Catalogue no. 12-001-XWE
ISSN: 1492-0921

Survey Methodology

June 2014

SURVEY
METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Statistics Canada Statistique Canada

Canada

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca,

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-877-287-4369

Depository Services Program

• Inquiries line	1-800-635-7943
• Fax line	1-800-565-7757

To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by "Key resource" > "Publications."

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "About us" > "The agency" > "Providing services to Canadians."

Published by authority of the Minister responsible for Statistics Canada

Cette publication est aussi disponible en français.

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is indexed in The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods. It is also covered by SCOPUS in the Elsevier Bibliographic Databases.

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year in electronic format. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (smj@statcan.gc.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the journal and on the web site (www.statcan.gc.ca/SurveyMethodology).

Survey Methodology

A Journal Published by Statistics Canada

Volume 40, Number 1, June 2014

Contents

Regular Papers

Short Notes

Survey Methodology, June 2014
Vol. 40, No. 1, pp. 1-13
Statistics Canada, Catalogue No. 12-001-X

Hierarchical Bayes Modeling of Survey-Weighted Small Area Proportions

Benmei Liu, Partha Lahiri and Graham Kalton [1]

Abstract

The paper reports the results of a Monte Carlo simulation study that was conducted to compare the effectiveness of four different hierarchical Bayes small area models for producing state estimates of proportions based on data from stratified simple random samples from a fixed finite population. Two of the models adopted the commonly made assumptions that the survey weighted proportion for each sampled small area has a normal distribution and that the sampling variance of this proportion is known. One of these models used a linear linking model and the other used a logistic linking model. The other two models both employed logistic linking models and assumed that the sampling variance was unknown. One of these models assumed a normal distribution for the sampling model while the other assumed a beta distribution. The study found that for all four models the credible interval design-based coverage of the finite population state proportions deviated markedly from the 95 percent nominal level used in constructing the intervals.

Key Words: Weighted proportions; Hierarchical Bayes modeling; Beta distribution; credible interval.

1 Introduction

Small area estimation methods are often used to estimate the proportions of units with a given characteristic for small areas. For example, small area estimation methods are used: in the Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program to estimate poverty rates for states, counties, and school districts (Citro and Kalton 2000; Maples and Bell 2005); with data from the National Survey on Drug Use and Health (NSDUH) to estimate substance rates for states (Wright, Sathe and Spagnola 2007); and with data from the National Assessment of Adult Literacy (NAAL) to estimate proportions at the lowest level of literacy for states and counties (Mohadjer, Rao, Liu, Krenzke and Van De Kerckhove 2012). In each case, the survey's sample sizes in the small areas are not large enough to support direct estimates of adequate precision. A wide variety of methods have been developed to address such small area estimation problems. See Rao (2003) and Jiang and Lahiri (2006a) for reviews, and Chattopadhyay, Lahiri, Larsen and Reimnitz (1999), Farrell, MacGibbon and Tomberlin (1997), Malec, Sedransk, Moriarity and LeClere (1997) and Malec, Davis and Cao (1999) for methods specifically for estimating small area proportions. The range of methods includes both empirical best prediction (EBP) and hierarchical Bayes (HB) approaches and models developed at both the area and unit levels. We focus on HB area level models in this paper.

When an HB area level model is used to produce estimates of proportions of units with a given characteristic for small areas, it is commonly assumed that the survey-weighted proportion for each sampled small area has a normal sampling distribution and that the sampling variance of this proportion is known. However, these assumptions are problematic when the small area sample size is small or when the true proportion is near 0 or 1. Reliance on the central limit theorem for approximate normality of the sampling distribution of a proportion requires reasonably large samples, particularly when the population

[1] Benmei Liu, Division of Cancer Control and Population Sciences, National Cancer Institute, 9609 Medical Center Drive Room 4E524, Bethesda, Maryland 20892; E-mail: liub2@mail.nih.gov; Partha Lahiri, JPSM, University of Maryland, 1218 Lefrak Hall, College Park, Maryland 20742; Graham Kalton, Westat, 1600 Research Boulevard, Rockville, Maryland 20850. The majority of this research took place while the first author was a graduate student in the Joint Program in Survey Methodology at the University of Maryland.

proportion is very small or very large (e.g., under 0.1 or over 0.9). Moreover, with very small or very large proportions, the sampling variance of a sample proportion is highly sensitive to the actual value of the proportion, thus making it difficult to establish a suitable value for the sampling variance. In an effort to overcome these problems, we propose two alternative models for small area proportions and compare them with two commonly used models. The models are described in Section 3. The four models are compared by means of a Monte Carlo simulation study in which stratified simple random samples are generated from a fixed finite population. The simulation study is described in Section 4 and the results are presented in Section 5. The paper finishes with some concluding remarks in Section 6. First, however, we introduce the notation for a stratified simple random sample design in Section 2.

2 Notation

Let N_{ih} denote the population size in stratum h in area i of a finite population $(i=1,\ldots,m; h=1,\ldots,H_i)$. Let y_{ihk} be the binary response for the characteristic of interest for unit k in stratum h in area i $(k=1,\ldots,N_{ih})$. The parameters to be estimated are the small area proportions $P_i = \sum_h \sum_k y_{ihk} / N_{ih}$.

With the stratified simple random sample design under study, n_{ih} units are selected from the N_{ih} units in stratum (ih). The standard direct survey estimator for P_i is:

$$p_{iw} = \frac{\sum_h^{H_i} \sum_k^{n_{ih}} w_{ih} y_{ihk}}{\sum_h^{H_i} \sum_k^{n_{ih}} w_{ih}}, \; i=1,\ldots,m; \tag{2.1}$$

where w_{ih} denotes the sampling weight given by $w_{ih} = N_{ih} / n_{ih}$.

The variance of p_{iw} can be expressed as

$$VAR_{st}(p_{iw}) = \frac{P_i(1-P_i)}{n_i} DEFF_i, \tag{2.2}$$

where $DEFF_i$ is the design effect reflecting the effect of the complex sample design (Kish 1965). For a stratified simple random sample with negligible sampling fractions in all strata, the design effect is given approximately by:

$$DEFF_i = \frac{\sum_h W_{ih}^2 P_{ih} \left(1-P_{ih}\right) / n_{ih}}{P\left(1-P_i\right) / n_i}, \tag{2.3}$$

where $W_{ih} = N_{ih} / N_i$, $N_i = \sum_h N_{ih}$, $n_i = \sum_h n_{ih}$ and P_{ih} is the population proportion in stratum h in area i.

The design effect $DEFF_i$ is a function of the P_{ih}, which are unknown. If $P_{ih}(1-P_{ih}) \approx P_i(1-P_i)$, $DEFF_i$ can be approximated by $deff_{iw} = n_i \sum_h W_{ih}^2 / n_{ih}$. The value of $deff_{iw}$ can be readily computed since it does not depend on any unknown parameters.

Small area estimation procedures can be used to address the problem that p_{iw} is very imprecise when the sample size n_i is small. Section 3 describes the HB area level models investigated in this study.

3 Models Studied

A general area-level small area model has two components. One—the sampling model—is a model for the sampling error of the direct survey estimates. The other—the linking model—relates the population value for an area to area-specific auxiliary variables $x_i = (x_{i1}, \ldots, x_{ip})'$.

Section 3.1 describes two area models that are often used for estimating small area proportions and Section 3.2 outlines some problems associated with these models. Section 3.3 describes two alternative models that may serve to address these problems.

3.1 Two Commonly Used Models

We study two commonly used models for comparison with the new models described in Section 3.4. The first is the Fay-Herriot model (Fay and Herriot 1979), which assumes known sampling variances and normal distributions for both the sampling and the linking models. The second is the normal-logistic model, which differs from the Fay-Herriot model only by the replacement of a logit-normal distribution for the normal distribution in the linking model.

Model 1: (Fay-Herriot normal-normal model)

Sampling model:

$$p_{iw} \mid P_i \overset{ind}{\sim} N(P_i,\ \psi_i) \tag{3.1}$$

Linking model:

$$P_i \mid \beta, \sigma_v^2 \overset{ind}{\sim} N(x_i^{'}\beta, \sigma_v^2) \tag{3.2}$$

Model 2: (normal-logistic model)

Sampling model:

$$p_{iw} \mid P_i \overset{ind}{\sim} N(P_i,\ \psi_i) \tag{3.3}$$

Linking model:

$$g(P_i) \mid \beta, \sigma_v^2 \overset{ind}{\sim} N(x_i^{'}\beta, \sigma_v^2) \tag{3.4}$$

In both models the sampling variance ψ_i is assumed to be known. Model 1 is referred as a matched model because the sampling and linking models can be combined to produce a relatively simple linear mixed model. However, a nonlinear linking model is often preferred for modeling proportions, leading to unmatched sampling and linking models, as in Model 2 (see, for example, You and Rao 2002). The link function $g(\cdot)$ can be empirically determined by checking the model fit. The *log* and *logit* link functions have been used. The $\text{logit}(P_i)$ linking model is chosen here in order to guarantee that the estimate of P_i always falls within the allowable range of (0,1).

3.2 Issues with Models 1 and 2

There are two main issues associated with Models 1 and 2. The first is that both models assume known sampling variances ψ_i, whereas in practice they have to be estimated. A simple approach is to use the direct variance estimate but that estimate is very imprecise when P_i is either very small or very large and when the sample size n_i is small. An alternative, more complex, approach is to develop an approximate estimate of P_i, say p_{isyn}, from a simple model such as a logistic model for p_{iw} in terms of the auxiliary variables, and then use that estimate in the following synthetic variance estimator:

$$\text{var}_{stsyn} = \frac{p_{isyn}(1 - p_{isyn})}{n_i} deff_{iw}. \tag{3.5}$$

When there are no auxiliary variables available, the overall sample proportion may be used for p_{isyn} in the computation of the synthetic variance estimator.

The second issue concerns the normality assumption in the sampling model, which is based on a large sample approximation. As noted in Section 1, when the sample size n_i is small and P_i is near 0 or 1, as is often the case with small area estimation, that assumption is problematic.

3.3 Two Alternative Models

Under Models 1 and 2, the unknown sampling variances ψ_i are estimated in some way, and then the resultant estimates are treated as if they were the known true values. A possible alternative approach is to treat the ψ_i as unknown parameters in the HB model, as has been done in a number of studies. For example, Arora and Lahiri (1997) applied an HB model to model the design-based variances for the sample estimates. Singh, Folsom and Vaish (2005) proposed the use of a generalized design effect model to smooth the sampling covariance matrix in small area modeling with survey data. Recently, You (2008) proposed the use of equal design effects over time to model the sampling variances in estimating small area unemployment rates using a cross-sectional and time series log-linear model. The approach of treating the sampling variances ψ_i as unknown is adopted in Model 3, as a variant of Model 2. One approach for addressing the non-normality of the sampling distributions of the survey-weighted small area proportions is to replace the normal distribution assumption by an alternative distribution. That approach is applied in Model 4 with the assumption of a beta sampling distribution, a distribution that has the desirable property of having a (0,1) range. In other regards Model 4 is the same as Model 3, including treating the ψ_i, $i = 1,...,m$ as unknown parameters. Model 4 was previously considered by Jiang and Lahiri (2006b) in an illustrative example to estimate finite population domain means using an EBP approach.

Model 3: (normal-logistic model with unknown sampling variance)

Sampling model:

$$p_{iw} \mid P_i \overset{ind}{\sim} N(P_i,\ \psi_i) \tag{3.6}$$

Linking model:

$$logit(P_i) \mid \beta, \sigma_v^2 \overset{ind}{\sim} N(x'\beta, \sigma_v^2) \qquad (3.7)$$

Model 4: (beta-logistic model with unknown sampling variance)

Sampling model:

$$p_{iw} \mid P_i \overset{ind}{\sim} beta(a_i, b_i) \qquad (3.8)$$

Linking model:

$$logit(P_i) \mid \beta, \sigma_v^2 \overset{ind}{\sim} N(x_i'\beta, \sigma_v^2) \qquad (3.9)$$

For both Model 3 and Model 4, the approximate variance function $\psi_i = [P_i(1 - P_i)/n_i]deff_{iw}$ is used. The parameters a_i and b_i in Model 4 are given by:

$$a_i = P_i\left(\frac{n_i}{deff_{iw}} - 1\right), \text{ and } b_i = (1 - P_i)\left(\frac{n_i}{deff_{iw}} - 1\right).$$

HB small area estimates can be computed from all four models using the Metropolis-Hastings algorithm within the Gibbs sampler. Details of the algorithm, which draws random samples based on the full conditional distributions of the unknown parameters starting with one or multiple sets of initial values, are given by Robert and Casella (1999) and Chen, Shao, and Ibraham (2000). You and Rao (2002) also describe in detail how the Metropolis-Hastings algorithm works within the Gibbs sampler for models similar to Models 1 and 2. The algorithm works for Models 3 and 4 in the same way as for Model 2. The full conditional distributions under each model are provided in Appendix A.

4 Simulation Study

4.1 The Study Population and the Sample Design

This section describes the simulation study that was conducted to compare the efficiency of the small area estimates produced by the four HB models. The simulation study was based on the 2002 Natality public-use data file that covered all births occurring within the United States in that calendar year. The file contained data obtained from the certificates filed for births occurring in each state and territory (for details see U.S. National Center for Health Statistics 2009).

The finite population studied was restricted to the 4,024,378 records of live births that occurred in 2002 in the 50 states of U.S. and the District of Columbia (DC) and that had birth weights reported. The parameter of interest is the state level low birthweight rate P_i, $i = 1, ..., 51$, where low birthweight is defined as less than 2,500 grams. The value of P_i varied from 5 percent to 11 percent across the states.

Within each state, a stratified SRS design was used to draw samples from the birth records. Mother's race (White, Black, and Other) was used as the stratification variable. The national sample size was set to be about 1,500 birth records for each race group. A uniform sampling fraction was used across the states for each race group, subjecting to the condition that at least two birth records were sampled within each race group in each state. The resultant national sample size turned out to be $n = 4,526$ birth records. The state sample sizes n_i ranged from 7 (for small states such as Vermont) to 690 (for California), with a median sample size of 61. This sampling procedure was repeated $R = 1,000$ times, creating 1,000 independent sample data sets. The sampling weights remained the same over different simulation runs.

4.2 Computation of the HB Estimates

For simplicity, the following assumptions were made for the HB models:

1. No auxiliary variables were used, so that $x_i{}'\beta = \mu$.
2. For Models 1 and 2, the sampling variances were taken to be given by $\psi_i = [p_w(1 - p_w)/n_i]deff_{iw}$, where $p_w = \sum\sum\sum w_{ih} y_{ihk} / \sum\sum n_i w_{ih}$ is the national estimate of the proportion of low birthweight live births. (A check on the use of $deff_{iw}$ as an approximation for $DEFF_i$ showed that the approximation was reasonable: the two quantities were close, with a product moment correlation of 0.96 and an average ratio of 1.08 between $deff_{iw}$ and $DEFF_i$.)
3. Flat prior for μ, i.e., $f(\mu) \propto 1$, and inverse gamma for σ_v^2, i.e., $\sigma_v^2 \sim IG(0.001,\ 0.001)$.

For each sample data set, the first step in the computations was to calculate the state direct sample estimates. The estimates for each sample data set were then used in turn as input to the WinBUGS software (Lunn, Thomas, Best and Spiegelhalter 2000), which was used to produce the HB estimates for all four models.

In a sizable number of the states with small n_i, the direct estimates were zero in some of the sample data sets. Since WinBUGS can handle direct estimates of zero only for Model 1, the zero direct estimates were perturbed to very small positive numbers for the other models.

For each WinBUGS run, three independent chains were used. For each chain, burn-ins of 10,000 samples were produced, with 10,000 samples after burn-in. The samples after burn-in were thinned by a factor of two to reduce auto-correlation of the MCMC samples. The resultant 15,000 MCMC samples from the three chains after burn-in were then used to compute the posterior mean and percentiles for each HB model based on each sample data set. The potential scale reduction factor $\hat{R}$ was used as the primary measure for convergence (see Gelman and Rubin 1992). The WinBUGS code is given in Appendix B.

5. Simulation Results

In Section 5.1 we report our main results for the credible intervals obtained for the state proportions of low birthweight live births from the application of each of the four models. Section 5.2 then examines the biases and root mean square errors of these estimates.

5.1 Model estimates and credible intervals

Let P_i^{HB} denote an HB estimator of P_i, the percentage of low birthweight live births in state i, and let $P_{i,q}^{HB}$ denote the q^{th} percentile of the posterior distribution of P_i. Based on the results from the 1,000 simulation data sets, Table 5.1 presents the following for each model: the noncoverage probability for the 95 percent credible intervals of P_i , i.e., the probability that the interval from $P_{i,.025}^{HB}$ to $P_{i,.975}^{HB}$ fails to cover P_i and the mean width of the credible intervals $P_{i,.975}^{HB}$ - $P_{i,.025}^{HB}$. The corresponding Monte Carlo simulation standard errors are also reported in the table in parentheses.

To examine the effect of state sample size on the simulation results, the 50 states and the District of Columbia are divided into three groups according to their sample sizes: the 15 states with small sample sizes ($n_i \leq 30$); the 24 states with medium sample sizes ($30 < n_i \leq 100$); and the 12 states with large sample sizes ($n_i > 100$). The results presented in Table 5.1 are overall averages across all states and averages for the three groups separately.

It can be seen from the upper half of Table 5.1 that the Fay-Herriot model (M1) credible intervals are very conservative, giving nearly zero noncoverage. The lower half of the table shows that this result is obtained at the cost of the largest average credible interval width among the four models. The M1 credible interval widths are very stable. A small proportion of the M1 credible intervals had negative lower bounds.

A possible explanation for the low level of noncoverage with M1 is that the sampling variances were overestimated, perhaps because $deff_{iw}$ was used in place of $DEFF_i$. To examine this possibility, we used $DEFF_i$ in computing the sampling variance and found virtually no difference in the noncoverage rate. We also ran the model with the true variance as defined in (2.2) and again found no appreciable difference in the noncoverage rates. The non-normality of the sampling distribution of p_{iw} could also be a source of this problem.

Table 5.1
Percentage of times that the 95 percent credible intervals fail to cover P_i, mean 95 percent credible interval width, along with the Monte Carlo simulation standard errors based on 1,000 simulations (in percentages)

State sample size n_i	M1*	M2	M3	M4
	Noncoverage percentage (Monte Carlo simulation standard error)			
Overall	0.40 (0.028)	8.24 (0.109)	6.52 (0.101)	4.36 (0.088)
$n_i \leq 30$ (15 states)	0.05 (0.019)	11.39 (0.239)	8.45 (0.216)	6.21 (0.190)
$30 < n_i \leq 100$ (24 states)	0.46 (0.043)	9.44 (0.167)	7.61 (0.156)	4.52 (0.132)
$n_i > 100$ (12 states)	0.70 (0.076)	1.91 (0.122)	1.94 (0.124)	1.74 (0.119)
	Mean width of the 95% credible interval (Monte Carlo simulation standard error)			
Overall	9.05 (0.004)	5.52 (0.009)	6.20 (0.009)	8.45 (0.014)
$n_i \leq 30$ (15 states)	10.27 (0.009)	5.94 (0.020)	6.78 (0.021)	9.30 (0.034)
$30 < n_i \leq 100$ (24 states)	9.16 (0.005)	5.60 (0.013)	6.28 (0.013)	8.71 (0.021)
$n_i > 100$ (12 states)	7.29 (0.004)	4.84 (0.012)	5.30 (0.013)	6.88 (0.017)

*Note: For Model 1, a small proportion of the credible intervals had negative lower bounds.

At 8.2 percent, the overall noncoverage rate of the credible intervals for the normal-logistic model (M2) is appreciably above the nominal rate of 5 percent. This model has the smallest average interval width. The noncoverage rate for the normal-logistic model with unknown variance (M3) is closer to the nominal rate, with an overall interval width that is somewhat larger than that for M2.

The noncoverage rate for the beta-logistic model (M4) of 4.4 percent overall is closest to the nominal noncoverage rate. However, the average width of the credible intervals is larger than those for M2 and M3 and the Monte Carlo standard error of the interval width is larger than that of the other three models. This instability may be due to the complexity of the full conditional distribution for the beta model. The large proportion of the 1,000 direct estimates that were 0 for some of the states with small sample sizes may also have caused significant problems in fitting the beta distribution.

As is to be expected, for all four models the mean width of the credible intervals declines with increasing state sample size and the variation in the widths also declines with increased sample size. Even with these declines, however, the noncoverage rates also decline with increasing sample size for Models 2, 3, and 4. The noncoverage rates are in fact very small for the states with large n_i, suggesting that the credible intervals are not adequately reflecting the effect of the greater precision of the direct estimates in the states with large sample sizes.

5.2 Biases and RMSEs of the model-based estimates

For further investigation of these results, we examined the bias and the root mean square errors (RMSEs) of the estimates P_i^{HB} for each model. The results are presented in Table 5.2 in the same format as Table 5.1. The biases for the estimates under M1, M2, and M3 exhibit a similar pattern: the biases are large and positive for the small states, and offset to some extent by relatively small negative biases for the medium and large states. The biases for the estimates for M4 have a very different pattern: they are almost zero for the small states and have large negative values for the medium and large states. This indicates that M4 would perform better than the other three models in terms of bias when the small area sample sizes are small.

Table 5.2
The biases and the root mean square errors of the estimates of P_i based on the four models (in percentages)

State sample size n_i	M1		M2		M3		M4	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
Overall	0.165	1.518	0.071	1.346	-0.009	1.411	-0.214	1.712
$n_i \leq 30$ (15 states)	0.621	1.651	0.572	1.630	0.466	1.652	0.009	1.922
$30 < n_i \leq 100$ (24 states)	-0.006	1.547	-0.123	1.386	-0.201	1.452	-0.319	1.775
$n_i > 100$ (12 states)	-0.063	1.294	-0.167	0.911	-0.219	1.026	-0.283	1.323

6. Discussion

In this paper, we report the results of a simulation study from a real finite population to evaluate the credible intervals obtained from four different hierarchical models in terms of their interval lengths and their design-based coverage properties. To the best of our knowledge, such a design-based evaluation of small area credible (or confidence) intervals has not previously been performed in the evaluation of small area estimates.

In the simulation study, we have compared the design-based coverage properties of credible intervals resulting from different hierarchical Bayes models for estimating small area proportions from a stratified simple random sample design. Overall, none of the models emerges as a clear winner and so we are not in a position to recommend any of the models studied.

The hierarchical Bayes version of the well-known Fay-Herriot model appears to produce overly conservative credible intervals. The non-normality of both the sampling and the linking models is a possible source of this problem. The credible intervals for the beta-logistic hierarchical model achieve almost the nominal coverage for the finite population proportions and the bias property for this model is the best among the four models being compared when the sample sizes are small. However, since one of the full conditionals for the beta-logistic model involves the survey-weighted proportions, there is a problem with the MCMC whenever the survey-weighted proportion is zero. The credible intervals for this model are also wider than those for the other two models with a logistic linking model. It may be possible to reduce the width of the credible interval for the beta-logistic model by modifying the model in some way, such as by employing a suitable two-part mixture random effect model that will avoid the problem of survey-weighted proportions of zero. Further investigation is needed. Also consideration could usefully be given to other possible models, possibly a discrete probability model for Level 1, to improve on interval estimation of small proportions for small areas.

The simulation study found that the coverage of the Bayesian credible intervals for the finite population proportions was far from the nominal 95 percent level for all four models, and a similar finding was also obtained for the design-based coverage of the widely-used Fay-Herriot model. In the light of these findings we carried out a number of further analyses in a search for an explanation. These analyses included: adding predictor variables to the models; using a uniform prior distribution for σ_v^2 (based on arguments made by Gelman 2006); the use of empirical best prediction approach for the M1 model; increasing the sample size in states with few births to a minimum of 50; and applying the methods to estimate the proportion of births in each state below the national median birthweight. Although there are some differences in the coverage properties for the state finite population proportions, none of these analyses produced coverage rates close to the nominal rates. The only case where the nominal rates coincided with the actual coverage rates was for a simulated dataset constructed under model M1 for the state proportions below the national median birthweight; the average coverage rates were 5.1 and 5.2 percent for the EBP and HB approaches, respectively.

This simulation study was restricted to a single stage sample design. In addition, for simplicity no auxiliary variables were included in the linking models in our main analyses, whereas in practice the inclusion of such variables is routine and almost essential. Further simulation studies are needed to cover different sample designs, different sample sizes, and to incorporate some auxiliary variables in the linking models. We hope that our study will encourage others to conduct similar design-based simulations to evaluate small area estimation methods. Based on our limited results, users of small area estimates need to be cautioned about the interpretation of the credible intervals associated with the estimates.

Acknowledgements

The authors wish to thank the associate editor and two referees for a number of constructive suggestions which led to substantial improvement of the original manuscript. The second author's research was supported by the National Science Foundation SES-085100.

Appendix A

A1. Full conditional distributions for the parameters of each model

Let $\vec{p}=(p_{1w},\ldots,p_{mw})^t$ and $r_i=\dfrac{\psi_i}{\psi_i+\sigma_v^2}$.

The full conditional distributions for the Fay-Herriot model (M1) are given as follows:

i) $\theta_i \mid \mu,\sigma_v^2,\vec{p} \sim N((1-r_i)p_{iw}+r_i\mu,\quad \psi_i(1-r_i))$;

ii) $\mu \mid \theta_i,\sigma_v^2,p \sim N\left(\dfrac{1}{m}\sum_{i=1}^{m}\theta_i,\dfrac{\sigma_v^2}{m}\right)$;

iii) $\sigma_v^2 \mid \mu,\theta_i,\vec{p} \sim ING\left(a+\dfrac{1}{2}m,b+\dfrac{1}{2}\sum_{i=1}^{m}(\theta_i-\mu)^2\right)$.

The full conditional distributions for the Normal-Logistic model (M2) are given as follows:

i) $\theta_i \mid \mu,\sigma_v^2,\vec{p} \propto \dfrac{1}{\theta_i(1-\theta_i)\sigma_v\sqrt{\psi_i}}\exp\left(-\dfrac{(p_{iw}-\theta_i)^2}{2\psi_i}-\dfrac{(\operatorname{logit}(\theta_i)-\mu)^2}{2\sigma_v^2}\right)$;

ii) $\mu \mid \theta_i,\sigma_v^2,p \sim N\left(\dfrac{1}{m}\sum_{i=1}^{m}\operatorname{logit}(\theta_i),\ \dfrac{\sigma_v^2}{m}\right)$;

iii) $\sigma_v^2 \mid \mu,\theta_i,\vec{p} \sim ING\left(a+\dfrac{1}{2}m,b+\dfrac{1}{2}\sum_{i=1}^{m}(\operatorname{logit}(\theta_i)-\mu)^2\right)$.

The full conditional distributions for the Normal-Logistic model with unknown variance (M3) are the same as those of M2 except that replacing ψ_i by $\theta_i(1-\theta_i)deff_{iw}/n_i$ for the distribution of θ_i given other parameters.

Let $\delta_{iw}=\dfrac{n_i}{deff_{iw}}-1$. The full conditional distributions for the Beta-Logistic model (M4) are given as follows:

i) $\theta_i \mid \mu,\sigma_v^2,\vec{p} \propto \dfrac{1}{\theta_i(1-\theta_i)\sigma_v}\dfrac{p_{iw}^{\theta_i\delta_{iw}-1}(1-p_{iw})^{(1-\theta_i)\delta_{iw}-1}}{\Gamma(\theta_i\delta_{iw})\Gamma((1-\theta_i)\delta_{iw})}\exp\left(-\dfrac{(\operatorname{logit}(\theta_i)-\mu)^2}{2\sigma_v^2}\right)$;

ii) $\mu \,|\, \theta_i, \sigma_v^2, p \sim N\left(\frac{1}{m}\sum_{i=1}^{m}\text{logit}(\theta_i),\ \frac{\sigma_v^2}{m}\right)$;

iii) $\sigma_v^2 \,|\, \mu, \theta_i, \vec{p} \sim ING\left(a + \frac{1}{2}m, b + \frac{1}{2}\sum_{i=1}^{m}(\text{logit}(\theta_i) - \mu)^2\right)$.

Appendix B

WinBUGS code for Model 1:

```
model {
      for ( i in 1:N)  {
         pobs[i] ~ dnorm(theta[i], D[i])
         D[i] <- 1/varhat[i]
         theta[i]<-u+v[i]
         v[i]~dnorm(0, tau)
                          }
        u~dflat()
       tau~dgamma(0.001, 0.001)
       sigma_v2<-1/tau
          }
```

WinBUGS code for Model 2:

```
model {
      for ( i in 1:N)  {
         pobs[i] ~ dnorm(theta[i], D[i])
         D[i] <- 1/varhat[i]
         logit(theta[i])<-u+v[i]
         v[i]~dnorm(0, tau)
                           }
       u~dflat()
       tau~dgamma(0.001, 0.001)
       sigma_v2<-1/tau
          }
```

WinBUGS code for Model 3:

```
model {
      for ( i in 1:N)  {
         pobs[i] ~ dnorm(theta[i], E[i])
         E[i] <- SAMPn[i]/(theta[i]*(1-theta[i])*DEFF_kish[i])
         logit(theta[i])<-u+v[i]
         v[i]~dnorm(0, tau)
```

```
          D[i]<-1/E[i]
                           }
        u~dflat()
        tau~dgamma(0.001, 0.001)
        sigma_v2<-1/tau
           }
```

WinBUGS code for Model 4:

```
model {
      for ( i in 1:N)  {
         pobs[i] ~ dbeta(a[i], b[i])
         a[i] <- theta[i]*(theta[i]*(1-theta[i])/D[i]-1)
         b[i] <- (1-theta[i])*(theta[i]*(1-theta[i])/D[i]-1)
         logit(theta[i])<-u+v[i]
         v[i]~dnorm(0, tau)
         D[i]<-theta[i]*(1-theta[i])*DEFF_kish[i]/SAMPn[i]
                           }
        u~dflat()
        tau~dgamma(0.001, 0.001)
        sigma_v2<-1/tau
           }
```

References

Arora, V. and Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica*, 7, 1053-1063.

Chattopadhyay, M., Lahiri, P., Larsen, M., and Reimnitz, J. (1999). Composite estimation of drug prevalence for sub-state areas. *Survey Methodology*, 25, 81-86.

Chen, M., Shao, Q., and Ibraham, J.G. (2000). *Monte Carlo Methods in Bayesian Computation.* New York: Springer-Verlag.

Citro, C., and Kalton, G. (Eds.) (2000). *Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond*. Washington, DC: National Academy Press.

Farrell, P.J., MacGibbon, B., and Tomberlin, T.J. (1997). Empirical Bayes estimators of small area proportions in multistage designs. *Statistical Sinica*, 7, 1065-1083.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515-533.

Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequence. *Statistical Science*, 7, 457-472.

Jiang, J., and Lahiri, P. (2006a). Mixed model prediction and small area estimation. *Test*, 15, 111-999.

Jiang, J., and Lahiri, P. (2006b). Estimation of finite population domain means: A model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 101, 301-311.

Kish, L. (1965). *Survey sampling*. New York: John Wiley.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS – A Bayesian modeling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10, 325-337.

Malec, D., Davis, W., and Cao, X. (1999). Small area estimates of overweight prevalence using sample selection adjustment. *Statistics in Medicine*, 18, 3189-3200.

Malec, D., Sedransk, J., Moriarity, C.L., and LeClere, F.B. (1997). Small area inference for binary variables in the National Health Interview Survey. *Journal of the American Statistical Association,* 92, 815-826.

Maples, J., and Bell, W.R. (2005). Evaluation of school district poverty estimates: Predictive models using IRS income tax data. *Proceedings of the Survey Research Methods Section, American Statistical Association,* 1322-1329.

Mohadjer, L., Rao, J.N.K., Liu, B., Krenzke, T., and Van De Kerckhove, W. (2012). Hierarchical Bayes small area estimates of adult literacy using unmatched sampling and linking models. *Journal of the Indian Society of Agricultural Statistics,* 66 (1), 55-63.

Rao, J.N.K. (2003). *Small area estimation*. New York: John Wiley and Sons.

Robert, C.P., and Casella, G. (1999). *Monte Carlo Statistical Methods*. New York: Springer-Verlag.

Singh, A.C., Folsom, R.E. JR., and Vaish, A.K. (2005). Small area modeling for survey data with smoothed error covariance structure via generalized design effects. *Federal Committee on Statistical Methods Statistical Working Paper*, No. 39 (http://www.fcsm.gov/05papers/Singh_etal_IIIC.pdf).

U.S. National Center for Health Statistics (2009). *National Vital Statistics System*. *Birth Data*. (http://www.cdc.gov/nchs/births.htm)

Wright, D., Sathe, N., and Spagnola, K. (2007). *State Estimates of Substance Use from the 2004-2005 National Surveys on Drug Use and Health.* (DHHS Publication No. SMA 07-4235, NSDUH Series H-31). Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

You, Y. (2008). An integrated modeling approach to unemployment rate estimation for subprovincial areas of Canada. *Survey Methodology*, 34, 19-27.

You, Y., and Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *Canadian Journal of Statistics*, 30, 3-15.

Survey Methodology, June 2014
Vol. 40, No. 1, pp. 15-28
Statistics Canada, Catalogue No. 12-001-X

Bayes linear estimation for finite population with emphasis on categorical data

Kelly Cristina M. Gonçalves, Fernando A.S. Moura and Helio S. Migon[1]

Abstract

Bayes linear estimator for finite population is obtained from a two-stage regression model, specified only by the means and variances of some model parameters associated with each stage of the hierarchy. Many common design-based estimators found in the literature can be obtained as particular cases. A new ratio estimator is also proposed for the practical situation in which auxiliary information is available. The same Bayes linear approach is proposed for obtaining estimation of proportions for multiple categorical data associated with finite population units, which is the main contribution of this work. A numerical example is provided to illustrate it.

Key Words: Exchangeability; Linear model; Bayesian linear prediction.

1 Introduction

Surveys have long been an important way of obtaining accurate information from a finite population. For instance, governments need to obtain descriptive statistics of the population for purposes of evaluating and implementing their policies. For those concerned with official statistics in the first third of the twentieth century, the major issue was to establish a standard of acceptable practice. Neyman (1934) created such a framework by introducing the role of randomization methods in the sampling process. He advocated the use of the randomization distribution induced by the sampling design to evaluate the frequentist properties of alternative procedures. He also introduced the idea of stratification with optimal sample size allocation and the use of unequal selection probabilities. His work was recognized as the cornerstone of design-based sample survey theory and inspired many other authors. For example, Horvitz and Thompson (1952) proposed a general theory of unequal probability sampling and the probability-weighted estimation method, the so-called “Horvitz and Thompson’s estimator”.

The design-based sample survey theory has been very appealing to official statistics agencies around the world. As pointed out by Skinner, Holt and Smith (1989), page 2, the main reason is that it is essentially distribution-free. Indeed, all advances in survey sampling theory from Neyman onwards have been strongly influenced by the descriptive use of survey sampling. The consequence of this has been a lack of theoretical developments related to the analytic use of surveys, in particular for prediction purposes. In some specific situations, the design-based approach has proved to be inefficient, providing inadequate predictors. For instance, estimation in small domains and the presence of the non-response cannot be dealt with by the design-based approach without some implicit assumptions, which is equivalent to assuming a model. Supporters of the design-based approach argue that model-based inference largely depends on the model assumptions, which might not be true. On the other hand, interval inference for target population parameters (usually totals or means) relies on the Central Limit Theorem, which cannot

[1] Kelly Cristina M. Gonçalves, Departamento de Estatística, Universidade Federal do Rio de Janeiro (UFRJ), RJ, Brazil. E-mail: kelly@im.ufrj.br; Fernando A.S. Moura, Departamento de Estatística, Universidade Federal do Rio de Janeiro (UFRJ), RJ, Brazil. E-mail: fmoura@im.ufrj.br; Helio S. Migon, Departamento de Estatística, Universidade Federal do Rio de Janeiro (UFRJ), RJ, Brazil. E-mail: migon@im.ufrj.br.

be applied in many practical situations, where the sample size is not large enough and/or independence assumptions of the random variables involved are not realistic.

Basu (1971) did not accept estimates of population quantities which depend on the sampling rule, like the inclusion probabilities. He argued that this estimation procedure does not satisfy the likelihood principle, at which he was adept. Basu (1971) created the circus elephant example to show that the Horvitz-Thompson estimator could lead to inappropriate estimates and proposed an alternative estimator. The question that arises is whether it is possible to conciliate both approaches. In the superpopulation model context, Zacks (2002) showed that some design-based estimators can be recovered by using a general regression model approach. Little (2003) claims that: "careful model specification sensitive to the survey design can address the concerns with model specifications, and Bayesian statistics provide a coherent and unified treatment of descriptive and analytic survey inference". He gave some illustrative examples of how standard design-based inference can be derived from the Bayesian perspective, using some models with non-informative prior distributions.

In the Bayesian context, another appealing proposal to conciliate the design-based and model-based approaches was proposed by Smouse (1984). The method incorporates prior information in finite population inference models by relying on Bayesian least squares techniques and requires only the specification of first and second moments of the distributions involved, describing prior knowledge about the structures present in the population. The approach is an alternative to the methods of randomization and appears midway between two extreme views: on the one hand the design-based procedures and on the other those based on superpopulation models. O'Hagan (1985), in an unpublished report, presented the Bayes linear estimators in some specific sample survey contexts and O'Hagan (1987) also derived Bayes linear estimators for some randomized response models. O'Hagan (1985) dealt with several population structures, such as stratification and clustering, by assuming suitable hypotheses about the first and second moments and showed how some common design-based estimators can be obtained as a particular case of his more general approach. He also pointed out that his estimates do not account for non-informative sampling. He quoted Scott (1977) and commented that informative sampling should be carried out by a full Bayesian analysis. An important reference about informative sampling dealing with hierarchical models can be found in Pfeffermann, Moura and Silva (2006).

The paper is organized as follows. Section 2 generally describes the Bayes linear estimation approach applied to a general linear regression model for finite population prediction and shows how to obtain some design-based estimators as particular cases. In Section 3, a new ratio estimator is proposed for practical situation in which auxiliary information is available. Section 4 extends the Bayes linear estimation approach to multiple categorical data. Finally, Section 5 offers some conclusions and suggestions for further research.

2 Bayes linear estimation for finite population

The Bayes approach has been found to be successful in many applications, particularly when the data analysis has been improved by expert judgements. But while Bayesian models have many appealing features, their application often involves the full specification of a prior distribution for a large number of parameters. Goldstein and Wooff (2007), section 1.2, argue that as the complexity of the problem increases, our actual ability to fully specify the prior and/or the sampling model in detail is impaired. They conclude that in such situations, there is a need to develop methods based on partial belief specification.

Hartigan (1969) proposed an estimation method, termed Bayes linear estimation approach, that only requires the specification of first and second moments. The resulting estimators have the property of minimizing posterior squared error loss among all estimators that are linear in the data and can be thought of as approximations to posterior means. The Bayes linear estimation approach is fully employed in this article and is briefly described below.

2.1 Bayes linear approach

Let $\mathbf{y}_s$ be the vector with observations and $\boldsymbol{\theta}$ be the parameter to be estimated. For each value of $\boldsymbol{\theta}$ and each possible estimate $\mathbf{d}$, belonging to the parametric space $\boldsymbol{\Theta}$, we associate a quadratic loss function $L(\boldsymbol{\theta}, \mathbf{d}) = (\boldsymbol{\theta} - \mathbf{d})'(\boldsymbol{\theta} - \mathbf{d}) = tr(\boldsymbol{\theta} - \mathbf{d})(\boldsymbol{\theta} - \mathbf{d})'$. The main interest is to find the value of $\mathbf{d}$ that minimizes $r(\mathbf{d}) = E[L(\boldsymbol{\theta}, \mathbf{d}) | \mathbf{y}_s]$, the conditional expected value of the quadratic loss function given the data.

Suppose that the joint distribution of $\boldsymbol{\theta}$ and $\mathbf{y}_s$ is partially specified by only their first two moments:

$$\begin{pmatrix} \boldsymbol{\theta} \\ \mathbf{y}_s \end{pmatrix} \sim \left[\begin{pmatrix} \mathbf{a} \\ \mathbf{f} \end{pmatrix}, \begin{pmatrix} \mathbf{R} & \mathbf{AQ} \\ \mathbf{QA}' & \mathbf{Q} \end{pmatrix} \right], \tag{2.1}$$

where $\mathbf{a}$ and $\mathbf{f}$, respectively, denote mean vectors and $\mathbf{R}$, $\mathbf{AQ}$ and $\mathbf{Q}$ the covariance matrix elements of $\boldsymbol{\theta}$ and $\mathbf{y}_s$.

The Bayes linear estimator (BLE) of $\boldsymbol{\theta}$ is the value of $\mathbf{d}$ that minimizes the expected value of this quadratic loss function within the class of all linear estimates of the form $\mathbf{d} = \mathbf{d}(\mathbf{y}_s) = \mathbf{h} + \mathbf{H}\mathbf{y}_s$, for some vector $\mathbf{h}$ and matrix $\mathbf{H}$. Thus, the BLE of $\boldsymbol{\theta}$, $\hat{\mathbf{d}}$, and its associated variance, $\hat{V}(\hat{\mathbf{d}})$, are respectively given by:

$$\hat{\mathbf{d}} = \mathbf{a} + \mathbf{A}(\mathbf{y}_s - \mathbf{f}) \text{ and } \hat{V}(\hat{\mathbf{d}}) = \mathbf{R} - \mathbf{AQA}'. \tag{2.2}$$

It should be noted that the BLE depends on the specification of the first and second moments of the joint distribution partially specified in (2.1). The issue of eliciting these quantities is dealt with in sections (2.3.1) and (4.1) for some particular cases.

2.2 Bayes linear approach to finite population

Consider $U = \{u_1, \ldots, u_N\}$ a finite population with N units. Let $\mathbf{y} = (y_1, \ldots, y_N)'$ be the vector with the values of interest of the units in U. The response vector $\mathbf{y}$ is partitioned into the known observed n-sample vector $\mathbf{y}_s$, and the non-observed vector $\mathbf{y}_{\bar{s}}$ of dimension $N - n$. The general problem is to predict a function of the vector $\mathbf{y}$, such as the total $T = \sum_{i=1}^{N} y_i = \mathbf{1}'_s \mathbf{y}_s + \mathbf{1}'_{\bar{s}} \mathbf{y}_{\bar{s}}$, where $\mathbf{1}_s$ and $\mathbf{1}_{\bar{s}}$ are the vectors of 1's of dimensions n and $N - n$, respectively. In the model-based approach, this is usually done by assuming a parametric model for the population values y_i's and then obtaining the Empirical Best Linear Unbiased Predictor (EBLUP) for the unknown vector $\mathbf{y}_{\bar{s}}$ under this model. Usually, the mean square error of the EBLUP of T is obtained by second order approximation, as well as an unbiased estimator of it. See Valliant, Dorfman and Royall (2000), chapter 2, for details.

The Bayesian approach to finite population prediction often assumes a parametric model, but it aims to find the posterior distribution of T given $\mathbf{y}_s$. Point estimates can be obtained by setting a loss function, although in many practical problems, the posterior mean is often considered and its associated variance is given by the posterior variance, *i.e.*:

$$E(T \mid \mathbf{y}_s) = \mathbf{1}'_s \mathbf{y}_s + \mathbf{1}'_{\bar{s}} E(\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s) \text{ and } V(T \mid \mathbf{y}_s) = \mathbf{1}'_{\bar{s}} V(\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s) \mathbf{1}_{\bar{s}}. \tag{2.3}$$

It is possible to obtain an approximation to the quantities in (2.3) by using a Bayes linear estimation approach. Here, we will particularly obtain the estimators by assuming a general two-stage hierarchical model for finite population, specified only by its mean and variance-covariance matrix, presented in Bolfarine and Zacks (1992), page 76. Particular cases describing usual population structures found in practice are easily derived from (2.4). The general model can be written as:

$$\mathbf{y} \mid \boldsymbol{\beta} \sim [\mathbf{X}\boldsymbol{\beta}, \mathbf{V}] \text{ and } \boldsymbol{\beta} \sim [\mathbf{a}, \mathbf{R}], \tag{2.4}$$

where $\mathbf{X}$ is a covariate matrix of dimension $N \times p$, with rows $\mathbf{X}_i = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, N$; $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of unknown parameters; and $\mathbf{y}$, given $\boldsymbol{\beta}$, is a random vector with mean $\mathbf{X}\boldsymbol{\beta}$ and known covariance matrix $\mathbf{V}$ of dimension $N \times N$. Analogously $\mathbf{a}$ and $\mathbf{R}$ are the respective $p \times 1$ prior mean vector and $p \times p$ prior covariance matrix of $\boldsymbol{\beta}$.

Since the response vector $\mathbf{y}$ is partitioned into $\mathbf{y}_s$, and $\mathbf{y}_{\bar{s}}$, the matrix $\mathbf{X}$, which is assumed to be known, is analogously partitioned into $\mathbf{X}_s$ and $\mathbf{X}_{\bar{s}}$, and $\mathbf{V}$ is partitioned into $\mathbf{V}_s$, $\mathbf{V}_{\bar{s}}$, $\mathbf{V}_{s\bar{s}}$ and $\mathbf{V}_{\bar{s}s}$. The first aim is to predict $\mathbf{y}_{\bar{s}}$ given the observed sample $\mathbf{y}_s$ and then the total T. We did this in the following steps: first, we used a joint prior distribution that is only partially specified in terms of moments, as follows:

$$\begin{pmatrix} \mathbf{y}_{\bar{s}} \\ \mathbf{y}_s \end{pmatrix} \Bigg| \boldsymbol{\beta} \sim \left[\begin{pmatrix} \mathbf{X}_{\bar{s}}\boldsymbol{\beta} \\ \mathbf{X}_s\boldsymbol{\beta} \end{pmatrix}, \begin{pmatrix} \mathbf{V}_{\bar{s}} & \mathbf{V}_{\bar{s}s} \\ \mathbf{V}_{s\bar{s}} & \mathbf{V}_s \end{pmatrix} \right].$$

Therefore, applying the general result in equation (2.2), the BLE of $E(\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s, \boldsymbol{\beta})$ and the minimum expected square loss (associated variance) are given by:

$$\hat{E}(\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s, \boldsymbol{\beta}) = \mathbf{X}_{\bar{s}}\boldsymbol{\beta} + \mathbf{V}_{\bar{s}s}\mathbf{V}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta}) \text{ and } \hat{V}(\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s, \boldsymbol{\beta}) = \mathbf{V}_{\bar{s}} - \mathbf{V}_{\bar{s}s}\mathbf{V}_s^{-1}\mathbf{V}_{s\bar{s}}. \tag{2.5}$$

Remark 1: It should be noted that if normality is assumed then $E(\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s, \boldsymbol{\beta})$ and $V(\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s, \boldsymbol{\beta})$ are respectively given by the right sides of (2.5). The BLE in (2.5) and its associated variance can be viewed respectively as approximations of $E(\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s, \boldsymbol{\beta})$ and $V(\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s, \boldsymbol{\beta})$ for non-normality cases.

Now, if we come back to model (2.4), we need to adapt the structure (2.1) and use the results in (2.2) to obtain the BLE of $\boldsymbol{\beta}$ and its associate variance, $\hat{V}(\hat{\boldsymbol{\beta}})$, respectively given by:

$$\hat{\boldsymbol{\beta}} = \mathbf{a} + \mathbf{R}\mathbf{X}'_s(\mathbf{X}_s\mathbf{R}\mathbf{X}'_s + \mathbf{V}_s)^{-1}(\mathbf{y}_s - \mathbf{X}_s\mathbf{a}) \text{ and } \hat{V}(\hat{\boldsymbol{\beta}}) = \mathbf{C} = \mathbf{R} - \mathbf{R}\mathbf{X}'_s(\mathbf{X}_s\mathbf{R}\mathbf{X}'_s + \mathbf{V}_s)^{-1}\mathbf{X}_s\mathbf{R}. \tag{2.6}$$

It is easy to show that the first equation in (2.6) can be rewritten as $\hat{\boldsymbol{\beta}} = \mathbf{C}\left(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{y}_s + \mathbf{R}^{-1}\mathbf{a}\right)$, where $\mathbf{C}^{-1} = \mathbf{R}^{-1} + \mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s$. It should be noted that if we place a vague prior distribution on $\boldsymbol{\beta}$, taking $\mathbf{R}^{-1} \to 0$, we obtain the minimum least square estimator of $\boldsymbol{\beta}$: $\hat{\boldsymbol{\beta}}_{LS} = \left(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{y}_s$.

Now, applying well known properties of conditional expectations and variances, we obtain:

$$E[\mathbf{y}_{\bar{s}}|\,\mathbf{y}_s] = E(E(\mathbf{y}_{\bar{s}}|\,\mathbf{y}_s,\boldsymbol{\beta})\,|\,\mathbf{y}_s) \text{ and } V[\mathbf{y}_{\bar{s}}|\,\mathbf{y}_s] = E(V(\mathbf{y}_{\bar{s}}|\,\mathbf{y}_s,\boldsymbol{\beta})\,|\,\mathbf{y}_s) + V(E(\mathbf{y}_{\bar{s}}|\,\mathbf{y}_s,\boldsymbol{\beta})\,|\,\mathbf{y}_s). \quad (2.7)$$

Replacing $E(\mathbf{y}_{\bar{s}}|\,\mathbf{y}_s,\boldsymbol{\beta})$ and $V(\mathbf{y}_{\bar{s}}|\,\mathbf{y}_s,\boldsymbol{\beta})$ in (2.7) with their respective BLE's in (2.5) and in turn, replacing $E(\boldsymbol{\beta}\,|\,\mathbf{y}_s)$ and $V(\boldsymbol{\beta}\,|\,\mathbf{y}_s)$ with $\hat{\boldsymbol{\beta}}$ and $\hat{V}(\hat{\boldsymbol{\beta}})$ in (2.6), we obtain the BLE of $E[\mathbf{y}_{\bar{s}}|\,\mathbf{y}_s]$ and its associated variance as:

$$\begin{aligned}\hat{E}[\mathbf{y}_{\bar{s}}|\,\mathbf{y}_s] &= \mathbf{X}_{\bar{s}}\hat{\boldsymbol{\beta}} + \mathbf{V}_{\bar{s}s}\mathbf{V}_s^{-1}\left(\mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}\right) \text{ and}\\ \hat{V}[\mathbf{y}_{\bar{s}}|\,\mathbf{y}_s] &= \mathbf{V}_{\bar{s}} - \mathbf{V}_{\bar{s}s}\mathbf{V}_s^{-1}\mathbf{V}_{s\bar{s}} + \left(\mathbf{X}_{\bar{s}} - \mathbf{V}_{\bar{s}s}\mathbf{V}_s^{-1}\mathbf{X}_s\right)\mathbf{C}\left(\mathbf{X}_{\bar{s}} - \mathbf{V}_{\bar{s}s}\mathbf{V}_s^{-1}\mathbf{X}_s\right)'.\end{aligned} \quad (2.8)$$

Remark 2: Analogously to the Remark 1, when normality is assumed we have that the right sides of (2.8) are respectively the values of $E[\mathbf{y}_{\bar{s}}|\,\mathbf{y}_s]$ and $V[\mathbf{y}_{\bar{s}}|\,\mathbf{y}_s]$.

