Hajek(y1[s==1], pik.U[s==1], pikl.s, N)
#Computes the var. est. of the Hajek total point estimator using y2
VE.Jk.CBS.HT.Total.Hajek(y2[s==1], pik.U[s==1], pikl.s, N)
```

VE.Jk.CBS.SYG.Corr.Hajek

The Campbell-Berger-Skinner unequal probability jackknife variance estimator for the estimator of a correlation coefficient using the Hajek point estimator (Sen-Yates-Grundy form)

Description

Computes the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator for the estimator of a correlation coefficient of two variables using the Hajek (1971) point estimator. It uses the Sen (1953); Yates-Grundy(1953) variance form.

Usage

```
VE.Jk.CBS.SYG.Corr.Hajek(VecY.s, VecX.s, VecPk.s, MatPkl.s)
```

Arguments

`VecY.s` vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be any missing value.

`VecX.s` vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be any missing value.

`VecPk.s` vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.

`MatPkl.s` matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in MatPkl.s must be greater than zero and less than or equal to one. There must not be any missing value.

Details

For the population correlation coefficient of two variables y and x:

$$C = \frac{\sum_{k\in U}(y_k - \bar{y})(x_k - \bar{x})}{\sqrt{\sum_{k\in U}(y_k - \bar{y})^2}\sqrt{\sum_{k\in U}(x_k - \bar{x})^2}}$$

the point estimator of C, assuming that N is unknown (see Sarndal et al., 1992, Sec. 5.9), is:

$$\hat{C}_{Hajek} = \frac{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sqrt{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{Hajek})^2}\sqrt{\sum_{k\in s} w_k(x_k - \hat{\bar{x}}_{Hajek})^2}}$$

where $\hat{\bar{y}}_{Hajek}$ is the Hajek (1971) point estimator of the population mean $\bar{y} = N^{-1}\sum_{k\in U} y_k$,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. The variance of $\hat{C}_{Hajek}$ can be estimated by the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{C}_{Hajek}) = \frac{-1}{2}\sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}(\varepsilon_k - \varepsilon_l)^2$$

where

$$\varepsilon_k = (1 - \tilde{w}_k)\left(\hat{C}_{Hajek} - \hat{C}_{Hajek(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l\in s} w_l}$$

and where $\hat{C}_{Hajek(k)}$ has the same functional form as $\hat{C}_{Hajek}$ but omitting the k-th element from the sample s. The Sen-Yates-Grundy form for the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator is proposed in Escobar-Berger (2013) under less-restrictive regularity conditions.

Value

The function returns a value for the estimated variance.

References

Campbell, C. (1980) A different view of finite population estimation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 319–324.

Berger, Y. G. and Skinner, C. J. (2005) A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, **67**, 79–89.

Escobar, E. L. and Berger, Y. G. (2013) A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica* (to appear).

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

Sen, A. R. (1953) On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **5**, 119–127.

Yates, F. and Grundy, P. M. (1953) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, **15**, 253–261.

See Also

VE.Jk.Tukey.Corr.Hajek
VE.Jk.CBS.HT.Corr.Hajek
VE.Jk.B.Corr.Hajek
VE.Jk.EB.SW2.Corr.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s      <- oaxaca$sHOMES00 #Defines the sample to be used for the example
y1     <- oaxaca$POP10    #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10 #Defines the variable of interest y2
x      <- oaxaca$HOMES10  #Defines the variable of interest x
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1]) #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the corr. coeff. point estimator using y1
VE.Jk.CBS.SYG.Corr.Hajek(y1[s==1], x[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the corr. coeff. point estimator using y2
VE.Jk.CBS.SYG.Corr.Hajek(y2[s==1], x[s==1], pik.U[s==1], pikl.s)
```

`VE.Jk.CBS.SYG.Mean.Hajek`

The Campbell-Berger-Skinner unequal probability jackknife variance estimator for the Hajek (1971) estimator of a mean (Sen-Yates-Grundy form)

Description

Computes the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator for the Hajek estimator of a mean. It uses the Sen (1953); Yates-Grundy(1953) variance form.

Usage

```
VE.Jk.CBS.SYG.Mean.Hajek(VecY.s, VecPk.s, MatPkl.s)
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s. There must not be any missing value.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.
`MatPkl.s`	matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in MatPkl.s must be greater than zero and less than or equal to one. There must not be any missing value.

Details

For the population mean of the variable y:

$$\bar{y} = \frac{1}{N}\sum_{k\in U} y_k$$

the approximately unbiased Hajek (1971) estimator of $\bar{y}$ is given by:

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{\bar{y}}_{Hajek}$ can be estimated by the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\bar{y}}_{Hajek}) = \frac{-1}{2}\sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}(\varepsilon_k-\varepsilon_l)^2$$

where

$$\varepsilon_k = (1 - \tilde{w}_k)\left(\hat{\bar{y}}_{Hajek} - \hat{\bar{y}}_{Hajek(k)}\right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l \in s} w_l}$$

and

$$\hat{\bar{y}}_{Hajek(k)} = \frac{\sum_{l \in s, l \neq k} w_l y_l}{\sum_{l \in s, l \neq k} w_l}$$

The Sen-Yates-Grundy form for the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator is proposed in Escobar-Berger (2013) under less-restrictive regularity conditions.

Value

The function returns a value for the estimated variance.

References

Campbell, C. (1980) A different view of finite population estimation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 319–324.

Berger, Y. G. and Skinner, C. J. (2005) A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, **67**, 79–89.

Escobar, E. L. and Berger, Y. G. (2013) A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica* (to appear).

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sen, A. R. (1953) On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **5**, 119–127.

Yates, F. and Grundy, P. M. (1953) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, **15**, 253–261.

See Also

VE.Jk.Tukey.Mean.Hajek
VE.Jk.CBS.HT.Mean.Hajek
VE.Jk.B.Mean.Hajek
VE.Jk.EB.SW2.Mean.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s      <- oaxaca$sHOMES00 #Defines the sample to be used for the example
y1     <- oaxaca$POP10    #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10 #Defines the variable of interest y2
```

```
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1]) #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the Hajek mean point estimator using y1
VE.Jk.CBS.SYG.Mean.Hajek(y1[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the Hajek mean point estimator using y2
VE.Jk.CBS.SYG.Mean.Hajek(y2[s==1], pik.U[s==1], pikl.s)
```

VE.Jk.CBS.SYG.Ratio *The Campbell-Berger-Skinner unequal probability jackknife variance estimator for the estimator of a ratio (Sen-Yates-Grundy form)*

Description

Computes the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator for the estimator of a ratio of two totals/means. It uses the Sen (1953); Yates-Grundy(1953) variance form.

Usage

```
VE.Jk.CBS.SYG.Ratio(VecY.s, VecX.s, VecPk.s, MatPkl.s)
```

Arguments

VecY.s	vector of the numerator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be any missing value.
VecX.s	vector of the denominator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be any missing value. All values of VecX.s must be greater than zero.
VecPk.s	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.
MatPkl.s	matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in MatPkl.s must be greater than zero and less than or equal to one. There must not be any missing value.

Details

For the population ratio of two totals/means of the variables y and x:

$$R = \frac{\sum_{k\in U} y_k/N}{\sum_{k\in U} x_k/N} = \frac{\sum_{k\in U} y_k}{\sum_{k\in U} x_k}$$

the ratio estimator of R is given by:

$$\hat{R} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k x_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{R}$ can be estimated by the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{R}) = \frac{-1}{2} \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} (\varepsilon_k - \varepsilon_l)^2$$

where

$$\varepsilon_k = (1 - \tilde{w}_k) \left(\hat{R} - \hat{R}_{(k)} \right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l \in s} w_l}$$

and

$$\hat{R}_{(k)} = \frac{\sum_{l \in s, l \neq k} w_l y_l / \sum_{l \in s, l \neq k} w_l}{\sum_{l \in s, l \neq k} w_l x_l / \sum_{l \in s, l \neq k} w_l} = \frac{\sum_{l \in s, l \neq k} w_l y_l}{\sum_{l \in s, l \neq k} w_l x_l}$$

The Sen-Yates-Grundy form for the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator is proposed in Escobar-Berger (2013) under less-restrictive regularity conditions.

Value

The function returns a value for the estimated variance.

References

Campbell, C. (1980) A different view of finite population estimation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 319–324.

Berger, Y. G. and Skinner, C. J. (2005) A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, **67**, 79–89.

Escobar, E. L. and Berger, Y. G. (2013) A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica* (to appear).

Sen, A. R. (1953) On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **5**, 119–127.

Yates, F. and Grundy, P. M. (1953) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, **15**, 253–261.

See Also

`VE.Jk.Tukey.Ratio`
`VE.Jk.CBS.HT.Ratio`
`VE.Jk.B.Ratio`
`VE.Jk.EB.SW2.Ratio`

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s      <- oaxaca$sHOMES00 #Defines the sample to be used for the example
y1     <- oaxaca$POP10    #Defines the numerator variable of interest y1
y2     <- oaxaca$POPMAL10 #Defines the numerator variable of interest y2
x      <- oaxaca$HOMES10  #Defines the denominator variable of interest x
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1]) #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the ratio point estimator using y1
VE.Jk.CBS.SYG.Ratio(y1[s==1], x[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the ratio point estimator using y2
VE.Jk.CBS.SYG.Ratio(y2[s==1], x[s==1], pik.U[s==1], pikl.s)
```

VE.Jk.CBS.SYG.RegCo.Hajek

The Campbell-Berger-Skinner unequal probability jackknife variance estimator for the estimator of the regression coefficient using the Hajek point estimator (Sen-Yates-Grundy form)

Description

Computes the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator for the estimator of the regression coefficient using the Hajek (1971) point estimator. It uses the Sen (1953); Yates-Grundy(1953) variance form.

Usage