The general expression of BLE for the total T and its associated variance are respectively obtained by replacing $E(\mathbf{y}_{\bar{s}}|\,\mathbf{y}_s)$ and $V(\mathbf{y}_{\bar{s}}|\,\mathbf{y}_s)$ in equations in (2.3) with their respective counterparts $\hat{E}[\mathbf{y}_{\bar{s}}|\,\mathbf{y}_s]$ and $\hat{V}[\mathbf{y}_{\bar{s}}|\,\mathbf{y}_s]$:

$$\hat{T} = \mathbf{1}_s'\mathbf{y}_s + \mathbf{1}_{\bar{s}}'\hat{E}[\mathbf{y}_{\bar{s}}|\,\mathbf{y}_s] \text{ and } \hat{V}\left(\hat{T}\right) = \mathbf{1}_{\bar{s}}'\hat{V}[\mathbf{y}_{\bar{s}}|\,\mathbf{y}_s]\mathbf{1}_{\bar{s}}. \quad (2.9)$$

It should be noted that in many applications of (2.9), the matrix $\mathbf{V}$ is assumed diagonal, which implies $\mathbf{V}_{\bar{s}s} = \mathbf{0}$ and then we have:

$$\hat{T} = \mathbf{1}_s'\mathbf{y}_s + \mathbf{1}_{\bar{s}}'\mathbf{X}_{\bar{s}}\hat{\boldsymbol{\beta}} \text{ and } \hat{V}\left(\hat{T}\right) = \mathbf{1}_{\bar{s}}'\left[\mathbf{V}_{\bar{s}} + \mathbf{X}_{\bar{s}}\mathbf{C}\mathbf{X}_{\bar{s}}'\right]\mathbf{1}_{\bar{s}}. \quad (2.10)$$

For the sake of illustration, we consider some examples discussed by O'Hagan (1985) and propose a new ratio estimator, which is one of the contributions of our work. All of them can be treated as special cases of the model (2.4).

2.3 Revisiting some common survey designs

2.3.1 Simple random sampling without replacement: Second order exchangeability

O'Hagan (1985) considered the simple case where the population has no relevant structure, which can be done by setting up:

$$E(y_i) = m, V(y_i) = v \text{ and } \operatorname{Cov}\left(y_i, y_j\right) = c, i, j = 1, \ldots, N, \forall i \neq j. \quad (2.11)$$

Remark 3: The correlation introduced in model (2.11) can be justified to mimic simple random sampling without replacement.

Applying the general result established in (2.10) to (2.11) with $\boldsymbol{\beta}$ of dimension 1, $\mathbf{X} = \mathbf{1}_N$, $\mathbf{a} = m$, $\mathbf{R} = c$ and $\mathbf{V} = \sigma^2\mathbf{I}$, where $\sigma^2 = v - c$, we obtain the BLE of T and its respective associated variance:

$$\hat{T}_{srs} = n\bar{y}_s + (N - n)\hat{\mu} \text{ and } \hat{V}\left(\hat{T}_{srs}\right) = (N - n)\sigma^2 + (N - n)^2 c\sigma^2\left(\sigma^2 + nc\right)^{-1}, \qquad (2.12)$$

where

$\bar{y}_s = n^{-1}\mathbf{1}'_s\mathbf{y}_s$ is the sample mean,

$\hat{\mu} = \omega\bar{y}_s + (1 - \omega)m$ is the expected value of the non-observed values of $\mathbf{y}$ and

$\omega = \dfrac{n\sigma^{-2}}{c^{-1} + n\sigma^{-2}}$, where $\sigma^2 = v - c$.

It should be noted that $\hat{\mu}$ is a weighted average of the prior mean m and the sample mean $\bar{y}_s$, where ω is the ratio between two population quantities. The mean m can be viewed as the investigator's prior of the true population mean $\bar{y}$. The uncertainty about y_i is split into two components: the uncertainty about the overall level of the y_i's (between variation) and the one with respect to how much each y_i may vary from that overall level (within variation). A useful measure of variability of units within the population is given by

$$S^2 = \frac{1}{N - 1}\sum_{i=1}^{N}(y_i - \bar{y})^2.$$

It is not difficult to show that $E\left(S^2\right) = v - c = \sigma^2$. Therefore, σ^2 can be interpreted as a prior estimate of variability within the population. We also obtain $V(\bar{y}) = c + N^{-1}\sigma^2$. In many applications, N is large and thus the constant c can be viewed as the between variation.

Letting $v \rightarrow \infty$ and keeping σ^2 fixed, that is, assuming prior ignorance, the estimates in (2.12) yield:

$$\hat{T}_{srs} = N\,\bar{y}_s \text{ and } \hat{V}\left(\hat{T}_{srs}\right) = N^2\left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n}.$$

These expressions are very similar to the well-known total estimate and its variance in the design-based context for the simple random sampling case. O'Hagan (1985) discussed some possibilities to avoid the difficult task of assigning a value for σ^2. The most natural way to do this is to find the BLE of it, but linear in the squares and cross product variance terms. However, it requires to specify fourth order moments of the y_i's. Goldstein (1979) proposed a BLE for the variance, which uses only linear functions of data. Nevertheless, it results in a complicated expression to its associated variance of his modified BLE. O'Hagan (1985) argued that if prior information about variance components is weak, any posterior estimate is close to the standard non-Bayesian estimates using only the data, wherever such estimate is available. Therefore, he suggested, as an approximate Bayesian procedure, substituting these standard variance estimates into the BLE and its associated variance wherever appropriate. For this case, we can replace σ^2 with $s^2 = (n - 1)^{-1}\sum_{i=1}^{n}(y_i - \bar{y}_s)^2$, which is design-based unbiased for S^2.

2.3.2 Stratified simple random sampling without replacement

Denote by y_{hi} the i^{th} unit, $i = 1, \ldots, N_h$ belonging to stratum h, $h = 1, .., H$. It is assumed that the stratum sizes, N_h, are known for all strata. The second-order exchangeability within each stratum is stated in O'Hagan (1985) as:

$$E(y_{hi}) = m_h, V(y_{hi}) = v_h, \operatorname{Cov}(y_{hi}, y_{hj}) = c_h, i \neq j \text{ and } \operatorname{Cov}(y_{hi}, y_{lj}) = d_{hl}, h \neq l.$$

Remark 4: It is reasonable to assume that the information gained about one stratum could change the beliefs about other strata in some special applications. However, if we want to mimic the stratified simple random sampling, we should assume that observations in different strata are uncorrelated, letting $d_{hl} = 0$.

The general model (2.4) can be applied to this case by setting $\mathbf{X} = \operatorname{diag}(\mathbf{X}_1, \ldots, \mathbf{X}_H)$ and $\mathbf{V} = \operatorname{diag}(\mathbf{V}_1, \ldots, \mathbf{V}_H)$, with $\mathbf{X}_h = \mathbf{1}_{N_h}$ and $\mathbf{V}_h = \sigma_h^2 \mathbf{I}_{N_h}$, where $\sigma_h^2 = v_h - c_h$, $\forall\ h = 1, \ldots, H$, $\mathbf{a} = (m_1, \ldots, m_H)'$, $\mathbf{R}$ is an $H \times H$ matrix with $R_{hl} = c_h$, if $h = l$ and $R_{hl} = d_{hl}$ otherwise. The BLE of T and its associated variance are obtained from (2.10) and can be found in O'Hagan (1985). Models for cluster sampling can be found in Bolfarine and Zacks (1992), page 11. The BLE for cluster models can be seen in O'Hagan (1985).

3 Auxiliary information: Ratio estimator

In many practical situations, it is possible to have information about an auxiliary variate x_i (correlated with y_i) for all the population units, or at least for each unit in the sample, plus the population mean, $\bar{X}$. In practice, x_i is often the value of y_i at some previous time when a complete census was taken. This approach is used in situations where the expected value and the variance of y_i is proportional to x_i, so in the BLE setup, we replace some hypotheses about the y's with ones about the first two moments of the rate y_i / x_i. To the best of our knowledge, the new ratio estimator proposed below is a novel contribution in sampling survey theory.

The new ratio estimator is obtained as a particular case of model (2.4) and with the hypothesis of exchangeability, used in Bayes linear approach, applied to the rate y_i / x_i for all $i = 1, \ldots, N$, as described below:

$$E\left(\frac{y_i}{x_i}\right) = m, V\left(\frac{y_i}{x_i}\right) = v \text{ and } \operatorname{Cov}\left(\frac{y_i}{x_i}, \frac{y_j}{x_j}\right) = c, i, j = 1, \ldots, N, \forall i \neq j. \tag{3.1}$$

Applying the general result established in (2.10) to (3.1) with $\mathbf{X} = (x_1, \ldots, x_N)'$, the vector $N \times 1$ of auxiliary variables, $\mathbf{a} = m$, $\mathbf{R} = c$ and $\mathbf{V} = \sigma^2 \operatorname{diag}(x_1, \ldots, x_N)$, where $\sigma^2 = v - c$, we obtain the BLE of T and its associated variance as follows:

$$\hat{T}_{ra} = n\bar{y}_s + (N - n)\hat{\mu}\bar{x}_{\bar{s}} \text{ and}$$

$$\hat{V}(\hat{T}_{ra}) = (N - n)\bar{x}_{\bar{s}}\sigma^2 + (N - n)^2 \bar{x}_{\bar{s}}^{\,2} \left(c^{-1} + \sigma^{-2} n\bar{x}_s\right)^{-1}, \text{ where}$$

$$\hat{\mu} = \omega\frac{\bar{y}_s}{\bar{x}_s} + (1 - \omega)m \text{ and } \omega = \frac{\sigma^{-2} n\bar{x}_s}{\left(c^{-1} + \sigma^{-2} n\bar{x}_s\right)},$$

where $\bar{x}_{\bar{s}} = (N\bar{X} - n\bar{x}_s)/(N - n)$ is mean of x's for the non-sample units. Letting $v \to \infty$ and $n \to \infty$, but keeping σ^2 fixed, we recover the ratio type estimator, found in the design-based approach: $\hat{T}_{ra} = N\bar{X}(\bar{y}_s / \bar{x}_s)$.

4 Bayes linear method for categorical data

Often one may be interested in cases where the observed characteristic is whether or not the population unit possesses some attribute of interest. We can define a dichotomized variable $y_i = 1$, if the i^{th} unit has that attribute, and refer to this as a success, and $y_i = 0$ otherwise. For the binary case when the sample size is not large enough to rely on the Central Limit Theorem, the design-based approach could use the randomization introduced by the sampling design to justify the distribution of the binary random quantities. For instance, Cochran (1977), sections 3.4 and 3.5, shows how to apply hypergeometric and binomial distributions to obtain confidence intervals for population proportions when, respectively, simple random sampling with and without replacement designs are employed. On the other hand, model-dependent approaches have also been advanced and applied for predicting totals or means in the categories of interest. Malec, Sedransk, Moriarity and LeClere (1997) considered a logistic hierarchical model with two levels, where the clusters are the second one. They also compared the full hierarchical Bayes estimates with empirical Bayes estimates and standard methods. Moura and Migon (2002) presented a logistic hierarchical model approach for small area prediction of proportions, taking into account both possible spatial and unstructured heterogeneity effects. Nandram and Choi (2008) proposed a time-dependent multinomial-Dirichlet model to predict the results of an election under ignorable and non-ignorable non-response. They also used a Bayesian approach to allocate the undecided voters to the candidates.

Here again, we do not need to make any use of full model assumptions or a randomization approach, but we do need to make some assumptions about the first and the second moments of the random quantities involved. The BLE for binary data was briefly introduced by O'Hagan (1985), but here we develop it more generally for the case where we are interested in analyzing more than one attribute in a population. The purpose is to describe the estimation of the proportion of successes with categorical data. Let y_{ij} be the variable that indicates that unit i, $i = 1, \ldots, N$ is in category j, $j = 1, \ldots, k$ given by

$$y_{ij} = \begin{cases} 1, \text{ if } i^{\text{th}} \text{ unit has } j^{\text{th}} \text{ attribute;} \\ 0, \text{ otherwise.} \end{cases}$$

The main aim is to estimate a vector $\mathbf{p} = (p_1, \ldots, p_k)'$, where $p_j = N^{-1}\sum_{i=1}^{N} y_{ij}$, $j = 1, \ldots, k$, is the proportion of units in category j, given $\mathbf{y}_s$, a vector of dimension nk, defined as $\mathbf{y}_s = (y_{11}, y_{21}, \ldots, y_{n1}, \ldots, y_{1k}, y_{2k}, \ldots, y_{nk})'$. As we are dealing with situations in which for each unit it is only possible to associate a unique attribute, we have $\sum_{j=1}^{k} p_j = 1$. Thus, we only need to estimate $k - 1$ parameters, since it follows that $\hat{p}_k = 1 - \sum_{j=1}^{k-1} \hat{p}_j$ and the variance estimate is also analogously obtained by $\hat{V}(\hat{p}_k) = \sum_{j=1}^{k-1} \hat{V}(\hat{p}_j) + \sum_{l \neq j=1}^{k-1} \hat{\text{C}}\text{ov}(\hat{p}_j, \hat{p}_l)$.

In the absence of any other structural information, we suppose that the units in any given category are second-order exchangeable, but we do not assume any exchangeability between units of different categories. Our prior beliefs are expressed for $i = 1,\ldots,N,\ j = 1,\ldots,k-1,$ as follows:

$$\begin{aligned} m_j &= E\left(y_{ij}\right) = P\left(y_{ij} = 1\right), v_j = V\left(y_{ij}\right) = m_j\left(1 - m_j\right) \text{ and} \\ \operatorname{cov}\left(y_{ij}, y_{i'j}\right) &= P\left(y_{i'j} = 1 \middle| y_{ij} = 1\right) P\left(y_{ij} = 1\right) - P\left(y_{ij} = 1\right) P\left(y_{i'j} = 1\right) \\ &= m_j\left(m_{jj} - m_j\right) = c_j,\ \forall i \neq i' \text{ and } \sigma_j^2 = v_j - c_j = m_j\left(1 - m_{jj}\right), \end{aligned}$$

where $m_{jj} = P\left(y_{i'j} = 1 \middle| y_{ij} = 1\right),$ for all $i \neq i'.$

For $j \neq j',$ we analogously obtain the covariance between these categories as

$$\operatorname{cov}\left(y_{ij}, y_{i'j'}\right) = \begin{cases} m_j\left(m_{j'j} - m_{j'}\right), & \text{if } i \neq i', \\ -m_j m_{j'}, & \text{if } i = i'. \end{cases}$$

Often, we do not have all the data $\mathbf{y}_s,$ but only a sufficient statistics, such as the sample proportion for each category, $\overline{\mathbf{y}}_s.$ Let $\overline{\mathbf{y}}_s$ be the $k-1$-vector whose j^{th} position is given by the sample mean for category $j.$ Using the general model in (2.4), we obtain:

$$E\left(\overline{\mathbf{y}}_s\right) = E\left(E\left(\overline{\mathbf{y}}_s \mid \boldsymbol{\beta}\right)\right) = \mathbf{a} \text{ and } \operatorname{Var}\left(\overline{\mathbf{y}}_s\right) = E\left(V\left(\overline{\mathbf{y}}_s \mid \boldsymbol{\beta}\right)\right) + V\left(E\left(\overline{\mathbf{y}}_s \mid \boldsymbol{\beta}\right)\right) = \mathbf{V}_s + \mathbf{R}.$$

Applying the general model in (2.4), where: the responde variable is given by $\overline{\mathbf{y}}_s$; the vector $\boldsymbol{\beta}$ has dimension $k-1$; $\mathbf{X}_s = \mathbf{I}_s$ and $\mathbf{V} = \operatorname{diag}\left(\mathbf{V}_{\bar{s}}, \mathbf{V}_s\right),$ we obtain from (2.10):

$$\hat{\mathbf{p}} = \frac{n\overline{\mathbf{y}}_s + (N-n)\hat{\boldsymbol{\beta}}}{N} \text{ and } \hat{V}\left(\hat{\mathbf{p}}\right) = \frac{(N-n)^2\left[\mathbf{V}_{\bar{s}} + \mathbf{C}\right]}{N^2}, \tag{4.1}$$

where $\mathbf{C}^{-1} = \mathbf{R}^{-1} + \mathbf{V}_s$ and $\hat{\boldsymbol{\beta}} = \mathbf{C}\left(\mathbf{V}_s^{-1}\overline{\mathbf{y}}_s + \mathbf{R}^{-1}\mathbf{a}\right),$ as stated in (2.6).

Let $\mathbf{Q} = \mathbf{V}_s + \mathbf{R}.$ The BLE of $\mathbf{p}$ and its associate variance given in (4.1) can be written in terms of the prior quantities $m_j, m_{jj'}$ and $j = 1,\ldots,k-1$ by noting that $\mathbf{a} = \left(m_1,\ldots,m_{k-1}\right)',\ Q_{jj} = c_j + \sigma_j^2/n$ and $Q_{jj'} = m_j\left(m_{j'j} - m_{j'}\right) - m_j m_{j'j}/n.$ Therefore, the matrix $\mathbf{R} = \left\{r_{jj'}\right\}, j, j' = 1,..,k-1$ with $r_{jj} = c_j$ and $r_{jj'} = m_j\left(m_{j'j} - m_{j'}\right)$ and $\mathbf{V}_s = 1/n\ \left\{v_{jj'}\right\},\ j, j' = 1,..,k-1$ with $v_{jj} = \sigma_j^2$ and $v_{jj'} = -m_j m_{j'j}.$ Analogously, we get $\mathbf{V}_{\bar{s}} = n/(N-n)\,\mathbf{V}_s.$

4.1 Prior elicitation

Elicitation is the process of formulating a person's knowledge and beliefs about one or more uncertain quantities into a probability distribution for those quantities. According to Garthwaite, Kadane and O'Hagan (2005), it is convenient to think of the elicitation task as involving a facilitator, who helps the expert formulate the expert's knowledge in probabilistic form. In the context of eliciting a prior distribution for a Bayesian analysis, it is the expert's prior knowledge that is being elicited, but in general the objective is to express the expert's current knowledge in probabilistic form. If the expert is a statistician, or is very familiar with statistical concepts, then there may be no formal need for a facilitator, but this is rare in practice. O'Hagan (1998) illustrated with a practical example how to elicitate first and second moments. In particular, he adopted the Bayes linear approach because it makes easy the application of the elicitation procedure by engineers.

In this section, some restrictions about the prior quantities and an alternative to facilitate the process of elicitation are presented to obtain the BLE for categorical data. Because m_j and $m_{jj'}$ are probabilities, and $\mathbf{R}$ and $\mathbf{V}_s$ are the covariance matrices in model (2.4), the following restrictions must be satisfied:

1. $0 < m_j < 1$ and $0 \leq m_{jj'} \leq 1,\ j, j' = 1, \ldots, k-1;$
2. $\mathbf{R}$ and $\mathbf{V}_s$ are positive-definite symmetric matrices.

In order to verify if condition (2.2) is satisfied, the following steps may be carried out:

i. verify if $\mathbf{R}$ and $\mathbf{V}_s$ are symmetric by checking if $m_j m_{jj'} = m_{j'} m_{j'j}$;

ii. verify if $\mathbf{R}$ and $\mathbf{V}_s$ are positive-definite matrices by finding the eigenvalues of $\mathbf{R}$ and $\mathbf{V}_s$. If the eigenvalues are positive, then the matrices are positive-definite.

It should be noted that the eigenvalues are the roots of the characteristic polynomial and if this polynomial is of degree $n, n \leq 4,$ it is possible to analytically get its roots by using Bhaskara, Cardan or Ferrari; see Jacobson (2009), chapter 4, for formulas. However, if $n \geq 5,$ we usually need to apply an iterative method to get them. Nevertheless, for matrices higher than $2 \times 2,$ it is not trivial to analytically obtain these restrictions based on eigenvalues. The next proposition presents the conditions that m_j and $m_{jj'},\ j = 1, \ldots, k-1,$ must satisfy in order to obtain a suitable prior for a multinomial model with three categories using the Bayesian linear estimation approach.

Proposition 1 Suppose that we elicit $m_j,$ such that $0 < m_j < 1,\ j = 1, 2.$ Then, given ρ_{11}, ρ_{12} and $\rho_{22},$ we obtain m_{11}, m_{12}, m_{21} and m_{22} by (4.2). The prior quantities m_j and $m_{jj'},$ for $j, j' = 1, 2$ must satisfy the following constraints for the matrices $\mathbf{R}$ and $\mathbf{V}_s$ to be positive-definite:

$$m_{11} > m_1 \text{ and } m_{22} > m_2, m_{11}m_{22} - m_{11} - m_{22} + 1 > m_{12}m_{21} \text{ and}$$
$$m_{11}m_{22} - m_{11}m_2 - m_1 m_{22} > m_{12}m_{21} - 2m_2 m_{12}.$$

The verification of the Proposition 1 requires some algebra. We check that the matrices $\mathbf{R}$ and $\mathbf{V}_s$ are positive-definite using (i) and (ii) above. We use the fact that the eigenvalues of a matrix with dimension 2×2 are positive if and only if its determinant is positive and then we obtain $m_{jj'},\ j, j' = 1, 2$ which satisfies this restriction for both matrices. For cases with more than three categories we must numerically verify if the matrices $\mathbf{R}$ and $\mathbf{V}_s$ are positive-definite when replacing the numerical values of m_j and $m_{jj'},\ j = 1, \ldots, k-1$ into them.

On the other hand, if an expert has some difficulty in specifying some of these conditional probabilities $m_{jj'},$ it may be simpler to assign a prior to the coefficient of correlation. Define $\rho_{jj'}$ as the prior of the coefficient of correlation between two different units within categories j and $j',$ that is:

$$\rho_{jj'} = \operatorname{corr}\left(y_{ij}, y_{i'j'}\right) = \begin{cases} \dfrac{m_{jj} - m_j}{1 - m_j}, & j = j', \\ \dfrac{m_j\left(m_{j'j} - m_{j'}\right)}{\sqrt{m_j\left(1 - m_j\right) m_{j'}\left(1 - m_{j'}\right)}}, & j \neq j', \end{cases}$$

for $i, i' = 1, \ldots, n,\ i \neq i',\ j, j' = 1, \ldots, k - 1.$

Therefore, given $\rho_{jj'},\ j, j' = 1, \ldots, k - 1,$ we get

$$m_{jj'} = \begin{cases} m_j + \rho_{jj}\left(1 - m_j\right) & j = j', \\ \dfrac{m_j m_{j'} + \rho_{j'j}\sqrt{m_j\left(1 - m_j\right) m_{j'}\left(1 - m_{j'}\right)}}{m_{j'}}, & j \neq j'. \end{cases} \tag{4.2}$$

It should be noted that if there is some past data obtained from a previous survey, it is possible for an expert to use this information. For instance, m_j can be obtained by estimating the proportion of units in category $j,\ j = 1, \ldots, k - 1$ from the previous survey. Analogously, $\rho_{jj'}$ can be obtained using previous survey data. As stated in restriction (2.1), m_j cannot assume the values 0 and 1, otherwise the correlations would not be defined.

4.2 Prior sensitivity analysis

It is worth checking how the estimator and its associated variance depend on the priors assigned. We deal with the simple case of only two categories. It should be noted that in the case with more than 2 categories the number of prior quantities to be elicited increases fast, but the conclusions obtained under this illustration can be extended. On the other hand, if there is no prior information available we can use non-informative priors and, as described in Section 2.2, the estimators from the design-based approach are recovered.

The BLE for proportion for binary data can be obtained as a particular case of the estimator in (4.1),

$$\hat{p}_1 = \frac{n\bar{y}_1 + (N - n)\hat{\mu}}{N},$$

where

$\hat{\mu} = \omega\bar{y}_1 + (1 - \omega) m_1$ is the expected value of the non-observed values in category 1,

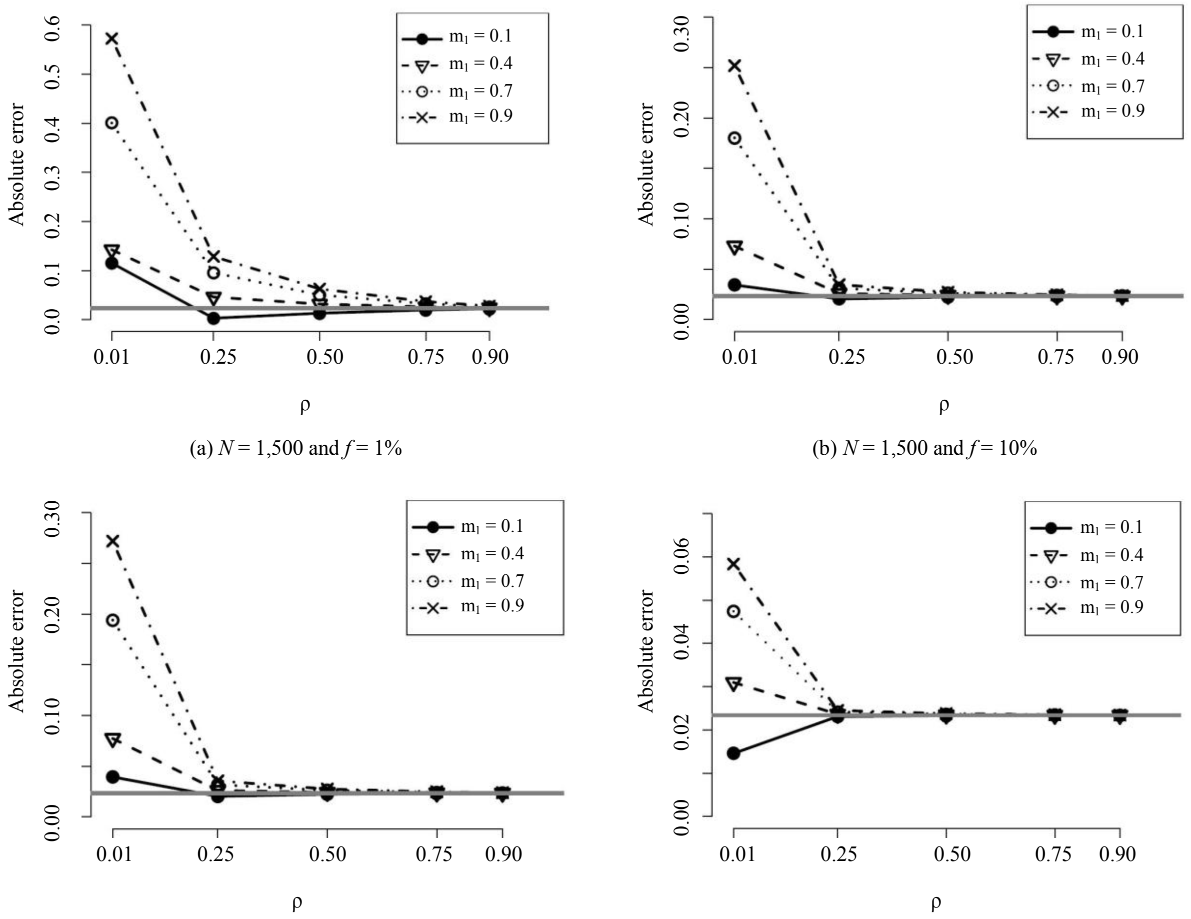
$$\omega = \frac{n\sigma_1^{-2}}{n\sigma_1^{-2} + c_1^{-1}},$$

and $\hat{p}_2 = 1 - \hat{p}_1$. Note that σ_1^2 and c_1 depend on $m_{11} = m_1 + \rho_{11}(1 - m_1)$, see page 13. We analyze how the estimates are affected by ρ_{11}.

1. If $\rho_{11} \to 0$, then $\omega \to 0$ and $\hat{\mu} \to m_1$. Thus, the estimator for the non-observed values largely depend on the value of the prior.
2. If $\rho_{11} \to 1$, then $\omega \to 1$ and $\hat{\mu} \to \bar{y}_1$. Thus, the estimator for the non-observed values does not depend on the value of the prior.

Moreover, it is trivial to see that if $n/N \rightarrow 1$, $\hat{p}_1 \rightarrow \bar{y}_1$. To illustrate these results, we created some artificial dataset by fixing the true proportion at $\mathbf{p} = (0.2380, 0.7620)'$ and the sample mean at $\bar{\mathbf{y}}_s = (0.2614, 0.7386)'$. These values were taken from Moura and Migon (2002). Then, we assessed how the values of m_1, N, $f = n/N$ and ρ_{11} affect the estimator $\hat{p}_1$. Figure 4.1 presents the two-dimensional plots of the absolute error of $\hat{p}_1$ *versus* ρ_{11} for some particular cases. The grey line represents the absolute error between the sample proportion $\bar{y}_1$ and the true p_1.

It should be noted that, as f or N increases, the absolute error decreases for any prior values. Moreover, when $\rho_{11} \rightarrow 0$ the absolute error increases when m_1 considerably differs from the true proportion p_1, but it decreases as the sample size increases. Finally, as $\rho_{11} \rightarrow 1$ we observe that the absolute error of $\hat{p}_1$ tends to the absolute error of the sample proportion $\bar{y}_1$. Thus, if we have good prior information, in terms of m_1, the estimator proposed performs well for all the values of ρ_{11}. But, if there is no prior information available, non-informative priors characterized by $\rho_{11} \rightarrow 1$ can be used and we obtain results similar to a design-based approach.



(a) N = 1,500 and f = 1%

(b) N = 1,500 and f = 10%

(c) N = 15,288 and f = 1%

(d) N = 15,288 and f = 10%

Figure 4.1 Absolute error for fixed $m_1 \in \{0.1, 0.4, 0.7, 0.9\}$, $N \in \{1{,}500; 15{,}288\}$ and $f \in \{1\%, 10\%\}$ and varying $\rho_{11} \in \{0.01, 0.25, 0.5, 0.75, 0.9\}$. The grey line represents the absolute error of the sample proportion $\bar{y}_1$

5 Conclusions

To elicit a full joint prior distribution in many dimensions would be an enormous task. The Bayes linear method only requires the elicitation of prior means, variances and covariances for the parameters. It is particularly useful when a statistical expert is not available to conduct a full elicitation. An example of a successful elicitation using this estimator can be found in O'Hagan (1998).

We derived the well-known design-based estimators using the structure of the BLE applied to a general regression model approach. We extended the estimator to categorical data and concluded that even if this estimator has many quantities to elicit, it is possible to reparameterize them or work with non-informative priors. The numerical example illustrates the behavior of the estimates as a function of the sample size and the specifications of the prior parameters. However, we are aware that eliciting priors for a large number of parameters is not an easy task if information from previous surveys is not available. Nevertheless, the examples discussed in the article show that even when prior information is not available, it is also possible to obtain the counterpart design-based estimators by setting sufficiently large variance to the priors. Furthermore, survey practitioners who need to obtain estimates for a large number of variables, would also realize that they would not be able to produce estimates with satisfactory accuracy for all variables, independently of which approach was employed. Finally, it is showed how BLE and design-based approaches can be conciliated.

Acknowledgements

This work is part of the master dissertation of Kelly C.M. Gonçalves under the supervision of Fernando Moura and Helio Migon, in the Graduate Program of UFRJ. Kelly has a scholarship from Coordenação de Aperfeiçoamento de Pessoal do Ensino Superior (CAPES). Fernando Moura and Helio Migon receive financial support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq-Brazil, BPPesq).

The authors would like to thank the editor, the associate editor and two referees for their very thoughtful and constructive comments.

References

Basu, D. (1971). An essay on the logical foundations of survey sampling, Part 1 (with discussion). In *Foundations of Statistical Inference*, (Eds., Godambe and Sprott), 203-242. Holt, Reinhart and Wilnston, Toronto.

Bolfarine, H., and Zacks, S. (1992). *Prediction Theory for Finite Populations*. New York: Springer-Verlag.

Cochran, W. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc.

Garthwaite, P., Kadane, J. and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100, 680-701.

Goldstein, M., and Wooff, D. (2007). *Bayes Linear Statistics: Theory and Methods*. Durham University, UK: Wiley series in probability and statistics.

Goldstein, M. (1979). The variance modified linear Bayes estimator. *Journal of the Royal Statistical Society*, 41, 96-100.

Hartigan, J. (1969). Linear bayesian methods. *Journal of the Royal Statistical Society*, Series B (Methodological), 446-454.

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

Jacobson, N. (2009). *Basic Algebra,* Vol 1. Dover Books on Mathematics.

Little, R.J. (2003). The Bayesian approch to sample survey inference. In *Analysis of Survey Data*, (Eds., R.L. Chambers and C.J. Skinner), chapter 4, 49-52. New York: John Wiley & Sons Inc.

Malec, D., Sedransk, J., Moriarity, C.L. and LeClere, F.B. (1997). Small area inference for binary variables in National Health Interview Survey. *Journal of the American Statistical Association*, 92, 815-826.

Moura, F., and Migon, H. (2002). Bayesian spatial models for small area estimation of proportions. *Statistical Modelling*, 2, 183-201.

Nandram, B., and Choi, J. (2008). A Bayesian allocation of undecided voters. *Survey Methodology*, 34, 1, 37-49.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625.

O'Hagan, A. (1985). Bayes linear estimators for finite populations. *Technical Report 58*, Department of Statistics - University of Warwick.

O'Hagan, A. (1987). Bayes linear estimators for randomized response models. *Journal of the American Statistical Association*, 82, 580-585.

O'Hagan, A. (1998). Eliciting expert beliefs in substantial practical applications. *The Statistician*, 47, 21-35.

Pfeffermann, D., Moura, F.A.S. and Silva, P.L.N. (2006). Multi-level modelling under informative sampling. *Biometrika*, 93, 943.

Scott, A.J. (1977). Large-sample posterior distributions for finite populations. *Annals of Mathematical Statistics*, 42, 1113-1117.

Skinner, C., Holt, D. and Smith, T. (1989). *Analysis of complex surveys*. New York: John Wiley & Sons, Inc.

Smouse, E. (1984). A note on bayesian least squares inference for finite population models. *Journal of the American Statistical Association*, 79, 390-392.

Valliant, R., Dorfman, A. and Royall, R. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons, Inc.

Zacks, S. (2002). In the footsteps of Basu: The predictive modelling approach to sampling from finite population. *Sankhyā: The Indian Journal of Statistics*, Series A, 64, 532-544.

A nonparametric method to generate synthetic populations to adjust for complex sampling design features

Qi Dong, Michael R. Elliott and Trivellore E. Raghunathan[1]

Abstract

Outside of the survey sampling literature, samples are often assumed to be generated by a simple random sampling process that produces independent and identically distributed (IID) samples. Many statistical methods are developed largely in this IID world. Application of these methods to data from complex sample surveys without making allowance for the survey design features can lead to erroneous inferences. Hence, much time and effort have been devoted to develop the statistical methods to analyze complex survey data and account for the sample design. This issue is particularly important when generating synthetic populations using finite population Bayesian inference, as is often done in missing data or disclosure risk settings, or when combining data from multiple surveys. By extending previous work in finite population Bayesian bootstrap literature, we propose a method to generate synthetic populations from a posterior predictive distribution in a fashion inverts the complex sampling design features and generates simple random samples from a superpopulation point of view, making adjustment on the complex data so that they can be analyzed as simple random samples. We consider a simulation study with a stratified, clustered unequal-probability of selection sample design, and use the proposed nonparametric method to generate synthetic populations for the 2006 National Health Interview Survey (NHIS), and the Medical Expenditure Panel Survey (MEPS), which are stratified, clustered unequal-probability of selection sample designs.

Key Words: Synthetic populations; Posterior predictive distribution; Bayesian bootstrap; Inverse sampling.

1 Introduction

Statistical methods outside the survey methodology setting have usually been developed without careful consideration for sample design, often implicitly assuming simple random samples, or, occasionally, one-stage cluster samples. Major efforts of modern survey statistics focus on extending methods to analyze complex survey data (Skinner, Holt and Smith 1989), accommodating issues such as stratification, unequal probability of selection, nonresponse bias or calibration. Hinkins, Oh and Scheuren (1997) proposed an inverse sampling design algorithm that connects the survey statistics and the classical statistics from another perspective. Their basic idea is to choose a subsample that has a simple random sample structure unconditionally. The subsample is often much smaller than the original sample, so they propose to repeat the process independently many times and average the results to increase the precision. They also described exact or approximate inverse sampling schemes for stratified simple random sampling, one-stage cluster sampling, and two-stage cluster sampling. However, this new idea is not used widely in practice, perhaps because it is extremely computionally intensive and the precision losses are often substantial. Similarly, generating synthetic populations from a posterior predictive distribution of a population conditional on complex sample data in a fashion that accounts for the complex sample design is not straightforward (Little 1991). However, in recent years demand for synthetic populations has increased, in order to deal with weight trimming or windorization problems (Lazzeroni and Little 1998; Elliott and Little 2000; Elliott 2007; Chen, Elliott and Little 2010), disclosure risk settings (Little 1993;

[1] Qi Dong, Netflix, Inc. 100 Winchester Cir, Los Gatos, CA 95032. E-mail: qidong@umich.edu; Michael R. Elliott, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109, Survey Methodology Program, Institute for Social Research, University of Michigan, 426 Thompson St., Ann Arbor, MI 48106. E-mail: mrelliot@umich.edu; Trivellore E. Raghunathan, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109, Survey Methodology Program, Institute for Social Research, University of Michigan, 426 Thompson St., Ann Arbor, MI 48106. E-mail: teraghu@umich.edu.

Raghunathan, Reiter and Rubin 2003; Reiter 2004, 2005), or combining data from multiple surveys (Raghunathan, Xie, Schenker, Parsons, Davis, Dodd and Feuer 2007; Dong 2012). Often the synthetic populations are generated under a distributional assumption (normal, binomial, Poisson), with the posterior distribution of the model parameters approximated by the asymptotic normal distribution. The mean and covariance matrix of the normal distribution are estimated after complex sampling design features are taken into account (Raghunathan *et al.* 2007).

A major weakness of model-based methods is that if the model is seriously misspecified, it may yield invalid inferences (Little 2004). In multivariate settings, we need to consider the relationships among the variables of interest and determine an appropriate model that fits the data, which may be hard if the data contains different types of variables. In this paper we propose a nonparametric method as a counterpart of the model-based method to generate synthetic populations. This work extends the finite population Bayesian bootstrap and related Pólya posterior models of Lo (1988), Ghosh and Meeden (1983), and Cohen (1997) to account for complex sample designs. Since it achieves the same goal of the inverse sampling technique, it can be treated as the Bayesian finite population version of inverse sampling. To make inference using this weighted finite population Bayesian bootstrap, we can either make use of the draws directly, or, for computational efficiency, use results previously derived in the disclosure risk and multiple imputation literature, since these non-parametrically-generated populations can be viewed as multiple imputations of the unobserved elements of the population.

This paper is organized as follows. Section 2 briefly discusses synthetic populations in the context of Bayesian finite population inference. Section 3 reviews and summarizes the Bayesian bootstrap method and its finite population extension, and shows that, for an unequal probability of selection sample, the distribution of synthetic populations generated under a variant of a Pólya urn scheme matches the posterior predictive distribution of a finite population Bayesian bootstrap. Section 4 presents the proposed method under stratified clustering sampling with unequal selection probabilities. Section 5 shows that inference from these non-parametrically-generated synthetic populations can be obtained using results from the disclosure risk and multiple imputation literature, where each synthetic population has zero "within-imputation" variance. Section 6 provides a simulation study to evaluate the performance of the nonparametric method in a repeated sampling context. Section 7 applies the method to generate synthetic populations than can be used to estimate health insurance coverage rates using the 2006 NHIS and MEPS data, and compares the result with a parametric (log-linear) modeling approach. Concluding remarks are provided in Section 8.

2 Generating synthetic populations from survey data

The basic concept of Bayesian finite population inference involves imputing the non-sampled values of the population from the posterior predictive distribution based on the observed data. Assume the population values are $Y = (Y_1, \dots, Y_N)$ and the observed data, $Y_{\text{obs}} = (y_1, \dots, y_n)$ is obtained in a survey with sampling indicators $I = (I_1, \dots, I_N)$. The Bayesian population inference allows for the use of parametric model $\Pr(Y \mid \theta)$ for population data based on the posterior predictive distribution for the unobserved elements of the population $\Pr(Y_{\text{nob}} \mid Y_{\text{obs}})$:

$$\Pr(Y_{\text{nob}} \mid Y_{\text{obs}}) = \int \Pr(Y_{\text{nob}} \mid Y_{\text{obs}}, \theta) \Pr(\theta \mid Y_{\text{obs}}) \, d\theta$$

(Ericson 1969; Little 1993; Rubin 1987; Scott 1977; Skinner *et al.* 1989). Here we use the model $\Pr(Y \mid \theta)$ to approximate the entire population distribution $\Pr(Y)$ and average over the posterior distribution based on the sampled data $\Pr(\theta \mid Y_{\text{obs}})$. In the case that there are design variables known for the entire population available, the above model can be naturally extended by conditioning on these variables.

Implicit in the derivation of above is that the sampling indicator I need not be modeled. This requires ignorable sampling (Rubin 1987) (the distribution of I does not depend on unobserved data), as well as a model for the data $\Pr(Y \mid \theta)$ that is attentive to design features and robust enough to sufficiently capture all relevant aspects of the distribution of Y of interest. Our goal here is to develop a method to generate draws from $\Pr(Y_{\text{nob}} \mid Y_{\text{obs}})$ that account for all the design features in Y_{obs} so that draws from the posterior distribution of $Y_{\text{nob}} \mid Y_{\text{obs}}$ can be treated as a simple random sample in analysis.

3 Weighted finite population Bayesian bootstrap

3.1 Finite Population Bayesian Bootstrap (FPBB)

Assume that the (scalar) population elements $Y_i, i = 1, \ldots, N$ are exchangeable and can take on $K \le N$ possible values $(b_1, \ldots, b_K)$; thus $Y_i \mid \theta \sim \text{MULTI}(1; \theta_1, \ldots, \theta_K)$. Further assuming a conjugate Dirichlet prior for $\theta \sim \text{DIR}(\alpha_1, \ldots, \alpha_K)$ yields (Ghosh and Meeden 1983)

$$\begin{aligned}
P(Y_{\text{nob}} \mid y) &= P\left(b_1^{\text{nob}} = N_1 - n_1, \ldots, b_K^{\text{nob}} = N_K - n_K \mid b_1^{\text{obs}} = n_1, \ldots, b_K^{\text{obs}} = n_K\right) \\
&= \frac{\int_0^1 \cdots \int_0^1 p(Y_{\text{nob}} \mid y, \theta)\, p(y \mid \theta)\, p(\theta)\, d\theta_1 \ldots d\theta_K}{\int_0^1 \cdots \int_0^1 p(y \mid \theta)\, p(\theta)\, d\theta_1 \ldots d\theta_K} \\
&= \frac{\int_0^1 \cdots \int_0^1 p(Y_{\text{nob}} \mid \theta)\, p(y \mid \theta)\, p(\theta)\, d\theta_1 \ldots d\theta_K}{\int_0^1 \cdots \int_0^1 p(y \mid \theta)\, p(\theta)\, d\theta_1 \ldots d\theta_K} \\
&= \frac{\int_0^1 \cdots \int_0^1 \prod_{i=1}^K \theta_i^{N_i - n_i} \prod_{i=1}^K \theta_i^{n_i} \prod_{i=1}^K \theta_i^{\alpha_i - 1} d\theta_1 \ldots d\theta_K}{\int_0^1 \cdots \int_0^1 \prod_{i=1}^K \theta_i^{n_i} \prod_{i=1}^K \theta_i^{\alpha_i - 1} d\theta_1 \ldots d\theta_K} \\
&= \frac{\left(\prod_{i=1}^K \Gamma(N_i + \alpha_i) / \Gamma(\alpha_i)\right) \Big/ \left(\Gamma(N + \alpha_0) / \Gamma(\alpha_0)\right)}{\prod_{i=1}^K \Gamma(n_i + \alpha_i) \Big/ \Gamma(n + \alpha_0)}
\end{aligned} \tag{3.1}$$

where $\alpha_0 = \sum_{i=1}^K \alpha_i$, $\sum_{i=1}^K N_i = N$, and $n_1, \ldots, n_K$ refers to the number of distinct values we observe from our sample $y = (y_1, \ldots, y_n)$, $\sum_{i=1}^K n_i = n$. If $\alpha_i \equiv 0$ then $p(Y_{\text{nob}} \mid y)$ reduces to

$$\left(\prod_{i=1}^K \Gamma(N_i) / \Gamma(n_i)\right) \Bigg/ \left(\Gamma(N) / \Gamma(n)\right).$$

To ease implementation, Lo (1988) proposed making draws from the FPBB posterior predictive distribution using a "Pólya urn scheme" procedure. Suppose an urn contains n balls, each of which have a distinct real number label $b_i, i = 1, \ldots, K.$ A Pólya sample of size m is selected by first selecting a ball at random from the urn and returning the selected ball into the urn, then putting one same ball into the urn and repeating this process until m balls have been selected. It can be shown that the probability of getting m_i balls of type b_i is given by

$$p(b_1 = m_1, \ldots, b_K = m_K) = \frac{\prod_{i=1}^{k} \Gamma(n_i + m_i) / \Gamma(n_i)}{\Gamma(n + m) / \Gamma(n)} \tag{3.2}$$

where n_i is the number of balls of type b_i originally in the urn. The distribution of the counts of type b_i is invariant under any permutation of the draws. Note that this corresponds directly to the posterior probability of a total of $(m_1, \ldots, m_K)$ elements of type $(b_1, \ldots, b_K)$ in a population, given that $(n_1, \ldots, n_K)$ elements were observed in a (simple random) sample of size $\sum_{i=1}^{K} n_i = n.$ Hence a FPBB replicate sample can be drawn from this Pólya posterior using the following steps:

Step 1. Draw a Pólya sample of size $m = N - n,$ denoted by $\left(y_1^*, \ldots, y_{N-n}^*\right)$ from the urn $\{y_1, \ldots, y_n\};$ by (3.2), with $m_k = N_k - n_k$ draws of value b_k^{obs} for $k = 1, \ldots, K,$ this corresponds to a draw of $P(Y_{\text{nob}} \mid y)$ from (3.1).

Step 2. Form the FPBB population $y_1, \ldots, y_n, y_1^*, \ldots, y_{N-n}^*.$

3.2 FPBB with unequal probabilities of selection

Cohen (1997) extended the FPBB procedure to adjust for the unequal probabilities of selection. Assume $(y_1, \ldots, y_n)$ is a sample from a finite population $(Y_1, \ldots, Y_N)$ with design weights $(w_1, \ldots, w_n),$ where

$$w_i = \frac{1}{P(I_i = 1)}$$

and I is the sampling indicator. The procedure has two steps:

Step 1. Draw a sample of size $N - n,$ denoted by $\left(y_1^*, \ldots, y_{N-n}^*\right),$ by drawing y_k^* from $(y_1, \ldots, y_n)$ in such a way that y_i is selected with probability

$$\frac{w_i - 1 + l_{i,k-1} * (N - n)/n}{N - n + (k - 1) * (N - n)/n},$$

where w_i is the weight of unit i and $l_{i,k-1}$ is the number of bootstrap selections of y_i among $y_1^*, \ldots, y_{k-1}^*.$ (The function wtpolyap in the R package *polypost* can be used to obtain draws from a weighted Pólya urn.)

Step 2. Form the FPBB population $y_1, \ldots, y_n,\ y_1^*, \ldots, y_{N-n}^*$.

Although Cohen (1997) did not provide theoretical proof for this procedure, it can be obtained as a straightforward extension of the standard FPBB and Pólya urn equivalency described in Section 3.1. First, we determine the posterior distribution of the FPBB sample with unequal probabilities of selection implied by the weighted FPBB procedure. The multinomial likelihood based on our weighted sample is given by

$$p\,(y_{\text{obs}} \,|\, \theta) = \prod_{i=1}^{K} \theta_i^{w_i^*},$$

where

$$w_i^* = \left(\frac{n}{N-n}\right)\sum_{j=1}^{n} I\left(y_j = b_i\right)\left(w_j - 1\right)$$

is the sum of the design weights minus one across all sampled elements with value $b_i, i = 1, \ldots, K,$ normalized to sum to n. (Note that this removes subjects sampled with weights equal to one – "certainty sample" elements – from the likelihood, as they have no chance to be part of the unobserved portion of the population, and thus contribute no information about these unobserved elements.) Assuming an improper Dirichlet prior $p\,(\theta) = \prod_{i=1}^{k} \theta_i^{-1}$, the weighted finite population Bayesian bootstrap posterior is given by

$$\begin{aligned}
P\,(Y_{\text{nob}} \mid y, w) &= P\left(b_1^{\text{nob}} = r_1, \ldots, b_K^{\text{nob}} = r_K \,|\, w_1^*, \ldots, w_K^*\right) \\
&= \frac{\int_0^1 \cdots \int_0^1 p\,(Y_{\text{nob}} \mid \theta)\, p\,(y \,|\, \theta)\, p\,(\theta)\, d\theta_1 \ldots d\theta_K}{\int_0^1 \cdots \int_0^1 p\,(y \,|\, \theta)\, p\,(\theta)\, d\theta_1 \ldots d\theta_K} \\
&= \frac{\int_0^1 \cdots \int_0^1 \prod_{i=1}^{K} \theta_i^{r_i} \prod_{i=1}^{K} \theta_i^{w_i^*} \prod_{i=1}^{K} \theta_i^{-1} d\theta_1 \cdots d\theta_K}{\int_0^1 \cdots \int_0^1 \prod_{i=1}^{K} \theta_i^{w_i^*} \prod_{i=1}^{K} \theta_i^{-1} d\theta_1 \cdots d\theta_K} \\
&= \prod_{i=1}^{K} \frac{\Gamma\left(w_i^* + r_i\right)}{\Gamma\left(w_i^*\right)} \bigg/ \frac{\Gamma\,(N)}{\Gamma\,(n)}
\end{aligned} \tag{3.3}$$

since $\sum_{j=1}^{n} r_i = N - n$ and $\sum_{j=1}^{n} w_i^* = n.$

Next, we show the distribution of samples obtained from the unequal probability of selection Pólya Urn scheme of Cohen (1997) is equal to the posterior distribution of the FPBB sample with unequal probabilities of selection. Given the observed data, the probability that we draw $N - n$ balls and that the first r_1 balls have value b_1 through the last r_k balls have value b_k is:

$$P(b_1 = r_1, \ldots, b_K = r_K) = \frac{w_1^*}{n} \times \frac{w_1^* + 1}{n + 1} \ldots \times \frac{w_1^* + r_1 - 1}{n + r_1 - 1} \times \ldots \times \frac{w_K^*}{n + \sum_{i=1}^{k-1} r_i} \times \ldots \times \frac{w_K^* + r_K - 1}{n + \sum_{i=1}^{k} r_i - 1}$$

$$= \prod_{i=1}^{K} \frac{\Gamma(w_i^* + r_i)}{\Gamma(w_i^*)} \Big/ \frac{\Gamma(N)}{\Gamma(n)}$$

where the first equality follows from the fact the distribution of the counts of type b_i is invariant under any permutation of the draws, as in the unweighted setting, and the second equality from the identity $\Gamma(x) = (x - 1)\Gamma(x)$ for $x > 0$. Thus, noting that

$$\frac{w_i - 1 + l_{i,k-1} * (N - n)/n}{N - n + (k - 1) * (N - n)/n} = \frac{w_i^* + l_{i,k-1}}{n + (k - 1)},$$

a draw from the unequal probability of selection Pólya Urn scheme yields a draw from $P(Y_{\text{nob}} \mid y, w)$ in (3.3).

4 Nonparametric method to generate synthetic populations

In this section, we extend the finite population Bayesian bootstrap methods to a stratified, clustered, unequal probability sample design setting to develop a nonparametric method to generate synthetic populations that adjusts for the complex sampling design features. The idea is to treat the unobserved part of the population as missing data and impute it by making draws from the actual data. We do the imputation in such a fashion that the resulting draws from the posterior distribution of the population will capture the complex design features and can be used in a standard fashion to compute posterior distributions of the population quantities of interest.

4.1 Use the Bayesian bootstrap to adjust for stratification and clustering

For a stratified clustering sampling, we first need to resample clusters within the strata. Denote c as the total number of clusters in the actual data, $c = \sum_{h=1}^{H} c_h$, and C as the number of clusters in the population, $C = \sum_{h=1}^{H} C_h$. One approach is to first apply FPBB Pólya urn scheme to impute the unobserved clusters within each stratum, $c_1^*, \ldots, c_{C_h - c_h}^*$, which together with the observed clusters provide the clusters in stratum h in the population. However, we typically do not know the number of clusters in a stratum from available public use data. Thus we suggest as an alternative to FPBB sample drawing a standard Bayesian bootstrap sample of the clusters within each stratum. Considering the equivalence between the classical bootstrap and Bayesian bootstrap, we follow Rao and Wu (1988), who suggested drawing a simple random sample with replacement (SRSWR) of m_h from the c_h clusters and within each stratum h calculating replicate weights for computation for each bootstrap sample as

$$w^{*(l)} = \left\{ w_{hik}^{*(l)},\ h = 1, \ldots, H,\ i = 1, \ldots, c_h,\ k = 1, \ldots, N_{hi} \right\},$$

where

$$w_{hik}^{*} = w_{hik}\left(\left(1-\sqrt{\frac{m_h}{c_h - 1}}\right)+\sqrt{\frac{m_h}{c_h - 1}}\frac{c_h}{m_h}\ m_{hi}^{*}\right)$$

and m_{hi}^{*} denotes the number of times that cluster $i, i = 1, \ldots, c_h$ is selected. To ensure all the replicate weights are non-negative, $m_h \leq (c_h - 1)$; here and below we take $m_h = (c_h - 1)$.

Note that, when clustering is not present, we simply draw a standard Bayesian bootstrap sample from the sampled data within each stratum (when stratification is present) or from the entire sample (if stratification not present, so that $H = 1$) and calculate the replicate weights as $w_{hik}^{*} = w_{hik} m_{hi}^{*}$.

This procedure is repeated L times to produce L Bayesian bootstrap (BB) samples denoted by $S_1, \ldots, S_L$. This step generates L Bayesian bootstrap samples which essentially are L draws from the posterior predictive distribution of the unobserved clusters given the actual data. However, the units for the L Bayesian bootstrap samples still have weights and cannot be analyzed as simple random samples.

4.2 Use weighted FPBB Pólya urn scheme to adjust for weighting

Once we have L BB samples with replicate weights, the second step imputes the unobserved units using the weighted FPBB Pólya urn scheme. In practice, the probability of selecting the k^{th} unit, y_k^{*}, depends on the selection of the first $k - 1$ units, $y_1^{*}, \ldots, y_{k-1}^{*}$. In other words, to determine the probability of selecting a new unit, we have to count the number of times that each unit in the sample has been selected among the previous selections. In settings where the population size is extremely large, we need only generate synthetic populations of size $T * n$, where T is sufficiently large to overwhelm the sample size (*e.g.*, 20-100). To further computational efficiency, we could also draw a moderate sized population $F > 1$ times and then pool these F populations to produce one synthetic population, S_l. The size of S_l then is $F * T * n$.

Note that our method only requires knowledge of the final weights in multistage cluster samples, since all stages of unequal probabilities of sampling will be corrected by use of the weighted FPBB Pólya urn scheme. This is a particularly useful feature of the proposed method, as in many public use datasets the components of the probabilities of selection (*e.g.*, cluster-level selection probabilities, non-response weights) are not available.

5 Inference from multiple nonparametric synthetic populations

Assume we generate L synthetic populations, $S_l, l = 1, \ldots, L$ using the nonparametric method described in Section 4, and that our inferential target is $Q \equiv Q(Y)$, a function of the population data (*e.g.*, population mean, correlation, population maximum likelihood estimator of a regression parameter, *etc.*). We can compute Q_l as the estimate of Q obtained from pooling the F synthetic populations that impute the unobserved units of S_l; since these are direct draws from the posterior predictive distribution of the population, we can compute posterior means, quantiles, and credible intervals from the corresponding empirical estimates from the draws, if L is sufficiently large.

However, in many settings, the computational effort required to impute the population may be very large, even if the full population is not required to be synthesized. Hence an alternative approach for inference is to approximate the posterior predictive distribution of a scalar population statistic Q via a t-distribution:

$$Q \mid S_1, \ldots, S_L \overset{\bullet}{\sim} t_{L-1}\left(\bar{Q}_L, \left(1 + L^{-1}\right) V_L\right)$$

where

$$\bar{Q}_L = \frac{\sum_{l=1}^{L} Q_l}{L} = \frac{\sum_{l=1}^{L} \sum_{f=1}^{F} Q_{lf}}{LF} \text{ and } V_L = \frac{1}{L} \sum_{l=1}^{L} (Q_l - \bar{Q}_L)^2 .$$

The result follows immediately from Section 4.1 of Raghunathan *et al.* 2003, and is based on the standard Rubin (1987) multiple imputation combining rules, treating the unobserved units of S_l as missing data and the sampled units as observed data. The average "within" imputation variance is zero, since the entire population is being synthesized; hence the posterior variance of Q is entirely a function of the between-imputation variance, and the degrees of freedom is simply given by the number of FPBB samples. (When the population is extremely large, we need only synthesize a draw sufficiently large for average "within" imputation variance to be trivial relative to the between imputation variance V_L.) The result assumes that $E\left(Q_{lf}\right) = Q$ - a result guaranteed by our weighted FPBB estimator - as well as a a sufficiently large sample size for Bayesian asymptotics to apply.

6 Simulation studies

In this section, we conduct two simulation studies to evaluate the repeated sampling properties of the population estimators constructed using the nonparametric method that generates synthetic populations while adjusting for the complex sampling design features. The first of these considers a one-stage, unequal probability of selection design where we vary the number of weighted FPBB draws for each synthetic population and the number of synthetic populations to assess the impact on inference. The second compares inferential properties from observed data and from the posterior distribution obtained from synthetic population in a stratified, multistage, unequal probability of selection sample, this time fixing the posterior sample size while considering both population means and population regression parameters as targets of inferences.

6.1 Single stage, unequal probability of selection sample design

We generated outcome data Y in a population of N subjects from a moderately skewed gamma distribution, conditional on uniformly distributed covariate X :

$$X_i \sim \text{UNI}\,(0.05; 0.65), i = 1, \ldots, N$$

$$Y_i \mid X_i = x_i \sim \text{GAMMA}\,(10 * x_i, 1)$$

We assume X is fully observed for the population, and that the probability of selection π is proportional to X, so that $\pi_i \doteq nx_i / \sum_i x_i$ in a without-replacement sample design as long as $n << N$. The estimand of interest is the population mean $\bar{Y} = N^{-1}\sum_{i=1}^{N} y_i = 3.564$. Note that $\operatorname{corr}(Y_i, X_i) = 0.6794$, so that unweighted sample means will be positively biased, and use of design weights $w_i = 1/\pi_i$ are required to obtained unbiased estimates of $\bar{Y}$. We generated a population of size $N = 1{,}000$ from which we sampled $n = 100$; bias, empirical and estimated variance, 95% interval length, and nominal 95% coverage are then estimated from 200 independent samples from the population. We varied the total number of simulated populations L as 5, 20, 100, and 1,000, and the number of FPBB draws F of size $N - n$ (so that $K = 9$) as 1, 20, and 100, in full factorial design. Variance, interval length, and interval coverage are obtained via the normal approximation; for $L = 100$ and 1,000, we also obtained variance, interval length, and interval coverage using the direct draws from the posterior predictive distribution, since a sufficient number of draws from the posterior were available to make such estimates.

Table 6.1 shows the results of the simulation study. In all cases the point estimate $\bar{Q}_L$ of the population mean was approximately unbiased, reflecting the ability of the weighted FPBB to "undo" the sampling weights in the generation of the synthetic population. Under the normal approximation, larger numbers of the synthetic population were associated with smaller variances and narrower interval lengths, as expected with larger numbers of degrees of freedom, although the difference between 20 and 100 was minimal, just as the t_{20} distribution begins to approximate a standard normal. Finally, using only a single FPBB draw of size $N - n$ appeared to overestimate the variance and lead to overcoverage, especially for small values of L. Values of L and F of 20 or greater appeared to yield reasonable results. Use of the direct draws for $L = 100$ and 1,000 yielded to variance and credible interval estimates that were very similar to that of the normal approximation, with slightly narrower interval lengths and somewhat less conservative coverage.

Table 6.1
Bias, empirical variance, mean of estimated variance, interval length and coverage of 95% nominal confidence interval of a population mean as a function of the number of synthetic populations (L) and the number of weighted finite Bayesian bootstraps that make up the synthetic population (F). Interval length and coverage obtained via t-approximation and empirically via direct simulation. One stage unequal probability of selection sample design. Results from 200 simulations.

L		**5**			**20**			**100**			**1,000**	
F	**1**	**20**	**100**	**1**	**20**	**100**	**1**	**20**	**100**	**1**	**20**	**100**
Bias	-0.020	0.009	-0.026	0.021	-0.030	0.010	-0.031	0.024	-0.028	-0.045	-0.070	0.079
Emp. Variance	0.126	0.099	0.106	0.088	0.092	0.120	0.093	0.079	0.085	0.084	0.093	0.078
Est. Variance: t	0.172	0.119	0.105	0.156	0.098	0.099	0.109	0.097	0.095	0.147	0.104	0.094
Interval Length: t	2.20	1.78	1.71	1.63	1.30	1.32	1.52	1.21	1.20	1.50	1.26	1.20
95% Coverage: t	97	95	96	99	94	92	98	96	95	98	96	98
Est. Variance: Empirical	0.138	0.095	0.084	0.148	0.093	0.094	0.108	0.096	0.094	0.084	0.093	0.078
Interval Length: Empirical	N/A	N/A	N/A	N/A	N/A	N/A	1.50	1.19	1.18	1.49	1.25	1.19
95% Coverage: Empirical	N/A	N/A	N/A	N/A	N/A	N/A	96	93	94	98	96	97

6.2 Stratified, multistage, unequal probability of selection sample design

We generated a population with strata and clusters within each stratum from the following bivariate normal distribution:

$$\begin{pmatrix} X_{1ijk} \\ X_{2ijk} \end{pmatrix} \sim N\left(\begin{pmatrix} 500 + 4.5 * i + u_{ij} \\ 500 + 4.5 * i + u_{ij} \end{pmatrix}, \begin{pmatrix} 100 & 50 \\ 50 & 100 \end{pmatrix}\right),$$

where

$i = 1 : 150$ denotes the stratum effect,

$u_{ij} \sim N(0, 10)$ denotes the random cluster effect,

$a_i \sim \text{uniform}(2, 52)$ is the number of clusters within stratum i,

$b_{ij} \sim \text{uniform}(10, 20)$ is the number of units within cluster j of stratum i.

The population for the simulation study has 61,324 subjects. We draw a stratified clustering sampling with unequal probabilities of selection. Specifically, we select two clusters from each stratum with probabilities proportional to cluster size (PPS) given by $b_{i\bullet} = \sum_{j=1}^{a_i} b_{ij}$. Within each selected cluster, we select approximately $1/5$ of the population. Thus, the probability that unit ij is selected is given by

$$\pi_{ij} = \frac{2b_{i\bullet}}{\sum_{j=1}^{a_i} b_{ij}} \times \frac{\lfloor b_{ij}/5 \rfloor}{b_{ij}}$$

for all j elements in cluster i with corresponding weight

$$w_{ij} = \frac{b_{ij} \sum_{j=1}^{a_i} b_{ij}}{2b_{i\bullet} \lfloor b_{ij}/5 \rfloor}.$$

Since the number of clusters and units are random, the complex sample size is slightly different across replications, averaging approximately 770.

Because of the large sample and population size, we focus on inference using t approximations. We generate $L = 100$ synthetic populations using F weighted FPBB samples of size $K = 100n$. The estimands of interest are the population marginal mean for x_1

$$\bar{X}_1 = N^{-1} \sum_{i=1}^{N} X_{1i}$$

and similarly for x_2, and the population regression coefficients of x_1 on x_2 given by

$$B_0 = \bar{X}_1 - B_1 \bar{X}_2, B_1 = \frac{\sum_{i=1}^{N} (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sum_{i=1}^{N} (X_{2i} - \bar{X}_2)^2}.$$

We drew 200 independent samples from the population and used the sample data directly to compute weighted sample means and linear regression coefficients along with associated variance estimates and

95% nominal confidence intervals using Taylor Series approximations, and compared these with the equivalent estimates obtained using the nonparametric synthetic data. Results are given in Table 6.2. (Since the marginal means have the same superpopulation value, we combine the results in Table 6.2.) Figure 6.1 displays the scatter plot of the pairs of estimated mean, intercept and slope from the actual samples and the corresponding synthetic populations along with a 45-degee line. The sampling distributions of the actual sample and synthetic population estimates closely correspond. The point estimates and standard errors for both the means and regression parameters closely correspond. The 95% confidence interval coverage rates for all three statistics also closely correspond, and are close to nominal values.

Table 6.2
Descriptive and analytic statistics estimated from the actual data and the synthetic populations in a simulation evaluation of the nonparametric method. Two-stage, unequal probability of selection stratified sample design. Results from 200 simulations.

Type	Actual Data				Synthetic Populations			
	Estimate	SE	SD	Coverage (%)	Estimate	SE	SD	Coverage (%)
Mean $\bar{X}$	836.701	0.461	0.491	93	836.793	0.476	0.493	94
Intercept B_0	1.013	1.768	1.848	94	1.014	1.775	1.846	92
Slope B_1	0.999	0.002	0.002	92	0.999	0.002	0.002	92

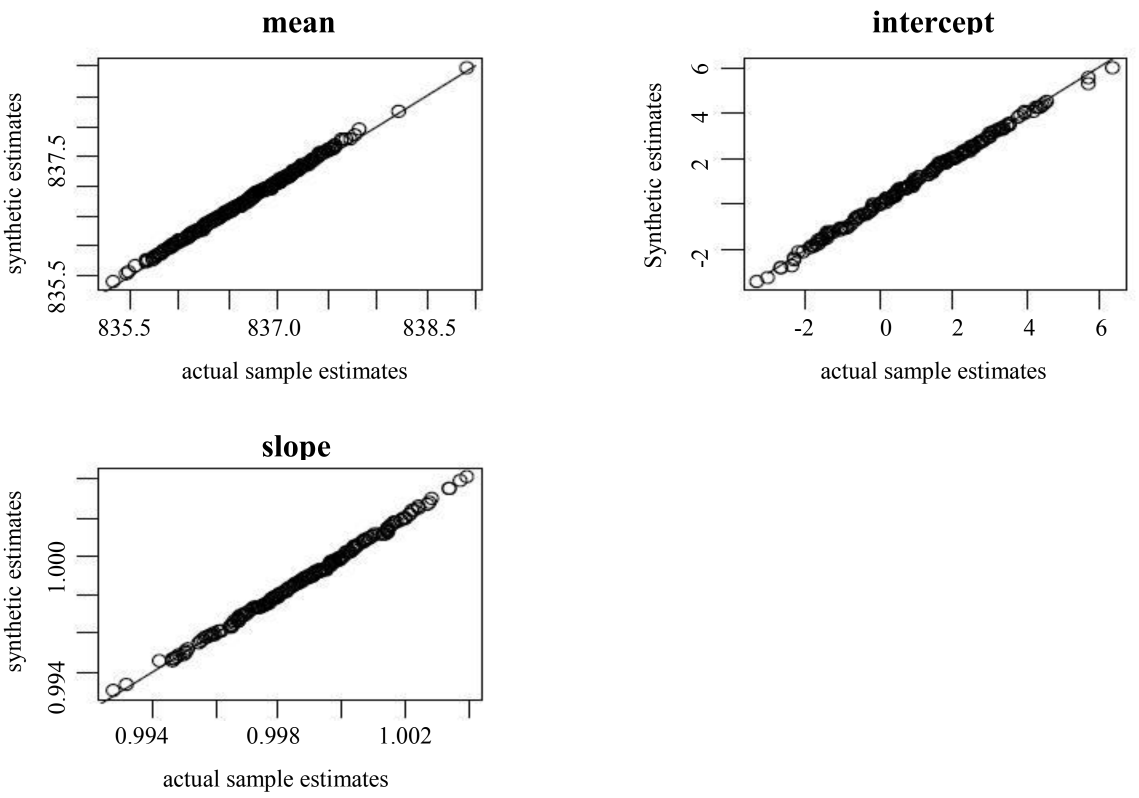


Figure 7.1 Scatter plot of the descriptive and analytic statistics from the actual and synthetic populations

7 Application

In this section, we use data from the 2006 National Health Interview Survey (NHIS) and the 2006 Medical Expenditure Panel Survey (MEPS) to evaluate the performance of the nonparametric method in a stratified clustering sampling design. The National Health Interview Survey (NHIS) is a nationwide, face-to-face health survey based on a stratified multistage design, with oversamples of black, Hispanic, and elderly populations. For confidentiality purposes, the true stratification and primary sampling unit (PSU) variables are not publicly-released; instead pseudo-strata and PSUs (two per stratum) are released. The MEPS is a subsample of the previous year's NHIS sample, and retains the same stratified multistage design.

Both NHIS and MEPS ask respondents whether they are covered by any health insurance and, if so, what type health insurance they are using (private *versus* government-sponsored such as Medicare or Medicaid). We estimate overall health insurance coverage rates as well as coverage rates in subpopulations defined by demographic variables such as gender, race, income level, or combinations thereof: specifically, we estimate health insurance coverage for males, non-Hispanic whites, and non-Hispanic whites with household income between $25,000 and $35,000 per year. We delete the cases with item-missing values and focus on our simulation on the complete cases. This results in 20,147 and 20,893 cases in the NHIS and MEPS data respectively.

7.1 Estimation of health insurance coverage from the NHIS and MEPS

In this simulation study, we will use the nonparametric method to adjust for the stratified clustering sampling used by the 2006 NHIS and MEPS and generate synthetic populations that can be analyzed as simple random samples. We also consider a model-based approach for generating synthetic populations using a log-linear model for the health insurance status by six independent demographic variables: gender, race, census region, education level, age (categorical), and income level (categorical). Then we evaluate the method by comparing the estimates of the health insurance coverage rate for the whole population and selected subdomains obtained from both the non-parametric and log-linear model synthetic populations to those obtained from the actual data.

7.1.1 Generating nonparametric synthetic populations

Using the nonparametric method developed in Section 3, we generate 200 synthetic populations for each survey. Specifically, we generate $B = 200$ BB samples and for each BB sample, we generate $F = 10$ FPBB of size $5n\,(K = 5)$. Thus, each synthetic population is 50 times as big as the actual sample (1,007,350 for NHIS, 1,044,650 for MEPS). Each synthetic population is analyzed as a simple random sample and the estimates are combined as described in Section 5.

7.1.2 Generating synthetic populations via log-linear models

In the common situation that the survey data of interest are in the form of a multidimensional contingency table, a log-linear model might be considered as a parametric approach to generate draws from a posterior predictive distribution. For simplicity of exposition, assume Y is the variable of our interest with m levels, and Z is a design variable with n levels (*e.g.*, gender, race, *etc.*) whose marginal distribution is known for the population. Assume $\pi_{ij}, i = 1, \ldots, m, j = 1, \ldots, n,$ represents the cell proportion of the ij^{th} cell, $\sum_{i=1}^{m} \sum_{j=1}^{n} \pi_{ij} = 1.$ A fully saturated log-linear model is given by (Agresti 2002):

$$\log\left(\pi_{ij}\right) = \lambda_0 + \lambda_i^Z + \lambda_j^Y + \lambda_{ij}^{ZY},\ i = 1, \ldots, m, j = 1, \ldots, n,$$

where $\log\left(\pi_{ij}\right)$ is the log of the probability that one observation falls in cell ij of the contingency table, λ_i^Z is the main effect for Z, λ_j^Y is the main effect for Y and λ_{ij}^{ZY} is the interaction effect for Z and Y. This model includes all possible one-way and two-way effects and thus is saturated as it has the same number of effects as cells in the contingency table. To avoid over-fitting the data in the example, we can consider non-saturated models that exclude some or all of the interaction terms, choosing the model based on likelihood ratio tests or AIC or BIC criteria.

The synthetic populations can be generated from the posterior predictive distribution from the model. However, when the data is collected under a complex sampling design, we are not aware of standard statistical software that can produce both the point estimate and covariance estimate of the regression coefficients. Instead, we have to use a jackknife replication method to adjust for stratification, clustering and weighting. Specifically, the parametric synthetic populations can be generated from the following steps:

1. Estimate coefficients and covariance matrix:

Under the selected model (assume the two-dimensional saturated model here just for illustration), estimate the coefficients $\lambda = \left(\lambda_0, \lambda_i^Z, \lambda_j^Y, \lambda_{ij}^{ZY}\right)'$, $i = 1, \ldots, m-1, j = 1, \ldots, n-1$ and the covariance matrix of the estimates $\hat{\lambda} = \left(\hat{\lambda}_0, \hat{\lambda}_i^Z, \hat{\lambda}_j^Y, \hat{\lambda}_{ij}^{ZY}\right)'$ after taking into account the complex design features using jackknife repeated replication (JRR):

- For each replication, withdraw one cluster, and inflate the weights for the respondents in the other clusters within the same stratum by $c_h / (c_h - 1)$ (replication weights), where c_h denotes the number of clusters within stratum h. Assume we have $\sum_{h=1}^{H} c_h = C$ clusters in total, then we have C replications. For each replication, we fit the log-linear model and obtain the maximum likelihood estimates (MLE) of the coefficients, $\lambda = \left(\lambda_0, \lambda_i^Z, \lambda_j^Y, \lambda_{ij}^{ZY}\right)'$, $i = 1, \ldots, m-1, j = 1, \ldots, n-1$.
- For each replication, use the replication weights to fit the log-linear model. Specifically, use the replication weights to calculate the size of each cell of the contingency table, which is used to fit the log-linear model. We denote the MLE for the r^{th} replication by a column vector, $\hat{\lambda}_r, r = 1, \ldots, c_h$ for stratum h. Notice that $\lambda = \left(\lambda_0, \lambda_i^Z, \lambda_j^Y, \lambda_{ij}^{ZY}\right)'$, $i = 1, \ldots, m-1,$ $j = 1, \ldots, n-1$ is a mn by 1 column vector. We denote $\lambda = \left(\lambda_0, \lambda_i^Z, \lambda_j^Y, \lambda_{ij}^{ZY}\right)' =$ $(\lambda_0, \lambda_1, \ldots, \lambda_{mn})'$. Similarly, $\hat{\lambda}_r, r = 1, \ldots, c_h, h = 1, \ldots, H$ are also mn by 1 column vectors denoted by $\left(\hat{\lambda}_0^{(r)}, \hat{\lambda}_1^{(r)}, \ldots, \hat{\lambda}_{mn}^{(r)}\right)'$.

The MLE of the coefficients $\lambda = \left(\lambda_0, \lambda_i^Z, \lambda_j^Y, \lambda_{ij}^{ZY}\right)'$, $i = 1, \ldots, m-1, j = 1, \ldots, n-1$ can be obtained by $\hat{\lambda}_{\text{MLE}} = \sum_{h=1}^{H} \sum_{r=1}^{c_h} \hat{\lambda}_r \big/ C$. For the mn by mn covariance matrix, the jackknife replication estimate of the pq^{th} $(p, q = 1, \ldots, mn)$ element is the covariance between the p^{th} and q^{th} coefficients, which is given by:

$$\sum_{h=1}^{H} \frac{c_h - 1}{c_h} \sum_{r=1}^{c_h} \left(\hat{\lambda}_p^{(r)} - \bar{\hat{\lambda}}_p\right)\left(\hat{\lambda}_q^{(r)} - \bar{\hat{\lambda}}_q\right),$$

where $\bar{\hat{\lambda}}_p = \sum_{h=1}^{H} \sum_{r=1}^{c_h} \hat{\lambda}_p^{(r)} \big/ C$ and $\bar{\hat{\lambda}}_q = \sum_{h=1}^{H} \sum_{r=1}^{c_h} \hat{\lambda}_q^{(r)} \big/ C$. This gives us the correct variance estimate of $\hat{\lambda}_{\text{MLE}}$.

2. Approximate the posterior distribution of the coefficients:

Let T denote the Cholesky decomposition such that $TT^t = \text{cov}\left(\hat{\lambda}_{\text{MLE}}\right)$. Generate a vector z of random normal deviates and define $\Lambda_* = \hat{\lambda}_{\text{MLE}} + Tz$.

3. Impute the unobserved values of the population:

Suppose L draws, $\Lambda_1, \ldots, \Lambda_L$, are made from the approximate posterior distribution of λ. For each

$$l = 1, \ldots, L, \Lambda_l = \left(\Lambda_0^{(l)}, \Lambda_i^{X(l)}, \Lambda_j^{Y(l)}, \Lambda_{ij}^{XY(l)}\right)', \; i = 1, \ldots, m-1, j = 1, \ldots, n-1,$$

we can generate one synthetic table using the assumed model:

$$\log\left(\pi_{ij}^{(l)}\right) = \Lambda_0^{(l)} + \Lambda_i^{X(l)} + \Lambda_j^{Y(l)} + \Lambda_{ij}^{XY(l)}, i = 1, \ldots, m-1, j = 1, \ldots, n-1.$$

Once the cell proportions are determined, we can generate the synthetic table of any size.

The results below are based on a seven-dimension contingency table (see Table 7.1 for the specific covariate categories). BIC measures indicated that a model with all 2-way but no 3-way interactions provided the most parsimonious fit.

Table 7.1
Variables and response categories for the 2006 NHIS and MEPS used in log-linear model.

Variables of Interest	Response Categories
Age	1: [18; 24]; 2: [25; 34]; 3: [35; 44]; 4: [45; 54]; 5: [55; 64]; 6: >= 65
Census Region	1: Northeast; 2: Midwest; 3: South; 4: West
Education	1: Less than high school; 2: High school; 3: Some college; 4: College
Gender	1: Male; 2: Female
Health Insurance Coverage	1: Any Private Insurance; 2: Public Insurance; 3: Uninsured
Income	1: (0; 10,000); 2: [10,000; 15,000); 3: [15,000; 20,000); 4: [20,000; 25,000); 5: [25,000; 35,000); 6: [35,000; 75,000); 7: >= 75,000
Race	1: Hispanic; 2: Non-Hispanic White; 3: Non-Hispanic Black; 4: Non-Hispanic All other race groups

7.2 Results

The results are summarized in Table 7.2. For the total population and the larger subpopulations, we can see that the point estimates (posterior mean) of health insurance rates are the same for both the nonparametric and log-linear approach, and are almost identical to those obtained from the actual data after complex sampling design features are accounted for. Both methods yield synthetic populations with slightly higher (posterior) variances than the actual data, reflecting the information loss in the synthesis. In the NHIS, the loss for the non-parametric estimator averaged a little over 20% and was slightly greater than for the log-linear model, which averaged around 10%. Both had losses of about 10% over the actual data in MEPS. However, for the smaller subpopulation (non-Hispanic whites earning $25,000-$35,000 per year), the log-linear model produced biased results, due to the fact that the log-linear model did not include all possible interactions. The nonparametric method yields estimates almost identical to those obtained from the actual data after complex sampling design features are accounted for. The log-linear model also substantially underestimated the variance of insurance coverage by 30-40% in these cells, *versus* an overestimation in the nonparametric approach of 10-40%.

Table 7.2
Estimates from actual data and from the synthetic populations (Nonparametric and log-linear model) for the 2006 NHIS and MEPS.

		Actual Data (Complex Design)		**Synthetic Populations**			
				Nonparametric		**Log-linear Model**	
Domain	**Types**	**NHIS**	**MEPS**	**NHIS**	**MEPS**	**NHIS**	**MEPS**
Whole Population				Proportion			
	Private	0.746	0.735	0.746	0.736	0.746	0.734
	Public	0.075	0.133	0.075	0.132	0.076	0.133
	Uninsured	0.179	0.132	0.179	0.132	0.178	0.132
				Variance			
	Private	2.46E-05	2.78E-05	3.15E-05	3.31E-05	2.66E-05	2.86E-05
	Public	6.29E-06	1.44E-05	8.06E-06	1.59E-05	7.99E-06	1.77E-05
	Uninsured	1.84E-05	1.41E-05	2.29E-05	1.71E-05	1.81E-05	1.56E-05
Male				Proportion			
	Private	0.740	0.735	0.740	0.736	0.740	0.735
	Public	0.060	0.101	0.060	0.100	0.060	0.102
	Uninsured	0.200	0.164	0.200	0.164	0.200	0.164
				Variance			
	Private	3.32E-05	3.87E-05	3.93E-05	4.31E-05	3.70E-05	3.52E-05
	Public	6.82E-06	1.53E-05	8.81E-06	1.63E-05	7.91E-06	1.91E-05
	Uninsured	2.94E-05	2.64E-05	3.29E-05	2.79E-05	3.19E-05	2.56E-05
Non-Hispanic White				Proportion			
	Private	0.805	0.788	0.804	0.788	0.804	0.788
	Public	0.062	0.116	0.062	0.116	0.062	0.117
	Uninsured	0.134	0.096	0.134	0.096	0.134	0.096
				Variance			
	Private	2.99E-05	3.35E-05	3.79E-05	4.12E-05	3.07E-05	3.98E-05
	Public	8.20E-06	1.81E-05	1.04E-05	2.00E-05	1.10E-05	2.45E-05
	Uninsured	2.02E-05	1.51E-05	2.35E-05	1.80E-05	1.82E-05	1.82E-05
Non-Hispanic White & Income [25,000; 35,000)				Proportion			
	Private	0.827	0.813	0.827	0.814	0.840	0.838
	Public	0.039	0.079	0.039	0.079	0.037	0.067
	Uninsured	0.134	0.108	0.134	0.107	0.122	0.096
				Variance			
	Private	1.00E-04	1.39E-04	1.48E-04	1.63E-04	6.80E-05	8.59E-05
	Public	2.82E-05	6.31E-05	3.86E-05	7.28E-05	1.79E-05	4.25E-05
	Uninsured	7.24E-05	8.92E-05	9.55E-05	1.11E-04	4.38E-05	5.79E-05

8 Discussion

In this paper, we propose and evaluate a nonparametric method to generate synthetic populations. This method adjusts for the complex sampling design features without assuming any models to the observed data so it is robust to model-misspecification. Also, unlike model-based methods that needs to develop separate imputation models for different variables of interest, the nonparametric method only uses the design variables to generate synthetic populations and thus is not variable-specific.

We considered the repeated sampling properties of our non-parametric synthetic estimators in a univariate gamma and bivariate normal setting, estimating means, slopes, and intercepts. Point estimates were unbiased, intervals had approximately nominal coverage, and losses of efficiency relative to the actual data were trivial. We also considered a "real world" setting, generating a predictive distribution for the 2006 NHIS and MEPS and estimating rates and associated variance estimates of health insurance coverage using both the nonparametric method and a fully parametric log-linear modeling approach. When the model fits the data well, the model-based method is more efficient than the nonparametric method. However, when the assumed model does not fit the data well, as was the case in certain small domains, the model-based method may produce invalid inference. In such situations, the nonparametric method is robust to model misspecfication.

In addition to robustness to model misspecification, another advantage is that the nonparametric method only uses the design variables such as stratum, cluster and weight to impute the unobserved part of the population. Unlike model-based methods, it does not need to model the complicated relationships among the variables of interest, which becomes impossible if there are item missing values in the actual data. The synthetic populations generated by the nonparametric method still preserve the item missing values in the actual data. This potentially fills in a gap in the multiple imputation area in that existing imputation methods typically ignore the complex sampling design features in the data and impute the missing values as if they are simple random samples. A related advantage is that, while design variables are used in the nonparametric generation of the synthetic populations, the synthetic populations themselves do not need to contain them, since they can be analyzed as simple random samples. Hence, disclosure risk associated with release of design variables can be eliminated (De Waal and Willenborg 1997; Mitra and Reiter 2006; Reiter and Mitra 2009).

A fourth practical advantage of the nonparametric method is that it is easier to implement in existing statistical software packages because it focuses on the design variables; thus specific strategies for various types of variables and data structures do not need to be developed.

Because use of the weighted FPBB does not require information about the number of clusters in the population or conditional probabilities of selection at each stage of selection in a multistage sample setting, we use an approximate Bayesian bootstrap method to adjust for stratification and clustering. We view this as advantageous in many ways, since public use datasets typically do not break out weights for each stage of the sample. However, it does have the disadvantage that, to ensure positive replicate weights, the Bayesian bootstrap method produces fewer clusters within strata than in the actual data. In the setting where the probabilities of selection are known for all stages of the sample, it seems likely that the weighted FPBB can be implemented at each stage, with the population of unobserved clusters and the population of elements within each cluster imputed in a two-stage fashion, paralleling Meeden (1999) just as the one-stage FPBB parallels Ghosh and Meeden (1983). This remains an area for future research.

Acknowledgements

This research was supported by NCI grant R01CA129101. The authors wish to thank the Editor, Associate Editor, and two anonymous reviewers for their comments. We are especially indebted to the reviewer that helped us to better understand and explain the links between the finite population Bayesian bootstrap and Pólya posterior discussed in Section 3.

References

Agresti, A. (2002). *Categorical Data Analysis*, New York: John Wiley & Sons, Inc.

Chen, Q., Elliott, M.R. and Little, R.J.A. (2010). Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling. *Survey Methodology*, 36, 1, 23-34.

Cohen, M.P. (1997). The Bayesian bootstrap and multiple imputation for unequal probability sample designs. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 635-638.

de Waal, A.G., and Willenborg, L.C.R.J. (1997). Statistical disclosure control and sampling weights. *Journal of Official Statistics*, 13, 417-434.

Dong, Q. (2012). Combining Information from Multiple Complex Surveys. Unpublished Thesis.

Elliott, M.R. (2007). Bayesian weight trimming for generalized linear regression models. *Survey Methodology*, 33, 1, 23-34.

Elliott, M.R., and Little, R.J.A. (2000). Model-based approaches to weight trimming. *Journal of Official Statistics*, 16, 191-210.

Ericson, W.A. (1969). Subjective Bayesian modeling in sampling finite populations. *Journal of the Royal Statistical Society*, B31, 195-234.

Ghosh, M., and Meeden, G. (1983). Estimation of the variance in finite population sampling. *Sankhyā: The Indian Journal of Statistics*, B45, 362-375.

Hinkins, S., Oh, H.L. and Scheuren, F. (1997). Inverse sampling design algorithms. *Survey Methodology*, 23, 1, 11-21.

Lazzeroni, L.C., and Little, R.J.A. (1998). Random effects models for smoothing poststratification weights. *Journal of Official Statistics*, 14, 61-78.

Little, R.J.A. (1991). Inference with survey weights. *Journal of Official Statistics*, 7, 405-424.

Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.

Little, R.J.A. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99, 546-556.

Lo, A.Y. (1988). A Bayesian bootstrap for a finite population. *Annals of Statistics*, 16, 1684-1695.

Meeden, G. (1999). A noninformative Bayesian approach for two-stage cluster sampling. *Sankhyā: The Indian Journal of Statistics*, B61, 133-144.

Mitra, R., and Reiter J.P. (2006). Adjusting survey weights when altering identifying design variables via synthetic data. Privacy in statistical databases: Lecture Notes in Computer Science, 4302, 177-188.

Raghunathan, T.E., Reiter, J.P. and Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19, 1-16.

Raghunathan, T.E., Xie, D.W., Schenker, N., Parsons, V.L., Davis, W.W., Dodd, K.W. and Feuer, D.J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening, *Journal of the American Statistical Association*,102, 474-486

Rao, J.N.K., and Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

Reiter, J.P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology*, 30, 2, 235-242.

Reiter, J.P. (2005). Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society*, A168, 185-205.

Reiter, J.P., and Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality*, 1, 1, Article 6.

Rubin, D.B (1987). *Multiple Imputation for Non-Response in Surveys*, New York: John Wiley & Sons, Inc.

Scott, A.J. (1977). Large sample posterior distributions in finite populations. *The Annals of Mathematical Statistics*, 42, 1113-1117.

Skinner, C., Holt, D. and Smith, T. (1989). *Analysis of Complex Surveys*, New York: John Wiley & Sons, Inc.

Using successive difference replication for estimating variances

Stephen Ash[1]

Abstract

Fay and Train (1995) present a method called successive difference replication that can be used to estimate the variance of an estimated total from a systematic random sample from an ordered list. The estimator uses the general form of a replication variance estimator, where the replicate factors are constructed such that the estimator mimics the successive difference estimator. This estimator is a modification of the estimator given by Wolter (1985). The paper furthers the methodology by explaining the impact of the row assignments on the variance estimator, showing how a reduced set of replicates leads to a reasonable estimator, and establishing conditions for successive difference replication to be equivalent to the successive difference estimator.

Key Words: Successive differences; Successive difference replication; Systematic random sampling.

1 Introduction

Fay and Train (1995) present a method called successive difference replication (SDR) that can be used to estimate the variance of an estimated total from a systematic random sample from an ordered list. The estimator uses the general form of a replication variance estimator where the replicate factors are constructed such that it mimics the successive difference (SD) estimator.

The paper establishes and uses new concepts to gain more understanding into the methodology originally proposed by Fay and Train (1995), hereafter referred to as F&T. The new concepts help to explain the impact of the row assignments on the variance estimator, show how a reduced set of replicates leads to a reasonable estimator, and establish conditions for successive difference replication to be equivalent to the successive difference estimator. It is our hope that this additional understanding of SDR will make it less mysterious and thereby more accessible to anyone estimating variances for a systematic random sample.

The paper begins by reviewing the SD estimator and how it is suited for variance estimation of systematic random samples. The main section of the paper presents two theorems that provide conditions for the SDR estimator to be equivalent to the SD estimator. The paper concludes with empirical examples that examine alternative row assignments and the suitability of using a reduced set of replicates.

For the remainder of the paper, sys will be used as shorthand for systematic random sampling from an ordered list. We abbreviate sys this way because systematic sampling from an unordered or randomly ordered list can be shown to be equivalent to simple random sampling (Madow and Madow 1944). For our discussion, we focus solely on equal probability selection and methods for selecting a sample in only one dimension. Excellent summaries of sys and estimating variances from sys can be found in Iachan (1982), Wolter (1985, chapter 7), Murthy and Rao (1988), and Bellhouse (1988).

[1] Stephen Ash, U.S. Census Bureau, 4600 Silver Hill Road, Washington DC 20233. E-mail: stephen.eliot.ash@census.gov.

1.1 Review of successive differences

Wolter (1984; estimator 2) provides a form of the successive difference estimator of the variance of an estimated mean $(\bar{y})$ for a *sys* sample design as

$$\hat{v}_{\text{SD1}}(\hat{\bar{y}}) = (1-f)\frac{1}{2n(n-1)}\sum_{k=2}^{n}(y_k - y_{k-1})^2,$$

where y_k is the variable of interest, k indexes the units of the ordered sample, and $f = n / N$ is the sampling fraction. The statistic of interest is Y or the total of y_k over the universe of interest and $\hat{Y}$ is an estimator of Y. Let N and n be the size of the universe and sample, respectively. The mean of y_k and its estimator are defined as $\bar{y} = Y/N$ and $\hat{\bar{y}}$, respectively. We also define the estimator of the total Y as $\hat{Y} = \sum_{k=1}^{n}\breve{y}_k$, where the weighted variable of interest with equal weights is $\breve{y}_k = (N/n)\,y_k$; with unequal sample weights w_k, it is defined as $\breve{y}_k = w_k y_k$. The estimator $\hat{v}_{\text{SD1}}(\hat{\bar{y}})$ has been described by Yates (1953; pages 229-231) and recommended by Wolter (1984). Murthy and Rao (1988, equation 32) provide a sketch of why the estimator works. The short version is that since *sys* only selects one unit within each implicit stratum, SD's solution is to collapse adjacent implicit strata. With two units, we can estimate the variance of an implicit stratum. Implicit strata are collapsed and averaged over all possible pairs and then multiplied by n, the number of implicit strata, to give the variance of all the implicit strata.

One SD variance estimator of a total from a *sys* sample is given by F&T as

$$\hat{v}_{\text{SD1}}(\hat{Y}) = (1-f)\frac{n}{2(n-1)}\sum_{k=2}^{n}(\breve{y}_k - \breve{y}_{k-1})^2.$$

Wolter (1985, equation 7.7.4) defined the same estimator where $w_k = (np_k)^{-1}$ and p_k is the with replacement probability of selection for unit k. F&T defined a second SD estimator

$$\hat{v}_{\text{SD2}}(\hat{Y}) = \frac{1}{2}(1-f)\left[\sum_{k=2}^{n}(\breve{y}_k - \breve{y}_{k-1})^2 + (\breve{y}_n - \breve{y}_1)^2\right],$$

which is "circular" in that it includes an extra squared difference that links the first and last unit from the sorted list.

We express the SD2 estimator more generally as a quadratic form as $\breve{\mathbf{y}}'\mathbf{C}\,\breve{\mathbf{y}}$, where $\breve{\mathbf{y}}' = [\breve{y}_1\breve{y}_2\ldots\breve{y}_n]$ is defined as the $n \times 1$ weighted observation vector and $\mathbf{C}$ is a square matrix with 2 for each value of the diagonal, -1 for every value of the superdiagonal and subdiagonal, and -1 for the bottom left and top right value. Here the superdiagonals are defined as the diagonals adjacent to the main diagonal. The exception is a 2×2 matrix.

2 Successive difference replication

2.1 Definition of successive difference replication

F&T present a method called successive difference replication (SDR) that estimates the variance from a sample selected with *sys* by mimicking $\hat{v}_{\text{SD2}}(\hat{Y})$, *i.e.*, SDR is equivalent or nearly equivalent to

$\hat{v}_{\text{SD2}}\left(\hat{Y}\right)$. We show how SDR can be used to produce replicate factors and weights for a general replicate variance estimator that is equivalent to the SD2 estimator. Before we define the SDR estimator in the first theorem, we first establish some terms and provide a lemma that is used by the theorem.

A row assignment scheme, or more simply RA, is an assignment of two rows of a matrix to each unit in the sample. We usually denote the pair of rows as (a_i, b_i) for unit $i.$ A connected loop is an RA that does not repeat any of the rows, *i.e.*, $a_i \neq a_j$ and $b_i \neq b_j$ for all i and j in the connected loop, and is circular, *i.e.*, $b_i = a_{i+1}$ for all $i < n$ and $b_n = a_1$. For example, one possible connected loop for three observations is (1,2), (2,3), (3,1).

A shift matrix $\mathbf{S}$ can be used to move either the rows or columns of a matrix. We will explain how to move rows, which is similar to columns. A shift matrix is a square matrix that has all 0s, except a single 1 in each column. If we wanted to move row p to row q, we would put a 1 in the q^{th} row of the p^{th} column and 0s elsewhere. We emphasize that order is important in applying a shift matrix to another matrix. The application of $\mathbf{S}$ to another square matrix $\mathbf{A}$ as $\mathbf{AS}$ shifts the columns of $\mathbf{A}$ and $\mathbf{SA}$ shifts the rows of $\mathbf{A}$.

Lemma: Let $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_c$ be shift matrices, then $\text{block}\left(\mathbf{S}_1'\mathbf{S}_1, \mathbf{S}_2'\mathbf{S}_2, \ldots, \mathbf{S}_C'\mathbf{S}_C\right) = \mathbf{I}$.

Proof. We first define a general block diagonal matrix $\mathbf{A}$ that is formed by the square matrices $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_C$ as

$$\mathbf{A} = \text{block}\left(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_C\right) = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \ldots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{A}_C \end{bmatrix}.$$

It can be shown that if both $\mathbf{A}$ and $\mathbf{B}$ are block diagonal matrices and the square matrices $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_C$ have the same dimensions as $\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_C$, respectively, then $\mathbf{AB} = \text{block}\left(\mathbf{A}_1\mathbf{B}_1, \mathbf{A}_2\mathbf{B}_2, \ldots, \mathbf{A}_C\mathbf{B}_C\right)$. For a given shift matrix, we also know that $\mathbf{S}'\mathbf{S} = \mathbf{I}$, since a one row down shift of a shift matrix is $\mathbf{I}$. With the two previous items, the lemma follows.

We also define a one row shift matrix as a shift matrix that either shifts all the rows of another matrix down one row and the last row moved to the first or shifts all the rows of another matrix up one row and the last row moved to the last. If $\mathbf{S}_D$ is a one row shift matrix that moves rows down then it has 1s along the upper superdiagonal and a 1 in the bottom left entry of the matrix, for example $\mathbf{S}_1$. Similarly, if $\mathbf{S}_U$ is a one row shift matrix that moves rows up, then it has 1s along the lower superdiagonal and a 1 in the top entry of the matrix, for example the subsequently defined $\mathbf{S}_2$. Note the property that $\mathbf{S}_D = \mathbf{S}_U'$ and $\mathbf{S}_U = \mathbf{S}_D'$; therefore $\mathbf{S}_U + \mathbf{S}_U' = \mathbf{S}_D + \mathbf{S}_D'$. We now present the main theorem of the paper that establishes the conditions under which SDR is equivalent to SD2.

Theorem 1: Let n be the sample size of a given *sys* sample and $\breve{\mathbf{y}}' = \left[\breve{y}_1 \breve{y}_2 \ldots \breve{y}_n\right]$ be defined as the $n \times 1$ weighted observation vector, where the order of the observations reflects the sort order of *sys*.

(a) Choose a Hadamard matrix of order k $\left(\mathbf{HH}' = k\,\mathbf{I}\right)$, where $n \leq k$.

(b) Choose a RA that assigns two rows (a_i, b_i) to each unit i in the sample. Let the RA define C connected loops of m_c units in each connected loop c.

(c) Choose the $m = n$ rows of $\mathbf{H}$ corresponding to the RA to make the $m \times k$ matrix $\mathbf{M}$. The order of the rows of $\mathbf{M}$ should correspond to the first row of the RA. For example, the first row of $\mathbf{M}$ should be row $a_{i=1}$ of $\mathbf{H}$, the second row should be row $a_{i=2}$ of $\mathbf{H}$, *etc*. Next define the $m \times m$ shift matrix as $\mathbf{S} = \text{block}(\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_C)$ where the $m_c \times m_c$ one row shift matrices $\mathbf{S}_c$ are defined to identify the position of the second row b_i of the RA in $\mathbf{M}$. In general, each shift matrix $\mathbf{S}_c$ will be a shift-up, shift-down, or a 2×2 shift matrix (see the subsequently defined $\mathbf{S}_4$).

Define the estimator for each replicate total r as $\hat{Y}_r = \sum_{i=1}^{n} f_{i,r} \breve{y}_i$, where the matrix of replicate factors is $\mathbf{F} = \mathbf{1}_m \mathbf{1}'_k + \left(2^{-3/2} \mathbf{I}_m - 2^{-3/2} \mathbf{S}\right) \mathbf{M}$ and individual values within the matrix are defined for each unit i (rows of $\mathbf{F}$) of replicate r (columns of $\mathbf{F}$) as $f_{i,r} = 1 + 2^{-3/2} h_{a_i,r} - 2^{-3/2} h_{b_i,r}$. $\mathbf{I}_m$ is a $m \times m$ identity matrix and $\mathbf{1}_m$ is a $m \times 1$ vector of 1s. Then the SDR variance estimator $\hat{v}_{\text{SDR}}\left(\hat{Y}\right) = (1 - f)\, 4/k \sum_{r=1}^{m} \left(\hat{Y}_r - \hat{Y}\right)^2$ is equivalent to the sum of C different SD2 estimators.

Proof. The SDR estimator can be written in matrix notation as

$$(1 - f)\frac{4}{k}\left(\breve{\mathbf{y}}'\left(\mathbf{1}_m \mathbf{1}'_k + \left(2^{-3/2} \mathbf{I}_m - 2^{-3/2} \mathbf{S}\right)\mathbf{M}\right) - \breve{\mathbf{y}}' \mathbf{1}_m \mathbf{1}'_k\right)\left(\breve{\mathbf{y}}'\left(\mathbf{1}_m \mathbf{1}'_k + \left(2^{-3/2} \mathbf{I}_m - 2^{-3/2} \mathbf{S}\right)\mathbf{M}\right) - \breve{\mathbf{y}}' \mathbf{1}_m \mathbf{1}'_k\right)'$$

$$= (1 - f)\frac{4}{k}\left(2^{-3/2}\right)^2 \breve{\mathbf{y}}' (\mathbf{I}_m - \mathbf{S})\, \mathbf{M}\mathbf{M}' (\mathbf{I}_m - \mathbf{S})'\, \breve{\mathbf{y}}$$

Because $\{\text{rows of } \mathbf{M}\} \subseteq \{\text{rows of } \mathbf{H}\}$, it can be shown that $\mathbf{M}\mathbf{M}' = k\,\mathbf{I}$. With this result, the variance becomes

$$(1 - f)\frac{1}{2k} \breve{\mathbf{y}}' (\mathbf{I}_m - \mathbf{S})(k\mathbf{I}_m)(\mathbf{I}_m - \mathbf{S})'\, \breve{\mathbf{y}} = \frac{1}{2}(1 - f)\, \breve{\mathbf{y}}' (\mathbf{I}_m - \mathbf{S})(\mathbf{I}_m - \mathbf{S})'\, \breve{\mathbf{y}}$$

$$= \frac{1}{2}(1 - f)\, \breve{\mathbf{y}}' (2\mathbf{I}_m - \mathbf{S} - \mathbf{S}')\, \breve{\mathbf{y}}$$

The last line follows from the lemma and has a constant value for any choice of $\mathbf{H}$. By noting the block diagonal structure of $\mathbf{S}$, we can write the estimator as

$$\frac{1}{2}(1 - f)\sum_{c=1}^{C} \breve{\mathbf{y}}'_c\, (2\mathbf{I}_m - \mathbf{S}_c - \mathbf{S}'_c)\, \breve{\mathbf{y}}_c,$$

where $\breve{\mathbf{y}}_c$ corresponds to the vector of the weighted observations in connected loop c, which is a result of partitioning the weighted observation vector as $\breve{\mathbf{y}}' = [\breve{\mathbf{y}}_{c=1} \breve{\mathbf{y}}_{c=2} \cdots \breve{\mathbf{y}}_{c=C}]$. The choice of the RA does not change the result since we know that $2\mathbf{I}_m - \mathbf{S}_c - \mathbf{S}'_c$ is constant for either an up or down one row shift matrix $\mathbf{S}_c$.