```
VE.Jk.CBS.SYG.RegCo.Hajek(VecY.s, VecX.s, VecPk.s, MatPkl.s)
```

Arguments

VecY.s	vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be any missing value.
VecX.s	vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be any missing value.
VecPk.s	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.
MatPkl.s	matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in MatPkl.s must be greater than zero and less than or equal to one. There must not be any missing value.

Details

From Linear Regression Analysis, for an imposed population model

$$y = \alpha + \beta x$$

the population regression coefficient β, assuming that the population size N is unknown (see Sarndal et al., 1992, Sec. 5.9), can be estimated by:

$$\hat{\beta}_{Hajek} = \frac{\sum_{k \in s} w_k (y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sum_{k \in s} w_k (x_k - \hat{\bar{x}}_{Hajek})^2}$$

where $\hat{\bar{y}}_{Hajek}$ is the Hajek (1971) point estimator of the population mean $\bar{y} = N^{-1} \sum_{k \in U} y_k$,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. The variance of $\hat{\beta}_{Hajek}$ can be estimated by the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\beta}_{Hajek}) = \frac{-1}{2} \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} (\varepsilon_k - \varepsilon_l)^2$$

where

$$\varepsilon_k = (1 - \tilde{w}_k) \left(\hat{\beta}_{Hajek} - \hat{\beta}_{Hajek(k)} \right)$$

with

$$\tilde{w}_k = \frac{w_k}{\sum_{l \in s} w_l}$$

and where $\hat{\beta}_{Hajek(k)}$ has the same functional form as $\hat{\beta}_{Hajek}$ but omitting the k-th element from the sample s. The Sen-Yates-Grundy form for the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator is proposed in Escobar-Berger (2013) under less-restrictive regularity conditions.

Value

The function returns a value for the estimated variance.

References

Campbell, C. (1980) A different view of finite population estimation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 319–324.

Berger, Y. G. and Skinner, C. J. (2005) A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, **67**, 79–89.

Escobar, E. L. and Berger, Y. G. (2013) A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica* (to appear).

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

Sen, A. R. (1953) On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **5**, 119–127.

Yates, F. and Grundy, P. M. (1953) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, **15**, 253–261.

See Also

VE.Jk.Tukey.RegCo.Hajek
VE.Jk.CBS.HT.RegCo.Hajek
VE.Jk.B.RegCo.Hajek
VE.Jk.EB.SW2.RegCo.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s      <- oaxaca$sHOMES00 #Defines the sample to be used for the example
y1     <- oaxaca$POP10    #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10 #Defines the variable of interest y2
x      <- oaxaca$HOMES10  #Defines the variable of interest x
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1]) #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the regression coeff. point estimator using y1
VE.Jk.CBS.SYG.RegCo.Hajek(y1[s==1], x[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the regression coeff. point estimator using y2
VE.Jk.CBS.SYG.RegCo.Hajek(y2[s==1], x[s==1], pik.U[s==1], pikl.s)
```

VE.Jk.CBS.SYG.Total.Hajek

The Campbell-Berger-Skinner unequal probability jackknife variance estimator for the Hajek (1971) estimator of a total (Sen-Yates-Grundy form)

Description

Computes the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator for the Hajek estimator of a total. It uses the Sen (1953); Yates-Grundy(1953) variance form.

Usage

```
VE.Jk.CBS.SYG.Total.Hajek(VecY.s, VecPk.s, MatPkl.s, N)
```

Arguments

`VecY.s` vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s. There must not be any missing value.

`VecPk.s` vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.

`MatPkl.s` matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in MatPkl.s must be greater than zero and less than or equal to one. There must not be any missing value.

`N` the population size.

Details

For the population total of the variable y:

$$t = \sum_{k\in U} y_k$$

the approximately unbiased Hajek (1971) estimator of t (implemented by the current function) is given by:

$$\hat{t}_{Hajek} = N\frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{t}_{Hajek}$ can be estimated by the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{t}_{Hajek}) = \frac{-1}{2}\sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}(\varepsilon_k - \varepsilon_l)^2$$

where

$$\varepsilon_k = (1 - \tilde{w_k})\left(\hat{t}_{Hajek} - \hat{t}_{Hajek(k)}\right)$$

with

$$\tilde{w_k} = \frac{w_k}{\sum_{l\in s} w_l}$$

and

$$\hat{t}_{Hajek(k)} = N\frac{\sum_{l\in s, l\neq k} w_l y_l}{\sum_{l\in s, l\neq k} w_l}$$

The Sen-Yates-Grundy form for the Campbell(1980); Berger-Skinner(2005) unequal probability jackknife variance estimator is proposed in Escobar-Berger (2013) under less-restrictive regularity conditions.

Value

The function returns a value for the estimated variance.

References

Campbell, C. (1980) A different view of finite population estimation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 319–324.

Berger, Y. G. and Skinner, C. J. (2005) A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, **67**, 79–89.

Escobar, E. L. and Berger, Y. G. (2013) A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica* (to appear).

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sen, A. R. (1953) On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **5**, 119–127.

Yates, F. and Grundy, P. M. (1953) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, **15**, 253–261.

See Also

VE.Jk.Tukey.Total.Hajek
VE.Jk.CBS.HT.Total.Hajek
VE.Jk.B.Total.Hajek
VE.Jk.EB.SW2.Total.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s      <- oaxaca$sHOMES00 #Defines the sample to be used for the example
N      <- dim(oaxaca)[1]  #Defines the population size
y1     <- oaxaca$POP10    #Defines the variable of interest y1
y2     <- oaxaca$POPMAL10 #Defines the variable of interest y2
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1]) #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the Hajek total point estimator using y1
VE.Jk.CBS.SYG.Total.Hajek(y1[s==1], pik.U[s==1], pikl.s, N)
#Computes the var. est. of the Hajek total point estimator using y2
VE.Jk.CBS.SYG.Total.Hajek(y2[s==1], pik.U[s==1], pikl.s, N)
```

VE.Jk.EB.SW2.Corr.Hajek

The self-weighted two-stage sampling Escobar-Berger (2013) jackknife variance estimator for the estimator of a correlation coefficient using the Hajek point estimator

Description

Computes the self-weighted two-stage sampling Escobar-Berger (2013) jackknife variance estimator for the estimator of a correlation coefficient of two variables using the Hajek (1971) point estimator.

Usage

```
VE.Jk.EB.SW2.Corr.Hajek(VecY.s, VecX.s, VecPk.s, nII, VecPi.s,
                        VecCluLab.s, VecCluSize.s)
```

Arguments

`VecY.s`	vector of the variable of interest Y; its length is equal to n, the total sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be any missing value.
`VecX.s`	vector of the variable of interest X; its length is equal to n, the total sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be any missing value.
`VecPk.s`	vector of the elements' first-order inclusion probabilities; its length is equal to n, the total sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.
`nII`	the second stage sample size, i.e. the fixed number of ultimate sampling units that were selected within each cluster. Its size must be less than or equal to the minimum cluster size in the sample.
`VecPi.s`	vector of the clusters' first-order inclusion probabilities; its length is equal to n, the total sample size. Hence values are expected to be repeated in the utilised sample dataset. Values in VecPi.s must be greater than zero and less than or equal to one. There must not be any missing value.
`VecCluLab.s`	vector of the clusters' labels for the elements; its length is equal to n, the total sample size. The labels must be integer numbers.
`VecCluSize.s`	vector of the clusters' sizes; its length is equal to n, the total sample size. Hence values are expected to be repeated in the utilised sample dataset. None of the sizes must be smaller than nII.

Details

For the population correlation coefficient of two variables y and x:

$$C = \frac{\sum_{k\in U}(y_k - \bar{y})(x_k - \bar{x})}{\sqrt{\sum_{k\in U}(y_k - \bar{y})^2}\sqrt{\sum_{k\in U}(x_k - \bar{x})^2}}$$

the point estimator of C, assuming that N is unknown (see Sarndal et al., 1992, Sec. 5.9), is:

$$\hat{C}_{Hajek} = \frac{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sqrt{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{Hajek})^2}\sqrt{\sum_{k\in s} w_k(x_k - \hat{\bar{x}}_{Hajek})^2}}$$

where $\hat{\bar{y}}_{Hajek}$ is the Hajek (1971) point estimator of the population mean $\bar{y} = N^{-1}\sum_{k\in U} y_k$,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. If s is a self-weighted two-stage sample, the variance of $\hat{C}_{Hajek}$ can be estimated by the Escobar-Berger (2013) jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{C}_{Hajek}) = v_{clu} + v_{obs}$$

$$v_{clu} = \sum_{i\in s}(1-\pi^*_{Ii})\varsigma^2_{(Ii)} - \frac{1}{\hat{d}}\left(\sum_{i\in s}(1-\pi_{Ii})\varsigma_{(Ii)}\right)^2$$

$$v_{obs} = \sum_{k\in s}\phi_k \varepsilon^2_{(k)}$$

where $\hat{d} = \sum_{i\in s}(1-\pi_{Ii})$, $\phi_k = I\{k \in s_i\}\pi^*_{Ii}(M_i - n_{II})/(M_i - 1)$, $\pi^*_{Ii} = \pi_{Ii}n_{II}(M_i - 1)/(n_{II} - 1)M_i$, with s_i denoting the sample elements from the i-th cluster, $I\{k \in s_i\}$ is an indicator that takes the value 1 if the k-th observation is within the i-th cluster and 0 otherwise, π_{Ii} is the inclusion probability of the i-th cluster in the sample s, M_i is the size of the i-th cluster, n_{II} is the sample size within each cluster, n_I is the number of sampled clusters, and where

$$\varsigma_{(Ii)} = \frac{n_I - 1}{n_I}(\hat{C}_{Hajek} - \hat{C}_{Hajek(Ii)})$$

$$\varepsilon_{(k)} = \frac{n-1}{n}(\hat{C}_{Hajek} - \hat{C}_{Hajek(k)})$$

where $\hat{C}_{Hajek(Ii)}$ and $\hat{C}_{Hajek(k)}$ have the same functional form as $\hat{C}_{Hajek}$ but omitting the i-th cluster and the k-th element, respectively, from the sample s. Note that this variance estimator utilises implicitly the Hajek (1964) approximations that are designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Escobar, E. L. and Berger, Y. G. (2013) A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica* (to appear).

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

See Also

VE.Jk.Tukey.Corr.Hajek
VE.Jk.CBS.HT.Corr.Hajek
VE.Jk.CBS.SYG.Corr.Hajek
VE.Jk.B.Corr.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
s         <- oaxaca$sSW_10_3 #Defines the sample to be used
SampData  <- oaxaca[s==1, ]  #Defines the sample dataset
nII       <- 3               #Defines the 2nd stage fixed sample size
#Defines the clusters' labels in the sample dataset
CluLab.s  <- SampData$IDDISTRI
#Defines the clusters' sizes in the sample dataset
CluSize.s <- SampData$SIZEDIST
#Reconstructs clusters' 1st order incl. probs. in the sample dataset
piIi.s    <- (10 * CluSize.s / 570)
#Reconstructs elements' 1st order incl. probs. in the sample dataset
pik.s     <- piIi.s * (nII/CluSize.s)
y1.s      <- SampData$POP10    #Defines the variable y1
y2.s      <- SampData$POPMAL10 #Defines the variable y2
x.s       <- SampData$HOMES10  #Defines the variable x
#Computes the var. est. of the corr. coeff. point estimator using y1
VE.Jk.EB.SW2.Corr.Hajek(y1.s, x.s, pik.s, nII, piIi.s, CluLab.s, CluSize.s)
#Computes the var. est. of the corr. coeff. point estimator using y2
VE.Jk.EB.SW2.Corr.Hajek(y2.s, x.s, pik.s, nII, piIi.s, CluLab.s, CluSize.s)
```

VE.Jk.EB.SW2.Mean.Hajek

The self-weighted two-stage sampling Escobar-Berger (2013) jackknife variance estimator for the Hajek (1971) estimator of a mean

---

Description

Computes the self-weighted two-stage sampling Escobar-Berger (2013) jackknife variance estimator for the Hajek estimator of a mean.