Note 1: Theorem 1 defines the SDR estimator in terms of replicate factors, but we can alternatively express the estimator in terms of replicate weights as

$$(1-f)\frac{4}{k}\mathbf{y}'(\mathbf{W}-\mathbf{1}_m\mathbf{1}_k')(\mathbf{W}-\mathbf{1}_m\mathbf{1}_k')'\,\mathbf{y}.$$

Here, $\mathbf{W}$ is the $m \times k$ matrix of replicate weights defined as $\mathbf{W} = \mathbf{w} * \mathbf{F}$, where $\mathbf{w} = (w_1, w_2, \ldots, w_n)$ is the vector of design weights for the n units of the sample and the operator $*$ multiplies element-wise the vector $\mathbf{w}$ by each of the columns of $\mathbf{F}$, *i.e.*, if $W_{i,r}$ and w_i are entries of $\mathbf{W}$ and $\mathbf{w}$, respectively, then the entries of $\mathbf{W}$ are defined as $W_{i,r} = w_i \times f_{i,r}$.

Note 2: Huang and Bell (2009) similarly defined SDR as a quadratic form and used it to establish some general properties of the estimator when y_k is i.i.d.(μ, σ^2). Our interest lies with the interpretation of how and how well SDR works. Defining the quadratic form with shift matrices and connected loops leads to insights into the row assignments and the efficiency of the estimator.

For a large sample size, it is not usually practical to use $\mathbf{H}$, where $n < k$. The second theorem shows one way that we can use $\mathbf{H}$ with $k < n$ to produce a larger Hadamard matrix $\tilde{\mathbf{H}}$ with $k \geq n$ that will result in the SDR estimator being equivalent to the SD2 estimator. The second theorem also builds upon and clarifies the instructions F&T give for the case of $n > k$. In F&T's instructions, they use the term cycle to denote every $m_d \leq k$ units of the sample. Theorem 2 does not make conditions on the RA, but otherwise it does follow the setup of F&T.

Theorem 2: Let n be the sample size of a given sys sample.

(a) Choose a Hadamard matrix $\mathbf{H}_A$ of order k_A, where $n > k_A$.

(b) Choose a RA that assigns rows to $\mathbf{H}_A$ to the sample. Retaining their original order, split the n sample units into D cycles. Each cycle d has $m_d \leq k_A$ units. Within each cycle, the RA defines one or more connected loops.

(c) Choose a seminormal Hadamard matrix $\mathbf{H}_B$ of order k_B and use it to define a larger Hadamard matrix $\tilde{\mathbf{H}}$ of order $\tilde{k}$ generated from the original $\mathbf{H}_A$. This can be done by applying a Welsch construction to $\mathbf{H}_A$, *i.e.*, $\tilde{\mathbf{H}} = \mathbf{H}_B \otimes \mathbf{H}_A$.

(d) Choose the $m = \sum_{d=1}^{D} m_d$ rows of $\tilde{\mathbf{H}}$ that correspond to the RA to make the $m \times \tilde{k}$ matrix $\tilde{\mathbf{M}}$. The order of the rows of $\tilde{\mathbf{M}}$ should correspond to the first row of the RA. Next define the $m \times m$ shift matrix as $\mathbf{S} = \text{block}(\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_D)$ where the $m_d \times m_d$ shift matrices $\mathbf{S}_d$ identify the position of the second row b_i of the RA in $\tilde{\mathbf{M}}$.

With this prescription, the SDR estimator is defined as

$$\hat{v}_{\text{SDR}}(\hat{Y}) = (1-f)\frac{4}{\tilde{k}}\sum_{r=1}^{\tilde{k}}(\hat{Y}_r - \hat{Y})^2$$

and is equivalent to the sum of at least D SD2 estimators.

Proof. The result follows by applying Theorem 1. The specific value of D follows from the fact that each of the D cycles can have one or more connected loops, so there will be a total of at least D connected loops.

Example 1: Let $n = 14$ and choose the nonnormal Hadamard $\mathbf{H}_A = \mathbf{H}_{4b}$ of order $k_A = 4$. The number of cycles will be $D = 4$ and the RA within each cycle is given in the second column of Table 2.1 for each unit. We define $\tilde{\mathbf{H}}$ of $\tilde{k} = 16$ using a Welsh construction of the original normal Hadamard matrix as

$$\mathbf{H}_{16} = \mathbf{H}_{4a} \otimes \mathbf{H}_{4b} = \begin{bmatrix} \mathbf{H}_{4b} & \mathbf{H}_{4b} & \mathbf{H}_{4b} & \mathbf{H}_{4b} \\ \mathbf{H}_{4b} & -\mathbf{H}_{4b} & \mathbf{H}_{4b} & -\mathbf{H}_{4b} \\ \mathbf{H}_{4b} & \mathbf{H}_{4b} & -\mathbf{H}_{4b} & -\mathbf{H}_{4b} \\ \mathbf{H}_{4b} & -\mathbf{H}_{4b} & -\mathbf{H}_{4b} & \mathbf{H}_{4b} \end{bmatrix}$$

where

$$\mathbf{H}_{4a} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \text{ and } \mathbf{H}_{4b} = \begin{bmatrix} 1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 \end{bmatrix}.$$

Using $\mathbf{H}_{16}$, we can calculate the replicate factors for 16 replicates as Table 2.1. In matrix notation, $\tilde{\mathbf{M}}$ includes all the rows of $\tilde{\mathbf{H}} = \mathbf{H}_{16}$ except rows 13 and 16. The rows of $\tilde{\mathbf{M}}$ are ordered by a_i, the first row assigned in the RA. The shift matrix is defined as $\mathbf{S} = \text{block}(\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3, \mathbf{S}_4)$, where the shift matrices corresponding to each cycle are

$$\mathbf{S}_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \mathbf{S}_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \mathbf{S}_3 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \mathbf{S}_4 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Table 2.1
Matrix of replicate factors $(f_{i,r})$ for example 1

Unit #	RA	RA		Replicate															
	$\mathbf{H}_A = \mathbf{H}_{4b}$	$\tilde{\mathbf{H}} = \mathbf{H}_{16}$	Cycle	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	(1,2)	(1,2)	1	1.7	1.0	1.7	1.0	1.7	1.0	1.7	1.0	1.7	1.0	1.7	1.0	1.7	1.0	1.7	1.0
2	(2,3)	(2,3)		0.3	1.0	1.0	1.7	0.3	1.0	1.0	1.7	0.3	1.0	1.0	1.7	0.3	1.0	1.0	1.7
3	(3,4)	(3,4)		1.0	0.3	1.0	0.3	1.0	0.3	1.0	0.3	1.0	0.3	1.0	0.3	1.0	0.3	1.0	0.3
4	(4,1)	(4,1)		1.0	1.7	0.3	1.0	1.0	1.7	0.3	1.0	1.0	1.7	0.3	1.0	1.0	1.7	0.3	1.0
5	(1,3)	(5,7)	2	1.0	1.0	1.7	1.7	1.0	1.0	0.3	0.3	1.0	1.0	1.7	1.7	1.0	1.0	0.3	0.3
6	(3,1)	(7,5)		1.0	1.0	0.3	0.3	1.0	1.0	1.7	1.7	1.0	1.0	0.3	0.3	1.0	1.0	1.7	1.7
7	(2,4)	(6,8)		0.3	0.3	1.0	1.0	1.7	1.7	1.0	1.0	0.3	0.3	1.0	1.0	1.7	1.7	1.0	1.0
8	(4,2)	(8,6)		1.7	1.7	1.0	1.0	0.3	0.3	1.0	1.0	1.7	1.7	1.0	1.0	0.3	0.3	1.0	1.0
9	(1,4)	(9,12)	3	1.0	0.3	1.7	1.0	1.0	0.3	1.7	1.0	1.0	1.7	0.3	1.0	1.0	1.7	0.3	1.0
10	(4,3)	(12,11)		1.0	1.7	1.0	1.7	1.0	1.7	1.0	1.7	1.0	0.3	1.0	0.3	1.0	0.3	1.0	0.3
11	(3,2)	(11,10)		1.7	1.0	1.0	0.3	1.7	1.0	1.0	0.3	0.3	1.0	1.0	1.7	0.3	1.0	1.0	1.7
12	(2,1)	(10,9)		0.3	1.0	0.3	1.0	0.3	1.0	0.3	1.0	1.7	1.0	1.7	1.0	1.7	1.0	1.7	1.0
13	(2,3)	(14,15)	4	0.3	1.0	1.0	1.7	1.7	1.0	1.0	0.3	1.7	1.0	1.0	0.3	0.3	1.0	1.0	1.7
14	(3,2)	(15,14)		1.7	1.0	1.0	0.3	0.3	1.0	1.0	1.7	0.3	1.0	1.0	1.7	1.7	1.0	1.0	0.3

Given the replicate factors in Table 2.1, the SDR estimator is equivalent to the sum of five different SD2 estimators, one for each connected loop of the RA, *i.e.*,

$$(1-f)\frac{4}{\tilde{k}}\sum_{r=1}^{\tilde{k}}\left(\hat{Y}_r-\hat{Y}\right)^2=\frac{1}{2}(1-f)\begin{bmatrix}\sum_{i=2}^{4}(y_i-y_{i-1})^2+(y_4-y_1)^2+2(y_6-y_5)^2\\ +2(y_8-y_7)^2+\sum_{i=10}^{12}(y_i-y_{i-1})^2+(y_{12}-y_9)^2\\ +2(y_{13}-y_{13})^2\end{bmatrix}. \qquad (2.1)$$

There are a few items to note with Example 1. First, the number of replicates needed is greater than the sample size. This happens when m_d is not constant across all cycles. The fourth cycle had only two sample units, but we had to use four replicates from each $\mathbf{H}_{4b}$ because at least one of the cycles used four rows.

To make the example more interesting, we chose a nonnormal Hadamard matrix $\mathbf{H}_{4b}$ for $\mathbf{H}_A$. This nonnormal Hadamard was generated by starting with the normal Hadamard $\mathbf{H}_{4a}$ and reversing the procedure for finding a normal Hadamard as described by Hedayat and Wallis (1978). Here we simply changed the sign of all units in the second row and then changed all the signs for the second column.

If we would have used the normal Hadamard matrix $\mathbf{H}_{4a}$ for both $\mathbf{H}_A$ and $\mathbf{H}_B$, the replicate factors for replicates 1, 5, 9, and 13 would have all been 1.0. We call a replicate a dead replicate when every unit gets a value of 1.0 and thereby the replicate estimate is equal to the original estimate. In SDR, there is nothing wrong with dead replicates, it is just the way the replicate factors are distributed by the Hadamard matrix. With a dead replicate, many of the values of 1.0 are in the dead replicate, and the other replicates are more mixed with values of 1.7 and 0.3. However, all the replicates, even the dead replicates, are needed in estimation.

The real value of Theorem 2 is in understanding F&T's original prescription for SDR when $n > k$. In F&T, the RA is applied repeatedly to the $m = k - 1$ rows of $\mathbf{H}_A$ (skipping the first row of $\mathbf{H}_A$), where $\mathbf{H}_A$ is chosen as a normal Hadamard matrix. Replicates are then formed using the k_A columns of $\mathbf{H}_A$. If we apply the larger framework of Theorem 2, we would say that they implicitly used a normal $\mathbf{H}_B$, which results in $\tilde{\mathbf{H}} = \mathbf{H}_B \otimes \mathbf{H}_A$ and only includes the first k_A replicates in the variance estimator. Since a subset of the replicates needed for SDR to be equivalent to SD2 is used, we say that the resultant estimator is an approximation of the SD2 estimator.

Example 1 (continued): If we only used the first four replicates of Table 2.1, the SDR estimator would be equivalent to (2.1) plus the remainder term R that is defined as

$$R=\begin{bmatrix}(y_1-y_2)(y_8-y_7)+(y_1-y_2)(y_{11}-y_{12})+(y_1-y_2)(-y_{13}+y_{14})\\ +(y_8-y_7)(y_{11}-y_{12})+(y_8-y_7)(y_{14}-y_{13})+(y_{11}-y_{12})(y_{14}-y_{13})\\ +(y_4-y_3)(y_8-y_7)+(y_4-y_3)(y_{10}-y_9)+(y_8-y_7)(y_{10}-y_9)\\ +(y_1-y_4)(y_5-y_6)+(y_1-y_4)(y_9-y_{12})+(y_5-y_6)(y_9-y_{12})\\ +(y_2-y_3)(y_5-y_6)+(y_2-y_3)(y_{10}-y_{11})+(y_2-y_3)(y_{13}-y_{14})\\ +(y_5-y_6)(y_{10}-y_{11})+(y_5-y_6)(y_{13}-y_{14})+(y_{10}-y_{11})(y_{13}-y_{14})\end{bmatrix}$$

Note that R includes the same number of positive and negative terms, which do not cancel exactly, but has the result that R is usually close to zero. Similarly, using replicates 1 to $q \times k_A$, where $q = 1, 2, \ldots, k_B$, will result in a R that has an equal number of positive and negative terms. Only with all the replicates of $\tilde{\mathbf{H}}$ will the remainder term R equal 0.

Example 2: The Current Population Survey (CPS) has a monthly sample size of $n = 72{,}000$ households per month (U.S. Census Bureau 2006). CPS has a two-stage sample design, where a first-stage sample of Primary Sample Units (PSUs), which are generally counties or groups of counties, are selected and then in the second-stage households are selected within the sample PSUs. Some PSUs, generally the metropolitan areas, are selected with certainty, *i.e.*, their first-stage probability of selection is 1.0. With the certainty PSUs, the sys sample can be treated as the first-stage sample design in variance estimation, *i.e.*, SDR is applied to produce replicates. In the noncertainty PSUs, Balanced Repeated Replication (BRR) [McCarthy 1966] is applied to produce replicates. Roughly 75% of the sample or 54,000 units are in SR PSUs, where SDR is applied.

The CPS application of SDR uses a Hadamard matrix with $k = 160$ and excludes two rows, *i.e.*, $m =$ 158. Replicate weights are produced for 160 replicates. Although it may seem like a logical conclusion of the paper, we do not suggest that CPS should use a Hadamard matrix of order $k = 54{,}000$ or produce 54,000 sets of replicate weights. That would result in an unreasonable number of replicates. Instead, we suggest that the subset of 160 replicates used by CPS is large and therefore provides a reasonable approximation to SD2. Later in the empirical examples, we examine the impact of using a reduced set of replicates.

2.2 Row assignment when $n > k$

Until this point we have assumed a given RA and have not discussed how to generate the RA for a given sample, where $n > k$. In this section, we review two RAs and discuss some considerations about RAs in general. The first RA is similar to the RA described by Sukasih and Jang (2003) and is intended for use with $k < n$ and Theorem 2.

RA1: The RA assigns a pair of rows a_i and b_i to every m_d units of the sample, which we call cycle d, where $m_d \leq k$. After $m_d - 1$ cycles, the RA is repeated until all units of the sample have been assigned a pair of rows.

Step 1: Sort the sample in the order in which it was sorted prior to sample selection.

Step 2: Initialize the cycle number as $d = 1$ and the number of connected loops as $c = 1$.

Step 3: Start the RA at the beginning of a cycle or a connected loop as $a_1 = c$.

Step 4: Repeat the following RA: $b_i = \mathrm{mod}(a_i + d, k)$ and $a_i = b_i$ until all m_d rows of the cycle have been used or the RA becomes a connected loop. Here, the modulo function or $\mathrm{mod}(a, b)$ is defined as the remainder of the division of a by b. If all m_d rows of the cycle have been used, start a new cycle: let

$d = d + 1$ and go back to step 3. Otherwise, (end of a connected loop, but not the end of a cycle) start a new connected loop: let $c = c + 1$ and go back to step 3.

Step 5: At the end of $d = m_d - 1$ cycles, start over with the first cycle – go back to step 2.

RA1 has the following characteristics:

- Each of the cycles $d = 1, 2, \ldots, m_d - 1$ of the RA, assigns m_d pairs of rows. This generates a total of $m_d\,(m_d - 1)$ pairs of rows.
- The RA repeats itself after $m_d - 1$ cycles. F&T suggest that after 10 cycles, the RA be restarted. We suggest that all $m_d - 1$ cycles be used before restarting the RA.
- The values of a_i and b_i are always c units apart.
- Halfway through the sequence, the pattern repeats itself in reverse order. If m is even, the cycles before and after the $(m_d + 1)/2^{\text{th}}$ cycle repeat themselves in reverse order.

RA1 differs from the RA of Sukasih and Jang (2003) in that we do not suggest that row 1 be skipped, nor that the RA be repeated after 10 cycles, or require that $k - 1$ be prime. First, a row of all 1s may seem odd, but it is not a problem. Similar to a column of all 1s in $\mathbf{M}$ which made a dead replicate, a row of all 1s will only effect the distribution of the replicate factors. The replicate factors for a unit i that are assigned row 1 (either $a_i = 1$ or $b_i = 1$) will have more replicate factors of 1.0 than otherwise. This is not wrong; it is just how the replicate factors are distributed by $\mathbf{H}_A$. The second difference is that we suggest repeating the assignment after m cycles, which is when the pattern repeat, instead of a fixed number of 10 cycles. Lastly, we do not require that $k - 1$ be prime but note that if $m_d = k - 1$ and $k - 1$ is prime, then every cycle is guaranteed to have only one connected loop.

We also provide a second simpler-to-implement RA called RA2 that will be compared with RA1 in the empirical examples.

RA2: No mixing of row assignments. Repeat the same simple RA for every m_d units, *i.e.*, $(1, 2), (2, 3), \ldots, (m_d, 1)$.

3 Empirical examples

The questions of interest for the empirical examples are:

Q1. How well does SDR perform with a subset of all the replicates needed for SDR to be equivalent to SD?

Q2. Which row assignment is better, RA1 or RA2?

Q3. Should we use more or fewer connected loops?

To address these questions, we applied the SDR variance estimator to several populations. With each population, we selected a *sys* sample of size $n = 64$. Table 3.1 outlines the three SDR estimators we applied.

Table 3.1
SDR estimators for the empirical examples

Estimator	k_A	$\mathbf{H}_A$	k_B	$\mathbf{H}_B$
1	4	$\mathbf{H}_{4a}$	16	$\mathbf{H}_{4a} \otimes \mathbf{H}_{4a}$
2	16	$\mathbf{H}_{4a} \otimes \mathbf{H}_{4a}$	4	$\mathbf{H}_{4a}$
3	64	$\mathbf{H}_{4a} \otimes \mathbf{H}_{4a} \otimes \mathbf{H}_{4a}$	1	1

With this construction, the SDR estimators had $k_B = 1, 4,$ or 16 cycles, but all used the same $\tilde{\mathbf{H}} = \mathbf{H}_{4a} \otimes \mathbf{H}_{4a} \otimes \mathbf{H}_{4a}$, which is the normal Hadamard matrix of order $\tilde{k} = 64$. For the three estimators of Table 3.1, we also varied the row assignment (RA1 and RA2) and the number of replicates used by each estimator is either 16, 32, 48, or 64. With both RA1 and RA2, there is only one connected loop within each cycle, so estimators 1, 2, and 3 had $k_B = 16, 4,$ and 1 connected loops, respectively. In the Appendix Section, the results for the SDR estimators are summarized in Table A1 and Table A2 includes the SD1, SD2, and the *srswor* variance estimators applied for comparison purposes.

Data sets used. The "A" populations are borrowed from the empirical example in Wolter (1984). For populations A1-A7, we generated 400 finite populations of size $N = 64{,}000$. From each population, there were $b = 100$ possible samples of size $n = 64$. The samples are indexed as $i = 1, 2, \ldots, b = 100$ and the units within each sample are indexed as $j = 1, 2, \ldots, n = 64$. Table 3.2 summaries how the variable of interest μ_{ij} is generated for each of the "A" populations.

Table 3.2
Description of Wolter's artificial populations

Population	Description	n	b	μ_{ij}	e_{ij}
A1	Random	20	50	0	e_{ij} iid $N(0,100)$
A2	Linear Trend	20	50	$i + (j-1)k$	e_{ij} iid $N(0,100)$
A3	Stratification Effects	20	50	j	e_{ij} iid $N(0,100)$
A4	Stratification Effects	20	50	$j + 10$	$e_{ij} = \begin{cases} \varepsilon_{ij,} & \text{if } \varepsilon_{ij} \geq -(j+10) \\ -(j+10), & \text{otherwise} \end{cases}$ ε_{ij} iid $N(0,100), \rho = 0.8$
A5	Autocorrelated	20	50	0	$e_{ij} = \rho e_{i-1,j} + \varepsilon_{ij}$ $e_{i1} \sim N\left(0, 100/\left(1-\rho^2\right)\right)$ ε_{ij} iid $N(0,100), \rho = 0.8$
A6	Autocorrelated	20	50	0	same as A5 with $\rho = 0.4$
A7	Periodic	20	50	$20\sin\{2\pi/50[i + (j-1)k]\}$	e_{ij} iid $N(0,100)$

Evaluation measures. We evaluated the different variance estimators with the three measures used by Wolter: expected relative bias (ERB), relative mean squared error (RMSE), and coverage ratios. The first measure, ERB, was used to examine the accuracy of the estimators and is defined for a specific estimator θ as $\text{ERB}(\hat{v}_\theta) = E_m\left(E_p\left(\hat{v}_\theta - v\right)\right)/E_m(v)$. In our notation, E_p and E_m refer to the design and model expectations, respectively. To examine the variance of the estimators, we also measured the RMSE, which is defined as $\text{RMSE}(\hat{v}_\theta) = E_m\left(E_p\left(\hat{v}_\theta - v\right)^2\right)/E_m(v)$. Coverage ratios were calculated as the percent of times the true population total fell within the confidence interval using the estimate, *i.e.*, $\left(\hat{Y} - z_\alpha\sqrt{\hat{v}_\alpha}, \hat{Y} + z_\alpha\sqrt{\hat{v}_\alpha}\right)$. Here z_α is the value from a normal distribution and was chosen to make 95% confidence intervals.

Results. With respect to Q1, columns 4-7 of Table A1 show that increasing the number of replicates had minimum impact on the bias. Only with the linear trend population (A2) did the SDR estimator with four connected loops show a consistent trend in reduced bias as the number of replicates increased. The other population and estimator combinations showed no significant decreasing or increasing trend as the number of replicates increased. This finding is a positive result because it indicates that reducing the set of replicates does not increase the bias. As expected, the RMSEs in columns 8-11 in Table A1 did increase as the number of replicates decreased, but surprisingly the increase was relatively minor. Similarly, the confidence intervals in columns 12-15 improved with increased replicates, except with populations A2 and A7.

When comparing RA1 and RA2 of Q2, the SDR estimator with four connected loops usually had smaller biases (columns 4-7 in Table A1) and variances (columns 8-11 in Table A1) with RA1 as compared to RA2. With 16 connected loops, both the biases and variances were similar for both RA1 and RA2. This evidence suggests that both the bias and variance are improved, but the impact reduces as the size of the connected loops decreases.

Addressing Q3, the biases diminished in columns 4-7 with an increasing number of connected loops. The exception was the periodic population (A7). When the RMSEs of SD1 and SD2 were not similar as in linear trend population (A2), increasing the number of connected loops also reduced the RMSEs. This result is not surprising. The estimator with one large connected loop is equivalent to SD2, so it can have the largest biases and RMSEs due to the term $(\hat{y}_1 - \hat{y}_{64})^2$. In the other direction, more connected loops effectively reduces the impact of the term $(\hat{y}_1 - \hat{y}_{64})^2$, so the estimator acts more like SD1, which generally has less bias and variance than SD2.

4 Concluding remarks

The paper provided the conditions for SDR to be equivalent to SD2 and showed how they are equivalent when the sample size is both smaller and larger than the chosen Hadamard matrix. When a smaller Hadamard matrix $\mathbf{H}_A$ is used and replicates are only derived from $\mathbf{H}_A$, the paper showed how the reduced set of replicates provides a reasonable approximation of the SD2 estimator. The empirical examples indicated that using a reduced set of replicates is reasonable since decreasing the number of replicates does not increase the bias of the estimates. Additionally, we saw that using many connected loops reduces the impact of the squared difference between the first and last unit in the sample. Since SD1 usually has larger biases and RMSEs than SD2, SDR estimators that use more rather than fewer connected loops will have smaller biases and RMSEs than SDR estimators.

Acknowledgements

The author would like to thank David Hornick and Brian Dumbacher for their review of the early draft and the referees and the editor for their comments that helped refine and clarify the paper.

Appendix

Table A1
SDR simulation results

Population	k_A	RA	Expected Relative Bias by # Replicates				Relative Mean Squared Errors				Coverage Ratios			
			16	**32**	**48**	**64**	**16**	**32**	**48**	**64**	**16**	**32**	**48**	**64**
A1	4	1	0.010	0.009	0.009	0.009	0.176	0.091	0.066	0.054	93	94	94	94
		2	0.010	0.010	0.010	0.009	0.176	0.095	0.064	0.048	92	94	94	95
	16	1	0.009	0.008	0.010	0.009	0.141	0.080	0.059	0.048	93	94	94	95
		2	0.009	0.010	0.010	0.009	0.194	0.096	0.065	0.049	92	94	94	95
	64	1 or 2	0.009	0.009	0.010	0.009	0.194	0.096	0.064	0.049	92	94	94	94
A2	4	1	-0.696	-0.840	-0.888	-0.907	0.485	0.706	0.789	0.823	62	45	38	35
		2	-0.538	-0.768	-0.845	-0.883	0.290	0.590	0.714	0.780	77	54	45	39
	16	1	0.113	-0.270	-0.500	-0.615	0.013	0.073	0.250	0.378	100	97	80	100
		2	1.302	0.152	-0.231	-0.423	1.695	0.023	0.054	0.179	100	100	99	100
	64	1 or 2	1.302	1.379	1.404	1.417	1.695	1.901	1.972	2.008	100	100	100	100
A3	4	1	0.049	0.031	0.025	0.021	0.195	0.095	0.068	0.054	93	94	94	95
		2	0.070	0.040	0.030	0.025	0.222	0.103	0.067	0.050	93	94	94	95
	16	1	0.155	0.105	0.075	0.060	0.207	0.106	0.070	0.055	95	95	95	95
		2	0.314	0.163	0.112	0.086	0.374	0.144	0.085	0.061	96	95	95	95
	64	1 or 2	0.314	0.324	0.327	0.327	0.374	0.245	0.199	0.176	96	97	97	97
A4	4	1	0.040	0.023	0.017	0.014	0.192	0.104	0.077	0.063	93	94	94	94
		2	0.060	0.030	0.021	0.017	0.217	0.110	0.075	0.058	93	94	94	95
	16	1	0.144	0.095	0.066	0.052	0.208	0.109	0.077	0.063	95	95	95	95
		2	0.291	0.146	0.098	0.075	0.357	0.144	0.090	0.067	96	95	95	95
	64	1 or 2	0.291	0.299	0.303	0.305	0.357	0.232	0.191	0.170	96	97	97	97
A5	4	1	0.063	0.063	0.063	0.065	0.192	0.106	0.076	0.063	94	94	95	95
		2	0.068	0.066	0.066	0.065	0.217	0.111	0.075	0.057	93	94	95	95
	16	1	0.063	0.063	0.063	0.065	0.161	0.093	0.068	0.057	94	95	95	95
		2	0.065	0.067	0.066	0.066	0.214	0.111	0.075	0.056	93	94	95	95
	64	1 or 2	0.065	0.066	0.066	0.065	0.214	0.110	0.074	0.056	93	94	95	95
A6	4	1	0.093	0.092	0.093	0.094	0.211	0.117	0.088	0.072	94	95	95	95
		2	0.092	0.096	0.095	0.094	0.229	0.120	0.086	0.067	94	95	95	95
	16	1	0.099	0.095	0.094	0.094	0.185	0.107	0.080	0.067	94	95	95	95
		2	0.093	0.094	0.094	0.093	0.226	0.117	0.085	0.067	94	95	95	95
	64	1 or 2	0.093	0.096	0.095	0.095	0.226	0.118	0.084	0.066	94	95	95	95
A7	4	1	0.105	0.069	0.112	0.253	0.219	0.106	0.091	0.143	94	95	95	97
		2	0.004	0.004	0.073	0.310	0.187	0.098	0.079	0.175	92	94	95	97
	16	1	0.177	0.168	0.462	0.847	0.229	0.137	0.351	0.828	95	96	98	99
		2	0.002	0.003	0.027	1.248	0.187	0.097	0.065	1.689	92	94	95	100
	64	1 or 2	0.002	0.003	0.030	0.115	0.187	0.097	0.065	0.062	92	94	95	96

Table A2
Comparison methods simulation results

Population	Expected Relative Bias by # Replicates			Relative Mean Squared Errors			Coverage Ratios		
	SD1	SD2	SRSWOR	SD1	SD2	SRSWOR	SD1	SD2	SRSWOR
A1	0.009	0.009	-0.001	0.049	0.049	0.032	94	94	97
A2	-0.960	1.417	25.317	0.921	2.008	640.916	23	100	100
A3	0.015	0.327	3.462	0.049	0.176	12.203	94	97	100
A4	0.006	0.305	3.284	0.057	0.170	11.109	94	97	100
A5	0.064	0.065	0.055	0.056	0.056	0.039	95	95	97
A6	0.093	0.095	0.084	0.065	0.066	0.046	95	95	98
A7	0.112	0.115	20.641	0.063	0.062	427.141	96	96	100

References

Bellhouse, D.R. (1988). Systematic sampling. Excerpt from *Handbook of Statistics*, 6, 125-145.

Fay, R.E., and Train, G.F. (1995). Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties. *Proceedings of the Section on Government Statistics*, American Statistical Association, 154-159.

Hedayat, A., and Wallis, W.D. (1978). Hadamard matrices and their applications. *The Annuals of Statistics*, 6, 1184-1238.

Huang, E.T., and Bell, W.R. (2009). A simulation study of the distribution of Fay's successive difference replication variance estimator. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 5294-5308.

Iachan, R. (1982). Systematic sampling: A critical review. *International Statistical Review*, 50, 293-303.

Madow, W.G., and Madow, L.H. (1944). On the theory of systematic sampling. *Annuals of Mathematical Statistics*, 15, 1-14.

McCarthy, P.J. (1966). Pseudo-replication: Half-samples. *Review of the International Statistical Institute*, 37, 239-264.

Murthy. M.N., and Rao, T.J. (1988). Systematic sampling with illustrative examples. Excerpt from *Handbook of Statistics*, 6, 147-185.

Sukasih, A.S., and Jang, D. (2003). Monte Carlo study on the successive difference replication method for non-linear statistics. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3608-3612.

Wolter, K.M. (1984). An investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association*, 781-790.

Wolter, K.M. (1985). *Introduction to Variance Estimation*, Springer-Verlag.

Yates, F. (1953). *Sampling Methods for Censuses and Surveys*, 2nd Edition, Hafner Publishing Company, New York, NY.

U.S. Census Bureau (2006). Technical Paper 66, "Design and Methodology: Current Population Survey," October 2006.

ELECTRONIC PUBLICATIONS AVAILABLE AT

PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À

www.statcan.gc.ca

Variance Estimation Using Linearization for Poverty and Social Exclusion Indicators

Eric Graf and Yves Tillé[1]

Abstract

We have used the generalized linearization technique based on the concept of influence function, as Osier has done (Osier 2009), to estimate the variance of complex statistics such as Laeken indicators. Simulations conducted using the R language show that the use of Gaussian kernel estimation to estimate an income density function results in a strongly biased variance estimate. We are proposing two other density estimation methods that significantly reduce the observed bias. One of the methods has already been outlined by Deville (2000). The results published in this article will help to significantly improve the quality of information on the precision of certain Laeken indicators that are disseminated and compared internationally.

Keywords: influence function; EU-SILC survey; non-linear statistics; poverty and inequality indicators.

1 Introduction

Deville (2000) proposed that the precision of non-linear statistics in sampling designs be estimated using the generalized linearization method, which relies on the concept of influence function proposed by Hampel (1974) in the field of robust statistics. Osier (2009) applied these theories to estimate the variance of complex statistics such as the Laeken indicators (Eurostat 2005) in the European Statistics on Income and Living Conditions (EU-SILC) survey. Goga, Deville and Ruiz-Gazen (2009) extend the theory of Deville (2000) to two-sample surveys. Verma and Betti (2011) provide a comprehensive list of traditional poverty indicators and associated linearized variables, and they also compare the performance of the linearization technique with that of the jackknife repeated replication method. In this article, we will limit ourselves to poverty indicators published in the EU-SILC survey and focus on the way to estimate the income density function at various points in the distribution.

In Section 2, we review the required theoretical foundations, the expressions for the poverty and inequality indicators being studied, and the linearized variables of those indicators. Some linearized variables are dependent on the density function of the variable of interest, which is usually estimated using the Gaussian kernel estimation method. Two alternative methods are presented in Section 3. The R-language simulations are described and discussed in Section 4. We show that Gaussian kernel estimation can generate strong bias in the estimated variance of indicators when an estimate of the income density function is being used. We also show that the other two density estimation methods proposed in Section 3 reduce the observed bias, and this is discussed in the findings in the last part of this article.

2 Review of given poverty indicators and their linearized variables

Let U be a finite population consisting of N identifiable units $u_1, \ldots, u_k, \ldots, u_N$. To simplify the notation, let unit u_k be denoted simply by the index k. In practice the population U is a sampling frame with acceptable coverage of a given population for which we wish to make inferences. To each unit k is

[1] Eric Graf and Yves Tillé, Institute of Statistics, Faculty of Economics and Business, University of Neuchâtel, Rue de la Pierre-à-Mazel 7, 2000 Neuchâtel, Switzerland. Email: eric.graf@unine.ch and yves.tille@unine.ch.

associated the value y_k for a given characteristic (in this case, income). Without loss of generality, to simplify the notation, assume that the values of y_k are distinct and sorted by order of magnitude, so that $y_k = y_{[k]}$. Data from sample surveys often contain duplicates, that is, a number of units with the same value y, as a result of rounding or range questions. In these cases and for this study, we can simply increase the values by a small (negligible), randomly selected, uniformly distributed amount so that the data may be sorted unambiguously.

Let S be a random sample of size n obtained using a sample design $p(s) = \mathrm{P}(S = s)$, for all $s \subset U$. In addition, let $\pi_k = \mathrm{P}(k \in s) > 0$ be the inclusion probability of unit k of U. As well, let $d_k = 1/\pi_k$ be the sampling weight, and let $w_k = w_k(s)$ be the estimation weight, which may be equal to d_k or may be more refined. For example, w_k may have been obtained after calibration (Deville and Särndal 1992) and therefore also reflect a non-response adjustment.

The estimators of poverty and inequality indicators are non-linear statistics that can't be expressed as regular functions of totals (that is, continuously differentiable up to the second order). In fact, they are rank statistics for the Gini coefficient and quantile statistics for the others. As Osier (2009) points out, their variance therefore can't be estimated using a Taylor linearization; the generalized linearization method is required instead (Deville 2000, Demnati and Rao 2004, and Osier 2009). An alternative for estimating variance would be to use bootstrap-type re-sampling techniques but, for the EU-SILC survey data, preference was given to the linearization technique, at least for a certain number of participating countries. Indeed, re-sampling methods often require more human and machine resources. As well, since Eurostat collaborates with some 30 countries that have different sampling designs and that may perform non-response adjustments and calibration to external sources, it seemed more appropriate to select an analytical solution for estimating variance. In addition, some countries might be using the existing SAS software POULPE (Ardilly and Osier 2007) to generate the required estimates. That was the case for initial tests using Swiss EU-SILC data. Here we use a procedure that, as Antal, Langel and Tillé (2011) point out, reconciles the approach introduced by Deville (2000) with that of Demnati and Rao (2004). Both approaches use the concept of *influence function* initially developed in the field of robust statistics (Hampel 1974). Antal *et al.* (2011) also state that the same linearized variables can be found by applying the method proposed by Graf (2011 and 2013) that constructs a linearized variable on the basis of a Taylor expansion with respect to sample inclusion indicators. Note also the work by Kovačević and Binder (1997) in which a linearization approach using estimating equations is developed.

Deville (2000) states that the influence of a unit k on a population parameter of interest θ is determined by an infinitesimal variation in the importance assigned to the unit. The parameter is expressed as a functional $\theta = T(M)$, where M is a measure allocating a mass of 1, $M(k) = M_k = 1$, only at points on the continuum corresponding to units $k \in U$. The specialization of the general measure M into a discrete measure turns the functional T, predefined on a continuum, into a discrete functional, in the same way as the total Y is defined as the sum of all y_k over the given finite population. The *influence function* of T, or the *linearized variable*, is defined as

$$I[T(M)]_k = z_k = \lim_{t \to 0} \frac{T(M + t\delta_k) - T(M)}{t}, \text{ for all } k \in U,$$

where δ_k is the Dirac measure for unit $k\,(\delta_k(i) = 1 \text{ if } i = k \text{ and } 0 \text{ otherwise})$. In practice, we have known data only from a sample S and Deville (2000) defines a linearized variable $\hat{z}_k$, or *empirical* *influence function,* by (1) determining the limit above using differential calculus and (2) replacing the unknowns in the evaluation with the corresponding estimated quantities using the sample. Deville justifies this procedure by showing that

$$T(\hat{M}) - T(M) \approx \left(\sum_{k \in S} w_k z_k - \sum_{k \in U} z_k\right).$$

The key result is that, under asymptotic conditions described by Deville (2000), which are in theory satisfied when the sample is "sufficiently large", the variance of the estimated total of the variable $\hat{z}_k$ is an approximation of the variance of the (complex) statistic $\hat{\theta}$:

$$\text{var}\left[\sum_{k \in s} \hat{z}_k w_k\right] \approx \text{var}(\hat{\theta}).$$

The starting point of Deville's approach is therefore the population parameter and not the estimator that is proposed to be used for the evaluation using the sample. When the estimator used follows naturally from the population parameter expression (for example, the total Y approached by the Horvitz-Thompson estimator), the procedure is unambiguous. However, imprecision arises if we estimate the same total Y using the ratio estimator with an auxiliary variable x. In that case, Deville's approach, which does not specify the form of the total estimator to use, will yield a constant influence function equal to 1, instead of bringing the unknown ratio of interest into play.

An alternative that avoids these problems is the approach by Demnati-Rao, when used in Deville's framework, as done in Antal *et al.* (2011). They present the Demnati-Rao approach as resulting from Deville's framework when the measure M used is not the discrete measure defined for U described above, but rather the following measure defined for S, the sample:

$$\hat{M}(k) = w_k, k \in S$$

where w_k is a weight. By defining the measure for S, the starting point becomes the estimator and not the parameter; it is the parameter that is initially expressed as a functional, and not the population parameter to be estimated. That is, the functional corresponds to the estimator for which we are seeking a variance estimate using generalized linearization. We then obtain the linearized variable based on that functional as follows:

$$I[T(\hat{M})]_k = \hat{z}_k = \lim_{t \to 0} \frac{T(\hat{M} + t\delta_k) - T(\hat{M})}{t}, \text{ for all } k \in S.$$

Antal *et al.* (2011) note that, to the extent that the functional in this limit is expressed as an explicit function of the variables that are the weights assigned by the measure $\hat{M}$ to the observations, this linearized variable is in fact a function of the partial derivatives with respect to the weights:

$$I[T(\hat{M})]_k = \frac{\partial T(\hat{M})}{\partial w_k}.$$

Antal *et al.* (2011) point out that the linearized variables that we will discuss below can be obtained using either approach. In fact, computing the limit using the Demnati-Rao approach does not necessarily result in the variance estimate suggested by Deville (2000). The practical approach used in this article might therefore be called the Deville-Demnati-Rao approach, in recognition of the theoretical framework provided by Deville (2000) and the practical algorithm for Deville's framework provided by Demnati and Rao (2004).

Using this method, the variance of $\hat{\theta}$ can be estimated for any sampling design, and a confidence interval can therefore be obtained by substituting the linearized variable in the variance formula for a total for the selected sampling design. If the sampling design is simple random sampling without replacement, the estimator of the variance of an inequality indicator $\hat{\theta}$ is defined as

$$\widehat{\operatorname{var}}_{\text{lin}}[\hat{\theta}] = \frac{N(N-n)}{n}\frac{1}{n-1}\sum_{k\in S}(\hat{z}_k - \bar{z})^2, \tag{2.1}$$

where

$$\bar{z} = n^{-1}\sum_{k\in S}\hat{z}_k.$$

Below, we review the empirical definitions of the inequality indicators considered with respect to population income measurement, as well as the expressions for the linearized variables of the indicators as we have implemented them.

2.1 Gini coefficient

The Gini coefficient, G, ranges from 0 (complete equality, that is, all individuals earn the same amount) to 1 (complete inequality, that is, one individual has all the income and the other individuals have no income). The coefficient G is expressed on the basis of the cumulative income of a given proportion of the poorest individuals. If $\mathcal{Y}$ is a random variable representing income, $f(y)$ its density function and $F(y)$ its distribution function, then the *Lorenz curve* (Lorenz 1905) can be written as

$$L(\alpha) = \frac{\int_0^{F^{-1}(\alpha)} yf(y)\,dy}{\int_0^{\infty} yf(y)\,dy} = \frac{1}{\operatorname{E}(\mathcal{Y})}\int_0^{\alpha} F^{-1}(u)\,du.$$

The Gini coefficient represents twice the area between the Lorenz curve and the line of complete equality (the diagonal line $f_{eg}(x) = x$), as shown in Figure 2.1. Therefore, the Gini coefficient can be defined as

$$G = 2\int_0^1 [\alpha - L(\alpha)]\,d\alpha.$$

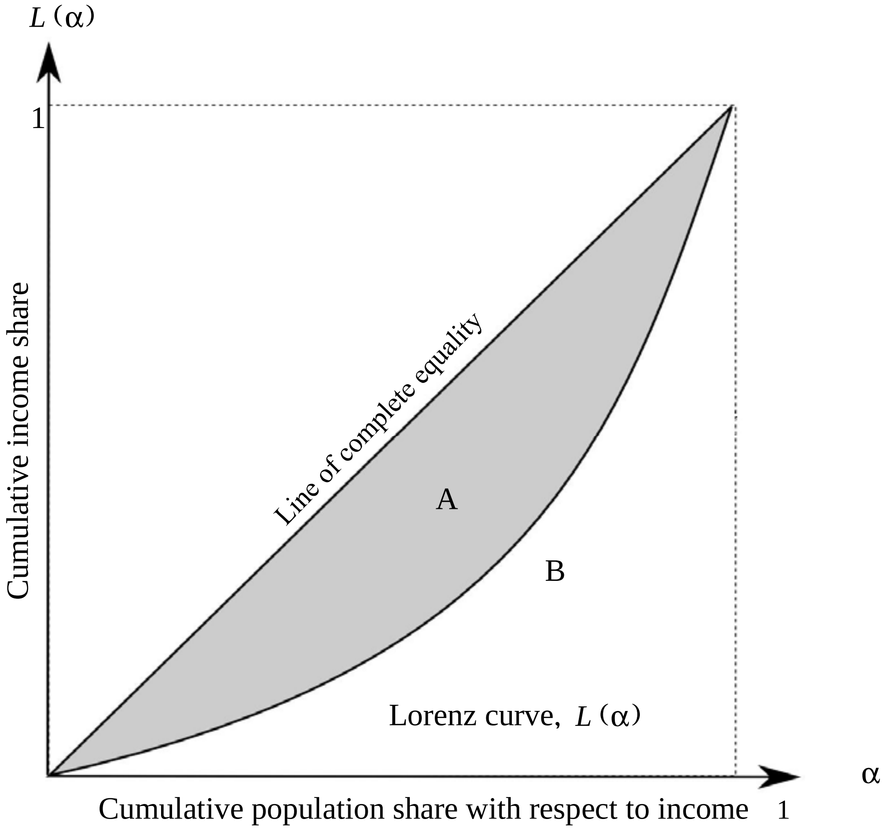

$L(\alpha)$
1
Cumulative income share
Line of complete equality
A
B
Lorenz curve, $L(\alpha)$
α
Cumulative population share with respect to income
1

Figure 2.1 Gini coefficient, G, and Lorenz curve, $L(\alpha)$. $G = 2A$, $A + B = 1/2$

If a population U is finite, then the values of y_k will not be random and the Gini coefficient can be calculated as

$$G = \frac{2\sum_{k\in U} ky_k}{N\sum_{k\in U} y_k} - \frac{N+1}{N},$$

where the values of y_k are sorted by rank. For a sample, the Gini coefficient can be estimated as

$$\begin{aligned}\hat{G} &= \frac{2}{\hat{N}\hat{Y}}\sum_{k\in S} w_k \hat{N}_k y_k - \left(1 + \frac{1}{\hat{N}\hat{Y}}\sum_{k\in S} w_k^2 y_k\right) \\ &= \frac{\sum_{k\in S}\sum_{\ell\in S} w_k w_\ell \left|y_k - y_\ell\right|}{2\hat{N}\hat{Y}},\end{aligned}$$

where $\hat{N}_k = \sum_{\ell\in S} w_\ell \mathbf{1}_{[y_\ell \le y_k]}$ is the sum of the weights w_k, $\hat{Y} = \sum_{k\in S} w_k y_k$ is the estimated total income of the population, and $\hat{N} = \sum_{k\in S} w_k$ is the estimated size of the population. The expression can be simplified as follows if all the weights are equal to N/n :

$$\hat{G} = \frac{2\sum_{k\in S} ky_k}{n\sum_{k\in S} y_k} - \frac{n+1}{n}.$$

Note that the definition may vary by a factor of $n/(n-1)$ depending on the author (Osier 2009 and Eurostat 2004b); however, this subtlety becomes negligible if the sample is large enough.

Langel and Tillé (2012) combine the various approaches to obtain the same estimated linearized variable of the Gini coefficient for the sample:

$$\hat{z}_k^{\text{GINI}} = \frac{1}{\hat{N}\hat{Y}}\left[2\hat{N}_k(y_k - \hat{\bar{Y}}_k) + \hat{Y} - \hat{N}y_k - \hat{G}\left(\hat{Y} + y_k\hat{N}\right)\right],$$

where $\hat{\bar{Y}}_k = \sum_{\ell=1}^{k} w_\ell y_\ell \big/ \hat{N}_k$, and the values of y_ℓ are sorted and distinct.

2.2 Quintile Share Ratio (QSR or S_{80}/S_{20})

A good overview of this indicator is provided by Langel and Tillé (2012). Let q_{80} and q_{20} be the 80th and 20th percentiles of the distribution function $\text{F}(y)$. The QSR is the ratio of the total income of the 20% of the population with the highest income to the total income of the 20% of the population with the lowest income. In the continuous case, the QSR can be defined as

$$\text{QSR} = \frac{\text{E}(\mathcal{Y}\,|\,\mathcal{Y} > q_{80})}{\text{E}(\mathcal{Y}\,|\,\mathcal{Y} < q_{20})} = \frac{1 - L(0.8)}{L(0.2)},$$

where $\mathcal{Y}$ is a random variable representing income. For finite populations, the QSR can be expressed and estimated for a sample on the basis of partial sums,

$$\widehat{\text{QSR}} = \frac{\hat{Y} - \hat{Y}_{0.8}}{\hat{Y}_{0.2}},$$

where, given the results obtained by Langel and Tillé (2011), we will use the following definition of the partial sum, which differs slightly from the official definition of Eurostat (2004a):

$$\hat{Y}_\alpha = \sum_{k\in S} w_k y_k H\left(\frac{\alpha\hat{N} - \hat{N}_{k-1}}{w_k}\right), \tag{2.2}$$

with

$$H(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \le x < 1 \\ 1 & \text{if } x \ge 1. \end{cases}$$

To obtain the linearized variable of the QSR, we must first calculate the linearized variable of the partial sum (2.2), which is

$$I(Y_\alpha)_k = y_k H(\alpha N - k + 1) + \left[\alpha - \mathbf{1}_{[y_k < Q_\alpha]}\right]Q_\alpha,$$

where $Q_\alpha = y_i$, with $\hat{N}_{i-1} < \alpha\hat{N} \le \hat{N}_i$, corresponds to the first definition of the quantile of a finite population in the article by Hyndman and Fan (1996). Osier (2009) obtains a linearized variable that is dependent on the density of the variable $\mathcal{Y}$. However, Langel and Tillé (2011) have shown that the problem of estimating this density for the QSR can be avoided through a simplification, so that it is not necessary to make a kernel approximation of income density as proposed by Osier (2009).

The influence function of the QSR is dependent on the influence functions of the partial sums:

$$I(\text{QSR})_k = z_k^{\text{QSR}} = \frac{y_k - I(Y_{0.8})}{Y_{0.2}} - \frac{(Y - Y_{0.8})\, I(Y_{0.2})}{Y_{0.2}^2}.$$

By making the necessary substitutions, we can see that the estimated linearized variable for a sample is

$$\hat{z}_k^{\text{QSR}} = \frac{y_k - \left\{ y_k H\left(\frac{0.8\hat{N} - \hat{N}_{k-1}}{w_k}\right) + \hat{Q}_{0.8}\left[0.8 - \mathbf{1}_{[y_k < \hat{Q}_{0.8}]}\right]\right\}}{\hat{Y}_{0.2}} - \frac{\left(\hat{Y} - \hat{Y}_{0.8}\right)\left\{ y_k H\left(\frac{0.2\hat{N} - \hat{N}_{k-1}}{w_k}\right) + \hat{Q}_{0.2}\left[0.2 - \mathbf{1}_{[y_k < \hat{Q}_{0.2}]}\right]\right\}}{\hat{Y}_{0.2}^2}. \tag{2.3}$$

2.3 Linearized variable of a quantile

Before we discuss poverty indicators, we should give a few details on the linearized variable of an α-order quantile, which can be expressed as

$$\hat{z}_k^{Q_\alpha} = -\frac{1}{f(\hat{Q}_\alpha)} \frac{1}{\hat{N}} \left[\mathbf{1}_{[y_k \le \hat{Q}_\alpha]} - \alpha\right],$$

where the weighted quantile can be defined in a manner similar to the partial sum (2.2), and $f(\cdot)$ is an income density function that will be discussed in details in Section 3. Note that Eurostat (2004a) recommends the second definition by Hyndman and Fan (1996). We could dispute the Eurostat definition and use another quantile definition, for example $Q_\alpha = y_{k-1} + (y_k - y_{k-1})[\alpha N - (k-1)]$, where $\alpha N < k \le \alpha N + 1$, which is the fourth definition according to Hyndman and Fan (1996). We then estimate the quantile for a sample as follows:

$$\hat{Q}_\alpha = y_{k-1} + (y_k - y_{k-1})\left(\frac{\alpha\hat{N} - \hat{N}_{k-1}}{w_k}\right).$$

The linearized variable of a quantile is dependent on the value of the income density function in that quantile. However, the actual income density is unknown and therefore must also be estimated using the sample. Deville (2000) and Osier (2009) suggest the use of Gaussian kernel estimation. We will discuss the problem of estimating f in more details in Section 3.

In addition to the problem of estimating the income density function, Croux (1998) shows that the empirical influence function of the median income is not a consistent estimator of the corresponding theoretical influence function. For a positive variable (such as income), the empirical influence function of the median income (the case discussed in Croux's article) converges toward an exponential distribution, the expectation of which is the influence function. It is not robust to large proportions of extreme values. It can be said to lack robustness in that the value of the estimator for a sample can differ greatly from the actual value for the population as a result of outliers (that is, values that are relatively very large) in the sample (see Hampel (1974) for a basic idea of robustness for infinite populations, and Beaumont, Haziza and Ruiz-Gazen (2013) for recent thoughts on this topic for finite population sampling).

2.4 Median income and at-risk-of-poverty threshold

Let $\hat{m} = \hat{Q}_{0.5}$ be the estimated median income of the sample. The At Risk of Poverty Threshold (ARPT) is defined as 60% of the median income:

$$\begin{aligned} \mathrm{ARPT} &= 0.6F^{-1}(0.5) \\ \widehat{\mathrm{ARPT}} &= 0.6\hat{Q}_{0.5} = 0.6\hat{m}. \end{aligned}$$

This is an absolute measure that is scale-dependent. The linearized variable of the ARPT is proportional to that of the median income:

$$\hat{z}_k^{\mathrm{ARPT}} = I(\mathrm{ARPT})_k = 0.6I(\mathrm{MED})_k = -\frac{0.6}{f(\hat{m})}\frac{1}{\hat{N}}\left[\mathbf{1}_{[y_k \le \hat{m}]} - 0.5\right].$$

2.5 At Risk of Poverty Rate

The At Risk of Poverty Rate (ARPR), where $\mathrm{ARPR} \in [0,1]$, is the share of the population with an income below the ARPT: $\mathrm{ARPR} = F(\mathrm{ARPT})$. The ARPR is scale-independent, like the Gini coefficient, QSR and relative median poverty gap (see Section 2.7). The official Eurostat definition (Eurostat 2004a) of the estimated ARPR for a sample is

$$\widehat{\mathrm{ARPR}} = \frac{\sum_{y_k < \widehat{\mathrm{ARPT}}} w_k}{\hat{N}}.$$

The linearized variable of the ARPR is defined by Osier (2009) as

$$\begin{aligned} \hat{z}_k^{\mathrm{ARPR}} &= \frac{1}{\hat{N}}\left(\mathbf{1}_{[y_k \le \widehat{\mathrm{ARPT}}]} - \widehat{\mathrm{ARPR}}\right) - \frac{f\left(\widehat{\mathrm{ARPT}}\right)}{f(\hat{m})}\frac{0.6}{\hat{N}}\left(\mathbf{1}_{[y_k \le \hat{m}]} - 0.5\right) \\ &= \frac{1}{\hat{N}}\left(\mathbf{1}_{\left[y_k \le \widehat{\mathrm{ARPT}}\right]} - \widehat{\mathrm{ARPR}}\right) + f\left(\widehat{\mathrm{ARPT}}\right)\hat{z}_k^{\mathrm{ARPT}}. \end{aligned}$$

Here, the income density function must be estimated at two points, namely the median income and the ARPT.

2.6 Median income of individuals below the ARPT

The median income of individuals below the ARPT is $m_p = F^{-1}(1/2\,F(\text{ARPT}))$. It is estimated in the same way as any other quantile, the exact definition of which may vary. The linearized variable of m_p (Osier 2009) is dependent on that of the ARPR:

$$\hat{z}_k^{m_p} = \frac{1}{f(\hat{m}_p)} \frac{\hat{z}_k^{\text{ARPR}}}{2} - \frac{1}{\hat{N}}\left(\mathbf{1}_{[y_k \le \hat{m}_p]} - F(\hat{m}_p)\right).$$

The estimated income density therefore appears three times, namely in the median income and ARPT for $\hat{z}_k^{\text{ARPR}}$, and in the median income of individuals below the ARPT, m_p.

2.7 Relative Median Poverty Gap

The relative median poverty gap (RMPG) is the relative difference between the ARPT and the median income of individuals below the ARPT. $\text{RMPG} = 0$ if the income of all “poor” individuals is equal to the ARPT, and $\text{RMPG} = 1$ if the income of all these individuals is zero. The RMPG is a measure of the extent to which the “poor” individuals are poor:

$$\text{RMPG} = \frac{\text{ARPT} - m_p}{\text{ARPT}}.$$

The estimated RMPG for a sample has already been described. The influence of each observation on the RMPG is defined by Osier (2009):

$$\hat{z}_k^{\text{RMPG}} = \frac{\hat{m}_p \hat{z}_k^{\text{ARPT}} - \widehat{\text{ARPT}}\,\hat{z}_k^{m_p}}{\widehat{\text{ARPT}}^2}.$$

The estimated income distribution density appears four times: once in the calculation of $\hat{z}_k^{\text{ARPT}}$ and three times in the calculation of $\hat{z}_k^{m_p}$.

3 Estimating the income density function

In a design-based approach with a finite population, inference is made in relation to the sampling design $\text{P}(S)$ used to select a sample S from a finite population U of size N. In this approach, only the sample inclusion indicators are random; all other quantities are fixed. The population income distribution function is then a step function: $F_y(x) = \sum_{k \in U} \mathbf{1}_{y_k \le x} / N$, and its derivative, the density function, does not exist due to discontinuities. If a model-based approach with a super-population model to justify the income density function term is not desired, then the distribution function must be artificially smoothed to make it differentiable. Therefore, our use of “density function” is not quite correct. For purposes of smoothing, Deville (2000) and Osier (2009) suggest using Gaussian kernel estimation to estimate the income density function:

$$\begin{aligned} K(u) &= \frac{1}{h\sqrt{2\pi}} e^{-u^2/2}, \qquad u = \frac{x - y_k}{h} \\ \hat{f}_1(x) &= \frac{1}{\hat{N}} \sum_{k\in S} w_k K\left(\frac{x - y_k}{h}\right) \\ &= \frac{1}{h\sqrt{2\pi}} \frac{1}{\hat{N}} \sum_{k\in S} w_k \exp\left[-\frac{(x - y_k)^2}{2h^2}\right] \end{aligned} \tag{3.1}$$

where h is the bandwidth that Osier estimates using $\hat{h} = \hat{\sigma}\hat{N}^{-0.2}$ and $\hat{\sigma}$ is the estimated standard deviation of the empirical income distribution:

$$\hat{\sigma} = \sqrt{\frac{\sum_{k\in S} w_k y_k^2}{\hat{N}} - \left(\frac{\sum_{k\in S} w_k y_k}{\hat{N}}\right)^2} = \sqrt{\frac{\sum_{k\in S} w_k y_k^2}{\hat{N}} - \bar{y}_w^2}.$$

Note that this estimate of σ is not robust, since it is very sensitive to the extreme values of y. Income data often have a distribution tail extending to the right with values that may be extremely high; these are "representative outliers" as defined by Chambers (1986) and Hulliger (1999). As the simulations in Section 4 will show, this can generate a strong bias in the variance estimates. Verma and Betti (2011) also use kernel estimation, recalling that Silverman (1986) states that the choice of kernel is not critical to ensure that $\hat{f}(y)$ converges toward $f(y)$, but the choice of bandwidth is. They use a value recommended by Silverman for distributions with a positive skewness coefficient, $h = 0.79(\hat{Q}_{75} - \hat{Q}_{25})\hat{N}^{-0.2}$. In their findings, they point out that the linearization method may be problematic because of irregularities in the empirical density function. They also state that these problems are all the more cause for concern because survey data often contain groups of observations with the same value (due to rounding or range questions), which can make estimating the density more complicated. The rest of this article describes the solutions we are proposing to reduce bias in variance estimates.

3.1 Using the logarithm

One solution that produces very good results, as shown below, is simply to use the logarithm to estimate the density of x. If $v = \log(x + a)$, where x is the income and a is a positive real number equals to, say, $(|\min_k(y_k)| + 1)$ where there may be negative or zero incomes (ignoring that a would be estimated), then

$$F_v(v) = \mathrm{P}(\mathcal{V} \le v) = \mathrm{P}(\log(\mathcal{Y} + a) \le v) = \mathrm{P}(\mathcal{Y} \le e^v - a) = F_y(e^v - a),$$

where $\mathcal{V}$ and $\mathcal{Y}$ would be random variables. Therefore,

$$f_v(v) = \frac{dF_v(v)}{dv} = \frac{dF_y(e^v - a)}{dv} = f_y(e^v - a)e^v.$$

That is, $f_v(v) = f_y(x)(x + a)$, which gives us the following estimator for the density of x:

$$\hat{f}_2(x) = \frac{\hat{f}_v(v)}{x + a} = \frac{\hat{f}_y(\log(x + a))}{x + a}. \tag{3.2}$$

The estimated density at x for $\mathcal{Y}$ can therefore be determined by estimating the density of the logarithm of the variable divided by the value of the variable at a given point. This property is valid for finite populations. Using the logarithm has the advantage of reducing the leveraging effect of large income values in the kernel density approximation calculation. Simulations show that this simple method significantly reduces bias.

3.2 Nearest neighbour with minimum bandwidth

Deville (2000) outlines another density estimation method that is a "nearest neighbour" method (see Silverman 1986) using the kernel

$$K_D(u) = \begin{cases} \dfrac{1}{b - a} & \text{if } a \le u < b \\ 0 & \text{otherwise,} \end{cases},$$

where $u = y_k$ and the choice of a and b, with $x \in [a, b]$, is to be determined and could depend on x. The distance $(b - a)$ represents the bandwidth h. The density estimate would therefore be

$$\begin{aligned} \hat{f}_D(x, a, b) &= \frac{1}{\hat{N}} \sum_{k \in S} K_D(y_k) \\ &= \frac{1}{\hat{N}} \sum_{k \in S} w_k \frac{1}{b - a} \mathbf{1}_{y_k \in [a, b[} \\ &= \frac{\hat{F}_y(b) - \hat{F}_y(a)}{b - a}, x \in [a, b[\end{aligned} \tag{3.3}$$

where $\hat{F}_y(x) = \sum_{k \in S} w_k \mathbf{1}_{y_k \le x} \Big/ \hat{N}$.

Note that the density estimate (3.3) is not a continuous function and would not be suitable for estimating density values at the end tails of the distribution. Since our work relies little on distribution tails, we shall consider this approach as an option.

Our second proposal for estimating the density of x is based on the idea above. It is a nearest neighbour method, but also imposes a minimum bandwidth. Specifically, our method requires the use of at least p observations nearest to point x with minimum bandwidth $h(p) \ge h_{\text{opt}}$ where

$$h_{\text{opt}} = \frac{0.9 \min(\hat{\sigma}, \hat{Q}_{75} - \hat{Q}_{25})}{1.34 \sqrt[5]{\hat{N}}}$$

is the rule of thumb (Silverman 1986) for determining the bandwidth. This is also the default bandwidth value in the R function `density`. This solution is more robust than (3.1) and avoids the problems that arise when a number of values y_k are very close to each other, which is often the case because the individuals interviewed tend to round their income.

As the values $y_k, k = 1, \ldots, n,$ are assumed to be ordered by rank, the width $h(p)$ of the window around x is initially determined by the p nearest observations, where $p \ll n.$ In the simulations discussed in the next section, after various trials, p was initially set at 30. The density of x is imputed to be the estimated density at the nearest observed point y_j that is less than or equal to $x,$ that is, $j = \max(k \mid y_k \leq x), k = 1, \ldots, n.$ The bandwidth at x in fact depends on the p_j nearest observations around $y_j,$ with $p_j \geq p,$ which will be denoted $h(p_j)$ in the rest of this article. The density is therefore estimated only at observed points, with no smoothing or interpolation between the values $\hat{f}(y_j).$ The algorithm for estimating $\hat{f}(y_j)$ is as follows (see also Figure 3.1):

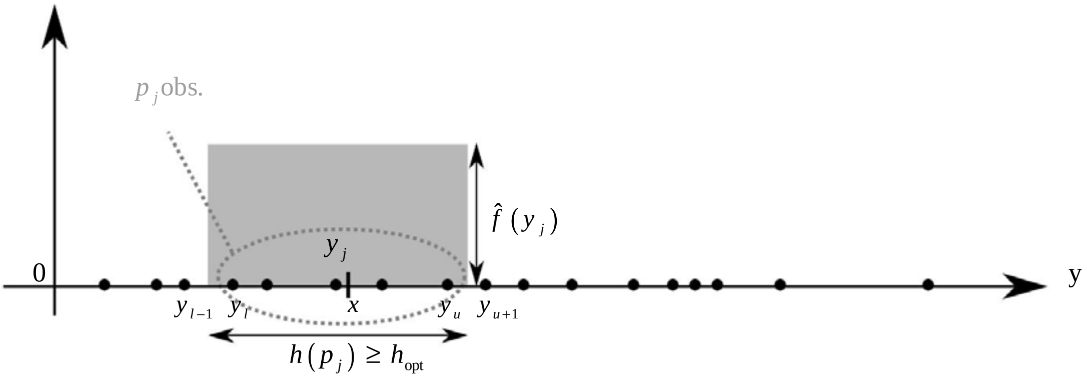


Figure 3.1 Window width $h(p_j)$

1. The initial width of the window around point $y_j,$ where $p_j = p,$ is defined as

$$h(p_j) = \frac{y_u + y_{u+1}}{2} - \frac{y_\ell + y_{\ell-1}}{2}; \; u = \begin{cases} j + p_j/2 - 1 & \text{if } p_j \text{ is even} \\ j + \lfloor p_j/2 \rfloor & \text{if } p_j \text{ is odd} \end{cases}$$

$$\ell = j - \lfloor p_j/2 \rfloor.$$

2. If the width of the resulting window $h(p_j)$ is less than h_{opt}, increment the two bounds:

   upper bound: $u \rightarrow u + 1,$ as long as $u < n,$
   lower bound: $l \rightarrow l - 1,$ as long as $l > 1,$
   which implies that $p_j \rightarrow p_j + 2,$ unless $u = n$ or $l = 1,$ in which case there is no longer the same number of points on each side of $y_j.$

3. Repeat step 2 until $h(p_j) \geq h_{\text{opt}}.$

4. The estimated density at x can then be written as

$$\hat{f}(x) = \hat{f}(y_j) = \begin{cases} \dfrac{p_j}{nh(p_j)} & \text{without weighting,} \\ \dfrac{\sum_{p_j \ closest \ to \, y_j} w_j^{\text{std}}}{nh(p_j)} & \text{with weighting,} \end{cases}$$

with standardized weights $w_k^{\text{std}} = w_k / \bar{w},\ k = 1, \ldots, n.$

The number of observations p_j used in the calculation may vary, and it depends on the local curvature of the empirical distribution function. The condition $h(p_j) \geq h_{opt}$ guarantees a minimum window width in places where numerous observations would be concentrated over a small interval. The procedure is made even more solid by combining this approach with the preceding approach, that is, by estimating the density of the logarithm of the variable divided by its (non-logarithmic) value:

$$\hat{f}_3(x) = \frac{\hat{f}(\log(x + a))}{x + a}. \tag{3.4}$$

3.3 Robustness of the linearized variable

As stated above, for the median or for other quantiles, Croux (1998) points out that the empirical influence function or the linearized variable estimated using the sample is not as robust as it appears to be, even if the density function is known. We confirmed this for the EU-SILC data used in the model simulations with a Generalized Beta distribution of the second kind (GB2) by means of the R function `profml.gb2` (Graf and Nedyalkova 2011). For small samples $(n \leq 100)$, the potential bias of the linearized variable resulting from too many outliers may also bias the variance estimate calculated using the linearized variable. For larger samples $(n \geq 1{,}000)$, a maximum relative bias in the variance estimated using the empirical *versus* theoretical linearized variable may reach 5%. However, it is below the percentage in absolute terms three times out of four.

4 Results

Simulations were conducted on three sets of real data to compare and assess the different density function estimation methods, $\hat{f}_1(x)$, see (3.1), $\hat{f}_2(x)$, see (3.2) and $\hat{f}_3(x)$, see (3.4). These methods are required to estimate the variance of certain poverty and inequality indicators.

1. The first dataset contains equivalent household incomes from the EU-SILC survey conducted by the Swiss Federal Statistical Office in 2009. It includes 17,534 individuals with a non-zero income.
2. The second dataset also comes from the 2009 EU-SILC survey, but is limited to salaried individuals. It contains salaries from the register of the Central Compensation Office that has

been linked with the survey respondents. We therefore have no non-response issues, and there are 7,922 individuals with a non-zero income.

3. The third test file, named *Ilocos*, comes with the R package `ineq` (Zeileis 2012). It contains 632 observations, which are household incomes in Ilocos, one of the 16 regions of the Philippines. The data come from two surveys by the National Statistics Office of the Philippines, in 1997 and in 1998.

The three datasets have a positive skewness coefficient, which is typical of income distributions. Each data set is considered to be one population, and we initially selected 10,000 simple random samples without replacement of various sizes. The values of the various indicators were calculated for each sample, giving us a Monte Carlo estimate of their variance, $\text{var}_{\text{sim}}(\hat{\theta})$, for a poverty or inequality indicator θ. The variance estimator using linearization is denoted $\widehat{\text{var}}_{\text{lin}}(\hat{\theta})$ and is calculated using the linearization variable $\hat{z}^{\hat{\theta}}$ estimated for each sample:

$$\widehat{\text{var}}_{\text{lin}}(\hat{\theta}) = \frac{N(N-n)}{n}\text{var}\left(\hat{z}_S^{\hat{\theta}}\right),$$

where n is the size of the sample used for the simulations and

$$\text{var}\left(\hat{z}_S^{\hat{\theta}}\right) = \frac{1}{n-1}\sum_{k\in S}\left(\hat{z}_{S,k}^{\hat{\theta}} - \overline{z}_S^{\hat{\theta}}\right)$$

where $\overline{z}_S^{\hat{\theta}} = n^{-1}\sum_S \hat{z}_{S,k}^{\hat{\theta}}$, see (2.1).

The quality of the variance estimator using linearization is assessed by comparing the expected Monte Carlo value of the variance estimated using linearization, denoted $E_{\text{sim}}\left[\widehat{\text{var}}_{\text{lin}}(\hat{\theta})\right]$, with the “true” Monte Carlo variance $\text{var}_{\text{sim}}(\hat{\theta})$ in terms of relative bias:

$$\text{RB}\left[\widehat{\text{var}}_{\text{lin}}(\hat{\theta})\right] = \frac{E_{\text{sim}}\left[\widehat{\text{var}}_{\text{lin}}(\hat{\theta})\right] - \text{var}_{\text{sim}}(\hat{\theta})}{\text{var}_{\text{sim}}(\hat{\theta})}. \tag{4.1}$$

For the second data set (EU-SILC 2009, income of salaried individuals) we also, in a second step, selected 10,000 random samples without replacement under a stratified sampling design, and then calibrated the sampling weights to agree with the eight known sociodemographic marginal totals for the population of 7,922 individuals. The five strata used correspond to the age groups of the salaried individuals (see Table 4.1).

The eight calibration cells were obtained by crossing the three following dichotomous variables (auxiliary calibration variables):

1. MARIÉ, which indicates whether or not the individual is married;
2. CHEF, which indicates whether or not the individual’s job is a management position; and
3. HOMME, which indicates the individual’s sex.

The totals for the population of 7,922 individuals for these calibration cells are shown in Table 4.2.

Table 4.1
Strata used in simulations with 2009 EU-SILC data and three sample sizes (income of salaried individuals, $N = 7,922$)

Stratum h	Description	N_h	%	n_h		
1	individuals under 25	1,187	15.0	75	112	150
2	26- to 35-year-olds	1,359	17.2	86	129	171
3	36- to 45-year-olds	2,137	27.0	135	202	270
4	46- to 55-year-olds	1,864	23.5	117	177	235
5	individuals over 55	1,375	17.4	87	130	174
	TOTAL	7,922	100.0	500	750	1,000

Table 4.2
Calibration margins in simulations with 2009 EU-SILC data (income of salaried individuals, $N = 7,922$)

Margin	MARIÉ	CHEF	HOMME	Population total	%
1	0	0	0	1,487	18.8
2	0	0	1	1,208	15.2
3	0	1	0	323	4.1
4	0	1	1	457	5.8
5	1	0	0	1,759	22.2
6	1	0	1	1,278	16.1
7	1	1	0	328	4.1
8	1	1	1	1,082	13.7
			TOTAL	7,922	100.0

For each stratified sample, a calibration (linear method) was performed to make the sums of the weights agree with the eight margins shown above. Point estimates of the indicators and their linearized variable were computed for each sample using the calibrated weights.

Variance was estimated using the method developed by Deville (2000), which consists of linearizing also with respect to the calibration by calculating the residuals $e^{\hat{\theta}}$ of the regression (weighted by the sampling weights) of the linearized variables of the indicators for the auxiliary calibration variables. The variance of the total of the residuals thus calculated, under a stratified random sampling plan without replacement is therefore an estimator of the variance of the estimated indicator; it is the quantity of interest:

$$\widehat{\text{var}}_{\text{lin}}(\hat{\theta}) = \sum_{h=1}^{H} \frac{N_h}{n_h} (N_h - n_h) s^2_{e_h^{\hat{\theta}}} \tag{4.2}$$

where

$$s^2_{e_h^{\hat{\theta}}} = \frac{1}{n_h - 1} \sum_{k \in S_h} \left(e_k^{\hat{\theta}} - \bar{e}^{\hat{\theta}}\right)^2$$

The quality of the variance estimator using linearization is assessed analogously to the procedure for simple random sampling, see (4.1).

Tables 4.3, 4.4 and 4.5 show the relative bias of the variance for the three data sets used and described above, using simple random sampling. Table 4.6 shows the relative bias of the variance using stratified random sampling with calibrated weights. The upper portions of the tables give the values for the Gini coefficient and QSR, which do not require estimating the income density function. The estimation of their variance works well. Note that there is a problem involving the underestimation of the variance of the Gini coefficient in the case of stratification with calibration (Table 4.6).

For the first data set, Table 4.3 does not reveal any major differences except that the estimation of income density using $\hat{f}_3(x)$ gives results that are more conservative. In fact, the relative bias remains of the same order of magnitude, but positive, while it is negative for the other two methods of estimating density. For the second data set, Table 4.4 shows that it is essential to use the logarithm or the nearest neighbour method with minimum bandwidth. The latter, all relative bias falls under 10% when the sample sizes are sufficiently large (see last column in the table). Simulations on the same data with a stratified sampling plan and calibration strengthen and confirm these results (see Table 4.6). For the third data set, Table 4.5 shows the same trends, although the results are less stable as a result of the small sample and population sizes. This is not surprising, since the minimum number of neighbours to consider is fixed at 30. In this case, for the Ilocos data set, simulations with a smaller value of p fixed at 10 makes no difference ultimately, because the condition $h(p_j) \geq h_{\text{opt}}$ automatically increases it above 30.

Furthermore, generally speaking, we can see that the greater the use of Gaussian kernel density estimation - $\hat{f}_1(x)$ - the greater the error. In fact, the relative bias of the variance for the median income of individuals below the ARPT and for the RMPG are almost systematically greater in absolute value that those for the other indicators. For the RMPG, the error may be offset (as in Table 4.3) if there are enough observations, since the density estimation appears in both the numerator and the denominator.

Table 4.3
Relative bias (4.1) of the variance obtained with 10,000 simple random samples without replacement from the 2009 EU-SILC data (equivalent household income, $N = 17,534$)

	Sample size (sampling rate)								
Indicator	$n = 500\,(2.9\ \%)$			$n = 750\,(4.3\ \%)$			$n = 1,000\,(5.7\ \%)$		
GINI	-0.02			-0.02			-0.02		
QSR	0.01			0.00			0.00		
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$
ARPT	-0.08	-0.06	0.04	-0.09	-0.07	0.03	-0.09	-0.07	0.04
ARPR	-0.05	-0.01	-0.00	-0.09	-0.06	-0.05	-0.08	-0.05	-0.03
RMPG	-0.09	-0.07	*0.15*	*-0.10*	-0.07	*0.12*	-0.09	-0.06	*0.14*
MEDP	*-0.16*	*-0.12*	0.09	*-0.19*	*-0.13*	0.05	*-0.18*	*-0.11*	0.07
MED	-0.08	-0.06	0.05	-0.08	-0.06	0.04	-0.08	-0.06	0.04

Table 4.4
Relative bias (4.1) of the variance obtained with 10,000 simple random samples without replacement from the 2009 EU-SILC data (income of salaried individuals, $N = 7,922$)

	Sample size (sampling rate)								
Indicator	$n = 500\,(6.3\ \%)$			$n = 750\,(9.5\ \%)$			$n = 1,000\,(12.6\ \%)$		
GINI	-0.03			-0.03			-0.02		
QSR	-0.00			0.00			0.00		
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$
ARPT	0.07	0.05	*0.13*	0.06	0.04	*0.10*	0.06	0.03	0.08
ARPR	-0.05	-0.04	-0.02	-0.05	-0.04	-0.01	-0.06	-0.05	-0.02
RMPG	*0.61*	*0.12*	*0.15*	*0.60*	*0.11*	0.08	*0.59*	0.09	0.05
MEDP	*0.73*	*0.17*	*0.18*	*0.72*	*0.16*	*0.10*	*0.72*	*0.15*	0.07
MED	0.07	0.04	*0.13*	0.06	0.04	*0.10*	0.05	0.03	0.07

Table 4.5
Relative bias (4.1) of the variance obtained with 10,000 simple random samples without replacement from Ilocos data (household income, $N = 632$)

Indicator	Sample size (sampling rate)					
	$n = 50\,(7.9\ \%)$			$n = 63\,(10.0\ \%)$		
GINI	*-0.16*			*-0.13*		
QSR	0.00			0.00		
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$
ARPT	-0.05	-0.06	-0.01	-0.03	-0.03	-0.01
ARPR	*-0.31*	-0.01	*-0.12*	*-0.33*	-0.03	*-0.18*
RMPG	*1.55*	*0.83*	*0.26*	*1.54*	*0.16*	*0.39*
MEDP	*1.02*	*0.28*	*-0.26*	*1.05*	0.07	*-0.11*
MED	0.04	0.03	0.08	0.07	0.07	0.09

Table 4.6
Relative bias (4.1) of the variance obtained with 10,000 stratified random samples without replacement, with weights calibrated to eight sociodemographic margins, from the 2009 EU-SILC data (income of salaried individuals, $N = 7,922$)

Indicator	Sample size (sampling rate)								
	$n = 500\,(6.3\ \%)$			$n = 750\,(9.5\ \%)$			$n = 1,000\,(12.6\ \%)$		
GINI	*-0.21*			*-0.20*			*-0.20*		
QSR	-0.06			-0.06			-0.07		
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$
ARPT	-0.07	-0.09	-0.01	-0.08	*-0.10*	-0.04	-0.09	*-0.11*	-0.06
ARPR	*-0.10*	*-0.10*	-0.08	-0.07	-0.06	-0.05	-0.06	-0.06	-0.05
RMPG	*0.63*	*0.13*	*0.13*	*0.61*	*0.11*	0.08	*0.59*	*0.10*	0.04
MEDP	*0.71*	*0.16*	*0.15*	*0.68*	*0.13*	0.09	*0.66*	*0.12*	0.04
MED	-0.07	-0.09	-0.01	-0.08	*-0.10*	-0.04	-0.08	*-0.11*	-0.06

In short, we see that the variance can be overestimated $\left(\text{RB}\left[\widehat{\text{var}}_{\text{lin}}(\hat{\theta})\right] > 0\right)$ or underestimated $\left(\text{RB}\left[\widehat{\text{var}}_{\text{lin}}(\hat{\theta})\right] < 0\right)$ depending on the indicator and the data set. The use of the logarithm $\left(\hat{f}_2(x)\right)$ provides significant improvement. The nearest neighbour method $\left(\hat{f}_3(x)\right)$ eliminates all problems if there is enough data (as in Tables 4.3, 4.4 and 4.6). Slight problems arise with this method when the samples are small (as in Table 4.5). Illogical variations and bias that persist in the tables may also be the result of a lack of robustness in the linearized variables for certain samples, as stated in Section 3.3.

5 Conclusions

In a number of countries, national sample surveys publish extrapolated values for the Laeken indicators (Eurostat 2005), since they are key indicators that make it possible to direct decision makers with regard to political and social matters. It is therefore critical that we be able to quantify the precision of these measures, which raises the issue of the appropriateness of the precision estimates available. This article shows that a substantial improvement may be made in the precision estimates for poverty and inequality indicators by using a (local) estimate of the income density or given monetary variable.

The simulations conducted show that the Gaussian kernel density estimation method currently implemented in most cases is not recommended without at least using the logarithm as proposed in

Section 3.1; otherwise, there may be significant bias in the estimated variance. The nearest neighbour method (Section 3.2), which also imposes a minimum bandwidth, may yield even better results, especially if there are agglomerations of observations with certain values in the given data. However, this method requires setting a minimum number p of neighbours on the basis of the data used. If few observations are available, the use of the logarithm is preferable instead. In all cases, we hope that this work will help raise awareness of the importance of being meticulous during the implementation of calculations for the linearized variable of any indicator involving quantiles.

6 Acknowledgments

This work was carried out under a collaboration agreement between the Institute of Statistics at the University of Neuchâtel and the Swiss Federal Statistical Office (SFSO). We would also like to specifically thank the Income, Consumption and Living Conditions unit of the SFSO for providing us with data from the Swiss portion of the EU-SILC survey. Thanks also to Matti Langel and Anne Massiani for their support during our investigations.

References

Antal, E., Langel, M. and Tilllé, Y. (2011). Variance estimation of inequality indices in complex sampling designs. *Proceedings 58th World Statistical Congress*, Dublin.

Ardilly, P., and Osier, G. (2007). Cross-sectional variance estimation for the French “Labour Force Survey”. *Survey Research Methods*, 1, 75-83.

Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika*, 100, 555-569.

Chambers, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.

Croux, C. (1998). Limit behaviour of the empirical influence function of the median. *Statistics & Probability Letters*, 37, 331-340.

Demnati, A., and Rao, J.N.K. (2004). Linearization variance estimators for survey data. *Survey Methodology*, 30, 1, 17-27.

Deville, J.-C. (2000). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25, 2, 219-230.

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

Eurostat (2004a). Common cross-sectional eu indicators based on eu-silc; the gender pay gap. Working and study documents of the Office for Official Publications of the European Communities, Luxembourg. EU-SILC 131-rev/04.

Eurostat (2004b). Theoretical study of the gini index. Working and study documents of the Office for Official Publications of the European Communities, Luxembourg. EU-SILC 131-A/04.

Eurostat (2005). The continuity of indicators during the transition between ECHP and EU-SILC. Working and study documents of the Office for Official Publications of the European Communities, Luxembourg.

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.

Graf, M. (2011). Use of survey weights for the analysis of compositional data. In *Compositional Data Analysis*, *Theory and Applications*, (Ed., V. Pawlosky-Glahn and A. Buccianti), Oxford: Wiley, chapter 9, 114-127.

Graf, M. (2013). A simplified approach to linerarization variance for surveys. *University of Neuchâtel*, working document.

Graf, M., and Nedyalkova, D. (2011). *GB2: Generalized Beta Distribution of the Second Kind: properties, likelihood, estimation.* R package version 1.0.

Hampel, F.R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69, 383-393.

Hulliger, B. (1999). Simple and robust estimators for sampling. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 54-63.

Hyndman, R.J., and Fan, Y. (1996). Sample quantiles in statistical packages. *American Statistician*, 50, 361-365.

Kovačević, M.S., and Binder, D.A. (1997). Variance estimation for measures of income inequality and polarization - The estimating equations approach. *Journal of Official Statistics*, 13, 41-58.

Langel, M., and Tillé, Y. (2011). Statistical inference for the quintile share ratio. *Journal of Statistical Planning and Inference*, 141, 2976-2985.

Langel, M., and Tillé, Y. (2012). Variance estimation of the Gini index: Revisiting a result several times published. In *Press in Journal of the Royal Statistical Society - Series A.*

Lorenz, M.O. (1905). Methods of measuring the concentration of wealth. *American Statistical Association*, 9, 209-219.

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods*, 3, 167-195.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, Londres.

Verma, V., and Betti, G. (2011). Taylor linearization sampling errors and design effects for poverty measures and other complex statistics. *Journal of Applied Statistics*, 38, 1549-1576.

Zeileis, A. (2012). *ineq: Measuring Inequality, Concentration, and Poverty*. R package version 0.2-10.

Survey Methodology, June 2014
Vol. 40, No. 1, pp. 81-104
Statistics Canada, Catalogue No. 12-001-X

Theoretical and empirical properties of model assisted decision-based regression estimators

Jun Shao, Eric Slud, Yang Cheng, Sheng Wang, and Carma Hogue[1]

Abstract

In 2009, two major surveys in the Governments Division of the U.S. Census Bureau were redesigned to reduce sample size, save resources, and improve the precision of the estimates (Cheng, Corcoran, Barth and Hogue 2009). The new design divides each of the traditional state by government-type strata with sufficiently many units into two sub-strata according to each governmental unit's total payroll, in order to sample less from the sub-stratum with small size units. The model-assisted approach is adopted in estimating population totals. Regression estimators using auxiliary variables are obtained either within each created sub-stratum or within the original stratum by collapsing two sub-strata. A decision-based method was proposed in Cheng, Slud and Hogue (2010), applying a hypothesis test to decide which regression estimator is used within each original stratum. Consistency and asymptotic normality of these model-assisted estimators are established here, under a design-based or model-assisted asymptotic framework. Our asymptotic results also suggest two types of consistent variance estimators, one obtained by substituting unknown quantities in the asymptotic variances and the other by applying the bootstrap. The performance of all the estimators of totals and of their variance estimators are examined in some empirical studies. The U.S. Annual Survey of Public Employment and Payroll (ASPEP) is used to motivate and illustrate our study.

Key Words: Asymptotic normality; Bootstrap; Decision-based estimator; Probability proportional to size; Stratification; Variance estimation.

1 Introduction

The U.S. Annual Survey of Public Employment and Payroll (ASPEP) provides current estimates for full- and part-time state and local government employment and payroll classified by government functions (such as: elementary and secondary education, higher education, police protection, fire protection, financial administration, judicial and legal, *etc*.). This survey covers state and local government units (89,526 according to the 2007 Census of Governments), which include counties, cities, townships, units called "special districts", and school districts. ASPEP is the only source of public employment data by government function and job category, providing data on numbers of full- and part-time employees and payroll, as well as on hours worked by part-time employees. Data collection usually begins in March and continues for about seven months, with the pay period containing March 12 as reference period.

Let U denote the finite population of N units stratified into H strata, $U_1, \ldots, U_H$, where U_h contains N_h units and $N_1 + \cdots + N_H = N$. The traditional sampling design for the ASPEP is a stratified probability proportional to size (PPS) design, where the strata are constructed using state and the government types, which are county, subcounty (city or town), special district, or school district. The size of each unit is the total payroll, and sampling across strata is independent. In 2009, a modified sampling design was developed, which cuts some strata U_h into two sub-strata, U_{h1} and U_{h2} with N_{h1} and N_{h2} units, respectively, where U_{h1} contains smaller-size units (Cheng *et al.* 2009). The idea was to save

[1] Jun Shao, Statistics Department University of Wisconsin, Madison WI , E-mail: shao@stat.wisc.edu; Eric Slud, Center for Statistical Research and Methodology, US Census Bureau, Washington DC and Mathematics Department, University of Maryland, College Park, MD, E-mail: eric.v.slud@census.gov; Yang Cheng, Demographic Statistical Methods Division, US Census Bureau, Washington DC, E-mail: yang.cheng@census.gov ; Sheng Wang, Mathematica Policy Research, Princeton NJ, E-mail: swang@mathematica-mpr.com; and Carma Hogue, Governments Division, US Census Bureau, Washington DC, E-mail: carma.ray.hogue@census.gov.

resources and reduce respondent burden by selecting a sample from U_{h1} with smaller sample size under the modified than under the traditional design. Let S_{hj} be a PPS sample of size n_{hj} from U_{hj}, $j = 1, 2$, $n_{h1} + n_{h2} = n_h$. Note that n_{h1} may still be larger than n_{h2} because N_{h1} is usually much larger than N_{h2}.

For unit $i \in U$, let y_i be a key survey variable (*e.g.*, the full-time employment, full-time payroll, part-time employment, part-time payroll, part-time hours), x_i be an auxiliary variable, say the same variable as y_i from the most recent census, and let z_i be the covariate used as the size variable in PPS sampling. The covariate values x_i and z_i are observed for all $i \in U$, whereas y_i is observed only for each sampled unit i.

The Horvitz-Thompson estimator of the unknown total $Y = \sum_{i \in U} y_i$ is

$$\hat{Y}_{\text{HT}} = \sum_h \sum_j \sum_{i \in S_{hj}} y_i / \pi_i , \tag{1.1}$$

where π_i is the first-order inclusion probability of unit i in S_{hj}, a known function of z_i's. To utilize the auxiliary variable x_i and increase the accuracy of estimation of Y, the model-assisted approach (Särndal, Swensson and Wretman 1992) is adopted. Applying regression within each S_{hj} leads to the regression estimator of Y as

$$\hat{Y}_{\text{reg},2} = \sum_h \sum_j \left[\frac{N_{hj} \hat{Y}_{hj}}{\hat{N}_{hj}} + \hat{\beta}_{hj} \left(X_{hj} - \frac{N_{hj} \hat{X}_{hj}}{\hat{N}_{hj}} \right) \right], \tag{1.2}$$

where $X_{hj} = \sum_{i \in U_{hj}} x_i$, $\hat{Y}_{hj} = \sum_{i \in S_{hj}} y_i / \pi_i$, $\hat{X}_{hj} = \sum_{i \in S_{hj}} x_i / \pi_i$, $\hat{N}_{hj} = \sum_{i \in S_{hj}} 1/\pi_i$, and

$$\hat{\beta}_{hj} = \frac{\sum_{i \in S_{hj}} \left(x_i - \hat{X}_{hj} / \hat{N}_{hj} \right) y_i / \pi_i}{\sum_{i \in S_{hj}} \left(x_i - \hat{X}_{hj} / \hat{N}_{hj} \right)^2 / \pi_i}.$$

Alternatively, combining the two sub-strata S_{h1} and S_{h2} results in the following regression estimator. (A referee correctly points out that $\hat{Y}_{\text{reg},1}$ in (1.3) is not the pooled estimator one would use if regression lines in stratum h were combined but the two sub-strata were not; however, it *is* the natural estimator when not only regression lines but also sub-strata are combined.)

$$\hat{Y}_{\text{reg},1} = \sum_h \left[\frac{N_h \hat{Y}_h}{\hat{N}_h} + \hat{\beta}_h \left(X_h - \frac{N_h \hat{X}_h}{\hat{N}_h} \right) \right], \tag{1.3}$$

where $\hat{Y}_h = \sum_j \hat{Y}_{hj}$, $\hat{X}_h = \sum_j \hat{X}_{hj}$, $\hat{N}_h = \sum_j \hat{N}_{hj}$, and

$$\hat{\beta}_h = \frac{\sum_j \sum_{i \in S_{hj}} \left(x_i - \hat{X}_h / \hat{N}_h \right) y_i / \pi_i}{\sum_j \sum_{i \in S_{hj}} \left(x_i - \hat{X}_h / \hat{N}_h \right)^2 / \pi_i}.$$

Since both $\hat{Y}_{\text{reg},1}$ and $\hat{Y}_{\text{reg},2}$ are model-assisted estimators, they are consistent with respect to repeated sampling, whether or not the regression model holds. If the least-squares regression lines in two sub-strata

U_{hj}'s are the same, $\hat{Y}_{\text{reg},1}$ may be more efficient than $\hat{Y}_{\text{reg},2}$. On the other hand, if the regression lines are different, $\hat{Y}_{\text{reg},2}$ may be more efficient than $\hat{Y}_{\text{reg},1}$.

A decision-based method was proposed in Cheng *et al.* (2010), which applies hypothesis testing to decide whether we combine S_{h1} and S_{h2}. Within stratum h, the slopes of the regression lines in U_{h1} and U_{h2} are tested for equality. Let

$$\hat{\alpha}_{hj} = \frac{\hat{Y}_{hj} - \hat{\beta}_{hj}\hat{X}_{hj}}{\hat{N}_{hj}}, \quad \hat{\sigma}^2_{xe,hj} = \frac{n_{hj}}{\hat{N}^2_{hj}} \sum_{i \in S_{hj}} \left(x_i - \frac{\hat{X}_{hj}}{\hat{N}_{hj}} \right)^2 \frac{\left(y_i - \hat{\alpha}_{hj} - \hat{\beta}_{hj} x_i \right)^2}{\pi_i^2},$$

$$\hat{\sigma}^2_{xhj} = \sum_{i \in S_{hj}} \frac{\left(x_i - \hat{X}_{hj} / \hat{N}_{hj} \right)^2}{\pi_i \hat{N}_{hj}}, \quad t_h = \sqrt{n_h - 4}\left(\hat{\beta}_{h1} - \hat{\beta}_{h2} \right) \Big/ \sqrt{n_h \sum_{j=1}^{2} \frac{\hat{\sigma}^2_{xe,hj}}{n_{hj}\hat{\sigma}^4_{xhj}}}.$$

If $|t_h| > t_{1-\tau/2, n_h - 4}$, where $t_{1-\tau/2,\nu}$ is the $1 - \tau/2$ quantile of the t-distribution with ν degrees of freedom, then we reject the hypothesis of common slope and use $\hat{\beta}_{hj}$ (and set $\zeta_h = 1$). Here τ is a nominal significance level set by default to 0.05, although we will consider other choices of τ in the simulations section. The test-statistic definition involving $n_h - 4$ degrees of freedom is a slightly artificial choice designed to make the moderate-sample rejection probabilities closer to nominal, but the large-sample asymptotic distribution theory justifying this test is given in part (c) of Theorem 1. If $|t_h| \le t_{1-\tau/2, n_h - 4}$, then we accept the hypothesis of common slope, combine sub-strata S_{h1} and S_{h2}, and use $\hat{\beta}_h$ (setting $\zeta_h = 0$). Tests are performed independently across strata $h = 1, \ldots, H$. The decision-based estimator of Y is then

$$\hat{Y}_{\text{dec}} = \sum_h \sum_j \zeta_h \left[\frac{N_{hj}\hat{Y}_{hj}}{\hat{N}_{hj}} + \hat{\beta}_{hj} \left(X_{hj} - \frac{N_{hj}\hat{X}_{hj}}{\hat{N}_{hj}} \right) \right] + \sum_h (1 - \zeta_h) \left[\frac{N_h \hat{Y}_h}{\hat{N}_h} + \hat{\beta}_h \left(X_h - \frac{N_h \hat{X}_h}{\hat{N}_h} \right) \right]. \quad (1.4)$$

Since two regression lines with a common slope can have different intercepts, one might test a further hypothesis regarding intercepts to decide whether to combine the two sub-strata. However, population points (x_i, y_i) falling on two parallel but not identical substratum regression lines would be discontinuous around the cut-off point between the two sub-strata U_{h1} and U_{h2}, which seems to occur only rarely in practical situations. In ASPEP, for example, Cheng *et al.* (2010) investigated the slopes and intercepts for sub-strata in 2002 and 2007 Census data sets, noting that the hypothesis of a common intercept could never be rejected when the hypothesis of a common slope could not be rejected. Thus, the decision-based estimator in (1.4) depends only on hypothesis testing for equality of sub-stratum regression slopes.

The two-stage estimators studied here are particular instances of procedures previously termed estimators following preliminary testing. There is a large literature on such procedures in surveys, including a bibliography by Bancroft and Han (1977), a book by Saleh (2006), and a treatment by Fuller (2009, Section 6.7). An idea from Saleh (2006) is to estimate coefficients by a convex combination of the estimated coefficients from the separate strata with proportions depending on a test statistic. Such smoothed estimators might be more efficient than our decision-based procedures. If the stratum-specific intercepts and slopes were regarded as random, then a model-based empirical-Bayes approach to survey estimation might also be tried.

The decision-based estimators (1.4) are novel because they are model-assisted design-consistent in the survey-sampling context, making explicit use of the known substratum population sizes. In a somewhat similar spirit, Rao and Ramachandran (1974) previously made an exact comparison of the separate and combined ratio estimators under a ratio model similar to the regression model of this paper.

The purpose of this paper is to show some asymptotic and empirical properties of the estimators of Y described above and their variance estimators. Consistency and asymptotic normality of $\hat{Y}_{\text{reg},1}$, $\hat{Y}_{\text{reg},2}$, and $\hat{Y}_{\text{dec}}$ are established in Section 2, in terms of either design-based or model-assisted asymptotic theory. Although the first-order asymptotics favor $\hat{Y}_{\text{reg},2}$, $\hat{Y}_{\text{reg},1}$ may be better when some substratum sample sizes n_{h2} are moderate, a second-order asymptotic effect. The virtue of the decision-based estimator $\hat{Y}_{\text{dec}}$ is in adapting to be close to the better of $\hat{Y}_{\text{reg},1}$ and $\hat{Y}_{\text{reg},2}$. As the discussion in paragraph (III) of Section 4.4 indicates, simulations show that the benefit of this adaptivity is to reduce MSE up to a few percent under reasonable parameter settings, and by larger amounts in stranger settings.

Variance estimation for the decision-based estimator is treated in Section 3. While the asymptotic theory in Section 2 suggests that consistent variance estimators are obtained by substituting for unknown quantities in the asymptotic variance formulas, we also study bootstrap variance estimators suggested in Cheng *et al.* (2010), which are generally found to have better finite sample performance than the substitution estimators. Empirical results are presented in Section 4, with Section 4.4 providing interpretations and concluding remarks. All technical proofs are given in the Appendix.

2 Consistency and asymptotic normality

To consider asymptotics, we view the population U as one of a sequence of populations $\{U^{(m)}, m = 1, 2, \ldots\}$, where the number of units in $U^{(m)}$ increases to infinity as $m \to \infty$. This paper treats only the case of strata in which a large sample n_h is drawn; that is, we assume that for each stratum h, the sample size n_h depends on m and increases to infinity as $m \to \infty$, but we omit the index m for simplicity. All limiting processes are considered as $m \to \infty$. Following authors such as Isaki and Fuller (1982) and Deville and Särndal (1992), we term this a *superpopulation* asymptotic framework. Under the design-based framework considered in Section 2.1, the attribute vectors in the underlying populations need not be viewed as random vectors. However, under the model-assisted framework considered in Section 2.2, regression models are assumed for attribute vectors.

Since each estimator is a sum of independent estimators constructed within each stratum, for simplicity we present asymptotic results for the case of $H = 1$. The results and conclusions immediately apply to the case of a fixed H and can also be extended to the situation where H increases to infinity. (It is typical for large-scale surveys to have many strata, although the number of ASPEP government-by-type strata that were split into substrata was somewhat less than 100.) Since we only consider $H = 1$, we omit the index h for stratum in this section, *e.g.*, $n_{hj} = n_j$, $n_h = n$, $N_{hj} = N_j$, and $N_h = N$. Also, for

$j = 1, 2,$ the estimators $\hat{\beta}_j$ and $\hat{\beta}$ are defined by the displayed formulas following equations (1.2) and (1.3), with subscript h suppressed, together with

$$\hat{\mu}_{xj} = \hat{X}_j \big/ \hat{N}_j, \quad \hat{\alpha}_j = \hat{Y}_j \big/ \hat{N}_j - \hat{\beta}_j \hat{\mu}_{xj}, \quad \hat{\sigma}^2_{xj} = \hat{N}_j^{-1} \sum_{i \in S_j} \pi_i^{-1} \left(x_i - \hat{\mu}_{xj}\right)^2$$

$$\hat{\sigma}^2_{xe,j} = n_j \sum_{i \in S_j} \left(x_i - \hat{\mu}_{xj}\right)^2 \left(y_i - \hat{\alpha}_j - \hat{\beta}_j x_i\right)^2 \Big/ \left(\pi_i^2 \hat{N}_j^2\right).$$

Furthermore, for simplicity we consider asymptotics only under with-replacement sampling. The results can be applied to the case of without replacement sampling if the sampling fraction n/N is negligible.

2.1 Design-based asymptotic framework

First, we establish the asymptotic normality of $\hat{Y}_{\text{reg},1}$ and $\hat{Y}_{\text{reg},2}$ under repeated sampling, that is, when y_i and x_i are fixed for $i \in U$, and S_j is a random PPS sample.

Theorem 1 Suppose that S_1 and S_2 are independent PPS samples with replacement from U_1 and U_2, respectively, where unit $i \in U_j$ has probability $p_{ij} = z_i \Big/ \sum_{i \in U_j} z_i > 0$ of being selected, and sampling weight $\pi_i^{-1} = 1 \big/ \left(n_j p_{ij}\right)$ for $j = 1, 2,$ and that the following four conditions hold, as the population sequence index m goes to ∞.

(C1) There exist constants φ_j and ω_j such that $\sqrt{n/n_j} \to \varphi_j$ and $N_j/N \to \omega_j$.

(C2) For $j = 1, 2,$ there exist constants μ_{yj}, μ_{xj} and β_j such that

$$\bar{Y}_j = Y_j / N_j = \sum_{i \in U_j} y_i / N_j \to \mu_{yj}, \bar{X}_j = X_j / N_j = \sum_{i \in U_j} x_i / N_j \to \mu_{xj}$$

exist, as do the limits $N_j^{-1} \sum_{i \in U_j} \left(x_i - \mu_{xj}\right)^2 \to \sigma^2_{xj} > 0,$ and in addition,

$$\left(\sqrt{n_j}/N_j\right) \sum_{i \in U_j} x_i \left(y_i - Y_j/N_j - \beta_j \left(x_i - X_j/N_j\right)\right) \to 0 \text{ as } n, N \to \infty.$$

(C3) The limits $D_{N_j} = \sum_{i \in U_j} p_{ij} b_{ij} b_{ij}^T / N_j^2 \to D_j$ exist, where for $i \in U_j$,

$$b_{ij} = \left[1/p_{ij} - N_j, x_i/p_{ij} - X_j, y_i/p_{ij} - Y_j\right]^T,$$

v^T denotes the vector transpose, and D_j is positive definite. The limit $\sigma^2_{xe,j} = \lim N_j^{-2} \sum_{i \in U_j} \left(x_i - \mu_{xj}\right)^2 \left(y_i - \alpha_j - \beta_j x_i\right)^2 \big/ p_{ij}$ also exists, for $\alpha_j = \mu_{yj} - \beta_j \mu_{xj}$.