Usage

```
VE.Jk.EB.SW2.Mean.Hajek(VecY.s, VecPk.s, nII, VecPi.s,
                            VecCluLab.s, VecCluSize.s)
```

Arguments

VecY.s — vector of the variable of interest; its length is equal to n, the total sample size. Its length has to be the same as the length of VecPk.s. There must not be any missing value.

`VecPk.s` vector of the elements' first-order inclusion probabilities; its length is equal to n, the total sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.

`nII` the second stage sample size, i.e. the fixed number of ultimate sampling units that were selected within each cluster. Its size must be less than or equal to the minimum cluster size in the sample.

`VecPi.s` vector of the clusters' first-order inclusion probabilities; its length is equal to n, the total sample size. Hence values are expected to be repeated in the utilised sample dataset. Values in VecPi.s must be greater than zero and less than or equal to one. There must not be any missing value.

`VecCluLab.s` vector of the clusters' labels for the elements; its length is equal to n, the total sample size. The labels must be integer numbers.

`VecCluSize.s` vector of the clusters' sizes; its length is equal to n, the total sample size. Hence values are expected to be repeated in the utilised sample dataset. None of the sizes must be smaller than nII.

Details

For the population mean of the variable y:

$$\bar{y} = \frac{1}{N}\sum_{k\in U} y_k$$

the approximately unbiased Hajek (1971) estimator of $\bar{y}$ is given by:

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. If s is a self-weighted two-stage sample, the variance of $\hat{\bar{y}}_{Hajek}$ can be estimated by the Escobar-Berger (2013) jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\bar{y}}_{Hajek}) = v_{clu} + v_{obs}$$

$$v_{clu} = \sum_{i\in s}(1-\pi^*_{Ii})\varsigma^2_{(Ii)} - \frac{1}{\hat{d}}\left(\sum_{i\in s}(1-\pi_{Ii})\varsigma_{(Ii)}\right)^2$$

$$v_{obs} = \sum_{k\in s}\phi_k\varepsilon^2_{(k)}$$

where $\hat{d} = \sum_{i\in s}(1-\pi_{Ii})$, $\phi_k = I\{k \in s_i\}\pi^*_{Ii}(M_i - n_{II})/(M_i - 1)$, $\pi^*_{Ii} = \pi_{Ii}n_{II}(M_i - 1)/(n_{II} - 1)M_i$, with s_i denoting the sample elements from the i-th cluster, $I\{k \in s_i\}$ is an indicator that takes the value 1 if the k-th observation is within the i-th cluster and 0 otherwise, π_{Ii} is the inclusion probability of the i-th cluster in the sample s, M_i is the size of the i-th cluster, n_{II} is the sample size within each cluster, n_I is the number of sampled clusters, and where

$$\varsigma_{(Ii)} = \frac{n_I - 1}{n_I}(\hat{\bar{y}}_{Hajek} - \hat{\bar{y}}_{Hajek(Ii)})$$

$$\varepsilon_{(k)} = \frac{n-1}{n}(\hat{\hat{y}}_{Hajek} - \hat{\hat{y}}_{Hajek(k)})$$

where $\hat{\hat{y}}_{Hajek(Ii)}$ and $\hat{\hat{y}}_{Hajek(k)}$ have the same functional form as $\hat{\hat{y}}_{Hajek}$ but omitting the i-th cluster and the k-th element, respectively, from the sample s. Note that this variance estimator utilises implicitly the Hajek (1964) approximations that are designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Escobar, E. L. and Berger, Y. G. (2013) A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica* (to appear).

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

See Also

VE.Jk.Tukey.Mean.Hajek
VE.Jk.CBS.HT.Mean.Hajek
VE.Jk.CBS.SYG.Mean.Hajek
VE.Jk.B.Mean.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
s          <- oaxaca$sSW_10_3 #Defines the sample to be used
SampData   <- oaxaca[s==1, ]  #Defines the sample dataset
nII        <- 3               #Defines the 2nd stage fixed sample size
#Defines the clusters' labels in the sample dataset
CluLab.s   <- SampData$IDDISTRI
#Defines the clusters' sizes in the sample dataset
CluSize.s <- SampData$SIZEDIST
#Reconstructs clusters' 1st order incl. probs. in the sample dataset
piIi.s     <- (10 * CluSize.s / 570)
#Reconstructs elements' 1st order incl. probs. in the sample dataset
pik.s      <- piIi.s * (nII/CluSize.s)
y1.s       <- SampData$POP10    #Defines the variable of interest y1
y2.s       <- SampData$POPMAL10 #Defines the variable of interest y2
#Computes the var. est. of the Hajek mean point estimator using y1
VE.Jk.EB.SW2.Mean.Hajek(y1.s, pik.s, nII, piIi.s, CluLab.s, CluSize.s)
#Computes the var. est. of the Hajek mean point estimator using y2
```

```
VE.Jk.EB.SW2.Mean.Hajek(y2.s, pik.s, nII, piIi.s, CluLab.s, CluSize.s)
```

`VE.Jk.EB.SW2.Ratio` *The self-weighted two-stage sampling Escobar-Berger (2013) jackknife variance estimator for the estimator of a ratio*

Description

Computes the self-weighted two-stage sampling Escobar-Berger (2013) jackknife variance estimator for the estimator of a ratio of two totals/means.

Usage

```
VE.Jk.EB.SW2.Ratio(VecY.s, VecX.s, VecPk.s, nII, VecPi.s,
                   VecCluLab.s, VecCluSize.s)
```

Arguments

`VecY.s`	vector of the numerator variable of interest; its length is equal to n, the total sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be any missing value.
`VecX.s`	vector of the denominator variable of interest; its length is equal to n, the total sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be any missing value. All values of VecX.s must be greater than zero.
`VecPk.s`	vector of the elements' first-order inclusion probabilities; its length is equal to n, the total sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.
`nII`	the second stage sample size, i.e. the fixed number of ultimate sampling units that were selected within each cluster. Its size must be less than or equal to the minimum cluster size in the sample.
`VecPi.s`	vector of the clusters' first-order inclusion probabilities; its length is equal to n, the total sample size. Hence values are expected to be repeated in the utilised sample dataset. Values in VecPi.s must be greater than zero and less than or equal to one. There must not be any missing value.
`VecCluLab.s`	vector of the clusters' labels for the elements; its length is equal to n, the total sample size. The labels must be integer numbers.
`VecCluSize.s`	vector of the clusters' sizes; its length is equal to n, the total sample size. Hence values are expected to be repeated in the utilised sample dataset. None of the sizes must be smaller than nII.

Details

For the population ratio of two totals/means of the variables y and x:

$$R = \frac{\sum_{k\in U} y_k/N}{\sum_{k\in U} x_k/N} = \frac{\sum_{k\in U} y_k}{\sum_{k\in U} x_k}$$

the ratio estimator of R is given by:

$$\hat{R} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k x_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. If s is a self-weighted two-stage sample, the variance of $\hat{R}$ can be estimated by the Escobar-Berger (2013) jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{R}) = v_{clu} + v_{obs}$$

$$v_{clu} = \sum_{i\in s}(1-\pi^*_{Ii})\varsigma^2_{(Ii)} - \frac{1}{\hat{d}}\left(\sum_{i\in s}(1-\pi_{Ii})\varsigma_{(Ii)}\right)^2$$

$$v_{obs} = \sum_{k\in s}\phi_k \varepsilon^2_{(k)}$$

where $\hat{d} = \sum_{i\in s}(1-\pi_{Ii})$, $\phi_k = I\{k \in s_i\}\pi^*_{Ii}(M_i - n_{II})/(M_i - 1)$, $\pi^*_{Ii} = \pi_{Ii}n_{II}(M_i - 1)/(n_{II} - 1)M_i$, with s_i denoting the sample elements from the i-th cluster, $I\{k \in s_i\}$ is an indicator that takes the value 1 if the k-th observation is within the i-th cluster and 0 otherwise, π_{Ii} is the inclusion probability of the i-th cluster in the sample s, M_i is the size of the i-th cluster, n_{II} is the sample size within each cluster, n_I is the number of sampled clusters, and where

$$\varsigma_{(Ii)} = \frac{n_I - 1}{n_I}(\hat{R} - \hat{R}_{(Ii)})$$

$$\varepsilon_{(k)} = \frac{n-1}{n}(\hat{R} - \hat{R}_{(k)})$$

where $\hat{R}_{(Ii)}$ and $\hat{R}_{(k)}$ have the same functional form as $\hat{R}$ but omitting the i-th cluster and the k-th element, respectively, from the sample s. Note that this variance estimator utilises implicitly the Hajek (1964) approximations that are designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Escobar, E. L. and Berger, Y. G. (2013) A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica* (to appear).

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

See Also

VE.Jk.Tukey.Ratio
VE.Jk.CBS.HT.Ratio
VE.Jk.CBS.SYG.Ratio
VE.Jk.B.Ratio

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
s         <- oaxaca$sSW_10_3 #Defines the sample to be used
SampData  <- oaxaca[s==1, ]  #Defines the sample dataset
nII       <- 3               #Defines the 2nd stage fixed sample size
#Defines the clusters' labels in the sample dataset
CluLab.s  <- SampData$IDDISTRI
#Defines the clusters' sizes in the sample dataset
CluSize.s <- SampData$SIZEDIST
#Reconstructs clusters' 1st order incl. probs. in the sample dataset
piIi.s    <- (10 * CluSize.s / 570)
#Reconstructs elements' 1st order incl. probs. in the sample dataset
pik.s     <- piIi.s * (nII/CluSize.s)
y1.s      <- SampData$POP10     #Defines the numerator variable y1
y2.s      <- SampData$POPMAL10 #Defines the numerator variable y2
x.s       <- SampData$HOMES10  #Defines the denominator variable x
#Computes the var. est. of the ratio point estimator using y1
VE.Jk.EB.SW2.Ratio(y1.s, x.s, pik.s, nII, piIi.s, CluLab.s, CluSize.s)
#Computes the var. est. of the ratio point estimator using y2
VE.Jk.EB.SW2.Ratio(y2.s, x.s, pik.s, nII, piIi.s, CluLab.s, CluSize.s)
```