(C4) The elements of $\Lambda_j = \sum_{i \in U_j} p_{ij} c_{ij} c_{ij}^T / N_j^4$ form a bounded sequence, where for $i \in U_j$,

$$c_{ij} = \left[\left(1/p_{ij} - N_j\right)^2, \left(x_i/p_{ij} - X_j\right)^2, \left(y_i/p_{ij} - Y_j\right)^2\right]^T.$$

Then, as $m \to \infty$, the following conclusions hold.

(a) For $j = 1, 2$, $\hat{\mu}_{xj} \to_P \mu_{xj}, \hat{\mu}_{yj} \to_P \mu_{yj}, \hat{\beta}_j \to_P \beta_j, \hat{\alpha}_j \to_P \alpha_j$, and $\hat{\sigma}^2_{xj} \to_P \sigma^2_{xj}$, where $\to_P$ denotes convergence in probability.

(b) The combined-stratum estimator $\hat{\beta}$ has the exact expression

$$\hat{\beta} = \frac{\sum_{j=1}^{2} \hat{\beta}_j \hat{\sigma}^2_{xj} \hat{N}_j + \left(\hat{X}_2 - \hat{X}_1\right)\left(\hat{Y}_2 - \hat{Y}_1\right) \hat{N}_1 \hat{N}_2 \big/ \left(\hat{N}_1 + \hat{N}_2\right)}{\sum_{j=1}^{2} \hat{\sigma}^2_{xj} \hat{N}_j + \left(\hat{X}_2 - \hat{X}_1\right)^2 \hat{N}_1 \hat{N}_2 \big/ \left(\hat{N}_1 + \hat{N}_2\right)} \tag{2.1}$$

and the in-probability limit

$$\beta = \frac{\sum_{j=1}^{2} \beta_j \sigma^2_{xj} \omega_j + (\mu_{x2} - \mu_{x1})(\mu_{y2} - \mu_{y1})\, \omega_1 \omega_2}{\sum_{j=1}^{2} \sigma^2_{xj} \omega_j + (\mu_{x2} - \mu_{x1})^2\, \omega_1 \omega_2}.$$

(c) $\sqrt{n_j}\left(\hat{\beta}_j - \beta_j\right) \to_d N\left(0, \sigma^2_{xe,j} \big/ \sigma^4_{x,j}\right)$, where $\to_d$ denotes convergence in distribution, and $\hat{\sigma}^2_{xe,j} \to_P \sigma^2_{xe,j}$.

(d) For $k = 1, 2$,

$$\sqrt{n}\left(\hat{Y}_{\text{reg},k} - Y\right) \big/ N \to_d N\left(0, \sigma^2_k\right) \tag{2.2}$$

where $\sigma^2_k = \sum_{j=1}^{2} a^T_{kj} D_j a_{kj}$ and

$$a_{1j} = \omega_j \varphi_j \left[-\left(\mu_y - \beta\mu_x\right), -\beta, 1\right]^T, \quad a_{2j} = \omega_j \varphi_j \left[-\left(\mu_{yj} - \beta_j \mu_{xj}\right), -\beta_j, 1\right]^T,$$

$\mu_x = \omega_1 \mu_{x1} + \omega_2 \mu_{x2}$, $\mu_y = \omega_1 \mu_{y1} + \omega_2 \mu_{y2}$, and D_j is given in condition (C3).

The conditions (C1)-(C4) of Theorem 1 provide a general formulation of the superpopulation framework for large-sample design-based statistical inference, within which the survey regression coefficients estimate well-defined frame-population descriptive parameters. The results in parts (a)-(b) show that the in-probability limits β_j, α_j of $\hat{\beta}_j, \hat{\alpha}_j$ have the standard interpretation as superpopulation least-squares slopes and intercepts. (These slope and intercept parameters also keep their usual model-based interpretations under the model (2.7) introduced in Section 2.2.) The asymptotic distribution theory for $\hat{\beta}_j$ in conclusion (c) allows us to deduce the large-sample behavior of $\hat{Y}_{\text{dec}}$ from that provided in (d) for $\hat{Y}_{\text{reg},k}$.

Under the further conditions

$$\beta_1 = \beta_2, \alpha_1 = \alpha_2, \tag{2.3}$$

it is clear from Theorem 1(b) that $\beta_j = \beta$, and $\sigma^2_1 = \sigma^2_2$ in (2.2), so that $\hat{Y}_{\text{reg},1}$ and $\hat{Y}_{\text{reg},2}$ and $\hat{Y}_{\text{dec}}$ are all asymptotically the same up to remainders of smaller order than $N/\sqrt{n}$, as we now show. Also, if

$\beta_1 \neq \beta_2$, then $\hat{Y}_{\text{reg},2} - \hat{Y}_{\text{dec}}$ continues to be $o_P\left(N/\sqrt{n}\right)$, and the test of equality of slopes rejects, *i.e.*, $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right) \to 1$, and therefore $\hat{Y}_{\text{dec}}$ has the same asymptotic distribution as $\hat{Y}_{\text{reg},2}$, which is more efficient than $\hat{Y}_{\text{reg},1}$ according to the result in Section 2.2.

Theorem 2 Assume the same hypotheses (C1)-(C4) as in Theorem 1.

(a) When (2.3) holds, then as $m \to \infty$

$$\sqrt{n}\left(\hat{\beta}_2 - \hat{\beta}_1\right) \to_d N\left(0, \sigma_d^2\right), \qquad \sigma_d^2 = \sum_{j=1}^{2} \frac{\sigma_{xe,j}^2}{\varphi_j^2 \sigma_{xj}^4}, \tag{2.4}$$

and the estimators $\hat{Y}_{\text{reg},1}, \hat{Y}_{\text{reg},2}$, and $\hat{Y}_{\text{dec}}$ are all asymptotically normally distributed and equivalent in the sense that

$$\frac{n}{N^2}\left[\left(\hat{Y}_{\text{reg},1} - \hat{Y}_{\text{reg},2}\right)^2 + \left(\hat{Y}_{\text{reg},2} - \hat{Y}_{\text{dec}}\right)^2\right] \to_P 0. \tag{2.5}$$

(b) When $\beta_1 \neq \beta_2$, $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right) \to 1$ and $\sqrt{n}\left(\hat{Y}_{\text{dec}} - Y\right)/N \to_d N\left(0, \sigma_2^2\right)$.

A more refined study of the asymptotic behavior of the estimators $\hat{Y}_{\text{dec}}$ can be undertaken in the spirit of Saleh (2006), as with contiguous or Pitman alternatives for non-survey statistical models, by assuming that $\sqrt{n}(\beta_1 - \beta_2) \to r$ for a constant r. Under this assumption, it can be shown that $\hat{Y}_{\text{reg},1} - \hat{Y}_{\text{reg},2} = o_P\left(N/\sqrt{n}\right)$ and, therefore, the three centered and scaled estimators $\sqrt{n}\left(\hat{Y}_{\text{dec}} - Y\right)$, $\sqrt{n}\left(\hat{Y}_{\text{reg},2} - Y\right)$, and $\sqrt{n}\left(\hat{Y}_{\text{reg},1} - Y\right)$ all have the same asymptotic normal distribution with mean 0. Furthermore,

$$P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right) \to \Phi(-z_{\tau/2} + r/\sigma_d) + \Phi(-z_{\tau/2} - r/\sigma_d), \tag{2.6}$$

where σ_d^2 is given in (2.4), and $z_{\tau/2}$ and Φ are respectively the standard normal percentage point and distribution function. Thus, $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right)$ has a limit different from 1. In particular, the limit in (2.6) equals τ when $\beta_1 = \beta_2$ (*i.e.*, when $r = 0$).

2.2 Model-assisted asymptotic setting

We elaborate in this section the behavior of estimators $\hat{Y}_{\text{reg},k}, \hat{Y}_{\text{dec}}$ under the assumed probabilistic model that the triples (x_i, y_i, z_i) in the finite population, $i \in U_j$, are independent and identically distributed (iid), where the size-variables $z_i > 0$ are used in defining PPS with-replacement draw probabilities $p_{ij} = z_i \big/ \sum_{i' \in U_j} z_{i'}$, and where x_i and y_i follow the model

$$y_i = \alpha_j + \beta_j x_i + \varepsilon_i,\ i \in U_j, \tag{2.7}$$

with α_j and β_j as unknown intercept and slope parameters for the regression within stratum U_j. The errors $\varepsilon_i, i \in U_j$, are assumed to be iid with mean 0 and finite variance σ_ε^2 and to be independent of

(x_i, z_i), and the variables x_i for $i \in U_j$ are assumed to have finite variance. Also, to enable PPS sampling, we assume that $\max_{i \in U_j} n_j p_{ij} < 1$ with probability approaching 1 for large m, *i.e.*, for large n_j, N_j.

In this section, asymptotic properties of estimators $\hat{Y}_{\text{reg},k}, \hat{Y}_{\text{dec}}$ are considered with respect to the regression model and repeated sampling. By Theorem 1, the model-assisted estimators $\hat{Y}_{\text{reg},1}$ and $\hat{Y}_{\text{reg},2}$ are still consistent and asymptotically normal for triples (x_i, y_i, z_i) iid within strata, since the conditions (C1)-(C4) are satisfied under moment assumptions on $z_i, 1/z_i$ even if model (2.7) is incorrect. However, the estimators $\hat{Y}_{\text{reg},k}$ are efficient when model (2.7) is correct.

Theorem 3 Assume model (2.7) along with (C1), with $E\left(x_i^4\right) < \infty, E\left(\varepsilon_i^4\right) < \infty, E\left(z_i\right) < \infty,$ and $E\left(\left(1 + x_i^4\right)/z_i^3\right) < \infty.$ Then all conclusions in Theorem 1 and Theorem 2 still hold. In particular, when $\beta_1 \neq \beta_2, \sigma_1^2,$ the asymptotic variance of $\sqrt{n}\left(\hat{Y}_{\text{reg},1} - Y\right)/N$, is larger than $\sigma_2^2,$ the asymptotic variance of $\sqrt{n}\left(\hat{Y}_{\text{reg},2} - Y\right)/N$. Furthermore,

$$\sqrt{n}\left(\hat{Y}_{\text{dec}} - Y\right)/N \rightarrow_d N\left(0, (1-\pi)\sigma_1^2 + \pi\sigma_2^2\right), \tag{2.8}$$

where π is the limit of $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right)$.

Note that π in (2.8) is equal to 1 when $\beta_1 \neq \beta_2$ and equal to τ when $\beta_1 = \beta_2$.

According to Theorem 3, under model (2.7), all three estimators defined in (1.2)-(1.4) have the same asymptotic efficiency when $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$ (condition (2.3)). Furthermore, $\hat{Y}_{\text{reg},1}$ is asymptotically worse than $\hat{Y}_{\text{reg},2}$ when $\beta_1 \neq \beta_2$. Thus, why would we not always use $\hat{Y}_{\text{reg},2}$?

The assertions in Theorem 3 are first-order asymptotic results. A more refined, second-order asymptotic result under the conditions in Theorem 3 and condition (2.3) when the sizes z_i are all equal is that, up to a term of order $n_1^{-2} + n_2^{-2}$,

$$\text{mse}\left(\frac{\hat{Y}_{\text{reg},1}}{N}\right) - \frac{\sigma_\varepsilon^2}{n} \leq \left[\text{mse}\left(\frac{\hat{Y}_{\text{reg},2}}{N}\right) - \frac{\sigma_\varepsilon^2}{n}\right]\left[1 - \frac{n_1 n_2 (\bar{X}_1 - \bar{X}_2)^2}{n D_n}\right], \tag{2.9}$$

where mse is the mean squared error conditional on x_i's, $\bar{X}_j = N_j^{-1} \sum_{i \in U_j} x_i$, and

$$D_n = \sum_{j=1}^{2} \sum_{i \in U_j} \left(x_i - \bar{X}_j\right)^2 + \frac{n_1 n_2 (\bar{X}_1 - \bar{X}_2)^2}{n}.$$

Result (2.9) indicates that, when weights are equal and $\beta_1 = \beta_2$ and $\alpha_1 = \alpha_2$, the finite sample performance of $\hat{Y}_{\text{reg},1}$ may be better than that of $\hat{Y}_{\text{reg},2}$ for moderate n_1 and n_2. See the simulation results in Section 4. The proof of (2.9) is a special case of a more general result in Slud (2012) and, thus, is omitted.

In applications, we do not know whether $\beta_1 = \beta_2$. Hence, the decision-based estimator $\hat{Y}_{\text{dec}}$ is an adaptive procedure to select a good estimator. In view of (2.8), the performance of $\hat{Y}_{\text{dec}}$ is close to (slightly worse than) that of $\hat{Y}_{\text{reg},2}$ when $\beta_1 \neq \beta_2$, and is close to (slightly worse than) that of $\hat{Y}_{\text{reg},1}$ when $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$. This is also supported by the simulation results in Section 4.

3 Variance estimation

It is common practice to report a variance estimate or standard error for each survey estimate. Variance estimation is also crucial for statistical inference when setting a confidence interval for an unknown parameter of interest.

The asymptotic results in Section 2 suggest a variance estimator for $\hat{Y}_{\text{reg},k}$ by substituting into (2.2) estimators for unknown quantities in σ_k^2. Since the total variance is a sum of H within-stratum variances, without loss of generality we consider one stratum $(H = 1)$. For $j = 1, 2,$ let

$$\hat{D}_{n_j} = \sum_{i \in S_j} \frac{\hat{b}_{ij}\hat{b}_{ij}^T}{(n_j - 1)\hat{N}_j}, \quad \hat{b}_{ij} = \left[1/p_{ij} - \hat{N}_j, x_i/p_{ij} - \hat{X}_j, y_i/p_{ij} - \hat{Y}_j\right]^T, \quad i \in S_j,$$

$$\hat{a}_{1j} = \frac{\hat{N}_j n^{1/2}}{\hat{N} n_j^{1/2}}\left[-(\bar{y}_j - \hat{\beta}_j \bar{x}_j), -\hat{\beta}_j, 1\right]^T, \quad \hat{a}_{2j} = \frac{\hat{N}_j n^{1/2}}{\hat{N} n_j^{1/2}}\left[-(\bar{y} - \hat{\beta}\bar{x}), -\hat{\beta}, 1\right]^T,$$

$$\bar{y}_j = \hat{Y}_j / \hat{N}_j, \quad \bar{x}_j = \hat{X}_j / \hat{N}_j, \quad \bar{y} = \sum_{j=1}^{2} \hat{Y}_j / (\hat{N}_1 + \hat{N}_2), \quad \bar{x} = \sum_{j=1}^{2} \hat{X}_j / (\hat{N}_1 + \hat{N}_2).$$

Then, under the conditions in Theorem 1,

$$\hat{\sigma}_k^2 = \sum_{j=1}^{2} \hat{a}_{kj}^T \hat{D}_{n_j} \hat{a}_{kj} \to_p \sigma_k^2, \quad k = 1, 2.$$

That is, $\hat{\sigma}_k^2$ is consistent for σ_k^2. The results in Theorems 2 and 3 also show that $\hat{\sigma}_2^2$ is a consistent variance estimator for the decision-based estimator $\hat{Y}_{\text{dec}}$, because we have either $\sigma_1^2 = \sigma_2^2$ or $P(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}) \to 1$.

These substitution variance estimators, however, may not perform well when one of n_1 and n_2 is moderate (see Section 4). An alternative method is the bootstrap as suggested by Cheng *et al.* (2010). Let $\hat{\theta}$ be the estimator under consideration. Its bootstrap variance estimator can be obtained as follows.

1. Draw a bootstrap sample S_j^* as a simple random sample of size n_j with replacement from S_j, where S_1^* and S_2^* are independently obtained. If there are k_j self-representing units in S_j, as discussed in Section 4.1 below, then with-replacement samples of sizes $n_j - k_j$ are drawn, $j = 1, 2$.

2. The survey weights and observed data from the original data set are used to form a bootstrap data set $S_1^* \cup S_2^*$. From this dataset, calculate the bootstrap analog $\hat{\theta}^*$ of $\hat{\theta}$.
3. Independently repeat the previous steps B times to obtain $\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B}$. The sample variance of $\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B}$ is the bootstrap variance estimator for $\hat{\theta}$.

Under the conditions in Theorems 1-2, the bootstrap variance estimators for $\hat{Y}_{\text{reg},1}, \hat{Y}_{\text{reg},2}$ and $\hat{Y}_{\text{dec}}$ are consistent estimators. The proof for the bootstrap is similar to the proofs of the theorems and is omitted.

4 Simulation results for $H = 1$

Large sample theory as presented above is not adequate to tell whether the asymptotic results adequately describe the behavior of the estimators $\hat{Y}_{\text{reg},1}, \hat{Y}_{\text{reg},2}$ and $\hat{Y}_{\text{dec}}$ and their variance estimators in moderate samples, or whether $\hat{Y}_{\text{reg},1}$ and $\hat{Y}_{\text{dec}}$ ever provide useful Mean-Squared-Error improvements in moderate sized samples. We present some simulation results to study these questions, as well as the small-sample issues arising in applying these methods in the context of the ASPEP survey.

In the simulations, values in the frame population U are either generated under some model or are taken from the 2002 and 2007 Government censuses with 2007 ASPEP sample weights. The first set of simulations (reported in Tables 4.1-4.6) summarizes average behavior over many model-generated frame populations. In the second set of artificial-data simulations, summarized in Table 4.8, the frame population remains fixed throughout the simulation. All frame populations consist of a single stratum $(H = 1)$ broken into two substrata $(j = 1, 2)$ according as a size variable falls below or above a specific quantile, usually the 0.8 quantile. Sampling from the frame populations is done PPS with-replacement in all simulations in this section.

4.1 Small sample considerations

Before proceeding to describe the simulations, we discuss some special features of PPS with-replacement (PPSWR) sampling which, when done in settings with small samples and unbalanced size variables, requires special computational handling. Numerically erratic results can arise when the small drawn samples are used stratumwise and then bootstrapped to estimate variances.

The weights $\pi_i^{-1} = 1\big/\left(n_j p_{ij}\right)$ in PPSWR are all greater than 1 only when the single-draw probabilities $p_{ij} = z_i \Big/ \sum_{i' \in U_j} z_{i'}$ are all below $1/n_j$. To avoid anomalous small-sample results, and to maintain the relevance of PPSWR designs in imitating PPS without-replacement designs, any units $i \in U_j$ with $n_j p_{ij} \geq 1$ are made *self-representing* (SR), *i.e.*, are sampled with certainty but only once, and if there are k_j such units, then the probabilities $\left\{p_{ij} : i \in U_j, n_j p_{ij} < 1\right\}$ are renormalized to draw a size $n_j - k_j$ PPSWR sample. If any of the remaining renormalized probabilities are $\geq 1/(n_j - k_j)$, then their units also become self-representing and a new renormalization is done. This is repeated as often as necessary. Thus, small samples with very unequal size-variable distributions may not be compatible with PPSWR sampling, a condition arising in some of the real-data ASPEP cases considered below.

Although a different choice could have been made, we conform with ASPEP practice in including all SR units in the fitting of the survey-weighted regression estimators $\hat{\beta}_2$ and $\hat{\beta}$. However, with this choice, PPSWR sampling followed by bootstrap resampling of small samples can lead to extremely erratic behavior, which must be recognized in summarizing the behavior of bootstrap variance estimators. The problem is that when a small number m of non-self-representing items are sampled PPSWR, in addition to a set of SR items, and then bootstrapped, the probability can be surprisingly large that there is only one unique non-SR item in the bootstrap sample, leading to very high bootstrap variability. This phenomenon was observed in the simulations reported below, with large-size substratum containing 20 or fewer elements and very skewed size-variables, either in the cases with lognormal or ASPEP x_i variables.

4.2 Artificial model-generated data

All of the artificial frame populations were generated with $N = 2{,}000$ iid triples (x_i, y_i, z_i) satisfying (2.7), for U_1 consisting of the $N_1 = 1{,}600$ for which x_i fell below their empirical 80'th percentile $c = \left(x_{(1,600)} + x_{(1,601)}\right)/2$, and U_2 consisting of the other 400 indices. In most cases, z_i were generated as $N(30 + x_i, 100)$ variates conditioned to be positive (which required occasional re-simulation in the lognormal-x_i models below) and were conditionally independent of y_i given x_i. (However, in some cases, unweighted samples were drawn by taking z_i identically equal.) PPS with-replacement stratified samples of sizes $(n_1, n_2) = (100, 50), (100, 20)$, or $(50, 20)$ were drawn in successive simulation runs, with size-variables z_i, from the same frame.

The models generating (x_i, y_i) are indexed as follows. In those with prefix **M1**, the predictors x_i are $\text{Gamma}(4, 0.1)$ distributed, with 0.8 quantile 55.2, while in the models **M2**, the x_i variables are Lognormal(1,6.25), with 0.8 quantile 22.3. The **M1** populations, and the **M2** models with suffix **E**, have conditional variance 100 for y_i given x_i, while the **M2** models without suffix **E** have conditional variance $20x_i$. Conditional means $E(y_i | x_i)$ are all linear, equal to $20 + 1.5x_i$ in models indexed **H0** and to $20 + x_i + 0.5(x_i - c)\, I_{[j=2]}$ within the substratum U_j in models **H1**. The intercepts of the regression models are so chosen that whether or not the slopes are the same, the lines intersect at $x = c$ (see the discussion in Section 1). Table 4.1 exhibits the average and standard deviation for the totals Y generated from the frame-population attributes $\{y_i\}_{i=1}^{2,000}$ under the various models. The variates x_i as well as the totals Y are much longer-tailed under the Lognormal models.

Table 4.1
Means and standard deviations for totals Y under simulation models.

	Gamma		Lognormal			
Model	**M1.H0**	**M1.H1**	**M2.H0**	**M2.H0E**	**M2.H1**	**M2.H1E**
E(Y)	160,000	123,177	225,603	225,603	173,485	173,485
SD(Y)	1,414.2	653.5	94,380	94,368	62,362	62,344

Simulated population models

M1.H0: $x_i \sim \text{Gamma}\,(4, 0.1)$ (shape parameter 4, scale 10),
$y_i \sim N\,(20 + 1.5x_i, 100)$ (variance 100), all $i \in U$.

M1.H1: $x_i \sim \text{Gamma}\,(4, 0.1)$, $y_i \sim N\left(20 + x_i + 0.5\,(x_i - c)\,I_{[x_i \geq c]}, 100\right)$, all i.

M2.H0: $\log\,(x_i) \sim N\,(1, 6.25), y_i \sim N\,(20 + 1.5x_i, 20x_i)$, all i.

M2.H0E: $\log\,(x_i) \sim N\,(1, 6.25), y_i \sim N\,(20 + 1.5x_i, 100)$, all i.

M2.H1: $\log\,(x_i) \sim N\,(1, 6.25), y_i \sim N\left(20 + x_i + 0.5\,(x_i - c)\,I_{[x_i \geq c]}, 20x_i\right)$, all i.

M2.H1E: $\log\,(x_i) \sim N\,(1, 6.25), y_i \sim N\left(20 + x_i + 0.5\,(x_i - c)\,I_{[x_i \geq c]}, 100\right)$, all i.

The simulation and bootstrap results in Tables 4.2-4.5 were generated by the following design and reporting scheme. For each population type, 60 distinct frame populations were generated, and 50 independent sampling experiments were conducted with each of those. In those cases where results of weighted and unweighted sampling were compared, these samples were drawn independently from the same set of 60 frame populations. Thus there were 3,000 independent replications for Monte Carlo averaging of statistical results, done for each of three different stratified sample sizes, and 400 bootstrap iterations were performed for each such generated sample.

Table 4.2
Empirical and estimated SD's and CI coverage, from model $\mathrm{M1}$ simulations.

		M1.H0			**M1.H1**		
Sizes	**Stat**	$\hat{Y}_{\text{reg},1}$	$\hat{Y}_{\text{reg},2}$	$\hat{Y}_{\text{dec}}$	$\hat{Y}_{\text{reg},1}$	$\hat{Y}_{\text{reg},2}$	$\hat{Y}_{\text{dec}}$
100,50	SD_{MC}	1,785.5	1,794.3	1,788.0	1,817.6	1,773.5	1,774.4
	$\widehat{\text{SD}}_S$	1,757.1	1,751.5	1,755.6	1,794.6	1,735.2	1,735.8
	$\widehat{\text{SD}}_B$	1,752.4	1,762.0	1,758.4	1,788.1	1,742.9	1,747.0
	CP_S	94.47	94.37	94.50	93.93	93.73	93.77
	CP_B	94.60	94.53	94.67	93.93	94.03	94.07
100,20	SD_{MC}	1,930.0	1,944.8	1,934.0	2,008.4	1,944.4	1,960.4
	$\widehat{\text{SD}}_S$	1,888.3	1,876.6	1,884.1	1,944.4	1,861.0	1,866.5
	$\widehat{\text{SD}}_B$	1,878.8	1,901.4	1,895.8	1,936.1	1,885.6	1,897.9
	CP_S	94.20	93.83	94.13	93.53	93.20	93.07
	CP_B	93.80	94.00	93.97	93.60	93.83	93.97
50,20	SD_{MC}	2,583.5	2,610.7	2,593.5	2,591.3	2,522.8	2,535.4
	$\widehat{\text{SD}}_S$	2,509.2	2,490.8	2,505.1	2,562.2	2,465.0	2,474.5
	$\widehat{\text{SD}}_B$	2,498.5	2,538.0	2,522.9	2,550.3	2,508.5	2,525.6
	CP_S	93.70	93.13	93.57	93.97	93.63	93.43
	CP_B	93.63	93.73	93.87	93.83	93.77	94.10

Table 4.3
Empirical and estimated SD's and CI coverage, from model $\mathrm{M2}$ simulations.

		M2.H0			M2.H1		
Sizes	**Stat**	$\hat{Y}_{\mathrm{reg},1}$	$\hat{Y}_{\mathrm{reg},2}$	$\hat{Y}_{\mathrm{dec}}$	$\hat{Y}_{\mathrm{reg},1}$	$\hat{Y}_{\mathrm{reg},2}$	$\hat{Y}_{\mathrm{dec}}$
100,50	SD_{MC}	3,400.1	3,475.4	3,406.8	3,481.9	3,483.8	3,482.2
	$\widehat{\mathrm{SD}}_S$	3,420.6	3,400.0	3,417.0	3,537.8	3,405.0	3,463.7
	$\widehat{\mathrm{SD}}_B$	3,590.0	3,715.2	3,623.4	3,852.0	3,921.9	3,898.4
	CP_S	95.10	93.43	94.83	95.03	93.40	94.13
	CP_B	95.67	95.77	95.77	95.63	95.77	95.70
100,20	SD_{MC}	5,655.2	6,184.0	5,698.6	5,853.0	6,181.1	5,955.6
	$\widehat{\mathrm{SD}}_S$	5,644.9	5,575.7	5,640.9	5,798.3	5,587.3	5,697.3
	$\widehat{\mathrm{SD}}_B$	5,565.1	6,687.3	5,857.8	5,907.8	6,838.0	6,466.6
	CP_S	93.83	88.47	93.40	92.77	88.30	90.70
	CP_B	92.33	93.67	93.37	92.63	94.33	94.17
50,20	SD_{MC}	5,773.2	6,319.2	5,833.9	5,934.2	6,230.6	6,009.8
	$\widehat{\mathrm{SD}}_S$	5,800.2	5,677.2	5,785.8	6,012.6	5,755.4	5,919.2
	$\widehat{\mathrm{SD}}_B$	5,728.5	6,825.2	6,086.0	6,102.2	6,978.1	6,522.1
	CP_S	94.60	88.67	93.97	94.07	89.37	92.27
	CP_B	93.40	94.23	94.27	93.47	95.03	94.80

Table 4.4
SD's for $\hat{Y}_{\mathrm{HT}}$ *vs.* $\hat{Y}_{\mathrm{dec}}$, and coverage for Bootstrap Percentile Confidence Intervals for $\hat{Y}_{\mathrm{dec}}$, for $\tau = 0.05$ *vs.* 0.20, for models $\mathrm{M1}$ and $\mathrm{M2}$, $\mathrm{H0}$ and $\mathrm{H1}$.

		$\hat{Y}_{\mathrm{dec}}, \tau = 0.05$		$\hat{Y}_{\mathrm{HT}}$	$\hat{Y}_{\mathrm{dec}}, \tau = 0.20$	
Model	**Samples**	SD_{MC}	CP_{BP}	$\mathrm{SD}_{\mathrm{HT}}$	SD_{MC}	CP_{BP}
M1.H0	100,50	1,788.0	94.23	2,774.0	1,745.5	94.60
	100,20	1,934.0	93.50	3,032.6	1,915.9	94.10
	50,20	2,593.5	93.17	3,000.7	2,500.1	94.43
M1.H1	100,50	1,774.4	93.70	2,387.3	1,737.3	94.43
	100,20	1,960.4	93.27	2,678.9	1,948.0	93.23
	50,20	2,535.4	93.90	3,035.0	2,509.8	94.23
M2.H0	100,50	3,406.8	95.20	4,160.0	3,398.8	94.83
	100,20	5,698.6	91.13	6,720.2	5,705.7	92.57
	50,20	5,833.9	92.60	7,080.0	5,979.8	92.17
M2.H1	100,50	3,482.2	95.13	4,393.6	3,423.9	94.03
	100,20	5,955.6	92.07	7,413.1	5,917.3	92.40
	50,20	6,009.8	92.33	7,840.4	6,105.6	92.17

Table 4.5
Comparisons of SD estimates and CI coverage for $H0$ and $H1$ for three lognormal settings, weighted (W) and unweighted (U) within $M2$, and weighted (E) within $M2.E$. CI % coverages are given for both the Bootstrap SD and Percentile Intervals.

Model	Size	Stat	SD	$\widehat{\text{SD}}_S$	$\widehat{\text{SD}}_B$	CP_S	CP_B	CP_{BP}
H0.W	100,50	$\hat{Y}_{\text{reg},1}$	3,400.1	3,420.6	3,590.0	95.10	95.67	94.93
		$\hat{Y}_{\text{reg},2}$	3,475.4	3,400.0	3,715.2	93.43	95.17	95.33
		$\hat{Y}_{\text{dec}}$	3,406.8	3,417.0	3,623.4	94.83	95.77	95.20
H0.U		$\hat{Y}_{\text{reg},1}$	5,481.6	3,674.8	5,571.9	81.43	93.50	92.07
		$\hat{Y}_{\text{reg},2}$	5,782.8	3,646.6	6,076.3	80.13	93.67	91.90
		$\hat{Y}_{\text{dec}}$	5,525.5	3,669.0	5,726.8	81.07	93.83	92.20
H0.E		$\hat{Y}_{\text{reg},1}$	1,888.8	1,930.1	1,904.7	94.73	94.53	94.23
		$\hat{Y}_{\text{reg},2}$	1,888.6	1,911.1	1,893.2	94.43	94.30	94.20
		$\hat{Y}_{\text{dec}}$	1,892.9	1,926.5	1,905.0	94.67	94.57	94.20
H0.W	50,20	$\hat{Y}_{\text{reg},1}$	5,773.2	5,800.2	5,728.5	94.60	93.40	92.00
		$\hat{Y}_{\text{reg},2}$	6,319.2	5,677.2	6,825.2	88.67	94.23	92.60
		$\hat{Y}_{\text{dec}}$	5,833.9	5,785.8	6,086.0	93.97	94.27	92.60
H0.U		$\hat{Y}_{\text{reg},1}$	10,000.3	5,136.5	9,905.6	71.10	90.73	89.80
		$\hat{Y}_{\text{reg},2}$	11,192.8	5,085.0	12,806.8	68.70	92.90	89.37
		$\hat{Y}_{\text{dec}}$	10,134.1	5,120.7	11,245.9	70.73	92.37	90.27
H0.E		$\hat{Y}_{\text{reg},1}$	2,811.4	2,831.6	2,769.5	94.13	94.00	93.93
		$\hat{Y}_{\text{reg},2}$	2,811.9	2,753.8	2,741.1	93.47	93.77	93.30
		$\hat{Y}_{\text{dec}}$	2,817.4	2,821.8	2,777.0	93.83	93.90	93.77
H1.W	100,50	$\hat{Y}_{\text{reg},1}$	3,481.9	3,537.8	3,852.0	95.03	95.63	95.27
		$\hat{Y}_{\text{reg},2}$	3,483.8	3,405.0	3,921.9	93.40	95.77	95.10
		$\hat{Y}_{\text{dec}}$	3,482.2	3,463.7	3,898.4	94.13	95.70	95.13
H1.U		$\hat{Y}_{\text{reg},1}$	5,631.4	3,774.8	5,614.6	80.90	92.33	91.07
		$\hat{Y}_{\text{reg},2}$	5,838.3	3,699.6	6,010.5	79.13	92.73	91.37
		$\hat{Y}_{\text{dec}}$	5,727.0	3,732.8	5,870.5	80.40	92.93	91.63
H1.E		$\hat{Y}_{\text{reg},1}$	2,005.5	2,094.2	2,019.1	95.60	94.97	94.60
		$\hat{Y}_{\text{reg},2}$	1,909.9	1,908.2	1,892.5	94.83	94.77	94.17
		$\hat{Y}_{\text{dec}}$	1,931.9	1,941.7	1,934.6	94.97	95.20	94.83
H1.W	50,20	$\hat{Y}_{\text{reg},1}$	5,934.2	6,012.6	6,102.2	94.07	93.47	91.97
		$\hat{Y}_{\text{reg},2}$	6,230.6	5,755.4	6,978.1	89.37	95.03	92.23
		$\hat{Y}_{\text{dec}}$	6,009.8	5,919.2	6,522.1	92.27	94.80	92.33
H1.U		$\hat{Y}_{\text{reg},1}$	9,315.8	5,350.9	10,040.0	74.17	93.10	90.57
		$\hat{Y}_{\text{reg},2}$	10,583.8	5,229.6	12,476.8	71.23	94.57	90.87
		$\hat{Y}_{\text{dec}}$	9,989.6	5,295.4	11,479.5	72.53	94.33	91.47
H1.E		$\hat{Y}_{\text{reg},1}$	3,096.1	3,137.7	2,795.6	94.63	93.43	93.37
		$\hat{Y}_{\text{reg},2}$	2,880.6	2,766.8	2,745.7	93.10	93.40	93.47
		$\hat{Y}_{\text{dec}}$	2,977.3	2,929.2	2,882.0	93.77	93.77	93.77

We calculated the following quantities for each combination of model, weighting, and sample size: the percentage biases of $\hat{Y}_{\text{reg},1}$, $\hat{Y}_{\text{reg},2}$, $\hat{Y}_{\text{dec}}$ (with $\tau = 0.05$ in all tables except Table 4.4, and $\tau = 0.05$ or 0.20 in Table 4.4) as estimators of Y; the Monte Carlo standard deviations (SD), SD_{MC}, of these three estimators; the estimated SD's of the estimators, respectively using the substitution $(\widehat{\text{SD}}_S)$ and bootstrap $(\widehat{\text{SD}}_B)$ SD estimators described in Section 3; the coverage probability, $\text{C}P_u$, of the nominal 95% confidence intervals for $Y : \hat{Y} \pm 1.960 \cdot \widehat{\text{SD}}_u$, where $\hat{Y}$ is one of the three estimators of Y, and $u = S$ or B; and the bootstrap percentile confidence intervals (and their coverage percentages $\text{C}P_{BP}$) obtained from the empirical 0.025 and 0.975 quantiles of the (400) bootstrapped values of each of the three estimators $\hat{Y}$ of Y. In addition, we calculated empirical biases of the Horvitz-Thompson estimates $\hat{Y}_{\text{HT}}$ in (1.1) and their empirical standard deviations SD_{HT}. (Of these calculated quantities, only the biases are not shown, since all of the biases were well below 0.5% except in the model **M2**.**H1**.**U**, and even there the largest magnitude of bias was about 1%.) Two further statistics, computed and displayed in Table 4.6 for each of the estimators $\hat{Y}$ of Y, are the standard errors across randomly generated frame populations of the Monte Carlo and Bootstrap within-population estimated SD's of estimators $\hat{Y}$.

Table 4.6
Cross-population Standard errors of Empirical and Bootstrap SD's estimated for the estimators $\hat{Y}_{\text{reg},1}$, $\hat{Y}_{\text{reg},2}$, and $\hat{Y}_{\text{dec}}$, for selected models and weighting.

		$\hat{Y}_{\text{reg},1}$		$\hat{Y}_{\text{reg},2}$		$\hat{Y}_{\text{dec}}$	
Model	**Sizes**	**SD**	$\widehat{\text{SD}}_B$	**SD**	$\widehat{\text{SD}}_B$	**SD**	$\widehat{\text{SD}}_B$
M1.**H0**	100,50	198	35	196	35	197	35
	50,20	210	52	208	51	210	51
M1.**H1**	100,50	204	39	183	40	184	41
	50,20	319	57	298	62	302	62
M2.**H0**	100,50	404	345	450	383	405	351
	50,20	825	518	1,075	916	889	631
M2.**H0**.**E**	100,50	187	49	185	45	184	47
	50,20	294	85	293	71	298	82
M2.**H1**	100,50	409	409	410	421	408	414
	50,20	767	624	946	929	841	730
M2.**H1**.**E**	100,50	208	59	196	46	204	50
	50,20	258	141	261	82	239	102
M2.**H1**.**U**	100,50	1,676	1,351	1,773	1,539	1,726	1,467
	50,20	2,397	2,543	3,425	3,454	3,102	3,159

4.3 Real government-census data

Our simulations based on repeated sampling from real-data frames rely on a national state-wise dataset assembled by Yang Cheng. For the ASPEP survey of governments for sample year 2007, which was also a census year, the ASPEP frame is the same as the 2007 Census of Governments file. Our dataset consists of the 2002 and 2007 ASPEP variable values (full- and part-time employees, payroll and hours) derived from the censuses in those years, plus the 2007 sample weights and in-sample indicators for ASPEP. Weights equal to 1 imply that governmental units were self-representing (SR), in the sense that they were chosen for inclusion with certainty in ASPEP. The size-variable z_i for PPS sampling within ASPEP is the sum of full- and part-time payroll from the most recent census, so we restrict attention to the 53,402 governmental units in the file for which this variable was positive. Table 4.7 gives the subcounty and special-district governmental types (the only ones that are subdivided into Small and Large unit substrata) in nine selected states, giving also the SR counts and numbers sampled in 2007. As mentioned in subsection 4.1, the final SR count for PPS with-replacement sampling can exceed the number of units initially chosen for certain inclusion, and the larger numbers, corresponding to the sample size actually drawn in 2007, are shown in the SR columns of Table 4.7. Inspection of this Table shows that several of the state by type combinations either have no population in a substratum or have too few non-SR units to be useful in simulating repeated samples. We take 15 as a rule-of-thumb minimum for the number of non-SR units, and suggest that substratum pairs with fewer non-SR units in the large-unit stratum should be collapsed without recourse to the decision-based strategy studied in this paper.

Table 4.7
Census population, ASPEP sample sizes and SR counts of Subcounty and Special-District governmental units by substratum in 2007, for 9 selected states.

	Subcounty					Special District				
	Small		Large			Small		Large		
	Pop	Samp	Pop	Samp	SR	Pop	Samp	Pop	Samp	SR
AL	378	15	55	45	26	0	0	400	102	64
CA	0	0	475	104	86	1,595	39	107	107	107
CO	0	0	265	34	18	627	16	65	55	33
FL	317	16	81	54	36	0	0	330	48	24
GA	461	17	49	36	20	0	0	293	70	32
MO	980	25	101	101	101	799	27	106	66	42
NY	1,473	25	69	69	69	606	16	33	23	4
PA	2,409	55	123	81	31	921	21	37	37	37
WI	1,702	36	129	71	44	281	16	61	40	20

For nine government-by-type combinations with 15 or more non-SR units and at least 17 non-sampled non-SR large-substratum units (except for AL, CO, and GA for which there were respectively 9, 10, and 11 non-sampled non-SR units), Table 4.8 displays results for the decision-based estimators and variance estimates in substratum pairs. In each of the state-type combinations, 3,000 stratified PPSWR samples of the indicated sizes were drawn from the ASPEP and government census frame described above, with x_i and y_i respectively the full-time payroll amount for the governmental unit as recorded in the 2002 and 2007 governmental censuses, and z_i the total (full-time plus part-time) payroll in 2002. Within each simulated sample, the estimators $\hat{Y}_{\text{reg},1}, \hat{Y}_{\text{reg},2}, \hat{Y}_{\text{dec}}$ were calculated, and the empirical variances estimated. The variance of $\hat{Y}_{\text{dec}}$ was also estimated by the substitution formula and bootstrap methods as in the artificial-data simulations. (But note that, as described above, the bootstrap samples were drawn only from the non-SR units in each substratum sample.) The results are shown in Table 4.8. The relative efficiencies between the combined and separate stratified regression estimators can be gleaned from the corresponding ratio of SD's given in column 5 of the table. The remaining SD's shown are the empirical, substitution, and bootstrap SD estimators of $\hat{Y}_{\text{dec}}$.

Table 4.8
Summary of repeated-sampling simulations from ASPEP 2007 frame. Total full-time pay (Y) given in multiple of \$100 million, and estimated SD's of $\hat{Y}_{\text{dec}}$ given in columns 6-8 in units of \$1 million. SD_1/SD_2 in column 5 is ratio of empirical SD of $\hat{Y}_{\text{reg},1}$ over that of $\hat{Y}_{\text{reg},2}$.

State	Stratum	Y	Size	SD_1/SD_2	$\widehat{\text{SD}}$	$\widehat{\text{SD}}_S$	$\widehat{\text{SD}}_B$
AL	SubCty	1.2	25,46	2.14	4.90	3.67	5.71
CA	SpcDst	4.3	30,90	0.98	29.4	21.2	26.8
CO	SpcDst	0.6	25,55	1.14	3.77	2.58	3.00
FL	SubCty	4.3	25,54	1.16	11.9	9.4	12.2
GA	SubCty	1.5	25,38	1.15	4.38	3.26	4.88
MO	SpcDst	0.6	40,70	2.13	2.99	2.20	2.99
NY	SubCty	23.6	35,52	1.53	13.6	12.0	14.1
PA	SubCty	3.0	40,70	1.12	7.28	5.79	7.60
WI	SubCty	1.4	40,70	2.06	5.00	4.45	5.17

4.4 Discussion of simulation results

The following is a summary and interpretation of the results in the Tables, as well as of other results not shown.

(I) Many of the artificial-data simulations serve to confirm the large-sample theoretical results of the Theorems. It has already been mentioned that in Tables 4.2 and 4.3 the biases for all three Y-estimators $(\hat{Y}_{\text{reg},1}, \hat{Y}_{\text{reg},2}, \hat{Y}_{\text{dec}})$ are generally small. Within Table 4.2, referring to models with predictors and weights related to the Gamma distribution in models $\mathbf{M1}$, the substitution and bootstrap variance estimators for each Y-estimator are quite accurate and close to one another, and the confidence intervals all have close to nominal coverage. Under both $\mathbf{M1}.\mathbf{H0}$ and $\mathbf{M1}.\mathbf{H1}$, there is a tendency with smaller n_2 sample size for the $\widehat{\text{SD}}_S$ and $\widehat{\text{SD}}_B$ estimators to be slight underestimates of the actual or empirical SD's, but $\widehat{\text{SD}}_B$ seems to track SD more closely than $\widehat{\text{SD}}_S$ for $\hat{Y}_{\text{reg},2}$ and $\hat{Y}_{\text{dec}}$.

(II) The lognormal x_i values in models $\mathbf{M2}$ are much more dispersed and skewed than the values in $\mathbf{M1}$, but the simulation results still support the asymptotic theory when $n_2 = 50$, although not when $n_2 = 20$. The substitution-estimator based confidence intervals for Y in terms of $\hat{Y}_{\text{reg},2}$ have coverage probability far too small when the substitution variance estimator is used. In Table 4.3, for each type of Y-estimator there is a pronounced tendency for the substitution variance estimator to underestimate the true (empirical) variance, and for the bootstrap estimator to overestimate.

Table 4.5 clarifies that the extreme behavior of variance estimators under models $\mathbf{M2}$ occurs partly because the predictors and y_i are dispersed and skewed, and partly because the size-variable used in PPS weighting shares these properties. The cases with suffix $\mathbf{W}$ in this Table are the same as in Table 4.3. The cases with suffix $\mathbf{E}$ have (x_i, z_i) the same as in Table 4.3, but the conditional variances of y_i given x_i have the constant value of 100; and with this change, the erratic behavior of SD estimators disappears. However, when the conditional y_i variances are as in the basic model $\mathbf{M2}$ but the PPSWR sampling is done *un*weighted, *i.e.*, with all z_i replaced by 1, the empirical and bootstrap SD estimators track each other and are very large, while the substitution variance estimator is too low by dramatic factors of $1/2$ to

$3/4$. This weird phenomenon applies equally to all three Y estimators. (However, an unweighted-sampling variant in model **M1** does not materially change the results from those shown in Table 4.2.)

(III) One objective of the simulations was to learn whether there can ever be any Mean-squared Error (MSE) benefit in using $\hat{Y}_{\text{reg},1}$ rather than $\hat{Y}_{\text{reg},2}$, without which there would be little motivation for $\hat{Y}_{\text{dec}}$. Indeed, the large-sample Theorems say that the main large-sample variance term is always optimal for $\hat{Y}_{\text{reg},2}$ (whether because it is the same as for $\hat{Y}_{\text{reg},1}$ under the null hypothesis or because it is strictly better under model (2.7) with distinct slopes). However, we indicated following Theorem 3, in the bound (2.9), that $\hat{Y}_{\text{reg},1}$ can have smaller second-order MSE than $\hat{Y}_{\text{reg},2}$, and the **H0** columns of Tables 4.2 and 4.3 do show a small but consistent SD advantage for $\hat{Y}_{\text{reg},1}$ *versus* $\hat{Y}_{\text{reg},2}$, an advantage which is more pronounced in **M2**. This advantage disappears under the fixed alternative **M1.H1** but interestingly, not under **M2.H1**. The slight but real conditional MSE advantage for $\hat{Y}_{\text{reg},1}$ when the substratum slopes are very close to equality is discussed further by Slud (2012).

The estimators $\hat{Y}_{\text{reg},1}, \hat{Y}_{\text{reg},2}, \hat{Y}_{\text{dec}}$ considered here are of regression type, and it may be of interest to compare their MSE behavior in the simulated populations with that of the simpler Horvitz-Thompson estimator $\hat{Y}_{\text{HT}}$ in (1.1). All of these estimators are nearly unbiased, so that MSE's are essentially the same as variances, and a comparison of the third and fifth columns of Table 4.4 shows that the $\hat{Y}_{\text{HT}}$ variances are considerably larger than those of $\hat{Y}_{\text{dec}}$. The difference is least pronounced with the larger sample sizes, but even there is 30-55%. The advantage of $\hat{Y}_{\text{dec}}$ is still very pronounced in model **M2**, where model variances and distributional skewness are larger, but less so than in model **M1**.

(IV) The definition of $\hat{Y}_{\text{dec}}$ contains the arbitrary nominal significance level τ, which in all tables other than Table 4.4 was taken to be 0.05. As the large-sample theory suggests, the properties of the decision-based estimator fall between those of $\hat{Y}_{\text{reg},1}$ and $\hat{Y}_{\text{reg},2}$, and larger values of τ make $\hat{Y}_{\text{dec}}$ more often equal to $\hat{Y}_{\text{reg},1}$. As can be seen from comparison of columns 6 and 7 of Table 4.4, the choice τ = 0.20 seems in the simulated models to lead to very slightly smaller SD of $\hat{Y}_{\text{dec}}$ under model **M1**, but in model **M2** the SD is if anything larger at the smaller sample sizes. The conclusion is weak because the differences are quite small compared to the differences between SD's from one frame population to another. Our preference is to let smaller τ dictate the frequent pooling of substrata except when there are pronounced differences in estimated slope between the substrata. This finding that larger significance levels τ do not improve performance of $\hat{Y}_{\text{dec}}$ differs from the finding of Saleh (2006) that larger significance levels are highly beneficial in other preliminary-testing contexts.

(V) Table 4.6 gives information about the variability across frame populations of SD estimators for the Y estimators. The bootstrap variance estimators appear less susceptible to variation across frame populations, because the bootstrap averaging stabilizes them. The key finding in this table seems to be that the variability across frame populations is moderate except in the unweighted **M2** setting, where it is remarkably large. This seems to account for the extreme inflation of variances under **M2.U** seen in Table 4.5.

(VI) In many bootstrap applications with approximately normally distributed statistics, failure of coverage of normal-theory-based confidence intervals due to nonnormality of the bootstrapped statistic can be mitigated by using the bootstrap percentile (BP) intervals (Shao and Tu 1995, Section 4.1). In the present simulations, Table 4.4 (columns 4 and 6) gives the coverage percentages of BP intervals for $\hat{Y}_{\text{dec}}$ in settings where Tables 4.2 and 4.3 give the coverages of normal-theory CI's based on the bootstrap-estimated SD. For whatever reason, the tables show that the normal-theory coverage CP_B tends systematically to be slightly below nominal and yet slightly larger than the BP interval coverage CP_{BP}. Thus, our simulations indicate the preference in this setting for the simpler interval $\hat{Y}_{\text{dec}} \pm 1.96 \cdot \widehat{\text{SD}}_B$.

(VII) It remains to draw lessons from the simulations with real government-census data in Section 4.3. The first necessary comment is that the spread and skewness of the full-time payroll predictors x_i and the total-payroll size-variable z_i are very large, much more like the lognormal models **M2** than the gamma models **M1**. Table 4.8 indicates (in column 5) a consistent MSE advantage for $\hat{Y}_{\text{reg},2}$ over $\hat{Y}_{\text{reg},1}$ except in the CA Special-district case, although the difference is small in the CO Special-district and the FL, GA and PA Subcounty cases. It is notable in almost all of these examples that the bootstrap SD estimator for $\hat{Y}_{\text{dec}}$ is more accurate than the substitution-formula estimator, despite the rather small numbers of sampled and unsampled non-SR units and (in several cases, as shown in Table 4.7) relatively large numbers of SR units. The substitution SD estimates are consistently too small while the bootstrap estimates are usually slightly high (*i.e.*, generally $\widehat{\text{SD}}_S < \widehat{\text{SD}} < \widehat{\text{SD}}_B$). The relative error of $\widehat{\text{SD}}_B$ *versus* $\widehat{\text{SD}}$ is no more than about 5% in these examples, except in the cases (AL, CO, GA) where there are particularly few non-sampled non-SR units in the large-unit substratum.