VE.Jk.EB.SW2.RegCo.Hajek

The self-weighted two-stage sampling Escobar-Berger (2013) jackknife variance estimator for the estimator of the regression coefficient using the Hajek point estimator

Description

Computes the self-weighted two-stage sampling Escobar-Berger (2013) jackknife variance estimator for the estimator of the regression coefficient using the Hajek (1971) point estimator.

Usage

```
VE.Jk.EB.SW2.RegCo.Hajek(VecY.s, VecX.s, VecPk.s, nII, VecPi.s,
                            VecCluLab.s, VecCluSize.s)
```

Arguments

`VecY.s` vector of the variable of interest Y; its length is equal to n, the total sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be any missing value.

`VecX.s` vector of the variable of interest X; its length is equal to n, the total sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be any missing value.

`VecPk.s` vector of the elements' first-order inclusion probabilities; its length is equal to n, the total sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.

`nII` the second stage sample size, i.e. the fixed number of ultimate sampling units that were selected within each cluster. Its size must be less than or equal to the minimum cluster size in the sample.

`VecPi.s` vector of the clusters' first-order inclusion probabilities; its length is equal to n, the total sample size. Hence values are expected to be repeated in the utilised sample dataset. Values in VecPi.s must be greater than zero and less than or equal to one. There must not be any missing value.

`VecCluLab.s` vector of the clusters' labels for the elements; its length is equal to n, the total sample size. The labels must be integer numbers.

`VecCluSize.s` vector of the clusters' sizes; its length is equal to n, the total sample size. Hence values are expected to be repeated in the utilised sample dataset. None of the sizes must be smaller than nII.

Details

From Linear Regression Analysis, for an imposed population model

$$y = \alpha + \beta x$$

the population regression coefficient β, assuming that the population size N is unknown (see Sarndal et al., 1992, Sec. 5.9), can be estimated by:

$$\hat{\beta}_{Hajek} = \frac{\sum_{k\in s} w_k (y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sum_{k\in s} w_k (x_k - \hat{\bar{x}}_{Hajek})^2}$$

where $\hat{\bar{y}}_{Hajek}$ is the Hajek (1971) point estimator of the population mean $\bar{y} = N^{-1}\sum_{k\in U} y_k$,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. If s is a self-weighted two-stage sample, the variance of $\hat{\beta}_{Hajek}$ can be estimated by the Escobar-Berger (2013) jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\beta}_{Hajek}) = v_{clu} + v_{obs}$$

$$v_{clu} = \sum_{i\in s} (1 - \pi^*_{Ii}) \varsigma^2_{(Ii)} - \frac{1}{\hat{d}} \left(\sum_{i\in s} (1 - \pi_{Ii}) \varsigma_{(Ii)} \right)^2$$

$$v_{obs} = \sum_{k\in s} \phi_k \varepsilon^2_{(k)}$$

where $\hat{d} = \sum_{i\in s}(1-\pi_{Ii})$, $\phi_k = I\{k \in s_i\}\pi^*_{Ii}(M_i - n_{II})/(M_i-1)$, $\pi^*_{Ii} = \pi_{Ii}n_{II}(M_i - 1)/(n_{II}-1)M_i$, with s_i denoting the sample elements from the i-th cluster, $I\{k \in s_i\}$ is an indicator that takes the value 1 if the k-th observation is within the i-th cluster and 0 otherwise, π_{Ii} is the inclusion probability of the i-th cluster in the sample s, M_i is the size of the i-th cluster, n_{II} is the sample size within each cluster, n_I is the number of sampled clusters, and where

$$\varsigma_{(Ii)} = \frac{n_I - 1}{n_I}(\hat{\beta}_{Hajek} - \hat{\beta}_{Hajek(Ii)})$$

$$\varepsilon_{(k)} = \frac{n-1}{n}(\hat{\beta}_{Hajek} - \hat{\beta}_{Hajek(k)})$$

where $\hat{\beta}_{Hajek(Ii)}$ and $\hat{\beta}_{Hajek(k)}$ have the same functional form as $\hat{\beta}_{Hajek}$ but omitting the i-th cluster and the k-th element, respectively, from the sample s. Note that this variance estimator utilises implicitly the Hajek (1964) approximations that are designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Escobar, E. L. and Berger, Y. G. (2013) A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica* (to appear).

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

See Also

VE.Jk.Tukey.RegCo.Hajek
VE.Jk.CBS.HT.RegCo.Hajek
VE.Jk.CBS.SYG.RegCo.Hajek
VE.Jk.B.RegCo.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
s         <- oaxaca$sSW_10_3 #Defines the sample to be used
SampData  <- oaxaca[s==1, ]  #Defines the sample dataset
nII       <- 3               #Defines the 2nd stage fixed sample size
#Defines the clusters' labels in the sample dataset
```

```
CluLab.s  <- SampData$IDDISTRI
#Defines the clusters' sizes in the sample dataset
CluSize.s <- SampData$SIZEDIST
#Reconstructs clusters' 1st order incl. probs. in the sample dataset
piIi.s    <- (10 * CluSize.s / 570)
#Reconstructs elements' 1st order incl. probs. in the sample dataset
pik.s     <- piIi.s * (nII/CluSize.s)
y1.s      <- SampData$POP10    #Defines the variable y1
y2.s      <- SampData$POPMAL10 #Defines the variable y2
x.s       <- SampData$HOMES10  #Defines the variable x
#Computes the var. est. of the regression coeff. point estimator using y1
VE.Jk.EB.SW2.RegCo.Hajek(y1.s, x.s, pik.s, nII, piIi.s, CluLab.s, CluSize.s)
#Computes the var. est. of the regression coeff. point estimator using y2
VE.Jk.EB.SW2.RegCo.Hajek(y2.s, x.s, pik.s, nII, piIi.s, CluLab.s, CluSize.s)
```

VE.Jk.EB.SW2.Total.Hajek

The self-weighted two-stage sampling Escobar-Berger (2013) jackknife variance estimator for the Hajek (1971) estimator of a total

Description

Computes the self-weighted two-stage sampling Escobar-Berger (2013) jackknife variance estimator for the Hajek estimator of a total.

Usage

```
VE.Jk.EB.SW2.Total.Hajek(VecY.s, VecPk.s, nII, VecPi.s,
                         VecCluLab.s, VecCluSize.s, N)
```

Arguments

`VecY.s` vector of the variable of interest; its length is equal to n, the total sample size. Its length has to be the same as the length of VecPk.s. There must not be any missing value.

`VecPk.s` vector of the elements' first-order inclusion probabilities; its length is equal to n, the total sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.

`nII` the second stage sample size, i.e. the fixed number of ultimate sampling units that were selected within each cluster. Its size must be less than or equal to the minimum cluster size in the sample.

`VecPi.s` vector of the clusters' first-order inclusion probabilities; its length is equal to n, the total sample size. Hence values are expected to be repeated in the utilised sample dataset. Values in VecPi.s must be greater than zero and less than or equal to one. There must not be any missing value.

`VecCluLab.s` vector of the clusters' labels for the elements; its length is equal to n, the total sample size. The labels must be integer numbers.

`VecCluSize.s` vector of the clusters' sizes; its length is equal to n, the total sample size. Hence values are expected to be repeated in the utilised sample dataset. None of the sizes must be smaller than nII.

`N` the population size.

Details

For the population total of the variable y:

$$t = \sum_{k \in U} y_k$$

the approximately unbiased Hajek (1971) estimator of t (implemented by the current function) is given by:

$$\hat{t}_{Hajek} = N \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. If s is a self-weighted two-stage sample, the variance of $\hat{t}_{Hajek}$ can be estimated by the Escobar-Berger (2013) jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{t}_{Hajek}) = v_{clu} + v_{obs}$$

$$v_{clu} = \sum_{i \in s} (1 - \pi^*_{Ii}) \varsigma^2_{(Ii)} - \frac{1}{\hat{d}} \left(\sum_{i \in s} (1 - \pi_{Ii}) \varsigma_{(Ii)} \right)^2$$

$$v_{obs} = \sum_{k \in s} \phi_k \varepsilon^2_{(k)}$$

where $\hat{d} = \sum_{i \in s} (1 - \pi_{Ii})$, $\phi_k = I\{k \in s_i\} \pi^*_{Ii} (M_i - n_{II})/(M_i - 1)$, $\pi^*_{Ii} = \pi_{Ii} n_{II} (M_i - 1)/(n_{II} - 1) M_i$, with s_i denoting the sample elements from the i-th cluster, $I\{k \in s_i\}$ is an indicator that takes the value 1 if the k-th observation is within the i-th cluster and 0 otherwise, π_{Ii} is the inclusion probability of the i-th cluster in the sample s, M_i is the size of the i-th cluster, n_{II} is the sample size within each cluster, n_I is the number of sampled clusters, and where

$$\varsigma_{(Ii)} = \frac{n_I - 1}{n_I} (\hat{t}_{Hajek} - \hat{t}_{Hajek(Ii)})$$

$$\varepsilon_{(k)} = \frac{n - 1}{n} (\hat{t}_{Hajek} - \hat{t}_{Hajek(k)})$$

where $\hat{t}_{Hajek(Ii)}$ and $\hat{t}_{Hajek(k)}$ have the same functional form as $\hat{t}_{Hajek}$ but omitting the i-th cluster and the k-th element, respectively, from the sample s. Note that this variance estimator utilises implicitly the Hajek (1964) approximations that are designed for large-entropy sampling designs, large samples and large populations, i.e. care should be taken with highly-stratified samples, e.g. Berger (2005).

Value

The function returns a value for the estimated variance.

References

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics*, **47**, 365–373.

Escobar, E. L. and Berger, Y. G. (2013) A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica* (to appear).

Hajek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 4, 1491–1523.

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

See Also

VE.Jk.Tukey.Total.Hajek
VE.Jk.CBS.HT.Total.Hajek
VE.Jk.CBS.SYG.Total.Hajek
VE.Jk.B.Total.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
s         <- oaxaca$sSW_10_3 #Defines the sample to be used
N         <- dim(oaxaca)[1]  #Defines the population size
SampData  <- oaxaca[s==1, ]  #Defines the sample dataset
nII       <- 3               #Defines the 2nd stage fixed sample size
#Defines the clusters' labels in the sample dataset
CluLab.s  <- SampData$IDDISTRI
#Defines the clusters' sizes in the sample dataset
CluSize.s <- SampData$SIZEDIST
#Reconstructs clusters' 1st order incl. probs. in the sample dataset
piIi.s    <- (10 * CluSize.s / 570)
#Reconstructs elements' 1st order incl. probs. in the sample dataset
pik.s     <- piIi.s * (nII/CluSize.s)
y1.s      <- SampData$POP10    #Defines the variable of interest y1
y2.s      <- SampData$POPMAL10 #Defines the variable of interest y2
#Computes the var. est. of the Hajek total point estimator using y1
VE.Jk.EB.SW2.Total.Hajek(y1.s, pik.s, nII, piIi.s, CluLab.s, CluSize.s, N)
#Computes the var. est. of the Hajek total point estimator using y2
VE.Jk.EB.SW2.Total.Hajek(y2.s, pik.s, nII, piIi.s, CluLab.s, CluSize.s, N)
```

VE.Jk.Tukey.Corr.Hajek

The Tukey (1958) jackknife variance estimator for the estimator of a correlation coefficient using the Hajek point estimator

Description

Computes the Quenouille(1956); Tukey (1958) jackknife variance estimator for the estimator of a correlation coefficient of two variables using the Hajek (1971) point estimator.

Usage

```
VE.Jk.Tukey.Corr.Hajek(VecY.s, VecX.s, VecPk.s, N)
```

Arguments

`VecY.s`	vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be any missing value.
`VecX.s`	vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be any missing value.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.
`N`	the population size. Note that this information is utilised for the finite population correction only.

Details

For the population correlation coefficient of two variables y and x:

$$C = \frac{\sum_{k\in U}(y_k - \bar{y})(x_k - \bar{x})}{\sqrt{\sum_{k\in U}(y_k - \bar{y})^2}\sqrt{\sum_{k\in U}(x_k - \bar{x})^2}}$$

the point estimator of C, assuming that N is unknown (see Sarndal et al., 1992, Sec. 5.9), is:

$$\hat{C}_{Hajek} = \frac{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sqrt{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{Hajek})^2}\sqrt{\sum_{k\in s} w_k(x_k - \hat{\bar{x}}_{Hajek})^2}}$$

where $\hat{\bar{y}}_{Hajek}$ is the Hajek (1971) point estimator of the population mean $\bar{y} = N^{-1}\sum_{k\in U} y_k$,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. The variance of $\hat{C}_{Hajek}$ can be estimated by the Quenouille(1956); Tukey (1958) jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{C}_{Hajek}) = \left(1 - \frac{n}{N}\right)\frac{n-1}{n}\sum_{k\in s}\left(\hat{C}_{Hajek(k)} - \hat{C}_{Hajek}\right)^2$$

where $\hat{C}_{Hajek(k)}$ has the same functional form as $\hat{C}_{Hajek}$ but omitting the k-th element from the sample s. Note that we are implementing the Tukey (1958) jackknife variance estimator using the 'ad hoc' finite population correction $1 - n/N$ (see Shao and Tu, 1995; Wolter, 2007).

Value

The function returns a value for the estimated variance.

References

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Quenouille, M. H. (1956) Notes on bias in estimation. *Biometrika*, **43**, 353–360.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap*. Springer-Verlag, Inc.

Tukey, J. W. (1958) Bias and confidence in not-quite large samples (abstract). *The Annals of Mathematical Statistics*, **29**, 2, p. 614.

Wolter, K. M. (2007) *Introduction to Variance Estimation*. 2nd Ed. Springer, Inc.

See Also

VE.Jk.CBS.HT.Corr.Hajek
VE.Jk.CBS.SYG.Corr.Hajek
VE.Jk.B.Corr.Hajek
VE.Jk.EB.SW2.Corr.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s     <- oaxaca$sHOMES00 #Defines the sample to be used for the example
N     <- dim(oaxaca)[1]  #Defines the population size
y1    <- oaxaca$POP10    #Defines the variable of interest y1
y2    <- oaxaca$POPMAL10 #Defines the variable of interest y2
x     <- oaxaca$HOMES10  #Defines the variable of interest x
#Computes the var. est. of the corr. coeff. point estimator using y1
VE.Jk.Tukey.Corr.Hajek(y1[s==1], x[s==1], pik.U[s==1], N)
#Computes the var. est. of the corr. coeff. point estimator using y2
VE.Jk.Tukey.Corr.Hajek(y2[s==1], x[s==1], pik.U[s==1], N)
```

VE.Jk.Tukey.Corr.NHT — *The Tukey (1958) jackknife variance estimator for the estimator of a correlation coefficient using the Narain-Horvitz-Thompson point estimator*

Description

Computes the Quenouille(1956); Tukey (1958) jackknife variance estimator for the estimator of a correlation coefficient of two variables using the Narain (1951); Horvitz-Thompson (1952) point estimator.

Usage

```
VE.Jk.Tukey.Corr.NHT(VecY.s, VecX.s, VecPk.s, N)
```

Arguments

`VecY.s` vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be any missing value.

`VecX.s` vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be any missing value.

`VecPk.s` vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.

`N` the population size.

Details

For the population correlation coefficient of two variables y and x:

$$C = \frac{\sum_{k\in U}(y_k - \bar{y})(x_k - \bar{x})}{\sqrt{\sum_{k\in U}(y_k - \bar{y})^2}\sqrt{\sum_{k\in U}(x_k - \bar{x})^2}}$$

the point estimator of C is given by:

$$\hat{C} = \frac{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{NHT})(x_k - \hat{\bar{x}}_{NHT})}{\sqrt{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{NHT})^2}\sqrt{\sum_{k\in s} w_k(x_k - \hat{\bar{x}}_{NHT})^2}}$$

where $\hat{\bar{y}}_{NHT}$ is the Narain (1951); Horvitz-Thompson (1952) estimator for the population mean $\bar{y} = N^{-1}\sum_{k\in U} y_k$,

$$\hat{\bar{y}}_{NHT} = \frac{1}{N}\sum_{k\in s} w_k y_k$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. The variance of $\hat{C}$ can be estimated by the Quenouille(1956); Tukey (1958) jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{C}) = \left(1 - \frac{n}{N}\right)\frac{n-1}{n}\sum_{k\in s}\left(\hat{C}_{(k)} - \hat{C}\right)^2$$

where $\hat{C}_{(k)}$ has the same functional form as $\hat{C}$ but omitting the k-th element from the sample s. Note that we are implementing the Tukey (1958) jackknife variance estimator using the ‘ad hoc’ finite population correction $1 - n/N$ (see Shao and Tu, 1995; Wolter, 2007).

Value

The function returns a value for the estimated variance.

References

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

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

Quenouille, M. H. (1956) Notes on bias in estimation. *Biometrika*, **43**, 353–360.

Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap*. Springer-Verlag, Inc.

Tukey, J. W. (1958) Bias and confidence in not-quite large samples (abstract). *The Annals of Mathematical Statistics*, **29**, 2, p. 614.

Wolter, K. M. (2007) *Introduction to Variance Estimation*. 2nd Ed. Springer, Inc.

See Also

Est.Corr.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s     <- oaxaca$sHOMES00 #Defines the sample to be used for the example
N     <- dim(oaxaca)[1]  #Defines the population size
y1    <- oaxaca$POP10    #Defines the variable of interest y1
y2    <- oaxaca$POPMAL10 #Defines the variable of interest y2
x     <- oaxaca$HOMES10  #Defines the variable of interest x
#Computes the var. est. of the corr. coeff. point estimator using y1
VE.Jk.Tukey.Corr.NHT(y1[s==1], x[s==1], pik.U[s==1], N)
#Computes the var. est. of the corr. coeff. point estimator using y2
VE.Jk.Tukey.Corr.NHT(y2[s==1], x[s==1], pik.U[s==1], N)
```

VE.Jk.Tukey.Mean.Hajek

The Tukey (1958) jackknife variance estimator for the Hajek estimator of a mean

---

Description

Computes the Quenouille(1956); Tukey (1958) jackknife variance estimator for the Hajek (1971) estimator of a mean.

Usage

```
VE.Jk.Tukey.Mean.Hajek(VecY.s, VecPk.s, N)
```

Arguments

`VecY.s` vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s. There must not be any missing value.

`VecPk.s` vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.

`N` the population size.

Details

For the population mean of the variable y:

$$\bar{y} = \frac{1}{N}\sum_{k\in U} y_k$$

the approximately unbiased Hajek (1971) estimator of $\bar{y}$ is given by:

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{\bar{y}}_{Hajek}$ can be estimated by the Quenouille(1956); Tukey (1958) jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\bar{y}}_{Hajek}) = \left(1 - \frac{n}{N}\right)\frac{n-1}{n}\sum_{k\in s}\left(\hat{\bar{y}}_{Hajek(k)} - \hat{\bar{y}}_{Hajek}\right)^2$$

where

$$\hat{\bar{y}}_{Hajek(k)} = \frac{\sum_{l\in s, l\neq k} w_l y_l}{\sum_{l\in s, l\neq k} w_l}$$

Note that we are implementing the Tukey (1958) jackknife variance estimator using the 'ad hoc' finite population correction $1 - n/N$ (see Shao and Tu, 1995; Wolter, 2007).