The large-unit substrata in ASPEP usually have small total frame population and often have relatively large numbers of SR units. While we have seen in these simulations that this does not quite invalidate inferences drawn with $\hat{Y}_{\text{reg},1}, \hat{Y}_{\text{reg},2}$ or $\hat{Y}_{\text{dec}}$, these statistics have distributions rather different from those of large-sample theory, and perhaps future substratum splits should allow slightly larger large-unit substrata for well-behaved statistical inferences.

More broadly, the simulation results indicate that the decision-based estimator with interval estimator defined from bootstrap variances is well-behaved and can be recommended except in extremely dispersed and skewed populations or in populations with large-unit sample sizes less than 20-25.

Acknowledgements

This paper describes research and analysis of the authors, and is released to inform interested parties and encourage discussion. Results and conclusions are the authors' and have not been endorsed by the Census Bureau. We would like to thank three referees and an associate editor for helpful comments and suggestions which improved the paper. Jun Shao's research was partially supported by the NSF Grant DMS-1007454.

Appendix

Proof of Theorem 1. Under PPS sampling, $\pi_i = n_j p_{ij}$ for unit $i \in U_j$, and on each with-replacement draw, the sampled index $i_t \in U_j, t = 1, \ldots, n_j$ has $P(i_t = i) = p_{ij}$ for each $i \in U_j$. By calculating the means and variances (under repeated sampling) of $\hat{N}_j, \hat{X}_j, \hat{Y}_j, N_j^{-1}\sum_{i\in S_j} x_i y_i/\pi_i$ and $N_j^{-1}\sum_{i\in S_j} x_i^2/\pi_i$, we find the variances to be of order n_j^{-1} by means of the limits in (C2)-(C3) and the bounds in (C4). The assertions in part (a) follow directly.

For assertion (b), we have by definition of $\hat{\beta}$ that

$$\hat{\beta} = \frac{\sum_{j=1}^{2}\sum_{i\in S_j}\left(x_i - \hat{\mu}_{x,j} + \hat{\mu}_{x,j} - \left(\hat{X}_1 + \hat{X}_2\right)/\left(\hat{N}_1 + \hat{N}_2\right)\right) y_i/\pi_i}{\sum_{j=1}^{2}\sum_{i\in S_j}\left(x_i - \hat{\mu}_{x,j} + \hat{\mu}_{x,j} - \left(\hat{X}_1 + \hat{X}_2\right)/\left(\hat{N}_1 + \hat{N}_2\right)\right)^2/\pi_i}$$

$$= \frac{N^{-1}\left(\sum_{j=1}^{2}\hat{\beta}_j\hat{\sigma}_{xj}^2\hat{N}_j + \left(\hat{N}_1\hat{N}_2/\left(\hat{N}_1 + \hat{N}_2\right)\right)(\hat{\mu}_{x1} - \hat{\mu}_{x2})\left(\hat{\mu}_{y1} - \hat{\mu}_{y2}\right)\right)}{N^{-1}\left(\sum_{j=1}^{2}\hat{\sigma}_{xj}^2\hat{N}_j + \left(\hat{N}_1\hat{N}_2/\left(\hat{N}_1 + \hat{N}_2\right)\right)(\hat{\mu}_{x1} - \hat{\mu}_{x2})^2\right)},$$

from which the equality (2.1) in (b) follows immediately by substituting the limits in part (a) along with the limits $N_j/N \to \omega_j$.

Let Σ_N be the block diagonal matrix with two diagonal blocks D_{N_1} and D_{N_2}, and for $j = 1, 2,$ let

$$\begin{aligned}\Omega_{1j} &= \frac{1}{N_j\sqrt{n_j}}\sum_{i\in S_j}\left(\frac{1}{p_{ij}} - N_j\right), \quad \Omega_{2j} = \frac{1}{N_j\sqrt{n_j}}\sum_{i\in S_j}\left(\frac{x_i}{p_{ij}} - X_j\right),\\ \Omega_{3j} &= \frac{1}{N_j\sqrt{n_j}}\sum_{i\in S_j}\left(\frac{y_i}{p_{ij}} - Y_j\right), \quad \Omega_{4j} = \frac{1}{N_j\sqrt{n_j}}\sum_{i\in S_j}\frac{x_i - \mu_{x,j}}{p_{ij}}\left(y_i - \alpha_j - \beta_j x_i\right).\end{aligned} \tag{A.1}$$

Since S_1 and S_2 are independent, $\{\Omega_{k1}\}_{k=1}^4$ is independent of $\{\Omega_{k2}\}_{k=1}^4$. Note that, here and throughout this proof, sums over $i \in S_j$ used to define $\hat{X}_j, \hat{Y}_j, \Omega_{kj}$, and variance estimators should be understood as sums *with multiplicity* in view of the with-replacement PPS sampling framework. Condition (C4) makes Liapounov's Central Limit Theorem applicable to show that

$$\Sigma_N^{-1/2}\,[\Omega_{11},\Omega_{21},\Omega_{31},\Omega_{12},\Omega_{22},\Omega_{32}]^T \to_d N(0, I_6),\ \Omega_{4j} \to_d N\left(0, \sigma_{xe,j}^2\right), \tag{A.2}$$

where I_6 is the 6×6 identity matrix, and $\sigma_{xe,j}^2$ is given in the statement of (d). The limits defining the asymptotic variances in (A.2) exist according to (C3).

Proof of (c). It is straightforward to check from the definition that

$$\begin{pmatrix}\hat{\beta}_j - \beta_j\\ \hat{\alpha}_j - \alpha_j\end{pmatrix} = \frac{1}{\hat{N}_j\hat{\sigma}_{xj}^2}\sum_{i\in S_j}\begin{pmatrix} x_i - \hat{\mu}_{xj}\\ \hat{\sigma}_{xj}^2 - (x_i - \hat{\mu}_{xj})\hat{\mu}_{xj}\end{pmatrix}\frac{y_i - \alpha_j - \beta_j x_i}{\pi_i}.$$

Since it was established in (a) that $\hat{\sigma}_{xj}^2 \to_P \sigma_{xj}^2$ and $\hat{N}_j/N_j \to_P 1,$ it follows that the limiting distribution of $\sqrt{n_j}\left(\hat{\beta}_j - \beta_j\right)$ is the same as that of

$$\sqrt{n}_j \left(N_j \sigma_{xj}^2\right)^{-1} \sum_{i \in S_j} \left(x_i - \mu_{xj}\right)\left(y_i - \alpha_j - \beta_j x_i\right) \big/ \pi_i ,$$

which is clearly the same as that of $\sigma_{xj}^{-2}\Omega_{4j}$ in (A.1). The first assertion of (c) follows immediately from (A.2). The consistency of $\hat{\sigma}_{xe,j}^2$ follows by noting by (a) that

$$\hat{\sigma}_{xe,j}^2 - N_j^{-2} \sum_{i \in S_j} \frac{\left(x_i - \mu_{xj}\right)^2}{\pi_i p_{ij}} \left(y_i - \alpha_j - \beta_j x_i\right)^2 \to_P 0. \tag{A.3}$$

The second term on the left-hand side of (A.3) has PPS with-replacement sampling variance calculated to be bounded by $1/n_j$ according to (C4), and by (C3) has expectation converging to $\sigma_{xe,j}^2$.

Proof of (d). From (1.2) and (a), $\left(\hat{Y}_{\text{reg},2} - Y\right)/N \to_P 0,$ which can also be seen from the representation

$$\begin{aligned}
\sqrt{n}\left(\hat{Y}_{\text{reg},2} - Y\right)\big/N &= \frac{\sqrt{n}}{N}\sum_{j=1}^{2}\left[\frac{N_j \hat{Y}_j}{\hat{N}_j} - Y_j + \hat{\beta}_j\left(X_j - \frac{N_j \hat{X}_j}{\hat{N}_j}\right)\right] \\
&= \frac{\sqrt{n}N_1^2}{\sqrt{n_1}N\hat{N}_1}\left[\left(-\bar{Y}_1 + \hat{\beta}_1\bar{X}_1\right)\Omega_{11} - \hat{\beta}_1\Omega_{21} + \Omega_{31}\right] \\
&\quad + \frac{\sqrt{n}N_2^2}{\sqrt{n_1}N\hat{N}_2}\left[\left(-\bar{Y}_2 + \hat{\beta}_2\bar{X}_2\right)\Omega_{12} - \hat{\beta}_2\Omega_{22} + \Omega_{32}\right] \\
&= d_{n1}^T\bar{\Omega}_1 + d_{n2}^T\bar{\Omega}_2,
\end{aligned}$$

where the second equality follows from the notational definitions of Ω_{kj} along with $\pi_i = n_j p_{ij}$, $\hat{Y}_j = \sum_{i \in S_j} y_i/\pi_i$, $\hat{X}_j = \sum_{i \in S_j} x_i/\pi_i$, and the third from

$$d_{nj} = \frac{\sqrt{n}N_j^2}{\sqrt{n_j}N\hat{N}_j}\left[-\bar{Y}_j + \hat{\beta}_j\bar{X}_j, -\hat{\beta}_j, 1\right]^T, \quad \bar{\Omega}_1 = [\Omega_{11}, \Omega_{21}, \Omega_{31}]^T, \quad \bar{\Omega}_2 = [\Omega_{21}, \Omega_{22}, \Omega_{32}]^T.$$

By (A.2), $\bar{\Omega}_1 = O_p(1)$ and $\bar{\Omega}_2 = O_p(1)$. By condition (C2), $d_{nj}^T = a_{2j}^T + o_p(1)$. Therefore, by (A.2), condition (C3) and the delta method,

$$\sqrt{n}\left(\hat{Y}_{\text{reg},2} - Y\right)\big/N = a_{21}^T\bar{\Omega}_1 + a_{22}^T\bar{\Omega}_2 + o_p(1) \to_d N\left(0, \sigma_2^2\right),$$

where the asymptotic variance $\sigma_2^2 = \sum_{j=1}^2 a_{2j}^T D_j a_{2j}$ is consistently estimated by

$$\frac{n}{N^2}\sum_{j=1}^{2}\sum_{i \in S_j}\frac{1}{\pi_i^2}\left(y_i - \hat{\beta}_j x_i - \left(\hat{Y}_j - \hat{\beta}_j\hat{X}_j\right)\big/\hat{N}_j\right)^2,$$

which agrees with formula (9) of Cheng *et al.* (2010). The proof that $\sqrt{n}\left(\hat{Y}_{\text{reg},1} - Y\right)/N \to_d N\left(0, \sigma_1^2\right)$ is similar.

Proof of Theorem 2. By Theorem 1 conclusion (c),

$$\sqrt{n}\left(\hat{\beta}_2 - \hat{\beta}_1 - \beta_2 + \beta_1\right) \rightarrow_d N\left(0, \sum_{j=1}^{2} \sigma^2_{xe,j} \Big/ \left(\varphi_j^2 \sigma_{xj}^4\right)\right). \tag{A.4}$$

The conclusion (2.4) in (a) of this Theorem follows immediately.

In the proof of Theorem 1, we showed that

$$\sqrt{n}\left(\hat{Y}_{\text{reg},2} - Y\right) \Big/ N = a_{21}^T \bar{\Omega}_1 + a_{22}^T \bar{\Omega}_2 + o_p(1), \tag{A.5}$$

where the constant vectors a_{kj} (and μ_x, μ_y) were defined in Theorem 1 (d). Similarly,

$$\sqrt{n}\left(\hat{Y}_{\text{reg},1} - Y\right) \Big/ N = a_{11}^T \bar{\Omega}_1 + a_{12}^T \bar{\Omega}_2 + o_p(1). \tag{A.6}$$

When (2.3) holds, $\beta_j = \beta$ (by Theorem 1 (b)) and $\mu_y - \beta\mu_x = \sum_{j=1}^{2} \omega_j \left(\mu_{yj} - \beta_j \mu_{xj}\right) = \mu_{y2} - \beta\mu_{x2}$, so that $a_{1j} = a_{2j}$ for $j = 1, 2$. It follows immediately from (A.5)-(A.6) that $\sqrt{n}\left(\hat{Y}_{\text{reg},1} - \hat{Y}_{\text{reg},2}\right) \Big/ N \rightarrow_P 0$, and therefore that the estimators $\hat{Y}_{\text{reg},k}$ have the same asymptotic distribution, which was shown to be normal in Theorem 1 (d). Finally, the definition of $\hat{Y}_{\text{dec}}$ implies that $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},1} \text{ or } \hat{Y}_{\text{reg},2}\right) = 1$ and (A.5)-(A.6) imply

$$\sqrt{n}\left(\hat{Y}_{\text{dec}} - Y\right) \Big/ N = a_{21}^T \bar{\Omega}_1 + a_{22}^T \bar{\Omega}_2 + o_p(1), \tag{A.7}$$

which completes the proof of (2.5) in (a).

Proof of (b). If $\beta_1 \neq \beta_2$, then (A.4) implies that $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right) \rightarrow 1$, *i.e.*, that the t-test for equality of $\hat{\beta}_j$ rejects with certainty in the limit. Then (A.7) continues to hold, and the asymptotic distribution of $\hat{Y}_{\text{dec}}$ is still as same as that of $\hat{Y}_{\text{reg},2}$.

Proof of Theorem 3. In this Theorem, the hypotheses (C2)-(C4) are replaced by the assumptions that the iid triples (y_i, x_i, z_i) satisfy moment conditions and the model (2.7). The assertions in (C2)-(C4) are then results holding with probability tending to 1 with large n, N which are established with the aid of the (strong) law of large numbers.

Beyond the conclusions of Theorems 1-2, it remains to show that $\hat{Y}_{\text{reg},2}$ has a smaller asymptotic variance than $\hat{Y}_{\text{reg},1}$. Let $\vartheta = (\vartheta_1, \vartheta_2)$ and

$$F_j(\vartheta) = [-\vartheta_1, -\vartheta_2, 1] D_j [-\vartheta_1, -\vartheta_2, 1]^T .$$

According to the definition of σ_1^2 and σ_2^2 in (2.2), it suffices to show that $F_j(\vartheta)$ has its minimum value at $\vartheta = \left(\alpha_j, \beta_j\right)$. We now prove this for $j = 1$. The proof for $j = 2$ is similar. Let $m_{ii'}$ be the (i, i') element of D_1. Since D_1 is symmetric and positive definite under condition (C3), $m_{12} = m_{21}$ and there exists a unique $\theta^* = \left(\theta_1^*, \theta_2^*\right)$ such that $F_1\left(\theta^*\right) = \min_\vartheta F_1(\vartheta)$ and $\partial F_1(\vartheta)/\partial\vartheta^T|_{\vartheta=\theta^*} = 0$. This implies that θ^* is the solution to the following two equations:

$$m_{11}\vartheta_1 + m_{12}\vartheta_2 = m_{13}, \quad m_{12}\vartheta_1 + m_{22}\vartheta_2 = m_{23} \tag{A.8}$$

Therefore, it suffices to show that $\theta^* = (\alpha_1, \beta_1)$. Since D_1 is positive definite, the equation system (A.8) has a unique solution. By the definition of D_1,

$$\begin{aligned} m_{11}\alpha_1 + m_{12}\beta_1 &= \lim_{N_1\to\infty}\frac{1}{N_1^2}\left[\sum_{i\in U_1}\left(\frac{1}{p_{i1}} - N_1\right)^2 p_{i1}\alpha_1 + \sum_{i\in U_1}\left(\frac{1}{p_{i1}} - N_1\right)\left(\frac{x_i}{p_{i1}} - X_1\right)p_{i1}\beta_1\right] \\ &= \lim_{N_1\to\infty}\frac{1}{N_1^2}\left[\sum_{i\in U_1}\left(\frac{1}{p_{i1}} - N_1\right)(\alpha_1 - N_1\alpha_1 p_{i1} + \beta_1 x_i - \beta_1 p_{i1} X_1)\right], \end{aligned}$$

and

$$\begin{aligned} m_{13} &= \lim_{N_1\to\infty}\frac{1}{N_1^2}\left[\sum_{i\in U_1}\left(\frac{1}{p_{i1}} - N_1\right)\left(\frac{y_i}{p_{i1}} - Y_1\right)p_{i1}\right] \\ &= \lim_{N_1\to\infty}\frac{1}{N_1^2}\left[\sum_{i\in U_1}\left(\frac{1}{p_{i1}} - N_1\right)(\alpha_1 + \beta_1 x_i + \varepsilon_i - N_1\alpha_1 p_{i1} - \beta_1 p_{i1} X_1)\right] \\ &= \lim_{N1\to\infty}\frac{1}{N_1^2}\left[\sum_{i\in U_1}\left(\frac{1}{p_{i1}} - N_1\right)(\alpha_1 - N_1\alpha_1 p_{i1} + \beta_1 x_i - \beta_1 p_{i1} X_1)\right], \end{aligned}$$

where the last equality follows from the assumption that ε_i is independent of x_i and z_i and has mean 0 and a finite variance, and each of the sequences z_i, $1/z_i$, and x_i/z_i is iid with finite expectation. Therefore, $m_{11}\alpha_1 + m_{12}\beta_1 = m_{13}$. Similarly one proves that $m_{12}\alpha_1 + m_{22}\beta_2 = m_{23}$. Therefore, (α_1, β_1) is the unique solution to equation system (A.8), *i.e.*, $F_1(\vartheta)$ achieves its minimum value at $\vartheta = (\alpha_1, \beta_1)$. Hence, $\sigma_2^2 < \sigma_1^2$. This finishes the proof of Theorem 3.

References

Bancroft, T., and Han, C.-P. (1977). Inference based on conditional specifications: A note and a bibliography. *International Statistical Review*, 45, 117-127.

Cheng, Y., Corcoran, C., Barth, J. and Hogue, C. (2009). An estimation procedure for the new public employment survey design. Washington, DC: American Statistical Association. *Survey Research Methods Section*, American Statistical Association, 3032-3046.

Cheng, Y., Slud, E. and Hogue, C. (2010). Variance estimation for decision-based estimators with application to the annual survey of public employment and payroll. *Government Statistics Section of the American Statistical Association*. Vancouver: American Statistical Association.

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

Fuller, W.A. (2009). *Sampling Statistics*. New York: John Wiley & Sons, Inc.

Isaki, C., and Fuller, W. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Rao, J.N.K., and Ramachandran, V. (1974). Comparison of the separate and combined ratio estimators. *Sankhyā*, C, 36, 151-156.

Saleh, A.K. Md. (2006). *Theory of Preliminary Test and Stein-type Estimation, with Applications*. Hoboken: Wiley-Interscience.

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

Shao, J., and Tu, D. (1995) *The Jackknife and Bootstrap.* New York: Springer.

Slud, E.V. (2012). Moderate-sample behavior of adaptively pooled stratified regression estimators. U.S. Census Bureau preprint.

Survey Methodology, June 2014
Vol. 40, No. 1, pp. 105-123
Statistics Canada, Catalogue No. 12-001-X

The influence of sampling method and interviewers on sample realization in the European Social Survey

Natalja Menold[1]

Abstract

This article addresses the impact of different sampling procedures on realised sample quality in the case of probability samples. This impact was expected to result from varying degrees of freedom on the part of interviewers to interview easily available or cooperative individuals (thus producing substitutions). The analysis was conducted in a cross-cultural context using data from the first four rounds of the European Social Survey (ESS). Substitutions are measured as deviations from a 50/50 gender ratio in subsamples with heterosexual couples. Significant deviations were found in numerous countries of the ESS. They were also found to be lowest in cases of samples with official registers of residents as sample frame (individual person register samples) if one partner was more difficult to contact than the other. This scope of substitutions did not differ across the ESS rounds and it was weakly correlated with payment and control procedures. It can be concluded from the results that individual person register samples are associated with higher sample quality.

Key Words: Sampling methods; Substitutions by interviewers; Non-observation errors.

1 Introduction

Biases in survey statistics are described by the total survey error models (Groves, Fowler, Couper, Lepkowski, Singer and Tourangeau 2004; Smith 2007). Total survey error results from two types of errors, which are referred to as observation errors and non-observation errors. This article focuses on cross-cultural comparability influenced by factors related to non-observation errors, that is to say the failure of survey statistics to adequately represent the target population. These types of errors – also called representation errors – result from differences between the obtained net sample (number of survey participants, Haeder and Lynn 2007) and the target population.

Previous research in cross-cultural contexts has revealed highly pronounced differences in response rates between countries (Billiet, Phillipsen, Fitzgerald and Stoop 2007; Couper and de Leeuw 2003; de Heer 1999; de Heer and Israis 1992; de Leeuw and de Heer 2002; Hox and de Leeuw 2002; Johnson, O'Rourke, Burris and Owens 2002; Stoop 2005; Symons, Matsuo, Beullens and Billiet 2008), differences in field procedures (Billiet *et al.* 2007; Kohler 2007; Kreuter and Kohler 2009; Smith 2007; Stoop 2005; Symons *et al.* 2008) and differences in sampling methods (Lynn, Haeder, Gabler and Laaksonen 2007). The latter refer to procedures for constructing sampling frames and selecting sample elements. All differences listed may impede cross-cultural comparability. In particular, cross-cultural comparability might be limited due to varying sampling methods to obtain a probability sample since standardising sampling methods is restricted by local availability of sampling frames, by their quality and usability, and by the survey budget (Lynn *et al.* 2007).

Lynn *et al.* (2007) addressed design effects and the sample sizes necessary to achieve comparability of net samples in the ESS. In doing so, they focused on sample selection prior to the field work stage. However, comparability of samples could also be influenced by interviewers during their work in the

[1] Natalja Menold, GESIS - Leibniz Institute for the Social Sciences, Survey Design and Methodology, P.O.Box 12 21 55, D-68072 Mannheim. E-Mail: natalja.menold@gesis.org.

field. Interviewers' degree of freedom in substituting sampled individuals with persons who are not sampled (individuals who are easy to contact and are cooperative) differs based on sampling methods (Hoffmeyer-Zlotnik 2006; Kohler 2007; Sodeur 2007). "Field substitution occurs when a non-responding unit is replaced by a substitute (reserve) unit during the field work stage of the survey process" (Vehovar 1999, page 335). The substitutions addressed by Vehovar are legitimate substitutions that are allowed by protocol. In contrast, this article addresses illegitimate substitutions (referred to below simply as "substitutions") which occur without permission. According to the AAPOR (2003), deliberate substitutions made by interviewers represent a kind of falsification.

The aim of this article is to show whether the effect of interviewers, which is assumed to be associated with substitutions, varies across different sampling methods that are used to obtain probability samples in a cross-cultural context. In addition, it will be shown whether payment, control procedures, the data collector (institution that carries out data collection in the field) or time factors are associated with this interviewer effect. The results may help surveyors when deciding upon sampling methods – a highly relevant cost factor in surveys – and when deciding upon methods to foster interviewer motivation to not substitute. The results are also important for research on interviewer behaviour since they hint at errors associated with interviewer influence on cross-cultural comparability.

The next section (Section 2) provides the theoretical and empirical background of the study presented in this article. The hypotheses are described in Section 3. Section 4 provides information on the procedure and the method used for analysing the influence of the interviewer. The results are presented in Section 5. Finally, Section 6 discusses the results and provides conclusions.

2 Theoretical and empirical background

Substitutions may arise during tasks that interviewers conduct prior to the interview. Interviewers (1) build a sampling frame, for example by generating lists of addresses in surveys; then they (2) elicit cooperation with selected units (addresses, dwellings, households) and they also select individuals for the interview from these units. Finally interviewers (3) elicit cooperation from sampled individuals (Groves *et al.* 2004). In the case of different sampling methods interviewers perform different tasks as described below (figure 2.1).

The first sampling method refers to individual person register samples (denoted below as PRS). Official population registers of individuals are used as sampling frames for PRS. The selection of individuals is conducted prior to the field work stage, thus reducing interviewers' tasks to simply obtaining cooperation from sampled individuals (figure 2.1). In the case of PRS interviewers may influence non-response (*e.g.*, Couper and Groves 1992; de Leeuw and Hox 1996; Durrant, Groves, Staetsky and Steele 2010), but in a theoretical sense they have no influence on the sampling frame or on the selection of sample elements. This level of interviewer impact on non-representation in the case of PRS is shown by the arrow in figure 2.1.

However, as shown for example by Groves *et al.* (2004) selected elements (individuals in the case of PRS) may differ in terms of the probability of being contacted by an interviewer (contactability) and the probability of obtaining survey participation when a contact is given (cooperation). For example, it has been found that people living in urban areas or individuals who are young, single, without children, better educated and socially active are more difficult to contact (Stoop 2004). In contrast, older respondents, women, less-educated people and socially isolated individuals refuse cooperation more often than others (Dohrenwend and Dohrenwend 1968; Stoop 2004; Williams, Irvine, McGinnis, McMurdo and Crombie 2007). If difficulties arise when trying to contact and obtain cooperation from target persons then substitutions may occur. For example, Koch (1995) reported the number of substitutions in a survey in which PRS was used.

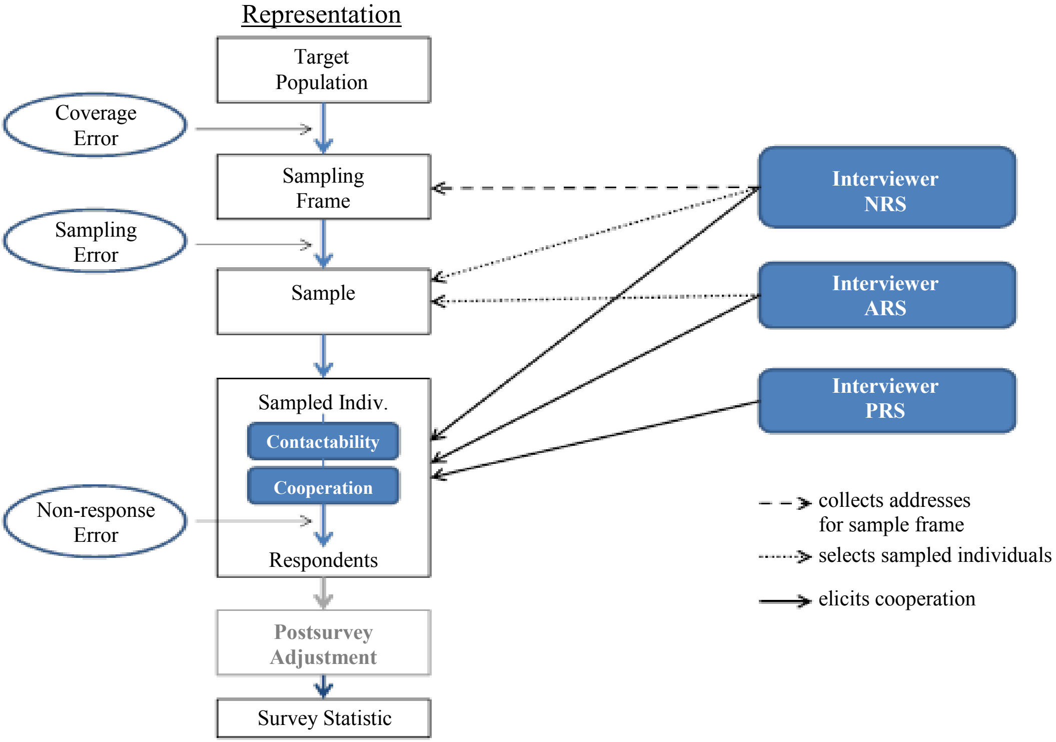


Figure 2.1 Interviewers' tasks in different sampling methods related to coverage, sampling and non-response errors. The path related with representation is adapted from Groves *et al.* 2004, page 48

The next sampling methods discussed are address/household register samples (ARS). In the case of ARS lists of households or addresses are employed as sampling frames. Households or addresses are selected by survey offices prior to the field work stage. In this case interviewers perform tasks two and three (see above): they contact selected units and select individuals for the interview if more than one eligible individual is living at one unit. Interviewers may deliberately deviate from the random selection rules and in this way they can have a negative impact on the selected sample (figure 2.1). Since interviewers have more freedom in selecting sampled individuals with ARS than with PRS, it is assumed that interviewer impact associated with substitutions is higher for ARS than for PRS (figure 2.2). Moreover, in the case of ARS the result of sample selection is not known beforehand and is therefore harder to control than in the case of PRS.

Non-register samples (NRS), in which neither lists of individuals nor lists of addresses/households are available as sampling frames, are described as the third sampling method. These include Random Route Samples (*e.g.*, Arber 2002; ESS Sampling Plans), and Address Listing and Sampling (ALS). In the case of

NRS interviewers themselves generate a sampling frame by listing or collecting addresses within a randomly selected geographical area. Interviewers have to strictly follow instructions concerning the procedures for collection of addresses. Interviewers perform this task in addition to selecting individuals at one address, as described for ARS, and to contacting and obtaining cooperation, as just described for both PRS and ARS (figure 2.1). With NRS interviewers can influence not only sample selection but also the sampling frame. An interviewer can deviate from instructions and chose only addresses where he or she expects to contact the target person and obtain cooperation. Substitutions are particularly likely to occur when using a procedure (Random Route) in which the interviewer conducts the interview at an address which they choose in an area following prescriptions regarding collecting of addresses and the direction of the route through the area. Another type of NRS is more restricted since the interviewer lists the addresses in a geographical area, but the actual selection is conducted by a coordination team (Address Listing and Sampling, ALS). The selected addresses are subsequently assigned to a different interviewer who then conducts the interviews. The degree of interviewer freedom in the case of ALS appears to be similar to that of ARS. However, the instructions for listing or collecting addresses can be ambiguous in the case of both types of NRS (Schnell, Hill and Esser 2011). Therefore, interviewers have more freedom to substitute with NRS than with ARS (figure 2.2).

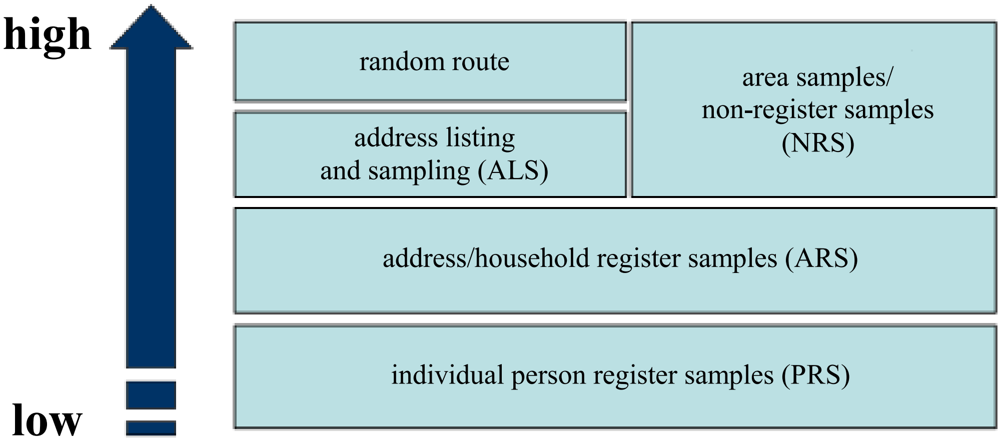


Figure 2.2 Degree of interviewers' freedom to substitute in various sampling methods

Deviations in obtained net samples, which are associated with deviations from the rules of random sample selection (*e.g.*, substitutions), can be empirically analysed using a method developed by Sodeur (1997). This method works by defining a subpopulation with one fixed and known parameter, then the statistics representing this parameter are observed in a subsample that is defined in a corresponding manner. The more the observed statistic deviates from the population parameter, the stronger the error from non-observation is. This article considers the gender ratio of heterosexual couples, which is known to be a 50/50 population parameter. Within the limits of random fluctuation any sample from the population of heterosexual couples should produce a proportion of males of around 50%. Significant deviations from this level of 50% indicate deviations from sample selection standards, such as through substitutions (see section 4.2 for details).

Using this method Sodeur (2007) and Hoffmeyer-Zlotnik (2006) found that deviations from the true parameter of a 50/50 gender ratio vary between different rounds of the German General Social Survey (ALLBUS), which also involved different sampling methods. The authors found that males who are difficult to contact are interviewed less often than females (since the males are the breadwinners in households with children). Aside from contactability, differences in cooperation between the partners can play a role (Hoffmeyer-Zlotnik 2006). If partners are retired they have comparable levels of contactability but they may differ in terms of cooperation. The retired man, now at home, feels responsible for providing information to an interviewer about the household (as its "head"). The woman can refuse to participate since the man likes to cooperate. An interviewer who contacts such households may interview men instead of women in order to avoid refusals (Hoffmeyer-Zlotnik 2006).

Kohler (2007) found larger deviations from the parameter of a 50/50 gender ratio in NRS samples as compared to other sampling methods in six cross-cultural surveys (Eurobarometer 62.1, European Quality of Life Survey EQLS'03, ESS 2002, ESS 2004, European Value Study 1999, International Social Survey Program, ISSP 2002). Unfortunately, the sampling method effect obtained by Kohler (2007) was survey-specific. The most poorly designed samples – area samples with a NRS – were used predominantly in one survey (EQLS). The differences that Kohler found between a random route and other sampling methods could thus be due to differences between EQLS and other surveys. Other researchers (Hoffmeyer-Zlotnik 2006; Souder 1997) addressed the effect of sampling methods on interviewer impact associated with substitutions while considering only a single German national survey; the results of this research are not applicable to cross-cultural contexts. Therefore, it is important to address the question regarding the relationship between sampling methods and interviewer impact associated with substitutions in cross-cultural surveys. Additionally, it is important to consider other explanatory factors which can affect substitutions. Substitutions made by interviewers can be affected not only by sampling methods but also by field procedures related to interviewer motivation to produce accurate survey data. Therefore, substitutions may vary based on the data collector (Hoffmeyer-Zlotnik 2006; Sodeur 1997; 2007) or controls used in a survey (Kohler 2007). Controls imply that a sample element is re-contacted to confirm the outcome produced by an interviewer. In addition to controls, methods of providing payment to interviewers can impact their performance. If interviewers are paid per completed interview they bear the risk of high costs due to long distances between selected addresses, numerous contact attempts or long interview times (Sodeur 2007). Consequently, a change of data collector, controls and payment should be considered when analysing the interviewer impact associated with substitutions. Apart from these factors it is interesting to see how such interviewer impact varies across time. For example, considering time in a cross-cultural context helps to indicate whether this interviewer impact is country-specific. A country-specific interviewer impact should be stable in a country across different survey rounds, even if the sampling method is changed.

3 Research hypotheses

If interviewer impact in terms of substitutions is operationalized using Sodeur's method, it is expected that it can be observed in survey statistics as deviations from the 50/50 gender ratio in subsamples with respondents as representatives of heterosexual couples. This interviewer impact is expected to differ in terms of varying contactability or cooperation on the part of the partners. Partners differ in terms of contactability in households comprised of couples with young children in which men are the breadwinners (Hoffmeyer-Zlotnik 2006; Sodeur 2007; Stoop 2004). If interviewers apply substitutions the proportion of men should be significantly lower than the true value (50%) in such households, since men are more difficult to contact than women. This pattern changes when taking retired partners into consideration. Here, as previously discussed by Hoffmeyer-Zlotnik (2006), partners have comparable levels of contactability but they may differ in terms of their cooperation. For subsamples of retired couples, the proportion of males is expected to be significantly higher than 50% in the case of interviewer impact due to substitutions. The hypothesis describing this impact in different types of households is:

Hypothesis 1 (H1): Deviations from the true gender ratio (50/50) vary depending on the type of household. In households comprised of couples with (young) children the proportion of males is lower than 50%, while in households comprised of retired partners this proportion is higher than 50%.

As was shown in section 2, different sampling methods can be associated with varying degrees of freedom on the part of interviewers to substitute (figure 2.2). Therefore, the following differences between the sampling methods are expected:

Hypothesis 2 (H2): Deviations from the true gender ratio (50/50) vary depending on the sampling method used in a survey. They are lowest with PRS and highest with NRS samples.

If deviations from the population parameter are caused by interviewers deviating from prescribed standards, then they should vary with the sampling method used or with the type of household, which in turn is associated with varying levels of contactability or cooperation on the part of the partners. The deviations should be stable across time when keeping sampling methods constant. However the deviations can correlate with interviewer payment and control procedures, or with the data collectors, who are expected to differ in terms of practices related to interviewers' work motivation.

Hypothesis 3 (H3): Except for changes in sampling method in a country, deviations from the 50/50 gender ratio are independent of the influence of other changes over time. Thus, they do not vary across different survey rounds. However, interviewer payment, control procedures and change of data collector are expected to correlate with deviations from the 50/50 gender ratio.

4 Methods

4.1 Data

To isolate any effects due to the sampling method from other survey-specific effects, one can use data from a multi-country survey in which the various countries applied different sampling methods. Many rounds of a survey should be available in order to be able to consider the time effect. Therefore, data from rounds one to four of the ESS were used (European Social Survey Round 1-4 Data 2011). The ESS was conducted with between 20 and more than 30 countries, which differ in terms of their sampling methods. In addition, the ESS sets high standards for survey organisations, such as strict random sampling and extended contact procedures, or regarding field control procedures (Koch, Blom, Stoop and

Kappelhof 2009; Philippens and Billiet 2004). The effectiveness of the standards used in the ESS has been demonstrated by Kohler (2007), who showed that round 1 of the ESS had the fewest deviations from the 50/50 gender ratio compared to other surveys. In addition, the ESS has consistently improved data collection methods (Koch *et al.* 2009). Furthermore, the ESS provides detailed documentation of sampling procedures as well as data collection (*cf.* ESS Documentation Reports), which allows for operationalization of variables of interest.

4.2 Method for evaluating interviewer impact

The method developed by Sodeur (1997) was used for the analysis. This method helps to evaluate the net sample quality in probability samples. The quality of the random sample selection has often been examined by means of other statistics available in a country (external criteria). However, these external statistics are often unknown, leading Sodeur to suggest the use of internal criteria - that is to say use of information from the net sample only. Sodeur (1997) describes the method as consisting of the following steps: (1) separating a subsample from the entire sample to focus on respondents as representatives of heterosexual couples: the partners should live together within one household and both partners should belong to the target population of the survey; (2) defining units which should be dropped from the subsample: singles, partners not living together within a household and households with other relatives who belong to the target population. Then, step three entails (3) defining a survey statistic – *e.g.*, the percentage of males – as the dependent variable which should be compared with the population parameter.

An analysis to determine the causes of deviations from the population parameter – for example, interviewer behaviour – requires additional specifications in steps 1 and 2 to ensure that interviewer behaviour (conceptually) varies with the contactability or cooperation of target persons. Such specifications have been made in this article in terms of definitions of different type of households (see hypothesis H1), whose selection is described in section 4.3.

The true 50/50 gender ratio in heterosexual couples is not related to any other gender ratios, such as that of the total population of residents in a country. Therefore, as Kohler (2007) argues, this gender ratio cannot be affected by any sort of measurement errors and it is unaffected by the household size since the analysis is restricted to two persons in the household and both persons belong to the target population.

Sodeur’s method has advantages over other methods since no additional external information or data are required. However, Sodeur’s method requires that the characteristics defined for selecting subsamples are known not only regarding the respondents but also regarding their partners (*e.g.*, gender of the partner). In addition, there should not be systematic gender differences in terms of refusal behaviour (differential refusal), which may occur even if interviewers work honestly. In practice, females have been found to be more reluctant than males (Pickery and Loosveldt 2002; Schnauber and Daschmann 2008; Stoop 2004; Williams *et al.* 2007). That also seems to be the case in the ESS, in which females were found to refuse more often than males. The author’s own analysis of ESS1-ESS4 data from contact forms shows that 30.3% males and 37.9% females refused cooperation in the ESS1 (in some countries no data regarding the gender variable was provided; therefore the missing data was 32.4%). In the ESS2 there were 30.8% males and 37.9% females who refused cooperation (31.3% data missing); in the ESS3 33.8% males and 39.0% females (27.2% of data missing) refused cooperation and in the ESS4 there were 38.4% males and 45.8% females who refused (with a reduced 15.8% of data missing). Therefore, males being present in a subsample of ESS data less than 50% of the time can be plausibly explained by substitutions, while a frequency of males higher than 50% can be alternatively explained by differential refusal. However, if the percentage of males varied with a sampling method – as expected in hypothesis H2 – it would be hard to explain such a result only by differential refusal, which seems to be a quite stable feature.

4.3 Procedure

The following section provides a description of the procedures used for testing hypotheses H1 to H3. First, separation of subsamples from the entire ESS sample is described. Deviations d from the true 50/50 gender ratio in a subsample represent the dependent variable for all subsequent analyses. The values of d are compared between different households to test hypothesis H1. Second, operationalization of the "sampling method" variable (to test H2) is described. Third, hypothesis H3 is related to the variables time, change of data collector, payment and interviewer controls, whose operationalization is described in the last section. H2 and H3 were tested with the help of a Multivariate Analysis of Covariance (MANCOVA) with subsequent individual Analyses of Covariance (ANCOVA) in which the sampling method was used as the independent variable and the ESS round, change of data collector, payment bonus and interviewer controls were used as covariates.

Separation of subsamples

The ESS target population "contains in each country persons 15 years or older who are resident within private households, regardless of nationality and citizenship, language or legal status" (*e.g.*, ESS-1 2002 Documentation Report, page 2). Respondents ($n = 88{,}375$) who live together with a partner of the opposite sex who is 15 years or older were selected from the total ESS1-ESS4 sample ($n = 184{,}988$). This reduced the data base of the analysis to about half of the entire sample. However, such a selection is required to ensure the expected percentage of males of 50%.

Three household types were distinguished among the selected subsample: couples with children aged between 0-6 ($< 7; n = 18{,}791$), couples with children between 7-14 ($n = 53{,}651$), and couples in which both partners are of retirement age (retirees, $n = 15{,}933$). To determine retirement age the current state pension age in each country was used (see appendix). The first two groups with children were formed since it was assumed that differences in contactability between partners are particularly high in these households. For the third group it was assumed that gender differences in contactability are fairly modest, while men and women might differ in terms of cooperation.

The fact that men are breadwinners within the two subsamples containing households with children is supported by the author's own analysis using data from the ESS. Upon looking at respondents' activities within the last seven days in households with children younger than seven, it was shown that 58% of males and 42% of females were in paid work. In terms of respondents' partners, 64% of males and 36% of females were in paid work. Similar results were found for respondents in households with children aged 7 to 14 (for respondents 54% males and 46% females were in paid work and in terms of partners there were 60.5% males and 39.5% females). For households with retired partners it was found that 80.6% of respondents were retired, 11.5% did housework and 1.3% was sick or disabled on a long term basis. With respect to respondents' partners, 84.4% were retired, 17% did housework and 2.1% were sick or disabled on a long-term basis.

Categorisation of sampling methods

Documentation Reports for each ESS round (European Social Survey (2011): ESS 1-4 Documentation Reports) were used to classify the sampling methods. Table 4.1 summarises the main characteristics of the sampling methods used in the ESS. Table 4.2 shows which sampling methods were used in each country in each of the rounds. For more details on selection procedures in the ESS see the Documentation Reports or Lynn *et al.* (2007).

For ARS it is important to see how multi-dwelling units at one address are dealt with since in this case interviewers also manage the situation. The survey documentation described this for only a few countries (Ireland, Israel, the Netherlands, and the United Kingdom). In Ireland, for example, interviewers listed the households and selected one of them using the Kish Grid (Kish 1965).

In Austria a NRS method was applied to only 50% of the sample, while the other 50% was selected based on an ARS. Since using NRS can lead to more substitutions compared to using only ARS it can be expected that the results in Austria are more similar to the results in countries with NRS than with ARS. Therefore, the author assigned Austria to NRS.

Table 4.1
Sampling methods in the countries of the EES (rounds 1-4)

	individual person register sample	**address/household register sample**	**non-register sample**
sampling frame	reliable lists of residents	reliable lists of addresses/households	regional areas (no lists of residents, addresses or households)
stage 1: Selection of PSUs			
definition of a unit	regional clusters, areas, municipalities	electoral sections, postal code areas	regional clusters, areas, municipalities
process of selection	systematic random sampling	systematic random sampling	systematic random sampling
result	community, municipality	electoral section, postal code section	geographical areas, municipalities
stage 2: Selection of households	not applicable		
definition of a unit		a household, an address	a household/dwelling unit
process of selection		simple or systematic random sampling	random route/ALS simple random sampling
result		addresses of households	a household/address/dwelling unit
stage 3: Selection of individuals			
definition of the unit	target person	target person	target person
process of selection	simple or systematic random sampling	random selection by interviewer by Kish Grid or last birth day method	random selection by interviewer by Kish Grid or last birth day method
result	name and address of sampled individuals	sampled individuals	sampled individuals

Table 4.2
Classification of ESS countries with respect to sampling methods

ESS round	individual person register sample (PRS)	address/household register sample (ARS)	non-register sample (NRS)
ESS 1	BE, DE, HU, NO, PL, SI, DK, FI, SE	Address: IE, IT, NL, GB, CH Household: CZ, LU, ES	FR, GR, PT, AT
ESS 2	BE, DE, HU, NO, PL, SI, DK, FI, SE, ES, EE, IS, SK	Address: IE, NL, GB, CH Household: LU, TR	FR, GR, PT, AT, CZ, UA
ESS 3	BE, DE, NO, PL, SI, DK, FI, SE, ES, EE, SK	Address: IE, NL, GB, CH, LV Household: CY, BG, HU	FR, PT, AT, UA, RU, RO
ESS 4	BE, DE, HU, NO, PL, SI, DK, FI, SE, ES, EE	Address: IE, NL, GB, CH, IL, LV Household: CZ, CY, LT, GR, KRO, TR, BG	FR, PT, SK, UA, RU, RO

Note Romania is not included in the ESS integrated data file; no information on sampling method has been provided for Italy in Documentation Reports ESS2-ESS4. Countries are labelled according to ISO 3166-1, see the appendix.

The kind of NRS method used in a country has rarely been described in the documentation. In the ESS1 it is evident that an ALS was used only in Greece. Usage of an ALS is described for the Czech Republic and Slovakia in later rounds. In the ESS4 Ukraine, Russia and Portugal report a procedure comparable to the ALS. However, in these countries interviewers (and not the offices) selected a fixed number of units from the lists generated by other interviewers.

Explanatory variables

Information related to the particular ESS round was used as a variable to control for the time effect. The Documentation Reports were used to obtain information related to other explanatory variables; change of data collector, payment and interviewer controls. Whether a country changed the data collector between rounds is shown in the appendix. Concerning payment, it was found that the ESS mainly employed a method involving payment per completed interview. An hourly rate of payment was used only in a few countries that also used PRS (in ESS1-2 in Norway and Sweden, as well as in ESS3-4 in Norway and Finland). Therefore, there was only a small variation in payment methods, and a corresponding data analysis was not possible. However, payment of bonuses varied across countries and rounds. Therefore, this information was used to generate a dichotomous control variable (bonus payment: yes/no).

Two variables are used to describe control procedures: the number of eligible sample elements selected for controls divided by the number of eligible sample elements (ratio selected), and the number of confirmed outcomes divided by the number of sample elements selected for controls (ratio confirmed). The first variable describes the number of controls in a country, while the second describes the effectiveness of these controls. The “ratio selected” varies between 10% for PRS, 13% for NRS and 16% for ARS. The “ratio confirmed” is somewhat higher for NRS ($M = 75.21$, $SD = 24.81$) than for the other two sampling methods (PRS: $M = 61.89$, $SD = 31.95$; ARS: $M = 66.49$; $SD = 32.56$).

5 Results

5.1 Differences between household types

Firstly, the results for testing hypothesis H1 are presented. This hypothesis expects deviations from the 50/50 gender ratio to vary according to the type of household. Figure 5.1 shows the differences (d) between the actual percentage of males and the expected true value of 50% in three subsamples. A 95% confidence interval (CI) was used to control for random fluctuation. As the expected proportion of men is $p = 0.5$, the variance averages $0.25/n$, whereby n is the number of cases in the subsample in a country. The 95% CI was calculated as follows (*cf.* Kohler 2007, page 59):

$$CI = 0.5 \pm 1.96 \times \sqrt{0.25/2}.$$

Figure 5.1 shows that for both subsamples covering households with children, significant values of d are negative in the majority of cases, meaning that the proportion of males in these subsamples is less than 50% (as expected by H1). Most of these d-values were approx. 10% or higher. Lower (approx. 5%) significant positive (unexpected) d-values are seen for three countries in which PRS was used (in the ESS1 in Belgium and Norway, in the ESS2 in Finland). However, these differences were not discernible in other rounds.

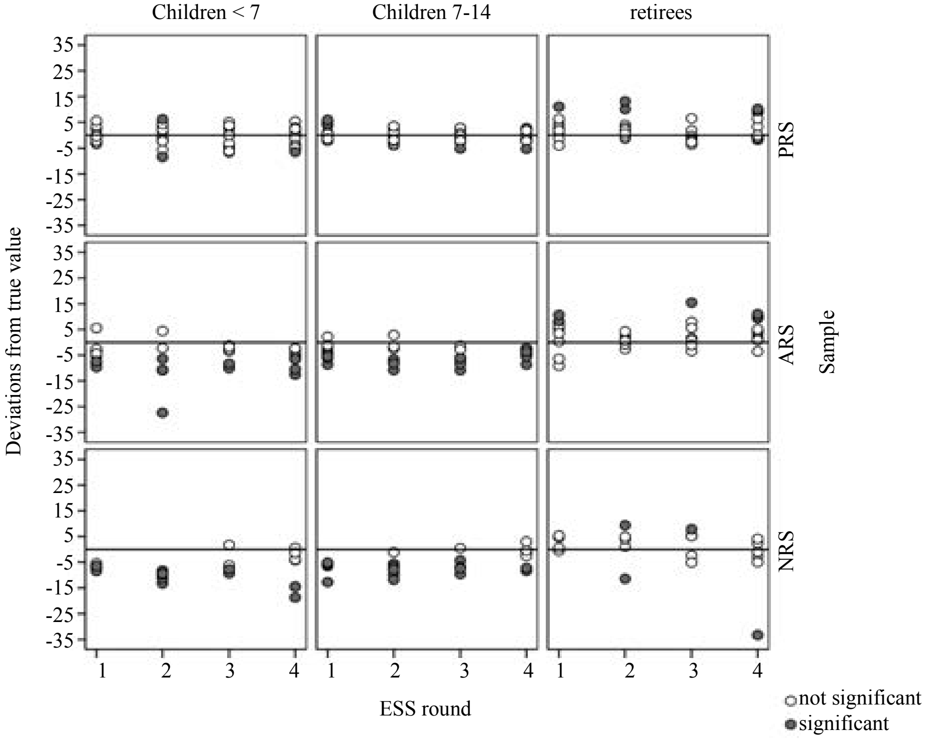


Figure 5.1 Deviations from the true value of 50% (d) in the percentages of males in different types of households of ESS1-ESS4

Regarding the results for the subsamples covering households with partners of retirement age (retirees) it is possible to see significantly high d-values (approx. 10% or higher) with the expected direction (positive, or that is to say the percentages of males are higher than 50%) for some countries across all sampling methods (in the ESS1 in Norway, the Czech Republic and the Netherlands; in the ESS 2 in Norway, Poland and France; in the ESS3 in Cyprus and Russia and in the ESS4 in Germany, Hungary, Cyprus and the United Kingdom). Interestingly, the proportion of men is markedly lower than 50% in Slovakia in the ESS4 (as low as approx. 33%) and in Portugal in the ESS2 (as low as approx. 11%). This result can be explained by specific patterns of role division between the partners. Here the woman appears to represent the household, even if the man is at home.