Value

The function returns a value for the estimated variance.

References

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Quenouille, M. H. (1956) Notes on bias in estimation. *Biometrika*, **43**, 353–360.

Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap*. Springer-Verlag, Inc.

Tukey, J. W. (1958) Bias and confidence in not-quite large samples (abstract). *The Annals of Mathematical Statistics*, **29**, 2, p. 614.

Wolter, K. M. (2007) *Introduction to Variance Estimation*. 2nd Ed. Springer, Inc.

See Also

VE.Jk.CBS.HT.Mean.Hajek
VE.Jk.CBS.SYG.Mean.Hajek
VE.Jk.B.Mean.Hajek
VE.Jk.EB.SW2.Mean.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s     <- oaxaca$sHOMES00 #Defines the sample to be used for the example
N     <- dim(oaxaca)[1]  #Defines the population size
y1    <- oaxaca$POP10    #Defines the variable of interest y1
y2    <- oaxaca$POPMAL10 #Defines the variable of interest y2
#Computes the var. est. of the Hajek mean point estimator using y1
VE.Jk.Tukey.Mean.Hajek(y1[s==1], pik.U[s==1], N)
#Computes the var. est. of the Hajek mean point estimator using y2
VE.Jk.Tukey.Mean.Hajek(y2[s==1], pik.U[s==1], N)
```

VE.Jk.Tukey.Ratio *The Tukey (1958) jackknife variance estimator for the estimator of a ratio*

Description

Computes the Quenouille(1956); Tukey (1958) jackknife variance estimator for the estimator of a ratio of two totals/means.

Usage

```
VE.Jk.Tukey.Ratio(VecY.s, VecX.s, VecPk.s, N)
```

Arguments

VecY.s	vector of the numerator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be any missing value.
VecX.s	vector of the denominator variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be any missing value. All values of VecX.s must be greater than zero.
VecPk.s	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.
N	the population size.

Details

For the population ratio of two totals/means of the variables y and x:

$$R = \frac{\sum_{k\in U} y_k/N}{\sum_{k\in U} x_k/N} = \frac{\sum_{k\in U} y_k}{\sum_{k\in U} x_k}$$

the ratio estimator of R is given by:

$$\hat{R} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k x_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{R}$ can be estimated by the Quenouille(1956); Tukey (1958) jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{R}) = \left(1 - \frac{n}{N}\right) \frac{n-1}{n} \sum_{k\in s} \left(\hat{R}_{(k)} - \hat{R}\right)^2$$

where

$$\hat{R}_{(k)} = \frac{\sum_{l\in s, l\neq k} w_l y_l}{\sum_{l\in s, l\neq k} w_l x_l}$$

Note that we are implementing the Tukey (1958) jackknife variance estimator using the 'ad hoc' finite population correction $1 - n/N$ (see Shao and Tu, 1995; Wolter, 2007).

Value

The function returns a value for the estimated variance.

References

Quenouille, M. H. (1956) Notes on bias in estimation. *Biometrika*, **43**, 353–360.

Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap*. Springer-Verlag, Inc.

Tukey, J. W. (1958) Bias and confidence in not-quite large samples (abstract). *The Annals of Mathematical Statistics*, **29**, 2, p. 614.

Wolter, K. M. (2007) *Introduction to Variance Estimation*. 2nd Ed. Springer, Inc.

See Also

VE.Jk.CBS.HT.Ratio
VE.Jk.CBS.SYG.Ratio
VE.Jk.B.Ratio
VE.Jk.EB.SW2.Ratio

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s     <- oaxaca$sHOMES00 #Defines the sample to be used for the example
```

```
N     <- dim(oaxaca)[1]  #Defines the population size
y1    <- oaxaca$POP10    #Defines the numerator variable of interest y1
y2    <- oaxaca$POPMAL10 #Defines the numerator variable of interest y2
x     <- oaxaca$HOMES10  #Defines the denominator variable of interest x
#Computes the var. est. of the ratio point estimator using y1
VE.Jk.Tukey.Ratio(y1[s==1], x[s==1], pik.U[s==1], N)
#Computes the var. est. of the ratio point estimator using y2
VE.Jk.Tukey.Ratio(y2[s==1], x[s==1], pik.U[s==1], N)
```

VE.Jk.Tukey.RegCo.Hajek

The Tukey (1958) jackknife variance estimator for the estimator of the regression coefficient using the Hajek point estimator

Description

Computes the Quenouille(1956); Tukey (1958) jackknife variance estimator for the estimator of the regression coefficient using the Hajek (1971) point estimator.

Usage

```
VE.Jk.Tukey.RegCo.Hajek(VecY.s, VecX.s, VecPk.s, N)
```

Arguments

VecY.s	vector of the variable of interest Y; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecX.s. There must not be any missing value.
VecX.s	vector of the variable of interest X; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s and VecY.s. There must not be any missing value.
VecPk.s	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.
N	the population size. Note that this information is utilised for the finite population correction only.

Details

From Linear Regression Analysis, for an imposed population model

$$y = \alpha + \beta x$$

the population regression coefficient β, assuming that the population size N is unknown (see Sarndal et al., 1992, Sec. 5.9), can be estimated by:

$$\hat{\beta}_{Hajek} = \frac{\sum_{k\in s} w_k(y_k - \hat{\bar{y}}_{Hajek})(x_k - \hat{\bar{x}}_{Hajek})}{\sum_{k\in s} w_k(x_k - \hat{\bar{x}}_{Hajek})^2}$$

where $\hat{\bar{y}}_{Hajek}$ is the Hajek (1971) point estimator of the population mean $\bar{y} = N^{-1}\sum_{k\in U} y_k$,

$$\hat{\bar{y}}_{Hajek} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}$$

and $w_k = 1/\pi_k$ with π_k denoting the inclusion probability of the k-th element in the sample s. The variance of $\hat{\beta}_{Hajek}$ can be estimated by the Quenouille(1956); Tukey (1958) jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{\beta}_{Hajek}) = \left(1 - \frac{n}{N}\right)\frac{n-1}{n}\sum_{k\in s}\left(\hat{\beta}_{Hajek(k)} - \hat{\beta}_{Hajek}\right)^2$$

where $\hat{\beta}_{Hajek(k)}$ has the same functional form as $\hat{\beta}_{Hajek}$ but omitting the k-th element from the sample s. Note that we are implementing the Tukey (1958) jackknife variance estimator using the 'ad hoc' finite population correction $1 - n/N$ (see Shao and Tu, 1995; Wolter, 2007).

Value

The function returns a value for the estimated variance.

References

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Quenouille, M. H. (1956) Notes on bias in estimation. *Biometrika*, **43**, 353–360.

Sarndal, C.-E. and Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, Inc.

Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap*. Springer-Verlag, Inc.

Tukey, J. W. (1958) Bias and confidence in not-quite large samples (abstract). *The Annals of Mathematical Statistics*, **29**, 2, p. 614.

Wolter, K. M. (2007) *Introduction to Variance Estimation*. 2nd Ed. Springer, Inc.

See Also

VE.Jk.CBS.HT.RegCo.Hajek
VE.Jk.CBS.SYG.RegCo.Hajek
VE.Jk.B.RegCo.Hajek
VE.Jk.EB.SW2.RegCo.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s     <- oaxaca$sHOMES00 #Defines the sample to be used for the example
N     <- dim(oaxaca)[1]  #Defines the population size
y1    <- oaxaca$POP10    #Defines the variable of interest y1
y2    <- oaxaca$POPMAL10 #Defines the variable of interest y2
```

```
x     <- oaxaca$HOMES10  #Defines the variable of interest x
#Computes the var. est. of the regression coeff. point estimator using y1
VE.Jk.Tukey.RegCo.Hajek(y1[s==1], x[s==1], pik.U[s==1], N)
#Computes the var. est. of the regression coeff. point estimator using y2
VE.Jk.Tukey.RegCo.Hajek(y2[s==1], x[s==1], pik.U[s==1], N)
```

VE.Jk.Tukey.Total.Hajek

The Tukey (1958) jackknife variance estimator for the Hajek estimator of a total

Description

Computes the Quenouille(1956); Tukey (1958) jackknife variance estimator for the Hajek (1971) estimator of a total.

Usage

```
VE.Jk.Tukey.Total.Hajek(VecY.s, VecPk.s, N)
```

Arguments

VecY.s	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s. There must not be any missing value.
VecPk.s	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.
N	the population size.

Details

For the population total of the variable y:

$$t = \sum_{k \in U} y_k$$

the approximately unbiased Hajek (1971) estimator of t (implemented by the current function) is given by:

$$\hat{t}_{Hajek} = N \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k}$$

where $w_k = 1/\pi_k$ and π_k denotes the inclusion probability of the k-th element in the sample s. The variance of $\hat{t}_{Hajek}$ can be estimated by the Quenouille(1956); Tukey (1958) jackknife variance estimator (implemented by the current function):

$$\hat{V}(\hat{t}_{Hajek}) = \left(1 - \frac{n}{N}\right) \frac{n-1}{n} \sum_{k \in s} \left(\hat{t}_{Hajek(k)} - \hat{t}_{Hajek}\right)^2$$

where

$$\hat{t}_{Hajek(k)} = N \frac{\sum_{l \in s, l \neq k} w_l y_l}{\sum_{l \in s, l \neq k} w_l}$$

Note that we are implementing the Tukey (1958) jackknife variance estimator using the 'ad hoc' finite population correction $1 - n/N$ (see Shao and Tu, 1995; Wolter, 2007).

Value

The function returns a value for the estimated variance.

References

Hajek, J. (1971) Comment on *An essay on the logical foundations of survey sampling* by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Quenouille, M. H. (1956) Notes on bias in estimation. *Biometrika*, **43**, 353–360.

Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap*. Springer-Verlag, Inc.

Tukey, J. W. (1958) Bias and confidence in not-quite large samples (abstract). *The Annals of Mathematical Statistics*, **29**, 2, p. 614.

Wolter, K. M. (2007) *Introduction to Variance Estimation*. 2nd Ed. Springer, Inc.

See Also

VE.Jk.CBS.HT.Total.Hajek
VE.Jk.CBS.SYG.Total.Hajek
VE.Jk.B.Total.Hajek
VE.Jk.EB.SW2.Total.Hajek

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s     <- oaxaca$sHOMES00 #Defines the sample to be used for the example
N     <- dim(oaxaca)[1]  #Defines the population size
y1    <- oaxaca$POP10    #Defines the variable of interest y1
y2    <- oaxaca$POPMAL10 #Defines the variable of interest y2
#Computes the var. est. of the Hajek total point estimator using y1
VE.Jk.Tukey.Total.Hajek(y1[s==1], pik.U[s==1], N)
#Computes the var. est. of the Hajek total point estimator using y2
VE.Jk.Tukey.Total.Hajek(y2[s==1], pik.U[s==1], N)
```

VE.SYG.Mean.NHT *The Sen-Yates-Grundy variance estimator for the Narain-Horvitz-Thompson point estimator for a mean*

Description

Computes the Sen (1953); Yates-Grundy(1953) variance estimator for the Narain (1951); Horvitz-Thompson (1952) point estimator for a population mean.

Usage

```
VE.SYG.Mean.NHT(VecY.s, VecPk.s, MatPkl.s, N)
```

Arguments

`VecY.s`	vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s. There must not be any missing value.
`VecPk.s`	vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.
`MatPkl.s`	matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in MatPkl.s must be greater than zero and less than or equal to one. There must not be any missing value.
`N`	the population size.

Details

For the population mean of the variable y:

$$\bar{y} = \frac{1}{N}\sum_{k\in U} y_k$$

the unbiased Narain (1951); Horvitz-Thompson (1952) estimator of $\bar{y}$ is given by:

$$\hat{\bar{y}}_{NHT} = \frac{1}{N}\sum_{k\in s}\frac{y_k}{\pi_k}$$

where π_k denotes the inclusion probability of the k-th element in the sample s. Let π_{kl} denotes the joint-inclusion probabilities of the k-th and l-th elements in the sample s. The variance of $\hat{\bar{y}}_{NHT}$ is given by:

$$V(\hat{\bar{y}}_{NHT}) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}(\pi_{kl} - \pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

which, if the utilised sampling design is of fixed-size, can therefore be estimated by the Sen-Yates-Grundy variance estimator (implemented by the current function):

$$\hat{V}(\hat{\bar{y}}_{NHT}) = \frac{1}{N^2}\frac{-1}{2}\sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2$$

Value

The function returns a value for the estimated variance.

References

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

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

Sen, A. R. (1953) On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **5**, 119–127.

Yates, F. and Grundy, P. M. (1953) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, **15**, 253–261.

See Also

VE.HT.Mean.NHT
VE.Hajek.Mean.NHT

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s      <- oaxaca$sHOMES00 #Defines the sample to be used for the example
N      <- dim(oaxaca)[1]  #Defines the population size
y1     <- oaxaca$POP10    #Defines the variable of interest y1
y2     <- oaxaca$HOMES10  #Defines the variable of interest y2
#This approx. is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1]) #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the NHT point estimator for y1
VE.SYG.Mean.NHT(y1[s==1], pik.U[s==1], pikl.s, N)
#Computes the var. est. of the NHT point estimator for y2
VE.SYG.Mean.NHT(y2[s==1], pik.U[s==1], pikl.s, N)
```

VE.SYG.Total.NHT *The Sen-Yates-Grundy variance estimator for the Narain-Horvitz-Thompson point estimator for a total*

Description

Computes the Sen (1953); Yates-Grundy(1953) variance estimator for the Narain (1951); Horvitz-Thompson (1952) point estimator for a population total.

Usage

```
VE.SYG.Total.NHT(VecY.s, VecPk.s, MatPkl.s)
```

Arguments

`VecY.s` vector of the variable of interest; its length is equal to n, the sample size. Its length has to be the same as the length of VecPk.s. There must not be any missing value.

`VecPk.s` vector of the first-order inclusion probabilities; its length is equal to n, the sample size. Values in VecPk.s must be greater than zero and less than or equal to one. There must not be any missing value.

`MatPkl.s` matrix of the second-order inclusion probabilities; its number of rows and columns is equal to n, the sample size. Values in MatPkl.s must be greater than zero and less than or equal to one. There must not be any missing value.

Details

For the population total of the variable y:

$$t = \sum_{k \in U} y_k$$

the unbiased Narain (1951); Horvitz-Thompson (1952) estimator of t is given by:

$$\hat{t}_{NHT} = \sum_{k \in s} \frac{y_k}{\pi_k}$$

where π_k denotes the inclusion probability of the k-th element in the sample s. Let π_{kl} denotes the joint-inclusion probabilities of the k-th and l-th elements in the sample s. The variance of $\hat{t}_{NHT}$ is given by:

$$V(\hat{t}_{NHT}) = \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

which, if the utilised sampling design is of fixed-size, can therefore be estimated by the Sen-Yates-Grundy variance estimator (implemented by the current function):

$$\hat{V}(\hat{t}_{NHT}) = \frac{-1}{2} \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2$$

Value

The function returns a value for the estimated variance.

References

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

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

Sen, A. R. (1953) On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **5**, 119–127.

Yates, F. and Grundy, P. M. (1953) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, **15**, 253–261.

See Also

VE.HT.Total.NHT
VE.Hajek.Total.NHT

Examples