To summarise, significant deviations from true value in different types of households were mainly in line with the expectations of hypothesis H1.

5.2 Differences between sampling methods

The effect of sampling methods (as expected by H2) was tested by means of MANCOVA. The d-values for the three types of households (three isolated subsamples) were considered as values of three dependent variables, which were simultaneously analysed in the MANCOVA. Since the MANCOVA is based on an analysis of means the absolute values of d were considered. Otherwise it would not have been possible to take into account differences with an unexpected direction, which would also be associated with the effect of sampling methods. Since most of the differences were negative in the subsamples with children, the absolute d-values represent a proportion of men that is lower than 50%. With respect to the subsamples with partners of retirement age, it should be taken into account that the proportion of men was not only higher than 50% but also lower than 50% in Portugal (ESS2) and in Slovakia (ESS4). In addition, significant and non-significant differences are considered in order to enable a comparison between countries with low and high d-values.

The MANCOVA revealed a high significant multivariate effect of the factor "sampling method" (Wilks Lambda (WL) $F_{(6,174)} = 6.87, p < 0.001,$ effect size $\eta^2 = 0.21$). In contrast, no significant results for explanatory variables were found ($p > 0.10,$ max $\eta^2 = 0.04$). In order to consider d-values in different household types univariate analyses of covariance (ANCOVAs) were employed. Variance homogeneity – as a presupposition for an ANCOVA – is given according to the Levene test in the subsample with retirees, and also according to the $F_{\max}$ test in the subsamples with children. Significant mean differences of d-values between sampling methods were found using the ANCOVAs in both subsamples with children (table 5.1). The variances explained in the ANCOVAs for these subsamples are quite high (see R^2 in table 5.1). On average the lowest d-value can be seen for the PRS, while the highest d-value is seen for the NRS (table 5.1 and figure 5.2). However, post-*hoc* single comparisons using subsamples with children show significant differences only between PRS and the other two sampling methods (table 5.2). Also, no remarkable differences in d-values were found between the countries with ALS and with Random Route samples.

Overall, the results show that hypothesis H2 can be partially supported if households with children are considered.

Table 5.1
Descriptive statistics $(M(SD))$ and results of the ANCOVAs for comparison of d in the three types of household

	types of household			
	children <7	**children 7-14**	**retirees**	***n* (countries)**
Sampling method (treatment)				
PRS	3.28(2.07)	2.21(1.37)	3.34 (3.35)	43
ARS	6.61(4.98)	4.87 (2.74)	4.94(3.83)	31
NRS	7.85 (4.4)	5.92 (3.55)	5.78(6.87)	21
$F(df_1 = 2, df_2 = 88)$	14.52***	20.9***	1.93	
Time: ESS round				
1	4.49(2.67)	4.08(2.94)	4.75(3.22)	22
2	6.92(5.73)	4.33(3.3)	3.63(3.71)	24
3	4.78(3.04)	4.02(3.18)	3.74(3.44)	23
4	5.23(4.41)	3.24(2.22)	5.39(6.66)	26
$F(df_1 = 1, df_2 = 88)$	0.00	1.18	0.02	
Payment bonus				
no	5.83(4.37)	4.41(3.10)	4.10(3.73)	54
yes	4.78(3.99)	3.23(2.52)	4.81(5.49)	41
$F(df_1 = 1, df_2 = 88)$	0.57	3.21^{+}	0.49	
Ratio controlled				
$F(df_1 = 1, df_2 = 88)$	0.11	0.51	1.09	
Ratio confirmed				
$F(df_1 = 1, df_2 = 88)$	3.11^{+}	0.11	0.00	
R^2	0.22	0.31	0.01	

Notes $***p < 0.001$, $^{+}p < 0.10$.

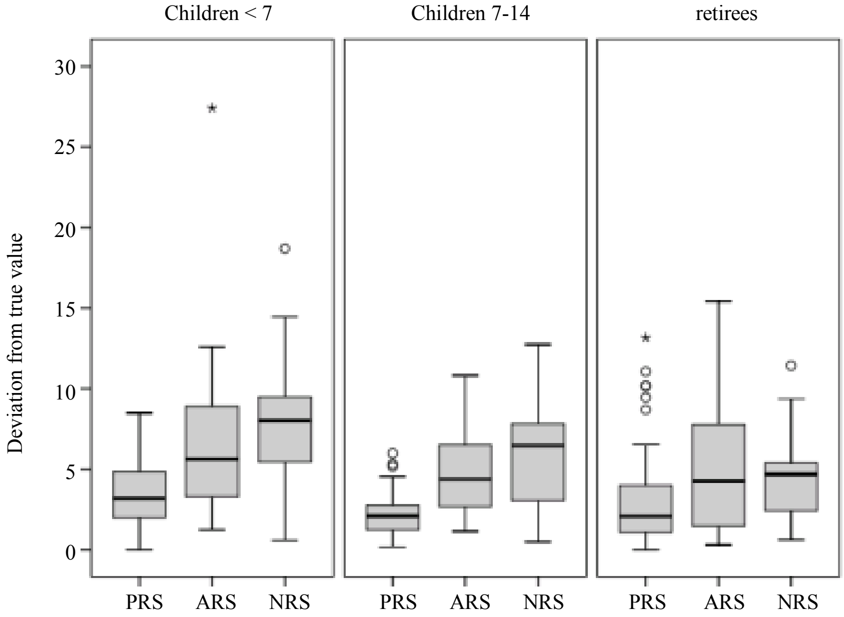


Figure 5.2 Box plots for absolute d-values are shown for different sampling methods in the three types of households

Table 5.2
Mean differences of $(d\,(SD))$ between sampling methods in subsamples with children

	children <7	children 7-14
differences between		
PRS and ARS	-3.34 (0.89)**	-2.66 (0.58)**
PRS and NRS	-4.58 (1.0)**	-3.71 (0.65)**
ARS and NRS	-1.24 (1.07)	-1.05 (0.7)

Note $**\ p < 0.01$. Single post *hoc* tests with Bonferroni correction.

5.3 The effect of explanatory variables

The effect of explanatory variables was analysed to test hypothesis H3, which expects deviations from the 50/50 gender ratio to be stable across time and to correlate with payment, interviewer controls and change of data collector.

Some countries in the ESS changed their sampling method procedures and/or data collector between the rounds (see appendix). The results showed that neither multivariate effects $\left(\text{WL}\,F_{(3,85)} = 0.81, p > 0.10\right)$ nor univariate effects are significant for the change of data collector. Thus, table 5.1 presents the ANCOVA results without this variable. If the "change of data collector" is included in the analyses, then the effect of the variable "ratio confirmed" is no longer significant, but this does not impact the effects of any of the other variables. This result shows that a change of data collectors may correlate with control procedures. The differences in d-values across the ESS rounds are not significant either, neither within multivariate analysis $\left(\text{WL}\,F_{(3,86)} = 0.51, p > 0.10\right)$ nor within the univariate analyses (for the latter see table 5.1).

Table 5.1 shows that in subsamples with children d-value means are lower if a payment bonus is used as compared to when it is not used. However, this difference is significant only on a 10% level $(p < 0.10)$ and only in households with older children. Hence, this result shows that payment methods may play a role, thereby reducing deviation from the true value in the case of higher payments.

Regarding control procedures, the number of controls ("ratio selected") is not related to the value of d (table 5.1). The success rate in controls ("ratio confirmed") is related to the value of d in the subsample with children younger than seven years old. This relationship is negative $(B = -0.06; SE = 0.04)$, meaning that the lower the confirmed control rates are, the higher the values of d are. However, this relationship is also significant only at a 10% level.

Concerning hypothesis H3, it has been shown that the effect of sampling methods is independent of the time effect. The results support the expectation of H3 concerning interviewer payment and controls. However, the results for these variables show that these effects are only weak and they can only be found in some household types.

6 Summary and conclusion

The results of the present study show that significant deviances from the population parameter (50% males) were seen in many ESS countries and that these were associated with the contactability or cooperation of partners in heterosexual couples (support for hypothesis H1). The magnitude of these deviances was found to differ between sampling methods when partners also differed in terms of their contactability (in subsamples with children). Thus, hypothesis H2 was partially supported. In subsamples with children, PRS was associated with the best data quality since the lowest deviances from the population parameter were found with this sampling method. However, the results for subsamples with retired partners show that highly pronounced deviances are also possible in the case of PRS.

The results for subsamples with children are in line with the explanation that interviewer behaviour related to substitutions is involved, since as expected deviations from the population parameter varied with the degree of interviewer freedom in influencing sample realization. Comparable results were reported by Sodeur (1997) and Kohler (2007). It is less plausible to explain interviewing males less than 50% of the time by differential refusal since in such cases the proportion of males is expected to be higher than 50%. Next, differential refusal is not expected to vary between different sampling methods. For retirees, interviewing males more than 50% of the time was found in several countries, but only in single rounds. This low stability of deviations from a 50% gender ratio can also be associated with interviewer impact instead of differential refusal, since the latter would be quite stable in a country over the period of time considered in the analysis. However, since the present study did not apply an experimental design it is important to address differential refusal and substitutions through further research in order to allow for better differentiation as well as for causal references.

Even though the deviations from a 50% population parameter varied in some countries across the rounds, overall their magnitude did not significantly change during the passage of time despite improved data collection procedures in the ESS (*cf.* Koch *et al.* 2009). Furthermore, deviations from the population parameter did not depend on the data collector nor were they country-specific.

The results also imply that interviewer payment and control procedures may reduce substitutions. However, it should be noted that only limited consideration of payment and control procedures was possible due to either their low variation in the data or limited information available in the survey documentation.

It should also be considered that the results presented here are based on specific subsamples and they cannot be used to generalise about the entire sample of the ESS. However, the "absence of a bias in the subsamples does not guarantee the absence of bias for the entire sample" (Kohler 2007, page 55). In addition, an analysis focused on special groups can often be of interest (*e.g.*, what are the opinions of parents with children or of employed people).

The results of the present study imply that PRS is associated with higher sample quality, meaning a lower non-representation bias in cross-cultural surveys than with other sampling methods. This is shown more clearly by the current study than by previous studies. Analyses using ALLBUS data by Sodeur (2007) and Hoffmeyer-Zlotnik (2006) only compared several rounds of one survey in a single country (Germany); in the analysis by Kohler (2007) a sampling method effect was confounded with the survey effect (see section 1). This has been avoided in the analysis presented here.

In conclusion, significant deviations from the population parameter, which appear to be associated with substitutions by interviewers, were observed in many countries of the ESS. In order to decrease this interviewer impact it is preferable to use sampling methods, such as PRS, with which the interviewers' degree of freedom in selecting respondents and in influencing sample quality is reduced. In addition, survey procedures that increase interviewers' motivation to produce accurate survey data are highly relevant and should be addressed by further research as well as by survey practices.

Appendix

Coding of ESS countries, change in sampling method and data collector, and state pension age for males and females in each country.

Country coding: ISO 3166-1	country	change sampling method (between rounds)	change data collector: between rounds	state pension age	
				males	females
BE	Belgium		1-2; 2-3	65	65
BG	Bulgaria		3-4	63	60
DE	Germany			65	65
DK	Denmark			65	65
EE	Estonia		2-3; 3-4	63	60
ES	Spain	ARS-NRS (1-2)	2-3	65	65
FI	Finland			65	65
HU	Hungary	PRS – ARS (2-3) and back (3-4)	2-3	62	62
NO	Norway			67	67
PL	Poland			65	60
SE	Sweden			65	65
SI	Slovenia			63	60
SK	Slovakia	PRS – NRS (3-4)	2-3; 3-4	62	59
CH	Switzerland			65	64
CZ	Czech Republic	ARS-NRS (1-2)	1-2	65	62
CY	Cyprus		3-4	65	65
GB	United Kingdom		1-2; 3-4	65	60
GR	Greece	NRS-ARS (2-4)		65	60
IE	Ireland		3-4	65	65
IL	Israel			67	64
IT	Italy			65	60
LU	Luxembourg			65	65
NL	Netherlands			65	65
TR	Turkey			47	44
AT	Austria			65	60
FR	France			60	60
PT	Portugal			65	65
RU	Russian Federation			60	55
UA	Ukraine			60	55

Notes Sources for state pension age:
1) http://www.oecd-ilibrary.org/finance-and-investment/pensions-at-a-glance-2011_pension_glance-2011-en
2) http://ec.europa.eu/employment_social/missoc/db/public/compareTables.do
3) Israel: http://www.btl.gov.il/

References

AAPOR (2003). *Interviewer Falsification in Survey Research: Current Best Methods for Prevention, Detection and Repair of Its Effects*. (http://www.aapor.org/pdfs/falsification.pdf; May 14 2009).

Arber, S. (2002). Design samples. In *Researching Social Life*, (Ed., N. Gilbert), Thousand Oaks: Sage, 58-84.

Billiet, J., Phillipsen, M., Fitzgerald, R. and Stoop, I. (2007). Estimation of nonresponse bias in the European Social Survey: Using information from reluctant respondents. *Journal of Official Statistics*, 23, 135-162.

Couper, M.P., and Groves, R.M. (1992). The role of the interviewer in survey participation. *Survey Methodology*, 18, 2, 263-277.

Couper, M.P., and De Leeuw, E.D. (2003). Nonresponse in cross-cultural and crossnational surveys. In *Cross-Cultural Survey Methods*, (Eds., J.A. Harkness, F.J.R. van de Vijver and P.Ph. Mohler), New York: John Wiley & Sons, Inc., 157-177.

De Heer, W. (1999). International response trends: Results of an international survey. *Journal of Official Statistics*, 15, 129-142.

De Heer, W., and Israis, A.Z. (1992). Response Trends in Europe. Paper presented to the American Statistical Association, August 1992.

De Leeuw, E., and De Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In *Survey Nonresponse*, (Eds., R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little), New York: John Wiley & Sons, Inc., 41-54.

De Leeuw, E.D., and Hox, J. (1996). The effect of the interviewer on the decision to cooperate in a survey of the elderly. In *International Perspectives on Nonresponse*, (Ed., S. Laaksonen). Helsinki: Statistics Finland, 46-52.

Dohrenwend, B.S., and Dohrenwend, B.P. (1968). Sources of refusals in surveys. *The Public Opinion Quarterly*, 32(1), 74-83.

Durrant, G.B., Groves, R.M., Staetsky, L. and Steele, F. (2010). Effects of interviewer attitudes and behaviors on refusal in household surveys. *Public Opinion Quarterly*, 74, 1-36.

ESS1-ESS4 data from Contact forms. Bergen, European Social Survey Data Archive, Norwegian Social Science Data Services. Retrieved June 8, 2013 from: http://ess.nsd.uib.no/ess/round1/download.html.

European Social Survey Round 1-4 Data (2011). Data file edition ESS1-4e01.0_F1. Norwegian Social Science Data Services, Norway – Data Archive and distributor of ESS data.

European Social Survey (2011). ESS-4 2008 Documentation Report. Edition 4.0. Bergen, European Social Survey Data Archive, Norwegian Social Science Data Services.

European Social Survey (2011). ESS-3 2006 Documentation Report. Edition 3.3. Bergen, European Social Survey Data Archive, Norwegian Social Science Data Services.

European Social Survey (2011). ESS-2 2004 Documentation Report. Edition 3.3. Bergen, European Social Survey Data Archive, Norwegian Social Science Data Services.

European Social Survey (2011). ESS-1 2002 Documentation Report. Edition 6.2. Bergen, European Social Survey Data Archive, Norwegian Social Science Data Services.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E. and Tourangeau, R. (2004). *Survey Methodology*. New Jersey: Wiley.

Haeder, S., and Lynn, P. (2007). How representative can a multi-nation survey be? In *Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey*, (Eds., R. Jowell, C. Roberts, R. Fitzgerald and E. Gillian), London *et al.*: Sage, 33-52.

Hoffmeyer-Zlotnik, J.H.P. (2006). Stichprobenziehung in der Umfragepraxis. Die unterschiedlichen Ergebnisse von Zufallsstichproben in face-to-face umfragen. In *Stichprobenqualität in Bevölkerungsumfragen*, (Eds., F. Faulbaum and Ch. Wolf). Informationszentrum Sozialwissenschaften: Bonn, 19-36.

Hox, J.J., and De Leeuw, E.D. (2002). The Influence of interviewers' attitude and behavior on household survey nonresponse: An international comparison. In *Survey Nonresponse*, (Eds., R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little), New York: John Wiley & Sons, Inc., 103-118.

Johnson, T.P., O'Rourke, D., Burris, J. and Owens, L. (2002). Culture and survey nonresponse. In *Survey Nonresponse*, (Eds., R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little), New York: John Wiley & Sons, Inc., 55-69.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Koch, A. (1995). Gefälschte Interviews: Ergebnisse der Interviewerkontrolle beim ALLBUS 1994. *ZUMA-Nachrichten*, 36, 89-105.

Koch, A., Blom, A.G., Stoop, I. and Kappelhof, J. (2009). Data collection quality assurance in cross-national surveys at the example of the ESS. *Methoden Daten Analysen*, 3, 219-247.

Kohler, U. (2007). Surveys from inside: An assessment of unit nonresponse bias with internal criteria. *Survey Research Methods*, 1, 2, 55-67.

Kreuter, F., and Kohler, U. (2009). Analyzing contact sequences in call record data. Potential and limitations of sequence indicators for nonresponse adjustments in the European Social Survey. *Journal of Official Statistics*, 25, 203-226.

Lynn, P., Haeder, S., Gabler, S. and Laaksonen, S. (2007). Methods for achiving equivalence of samples in cross-national surveys: The European Social Survey Experience. *Journal of Official Statistics*, 1, 107-124.

Pickery, J., and Loosveldt, G. (2002). A multilevel multinomial analysis of interviewer effects on various components of unit non response. *Quality and Quantity*, 36, 427-437.

Philippens, M., and Billiet, J. (2004). Monitoring and evaluating nonresponse issues and fieldwork efforts in the European Social Survey. Paper presented at the *European Conference on Quality and Methodology in Official Statistics*. Mainz, Germany.

Schnauber, A., and Daschmann, G. (2008). States oder traits? Was beeinflußt die Teilnahmebereitschaft an telefonischen Interviews? *Zeitschrift Für Empirische Sozialforschung*, 2, 97-123.

Schnell, R., Hill, P.B. and Esser, E. (2011). *Methoden der Empirischen Sozialforschung.* München: R. Oldenbourg Verlag.

Smith, T.W. (2007). Survey nonresponse procedures in cross-national perspective: The 2005 ISSP non-response survey. *Survey Research Methods*, 1, 45-54.

Sodeur, W. (1997). Interne kriterien zur Beurteilung von Wahrscheinlichkeitsauswahlen. *ZA-Information*, 41, 58-82.

Sodeur, W. (2007). Entscheidungsspielräume von Interviewern bei der Wahrscheinlichskeitsauswahl. *Methoden Daten Analysen*, 1, 2, 107-130.

Stoop, I.A.L. (2004). Surveying nonrespondents. *Field Methods*, 16, 23-54.

Stoop, I.A.L. (2005). *The Hunt for the Last Respondent*. The Hague: Social and Cultural Planning Office.

Symons, K., Matsuo, H., Beullens, K. and Billiet, J. (2008). *Response Based Quality Assessment in the ESS – Round 3: An Update for 19 countries*. London: Centre for Comparative Social Surveys, City University.

Vehovar, V. (1999). Field substitution and unit nonresponse. *Journal of Oficial Statistics*, 2, 335-350.

Williams, B., Irvine, L., McGinnis, A.R., McMurdo, M.E.T. and Crombie, I.K. (2007). When “no” might not quite mean “no”; the importance of informed and meaningful non-consent: results from a survey of individuals refusing participation in a health-related research project. *BMC Health Services Research*, 7, 59.

Survey Methodology, June 2014
Vol. 40, No. 1, pp. 125-134
Statistics Canada, Catalogue No. 12-001-X

Bayesian multiple imputation for large-scale categorical data with structural zeros

Daniel Manrique-Vallier and Jerome P. Reiter[1]

Abstract

We propose an approach for multiple imputation of items missing at random in large-scale surveys with exclusively categorical variables that have structural zeros. Our approach is to use mixtures of multinomial distributions as imputation engines, accounting for structural zeros by conceiving of the observed data as a truncated sample from a hypothetical population without structural zeros. This approach has several appealing features: imputations are generated from coherent, Bayesian joint models that automatically capture complex dependencies and readily scale to large numbers of variables. We outline a Gibbs sampling algorithm for implementing the approach, and we illustrate its potential with a repeated sampling study using public use census microdata from the state of New York, U.S.A.

Key Words: Latent class; Log-linear; Missing; Mixture; Multinomial; Nonresponse.

1 Introduction

Many agencies collect surveys comprising large numbers of exclusively categorical variables. Inevitably, these surveys suffer from item nonresponse that, when left unattended, can reduce precision or increase bias (Little and Rubin 2002). To handle item nonresponse, one approach is multiple imputation (Rubin 1987), in which the agency fills in the missing items by sampling repeatedly from predictive distributions. This creates $M > 1$ completed datasets that can be analyzed or disseminated to the public. When the imputation models meet certain conditions (Rubin 1987, Chapter 4), analysts of the M completed datasets can make valid inferences using complete-data statistical methods and software. For reviews of multiple imputation, see Rubin (1996), Barnard and Meng (1999), Reiter and Raghunathan (2007), and Harel and Zhou (2007).

Multiple imputation typically is implemented via one of two strategies. The first is to posit a joint model for all variables and estimate the model using Bayesian techniques, usually involving data augmentation and Markov chain Monte Carlo (MCMC) sampling. Common joint models include the multivariate normal for continuous data and log-linear models for categorical data (Schafer 1997). The second strategy is to use approaches based on chained equations (Van Buuren and Oudshoorn 1999; Raghunathan, Lepkowski, van Hoewyk and Solenberger 2001; White, Royston and Wood 2011). The analyst estimates a series of univariate conditional models and imputes missing values sequentially with these models. Typical conditional models include normal regressions for continuous dependent variables and logistic or multinomial logistic regressions for categorical dependent variables.

As noted by Vermunt, Ginkel, der Ark and Sijtsma (2008) and Si and Reiter (2013), chained equation strategies are not well-suited for large categorical datasets with complex dependencies. For any conditional (multinomial) logistic regression, the number of possible models is enormous once one considers potential interaction effects. Carefully specifying each conditional model is a very

[1] Daniel Manrique-Vallier is Assistant Professor at the Department of Statistics, Indiana University, Bloomington, IN 47408. E-mail: dmanriqu@indiana.edu; Jerome P. Reiter is Mrs. Alexander Hehmeyer Professor of Statistical Science, Duke University, Durham, NC 27708-0251. E-mail: jerry@stat.duke.edu.

time-consuming task with no guarantee of a theoretically coherent set of models; indeed, for this reason many practitioners of chained equations use default settings that include main effects only in the conditional models. By excluding interactions, analysts risk generating completed datasets that yield biased estimates. We note that similar model selection difficulties plague approaches based on log-linear models.

To avoid these issues, Si and Reiter (2013) propose a fully Bayesian, joint modeling approach to multiple imputation for high-dimensional categorical data based on latent class models. The idea is to model the implied contingency table of the categorical variables as a mixture of independent multinomial distributions, estimating the mixture distributions nonparametrically with Dirichlet process prior distributions. Mixtures of multinomials can describe arbitrarily complex dependencies and are computationally expedient, so that they are effective general purpose multiple imputation engines. For example, Si and Reiter (2013) applied their models to impute missing values in 80 categorical variables in the Trends in International Mathematics and Science Study.

The approach of Si and Reiter (2013) does not deal with an important and prevalent complication in survey data: certain combinations of variables may not be possible *a priori*. These are called structural zeros (Bishop, Fienberg and Holland 1975). For example, in the United States it is impossible for children under age 15 to be married. Structural zeros also can arise from skip patterns in surveys. The imputation algorithms of Si and Reiter (2013), if applied directly, allow non-zero probability for structural zeros, which in turn biases estimates of probabilities for feasible combinations.

In this article, we present a fully Bayesian, joint modeling approach to multiple imputation of large categorical datasets with structural zeros. Our approach blends the latent class imputation model of Si and Reiter (2013) with the approach to handling structural zeros developed by Manrique-Vallier and Reiter (forthcoming 2014). Using simulations, we show that the approach generates multiply-imputed datasets that do not violate structural zero conditions and can have well-calibrated repeated sampling properties.

2 Bayesian latent class imputation model with structural zeros

Suppose that we have a sample of n individuals measured on J categorical variables. Each individual has an associated response vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$, whose components take values from a set of L_j levels. For convenience, we label these levels using consecutive numbers, $x_{ij} \in \{1, \ldots, L_j\}$, so that $\mathbf{x}_i \in \mathcal{C} = \prod_{j=1}^{J} \{1, \ldots, L_j\}$. Note that $\mathcal{C}$ includes all combinations of the J variables, including structural zeros, and that each combination $\mathbf{x}$ can be viewed as a cell in the contingency table formed by $\mathcal{C}$. Let $\mathbf{x}_i = (\mathbf{x}_i^{\text{obs}}, \mathbf{x}_i^{\text{mis}})$, where $\mathbf{x}_i^{\text{obs}}$ includes the variables with observed values and $\mathbf{x}_i^{\text{mis}}$ includes the variables with missing values. Finally, let $S = \{s_1, \ldots, s_C\}$, where $s_c \in \mathcal{C}$ and $c = 1, \ldots, C < |S|$, be the set of structural zero cells, *i.e.*, $\Pr(\mathbf{x}_i \in S) = 0$.

2.1 Latent class models

As an initial step, we describe the Bayesian latent class model without any concerns for structural zeros and without any missing data, *i.e.*, $\mathbf{x}_i = \mathbf{x}_i^{\text{obs}}$. This model is a finite mixture of product-multinomial distributions,

$$p(\mathbf{x} \mid \boldsymbol{\lambda}, \boldsymbol{\pi}) = f^{\mathrm{LCM}}(\mathbf{x} \mid \boldsymbol{\lambda}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \lambda_{jk}\left[x_j\right], \tag{2.1}$$

where $\boldsymbol{\lambda} = \left(\lambda_{jk}[l]\right)$, with all $\lambda_{jk}[l] > 0$ and $\sum_{l=1}^{L_j} \lambda_{jk}[l] = 1$. Here, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ with $\sum_{k=1}^{K} \pi_k = 1$. This model corresponds to the generative process,

$$x_{ij} \mid z_i \overset{\text{indep}}{\sim} \text{Discrete}_{1:L_j}\left(\lambda_{jz_i}[1], \ldots, \lambda_{jz_i}\left[L_j\right]\right) \text{ for all } i \text{ and } j \tag{2.2}$$

$$z_i \mid \boldsymbol{\pi} \overset{\text{iid}}{\sim} \text{Discrete}_{1:K}(\pi_1, \ldots, \pi_K) \text{ for all } i. \tag{2.3}$$

As notation, let $(\mathcal{X}, \mathcal{Z})$ be a sample of n variates obtained from this process, with $\mathcal{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and $\mathcal{Z} = (z_1, \ldots, z_n)$. For K large enough, (2.1) can represent arbitrary joint distributions for $\mathbf{x}$ (Suppes and Zanotti 1981; Dunson and Xing 2009). And, using the conditional independence representation in (2.2) and (2.3), the model can be estimated and simulated from efficiently even for large J.

For prior distributions on $\boldsymbol{\pi}$, we follow Si and Reiter (2013) and Manrique-Vallier and Reiter (forthcoming 2014). We have

$$\lambda_{jk}[\cdot] \overset{\text{indep}}{\sim} \text{Dirichlet}\left(\mathbf{1}_{L_j}\right) \tag{2.4}$$

$$\pi_k = V_k \prod_{h<k} (1 - V_h) \tag{2.5}$$

$$V_k \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha) \text{ for } k = 1, \ldots, K-1; V_K = 1 \tag{2.6}$$

$$\alpha \sim \text{Gamma}(0.25, 0.25) \tag{2.7}$$

The prior distributions in (2.4) are equivalent to uniform distributions over the support of the $J \times K$ multinomial conditional probabilities and hence represent vague prior knowledge. The prior distribution for $\boldsymbol{\pi}$ in (2.5)-(2.7) is an example of a finite-dimensional stick-breaking prior distribution (Sethuraman 1994; Ishwaran and James 2001). As discussed in Dunson and Xing (2009) and Si and Reiter (2013), it typically allocates $\mathcal{Z}$ to fewer than K classes, thereby reducing computation and avoiding over-fitting. For further discussion and justification of this model as an imputation engine, see Si and Reiter (2013).

2.2 Truncated latent class models

The latent class model in (2.1) does not naturally specify cells with structural zeros *a priori*, because it assumes a positive probability for each cell. Thus, to represent tables with structural zeros, we need to truncate the model so that

$$f^{\mathrm{TLCM}}(\mathbf{x} \mid \boldsymbol{\lambda}, \boldsymbol{\pi}, S) \propto 1\{\mathbf{x} \notin S\} \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \lambda_{jk}\left[x_j\right]. \tag{2.8}$$

As Manrique-Vallier and Reiter (forthcoming 2014) show, obtaining samples from the posterior distribution of parameters $(\boldsymbol{\lambda}, \boldsymbol{\pi})$, conditional on a sample $\mathcal{X}^1 = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, can be greatly facilitated by adopting a sample augmentation strategy akin to those in Basu and Ebrahimi (2001) and O'Malley and Zaslavsky (2008). We consider $\mathcal{X}^1$ to be the portion of variates that did not fall into the set S from a larger sample, $\mathcal{X}$, generated directly from (2.1). Let n_0, $\mathcal{X}^0$, and $\mathcal{Z}^0$ be the the (unknown) sample size, response vectors, and latent class labels for the portion of $\mathcal{X}$ that did fall into S. Using a prior distribution from Meng and Zaslavsky (2002), Manrique-Vallier and Reiter (forthcoming 2014) show that if $p(N) \propto 1/N$, where $N = n_0 + n$, the posterior distribution of $(\boldsymbol{\lambda}, \boldsymbol{\pi})$ under the truncated model (2.8) can be obtained by integrating the posterior distribution under the augmented sample model over $(n_0, \mathcal{X}^0, \mathcal{Z}^0, \mathcal{Z}^1)$.

In doing so, Manrique-Vallier and Reiter (forthcoming 2014) develop a computationally efficient algorithm for dealing with large sets of structural zeros when they can be expressed as the union of sets defined by *margin conditions*. These are sets defined by fixing some levels of a subset of the categorical variables, for example, the set of all cells such that $\{\mathbf{x} \in \mathcal{C} : x_3 = 1, x_6 = 3\}$. Manrique-Vallier and Reiter (forthcoming 2014) introduce a vector notation to denote margin conditions, which we use here as well. Let $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_J)$ where, for $j = 1, \ldots, J$, we let $\mu_j = x_j$ whenever x_j is fixed at some level and $\mu_j = *$ otherwise, where $*$ is special notation for a placeholder. Using this notation and assuming $J = 8$, the conditions that define the example set above ($x_3 = 1$ and $x_6 = 3$) correspond to the vector $(*, *, 1, *, *, 3, *, *)$. To avoid cluttering the notation, we use the vectors $\boldsymbol{\mu}$ to represent both the margin conditions and the cells defined by those margin conditions, determined from context.

2.3 Estimation and multiple imputation

We now discuss how the model in Section 2.2 can be estimated, and subsequently converted into a multiple imputation engine, when some items are missing at random. The basic strategy is to use a Gibbs sampler. Given a completed dataset $(\mathbf{x}^{\text{obs}}, \mathbf{x}^{\text{mis}})$, we take a draw of the parameters using the algorithm from Manrique-Vallier and Reiter (forthcoming 2014). Given a draw of the parameters, we take a draw of $\mathbf{x}^{\text{mis}}$ as described below.

Formally, the algorithm proceeds as follows. Suppose that the set of structural zeros can be defined as the union of C disjoint margin conditions, $S = \cup_{c=1}^{C} \boldsymbol{\mu}_c$, and that we use the priors for α, $\boldsymbol{\lambda}$ and $\boldsymbol{\pi}$ defined in Section 2.1. Given $\mathbf{x}_i = (\mathbf{x}_i^{\text{obs}}, \mathbf{x}_i^{\text{mis}})$ for $i = 1, \ldots, n$, the algorithm of Manrique-Vallier and Reiter (forthcoming 2014) samples parameters as follows.

1. For $i = 1, \ldots, n$, sample $z_i^1 \sim \text{Discrete}_{1:K}(p_1, \ldots, p_k)$, with $p_k \propto \pi_k \prod_{j=1}^{J} \lambda_{jk}[x_{ij}^1]$.
2. For $j = 1, \ldots, J$ and $k = 1, \ldots, K$, sample $\lambda_{jk[\cdot]} \sim \text{Dirichlet}\left(\xi_{jk1}, \ldots, \xi_{jkL_j}\right)$, with $\xi_{jkl} = 1 + \sum_{i=1}^{n} 1\{x_{ij}^1 = l, z_i^1 = k\} + \sum_{i=1}^{n_0} 1\{x_{ij}^0 = l, z_i^0 = k\}$.

3. For $k = 1, \ldots, K - 1$ sample $V_k \sim \text{Beta}\left(1 + \nu_k, a + \sum_{h=k+1}^{K} \nu_k\right)$ where $\nu_k = \sum_{i=1}^{n} 1\{z_i^1 = k\} + \sum_{i=1}^{n_0} 1\{z_i^0 = k\}$. Let $V_K = 1$ and make $\pi_k = V_k \prod_{h<k} (1 - V_h)$ for all $k = 1, \ldots, K$.
4. For $c = 1, \ldots, C$, compute $\omega_c = \Pr(\mathbf{x} \in \boldsymbol{\mu}_c \mid \boldsymbol{\lambda}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \prod_{\mu_{cj} \neq *} \lambda_{jk} \left[\mu_{cj}\right]$.
5. Sample $(n_1, \ldots, n_C) \sim NM(n, \omega_1, \ldots, \omega_C)$, where NM is the negative multinomial distribution, and let $n_0 = \sum_{c=1}^{C} n_c$.
6. Let $\kappa \leftarrow 1$. Repeat the following for each $c = 1, \ldots, C$.
   (a) Compute the normalized vector $(p_1, \ldots, p_K)$, where $p_k \propto \pi_k \prod_{j:\mu_{cj} \neq *} \lambda_{jk} \left[\mu_{cj}\right]$.
   (b) Repeat the following three steps n_c times:
      i. Sample $z_\kappa^0 \sim \text{Discrete}(p_1, \ldots, p_k)$,
      ii. For $j = 1, \ldots, J$ sample

$$x_{\kappa j}^0 \sim \begin{cases} \text{Discrete}_{1:L_j}\left(\lambda_{jz_\kappa^0}[1], \ldots, \lambda_{jz_\kappa^0}\left[L_j\right]\right) & \text{if } \mu_{cj} = * \\ \delta_{\mu_{jc}} & \text{if } \mu_{cj} \neq * \end{cases}$$

      where $\delta_{\mu_{cj}}$ is a point mass distribution at μ_{cj},
      iii. Let $\kappa \leftarrow \kappa + 1$.
7. Sample $\alpha \sim \text{Gamma}(a - 1 + K, b - \log \pi_K)$.

Having sampled parameters, we now need to take a draw of $\mathbf{x}^{\text{mis}}$. For $i = 1, \ldots, n$, let $\mathbf{m}_i = (m_{i1}, \ldots, m_{iJ})$ be a vector such that $m_{ij} = 1$ if component j in $\mathbf{x}_i$ is missing and $m_{ij} = 0$ otherwise. Assuming that data are missing at random, we need to sample only the components of each $\mathbf{x}_i$ for which $m_{ij} = 1$, conditional on the components for which $m_{ij} = 0$. Thus, we add an eighth step to the algorithm.

8. For $i = 1, \ldots, n$, sample $\mathbf{x}_i^{\text{mis}}$ from its full conditional distribution,

$$p\left(\mathbf{x}_i^{\text{mis}} \middle| \ldots\right) \propto 1\{\mathbf{x}_i \notin S\} \prod_{j:m_{ij}=1} \lambda_{jz_i}\left[x_{ij}\right]. \tag{2.9}$$

In the absence of structural zeros, the x_{ij} to be imputed are conditionally independent given z_i, making the imputation task a routine multinomial sampling exercise (Si and Reiter 2013). However, the structural zeros in S induce dependency between the components. Thus, we cannot simply sample the components independently of one another. A naive approach is to use an acceptance-rejection scheme, sampling repeatedly from the proposal distribution $p\left(\mathbf{x}^{\text{mis}*}\right) = \prod_{j:m_{ij}=1} \lambda_{jz_i}\left[x_{ij}\right]$ until obtaining a variate such that $\mathbf{x}^{\text{mis}*} \notin S$. However, when the rejection region is large or has a high probability, this approach can be very inefficient.

Instead we suggest forming additional Gibbs sampling steps, computing the conditional distributions of all missing components so that they can be sampled individually. Let $\operatorname{Rep}(\mathbf{x}_i, j, l)$ be the vector that results from replacing component j in $\mathbf{x}_i$ by an arbitrary value $l \in \{1, 2, \ldots, L_j\}$. The full conditional distribution of missing component j of $\mathbf{x}_i$ (when $m_{ij} = 1$) is $p(x_{ij} \mid \ldots) \propto 1\{\operatorname{Rep}(\mathbf{x}_i, j, x_{ij}) \notin S\} \lambda_{jz_i}[x_{ij}]$. Thus, we replace step 8 in the algorithm with

8'. For each $(i, j) \in \{(i, j) : m_{ij} = 1\}$, sample $x_{ij} \sim \text{Discrete}_{1:L_j}(p_1, \ldots, p_{L_j})$, where $p_l \propto \lambda_{jz_i}[l] 1\{\operatorname{Rep}(\mathbf{x}_i, j, l) \notin S\}$.

The definition of p_l implies trimming the support of the full conditional distribution of x_{ij} from $\{1, \ldots, L_j\}$ to only values that avoid $\mathbf{x}_i \in S$, given current values of $\{x_{ij'} : \text{all } j' \neq j\}$.

To obtain M completed datasets for use in multiple imputation, analysts select M of the sampled $\mathbf{x}^{\text{mis}}$ after convergence of the Gibbs sampler. These datasets should be spaced sufficiently so as to be approximately independent (given $\mathbf{x}^{\text{obs}}$). This involves thinning the MCMC samples so that the autocorrelations among parameters are close to zero.

3 Simulation study

To illustrate empirically the performance of this imputation engine, we conducted a repeated sampling experiment using an extract of the 5% public use microdata sample from the 2000 U.S. census data for the state of New York (Ruggles, Alexander, Genadek, Goeken, Schroeder and Sobek 2010). The data include $H = 953{,}076$ individuals and ten categorical variables: ownership of dwelling (3 levels), mortgage status (4 levels), age (9 levels), sex (2 levels), marital status (6 levels), single race identification (5 levels), educational attainment (11 levels), employment status (4 levels), work disability status (3 levels), and veteran status (3 levels). These variables define a contingency table with 2,566,080 cells, of which 2,317,030 correspond to structural zeros.

We treat the H records as a population from which we take 500 independent samples of size $n =$ 1,000. For each sample, we impose missing data by randomly blanking 30% of the recorded item-level values of each variable. We then estimate the truncated latent class model of Section 2.3, using 10,000 MCMC iterates and discarding the first 5,000 as burn-in. From each remaining chain we create $M = 50$ completed datasets via a systematic sample of every 100 iterations. In all 500 simulation runs we use a maximum number of latent classes $K = 50$. The effective number of components, *i.e.*, those comprising at least one individual, are typically between 10 and 15 (depending on the particular sub-sample) and not larger than 26.

As estimands, we use all three-way probabilities with values exceeding 0.1 in the population (the $H =$ 953,076 individuals). This equates to 279 estimands. In each sample, we estimate 95% confidence intervals for each of the 279 probabilities using the multiple imputation combining rules of Rubin (1987). We also compute the corresponding intervals with the data before introducing missing values, which we call the complete data.

Figure 3.1 shows the percentages of the five hundred 95% confidence intervals that cover their population values. For the most part, the simulated coverage rates for multiple imputation are within Monte Carlo error of the nominal level. A few intervals based on multiple imputation have low coverage rates; in particular, three are below 85% while their counterparts with complete data are closer to the nominal level. However, as evident in Figure 3.2, the absolute magnitudes of the biases in the point estimates of these quantities tend to be modest. These encouraging results are in accord with the results in Si and Reiter (2013), whose simulations included up to 50 variables (without any structural zeros).

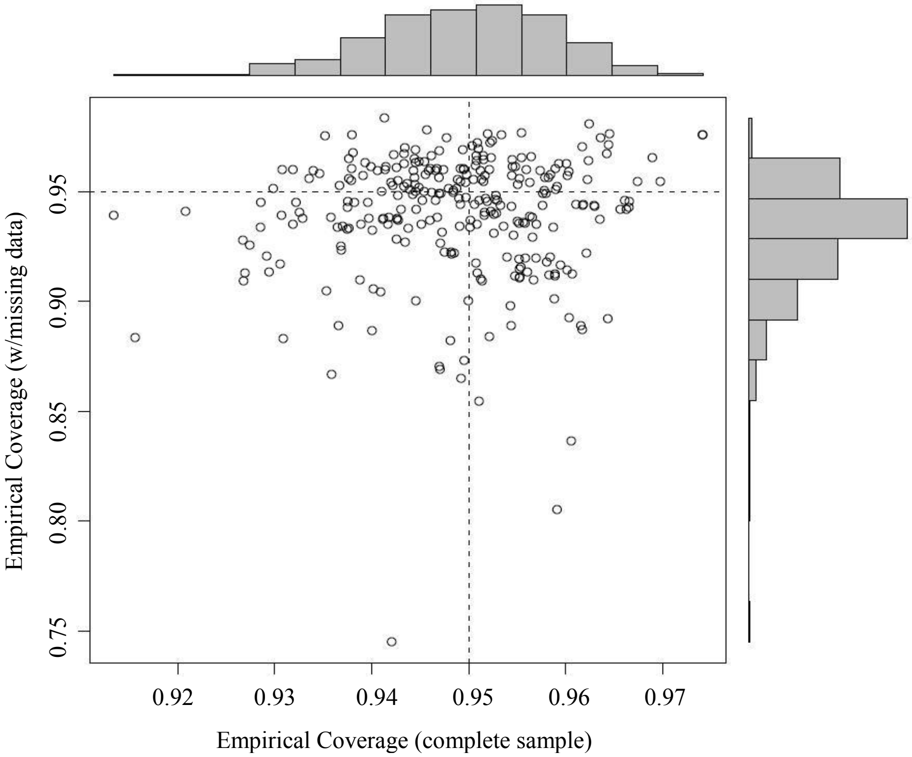


Figure 3.1 Comparison of empirical coverage rates (over 500 trials) of confidence intervals for three-way marginal probability estimates computed from the complete samples *vs.* multiply imputed datasets. Discontinuous lines indicate nominal coverage level. Random Unif(-0.004, 0.004) noise added for clarity.

For each estimand, we also compute the mean estimated fraction of missing information (FMI Rubin 1987, page 77) over the 500 trials. These are displayed in Figure 3.3. Most mean FMIs are close to the missing item rate of 30% that we imposed on every variable in the simulation design. However, many of the mean FMIs are significantly smaller than 30%, including four exactly equal to zero. The estimands with mean FMIs significantly below 0.30 correspond to entries of 3-way marginal probability tables where structural zeros severely restrict the possible imputations. In effect, the structural zeros reduce the information loss due to missingness. For example, the four estimands with mean $\text{FMI} = 0$ correspond to combinations of variables where restrictions leave only one possible imputation pattern to choose from; thus, no information is lost even though data values are actually missing. By incorporating the structural zeros, we automatically impute such cases appropriately and can take advantage of the information supplied by the structural zero restrictions.

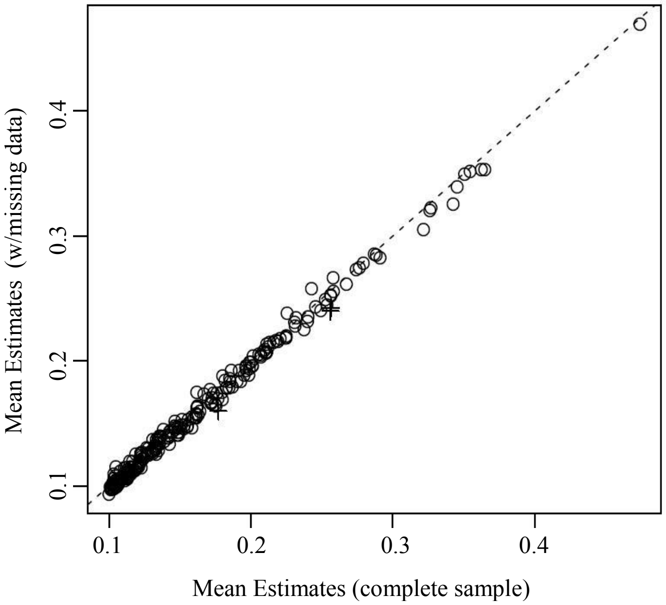


Figure 3.2 Mean (over 500 trials) three-way marginal probability estimates computed from the multiple imputed datasets *vs.* computed from the complete samples. Points marked with crosses are estimates for which the empirical coverage of the multiple-imputation based 95% confidence intervals fell below 85%.

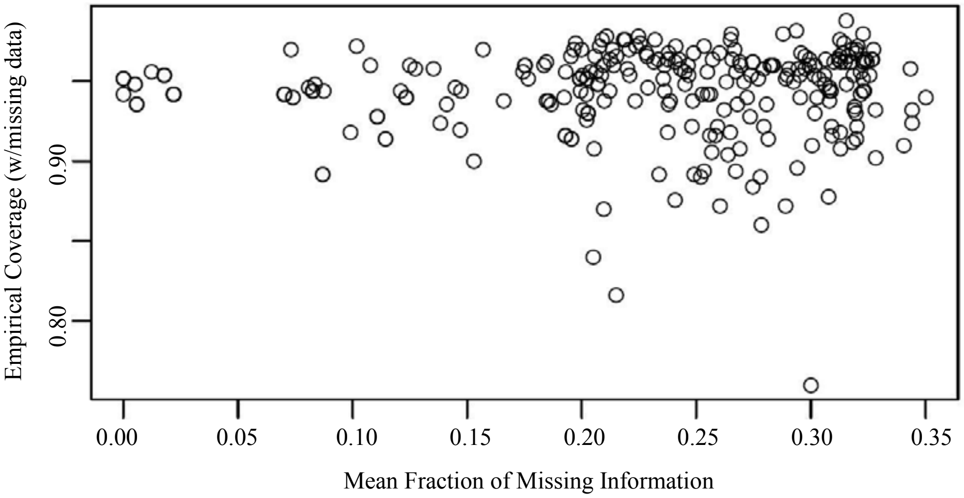


Figure 3.3 Empirical coverage rates (over 500 trials) of confidence intervals for 279 three-way marginal probability estimates computed from the multiply imputed datasets *vs.* their corresponding mean (over the 500 trials) estimated fraction of missing information.

4 Concluding remarks

Structural zero restrictions are an important feature of many surveys, *e.g.*, impossible combinations and skip patterns. They also play a key role in imputation. Ignoring structural zeros when estimating models can result in severe biases when estimating quantities that depend on joint or conditional probabilities. This translates to generating imputed values that do not accurately reflect the dependency structure in the data, and subsequently can lead to biased multiple imputation inferences. Additionally, structural zeros often function as consistency rules. Not enforcing them in imputation could result in completed datasets with inconsistent responses—like widowed toddlers or non-homeowners paying property taxes—that many agencies would be reluctant to release and many public users would find difficult to analyze. The approach suggested here based on Bayesian truncated latent class models offers survey researchers a way to avoid such problems, leading to multiple imputations from theoretically coherent and computationally expedient models that can capture complex dependencies, and simultaneously reducing the labor and guesswork in model specification that often accompanies traditional approaches to multiple imputation for categorical data. Computer code in C++ and R implementing the algorithms in this article can be obtained directly from the authors.

Acknowledgements

This research was supported by a grant from the National Science Foundation (SES 11-31897).

References

Barnard, J., and Meng, X. (1999). Applications of multiple imputation in medical studies: From AIDS to NHANES. *Statistical Methods in Medical Research*, 8, 17-36.

Basu, S., and Ebrahimi, N. (2001). Bayesian capture-recapture methods for error detection and estimation of population size: Heterogeneity and dependence. *Biometrika*, 88, 269-279.

Bishop, Y., Fienberg, S. and Holland, P. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press, reprinted in 2007, New York: Springer-Verlag.

Dunson, D., and Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104, 1042-1051.

Harel, O., and Zhou, X.H. (2007). Multiple imputation: review of theory, implementation and software. *Statistics in Medicine*, 26, 3057-3077.

Ishwaran, H., and James, L.F. (2001). Gibbs sampling for stick-breaking priors. *Journal of the American Statistical Association*, 96, 161-173.

Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.

Manrique-Vallier, D., and Reiter, J.P. (forthcoming 2014). Bayesian estimation of discrete multivariate truncated latent structure models. *Journal of Computational and Graphical Statistics*.

Meng, X.L., and Zaslavsky, A.M. (2002). Single observation unbiased priors. *The Annals of Statistics*, 30, 1345-1375.

O'Malley, A.J., and Zaslavsky, A.M. (2008). Domain-level covariance analysis for multilevel survey data with structured nonresponse. *Journal of the American Statistical Association*, 103, 1405-1418.

Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J. and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27, 1, 85-95.

Reiter, J.P., and Raghunathan, T.E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102, 1462-1471.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.

Ruggles, S., Alexander, T., Genadek, K., Goeken, R., Schroeder, M.B. and Sobek, M. (2010). Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database]. University of Minnesota, Minneapolis. http://usa.ipums.org.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4, 639-650.

Si, Y., and Reiter, J.P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, forthcoming.

Suppes, P., and Zanotti, M. (1981). When are probabilistic explanations possible? *Synthese*, 48, 191-199.

Van Buuren, S., and Oudshoorn, C. (1999). Flexible multivariate imputation by MICE. Technical report, Leiden: TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054.

Vermunt, J.K., Ginkel, J.R.V., der Ark, L.A.V. and Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology*, 38, 369-397.

White, I.R., Royston, P. and Wood, A.M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30, 377-399.