```
data(oaxaca) #Loads the Oaxaca municipalities dataset
#Reconstructs the 1st order incl. probs. for the example
pik.U  <- Pk.PropNorm.U(373, oaxaca$HOMES00)
s      <- oaxaca$sHOMES00 #Defines the sample to be used for the example
y1     <- oaxaca$POP10    #Defines the variable of interest y1
y2     <- oaxaca$HOMES10  #Defines the variable of interest y2
#This approximation is only suitable for large-entropy sampling designs
pikl.s <- Pkl.Hajek.s(pik.U[s==1]) #Approx. 2nd order incl. probs. from s
#Computes the var. est. of the NHT point estimator for y1
VE.SYG.Total.NHT(y1[s==1], pik.U[s==1], pikl.s)
#Computes the var. est. of the NHT point estimator for y2
VE.SYG.Total.NHT(y2[s==1], pik.U[s==1], pikl.s)
```

Chapter 5

Future research work

Abstract

Here we briefly discuss some possible extensions to the manuscripts of earlier chapters of this dissertation. Some future research work involve combining manuscripts.

Keywords and phrases: future research, variance estimation.

5.1 The importance of the research on variance estimation

Research on variance estimation is a crucial area within survey sampling. We can say that there is no doubt that estimating sampling variances is probably the most important way of improving a survey. For example, we need variance estimates to compute coefficients of variations and design effects. These two statistics are very important to improving survey sampling strategies as they indicate the gain or loss in precision when estimating certain attributes in a population.

5.2 Future work and extensions

5.2.1 On combining manuscripts from chapters 1 and 2

The variance estimator from chapter 1 is derived from the Campbell (1980); Berger and Skinner (2005) variance estimator that is defined for functions of Hájek (1971) means. One possible line of future research is to derive a similar variance estimator as the one in chapter 1 but this time from the proposed replication variance estimator in chapter 2 that is designed for functions of Narain (1951); Horvitz and Thompson (1952) totals.

5.2.2 On extending the manuscript from chapter 2 to account for imputation

In a similar fashion as the Berger and Skinner (2005) article was extended to account for imputation in Berger and Rao (2006), a possible extension for future research work is to extend the variance estimator proposed in chapter 2 to account for imputation.

5.2.3 On extending the manuscript from chapter 2 to ease its practical implementation (current research)

Another option of future research work, is to extend the variance estimator from chapter 2 in a similar way as the variance estimator from Berger and Skinner (2005) is extended in Berger (2007). This is current research already submitted.

5.2.4 On combining manuscripts from chapters 2 and 3

It would be interesting to combine the variance estimators utilised in chapters 2 and 3, to obtain a replication variance estimator for measures of change of complex statistics with rotating surveys.

Bibliography

Andersson, C., Andersson, K. and Lundquist, P. (2011a) Estimation of change in a rotation panel design. In *Proceeding of the 58th World Statistics Congress.* Dublin: International Statistical Institute.

Andersson, C., Andersson, K. and Lundquist, P. (2011b) *Variansskattningar avseende förändringsskattningar i panelundersökningar (Variance estimation of change in panel surveys.* Stockholm, Sweden: Methodology reports from Statistics Sweden (Statistiska Centralbyrån).

Andridge, R. R. and Little, R. J. A. (2010) A review of hot deck imputation for survey non-response. *International Statistical Review*, 78, 40–64.

Berger, Y. G. (2004) Variance estimation for measures of change in probability sampling. *Canadian Journal of Statistics*, 32, 451–467.

Berger, Y. G. (2007) A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika*, 94, 953–964.

Berger, Y. G. (2011) Asymptotic consistency under large entropy sampling designs with unequal probabilities. *Pakistan Journal of Statistics*, 27, 407–426.

Berger, Y. G. and Escobar, E. L. (2012) Variance estimation of imputed estimators of change over time from repeated surveys. In *Proceeding of the XIèmes Journées de Méthodologie Statistique de l'Insee.* Paris: Institut National de la Statistique et des Études Économiques (National Institute of Statistics and Economic Studies).

Berger, Y. G. and Escobar, E. L. (2013) Variance estimation of Hot-deck imputed estimators of change for rotating repeated surveys. (Submitted).

Berger, Y. G. and Priam, R. (2010) Estimation of correlations between cross-sectional estimates from repeated surveys: An application to the variance of change and the variance of the composite estimator. In *Proceeding of the 2010 International Methodology Symposium.* Ottawa: Statististics Canada.

Berger, Y. G. and Priam, R. (2012) Variance estimation of change in rotating repeated surveys. Submitted working paper available from the Authors.

Berger, Y. G. and Rao, J. N. K. (2006) Adjusted jackknife for imputation under unequal probability sampling without replacement. *Journal of the Royal Statistical Society B*, 68, 531–547.

Berger, Y. G. and Skinner, C. J. (2005) A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, 67, 79–89.

Berger, Y. G. and Tillé, Y. (2009) Sampling with unequal probabilities. In *Sample Surveys: Design, Methods and Applications* (eds. D. Pfeffermann and C. R. Rao), vol. 29A of *Handbook of Statistics*, 39–54. Amsterdam: Elsevier.

Binder, D. A. (1996) Linearization methods for single phase and two-phase samples: a cookbook approach. *Survey Methodology*, 22, 17–22.

Brewer, K. R. W. (1975) A simple procedure for πpswor. *Australian Journal of Statististics*, 17, 166–172.

Brewer, K. R. W. and Donadio, M. E. (2003) The high entropy variance of the Horvitz-Thompson estimator. *Survey Methodology*, 29, 189–196.

Brewer, K. R. W. and Gregoire, T. G. (2009) Introduction to survey sampling. In *Sample Surveys: Design, Methods and Applications* (eds. D. Pfeffermann and C. R. Rao), vol. 29A of *Handbook of Statistics*, 9–37. Amsterdam: Elsevier.

Brick, J. M. and Montaquila, J. M. (2009) Nonresponse and weighting. In *Sample Surveys: Design, Methods and Applications* (eds. D. Pfeffermann and C. R. Rao), vol. 29A of *Handbook of Statistics*, 163–185. Amsterdam: Elsevier.

Campbell, C. (1980) A different view of finite population estimation. *Proceeding of the Section on Survey Research Methods, American Statistical Association*, 319–324.

Chambers, R. L. and Clark, R. (2012) *An Introduction to Model-Based Survey Sampling with Applications.* Oxford: Oxford University Press.

Chambers, R. L. and Dunstan, R. (1986) Estimating distribution functions from survey data. *Biometrika*, 73, 597–604.

Chambers, R. L. and Skinner, C. J., eds. (2003) *Analysis of Survey Data.* Chichester: Wiley.

Chao, M. T. (1982) A general purpose unequal probability sampling plan. *Biometrika*, 69, 653–656.

Davison, A. C. and Hinkley, D. V. (1997) *Bootstrap Methods and Their Application.* Cambridge: Cambridge University Press.

Demnati, A. and Rao, J. N. K. (2004) Linearization variance estimators for survey data. *Survey Methodology*, 30, 17–26.

Demnati, A. and Rao, J. N. K. (2010) Linearization variance estimators for model parameters from complex survey data. *Survey Methodology*, 36, 193–201.

Deville, J. C. (1999) Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, 25, 193–203.

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

Deville, J. C. and Särndal, C. E. (1994) Variance estimation for the regression imputed horvitz-thompson estimator. *Journal of Official Statistics*, 10, 381–394.

Efron, B. (1979) Bootstrap methods: Another look at the Jackknife. *Annals of Statistics*, 7, 1–26.

Escobar, E. L. and Barrios, E. (2012) *samplingVarEst: Sampling Variance Estimation.* R package version 0.9-1.

Escobar, E. L. and Berger, Y. G. (2010) A novel jackknife variance estimator for two-stage without replacement sampling designs. In *Abstracts of Communications of the 10th International Vilnius Conference on Probability Theory and Mathematical Statistics*, 144. Vilnius, Lithuania: International Statistical Institute.

Escobar, E. L. and Berger, Y. G. (2011) Jackknife variance estimation for functions of Horvitz & Thompson estimators under unequal probability sampling without replacement. In *Proceeding of the 58th World Statistics Congress.* Dublin, Ireland: International Statistical Institute.

Escobar, E. L. and Berger, Y. G. (2013a) A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica*, 23, 595–613.

Escobar, E. L. and Berger, Y. G. (2013b) A new replicate variance estimator for unequal probability sampling without replacement. *Canadian Journal of Statistics*. (to appear).

Fay, B. E. (1991) A design-based perspective on missing data variance. In *Proceeding of the 1991 Annual Research Conference*, 429–440. U.S. Bureau of the Census.

Fay, B. E. (1994) Analyzing imputed survey datasets with model-assisted estimators. In *Proceedings of the Survey Research Methods Section*, 900–905. American Statistical Association.

Gambino, J. G. and Silva, P. L. N. (2009) Sampling and estimation in household surveys. In *Sample Surveys: Design, Methods and Applications* (eds. D. Pfeffermann and C. R. Rao), vol. 29A of *Handbook of Statistics*, 407–439. Amsterdam: Elsevier.

Goga, C., Deville, J. C. and Ruiz-Gazen, A. (2009) Use of functionals in linearization and composite estimation with application to two-sample survey data. *Biometrika*, 96, 691–709.

Hájek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35, 1491–1523.

Hájek, J. (1971) Comment on a paper by Basu, D. In *Foundations of Statistical Inference* (eds. V. P. Godambe and D. A. Sprott), 236. Toronto: Holt, Rinehart and Winston.

Hájek, J. (1981) *Sampling From a Finite Population.* New York: Dekker.

Haziza, D. (2009) Imputation and inference in the presence of missing data. In *Sample Surveys: Design, Methods and Applications* (eds. D. Pfeffermann and C. R. Rao), vol. 29A of *Handbook of Statistics*, 215–246. Amsterdam: Elsevier.

Haziza, D., Mecatti, F. and Rao, J. N. K. (2004) Comparison of variance estimators under Rao-Sampford method: a simulation study. In *Proceeding of the Survey Methods Section*, 3638–3643. American Statistical Association.

Haziza, D., Mecatti, F. and Rao, J. N. K. (2008) Evaluation of some approximate variance estimators under the Rao-Sampford unequal probability sampling design. *Metron - International Journal of Statistics*, LXVI, 91–108.

Holmes, D. J. and Skinner, C. J. (2000) *Variance Estimation for Labour Force Survey Estimates of Level and Change*. London, England: Technical report, Government Statistical Service Methodology Series, 21.

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

Huang, E. T. and Fuller, W. A. (1978) Nonnegative regression estimation for survey data. In *Proceedings of the Social Statistics Section*, 300–305. American Statistical Association.

Isaki, C. T. and Fuller, W. A. (1982) Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89–96.

Kalton, G. (2009) Design for surveys over time. In *Sample Surveys: Design, Methods and Applications* (eds. D. Pfeffermann and C. R. Rao), vol. 29A of *Handbook of Statistics*, 89–108. Amsterdam: Elsevier.

Kim, J. K. and Sitter, R. R. (2003) Efficient replication variance estimation for two-phase sampling. *Statistica Sinica*, 13, 641–653.

Kish, L. (1965) *Survey Sampling*. New York: Wiley.

Kish, L. (1995) The hunderd years' wars of survey sampling. *Statistics in Transition*, 2, 813–830.

Kish, L. and Frankel, M. R. (1974) Inference from complex samples. *Journal of the Royal Statistical Society B*, 36, 1–37.

Kovar, J. G., Rao, J. N. K. and Wu, C. F. J. (1988) Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics*, 16, 25–45.

Krewski, D. and Rao, J. N. K. (1981) Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010–1019.

Lehtonen, R. and Pahkinen, E. J. (2004) *Practical Methods for Design and Analysis of Complex Surveys*. Chichester: Wiley, 2nd edn.

Lohr, S. L. (1999) *Sampling: Design and Analysis*. London: Duxbury Press.

Midzuno, H. (1951) On the sampling system with probability proportionate to sum of sizes. *Annals of the Institute of Statistical Mathematics*, 3, 99–107.

Miller, R. G. (1964) A trustworthy jackknife. *The Annals of Mathematical Statistics*, 35, 1594–1605.

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, 169–174.

Nordberg, L. (2000) On variance estimation for measure of change when samples are coordinated by the use of permanent random numbers. *Journal of Official Statistics*, 16, 363–378.

Ohlsson, E. (1995) Coordination of samples using permanent random numbers. In *Business Survey Methods* (eds. B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge and P. S. Kott), chap. 9. Hoboken, New Jersey: Wiley.

Qualité, L. and Tillé, Y. (2008) Variance estimation of changes in repeated surveys and its application to the swiss survey of value added. *Survey Methodology*, 34, 173–181.

Quenouille, M. H. (1956) Notes on bias in estimation. *Biometrika*, 43, 353–360.

Rao, J. N. K. (1965) On two simple schemas of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3, 173–180.

Rao, J. N. K. and Shao, J. (1992) Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811–822.

Rao, J. N. K. and Sitter, R. R. (1995) Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453–460.

Rao, J. N. K., Wu, C. F. J. and Yue, K. (1992) Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 209–217.

Robinson, P. M. and Särndal, C. E. (1983) Asymptotic properties of the generalized regression estimator in probability sampling. *The Indian Journal Of Statistics: Sankhya B*, 45, 240–248.

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

Sampford, M. R. (1967) On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499–513.

Särndal, C. E. and Lundström, S. (2005) *Estimation in Surveys with Nonresponse.* Chichester: Wiley.

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

Sen, A. R. (1953) On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5, 119–127.

Shao, J. and Steel, P. (1999) Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254–265.

Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap.* New York: Springer.

Skinner, C. J. (2004) Comment on the paper: Linearization variance estimators for survey data, by Demnati, A. & Rao, J. N. K. *Survey Methodology*, 30, 17–26.

Smith, P., Pont, M. and Jones, T. (2003) Developments in business survey methodology in the office for national statistics, 1994–2000. *Journal of the Royal Statistical Society D (The Statistician)*, 52, 257–295.

Smith, T. M. F. (2001) Biometrika centenary: sample surveys. *Biometrika*, 88, 167–194.

Steel, P. and Fay, B. E. (1995) Variance estimation for finite populations with imputed data. In *Proceedings of the Survey Research Methods Section.* American Statistical Association.

Tam, S. M. (1984) On covariances from overlapping samples. *The American Statistician*, 38, 288–289.

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

Tukey, J. W. (1958) Bias and confidence in not-quite large samples (abstract). *The Annals of Mathematical Statistics*, 29, 614.

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

Wolter, K. M. (2007) *Introduction to Variance Estimation.* New York: Springer, 2nd edn.

Wood, J. (2008) On the covariance between related Horvitz-Thompson estimators. *Journal of Official Statistics*, 24, 53–78.

Yates, F. and Grundy, P. M. (1953) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, 15, 253–261.