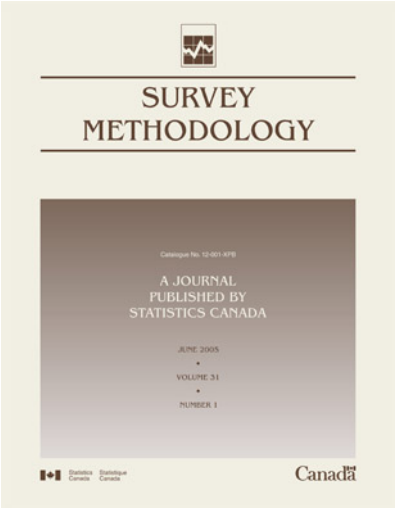
Catalogue no. 12-001-XIE

Survey Methodology

December 2006



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1-800-263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website at www.statcan.ca.

National inquiries line	1-800-263-1136
National telecommunications device for the hearing impaired	1-800-363-7629
Depository Services Program inquiries	1-800-700-1033
Fax line for Depository Services Program	1-800-889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Accessing and ordering information

This product, catalogue no. 12-001-XIE, is available for free in electronic format. To obtain a single issue, visit our website at www.statcan.ca and select Publications.

This product, catalogue no. 12-001-XPB, is also available as a standard printed publication at a price of CAN$30.00 per issue and CAN$58.00 for a one-year subscription.

The following additional shipping charges apply for delivery outside Canada:

	Single issue	Annual subscription
United States	CAN$6.00	CAN$12.00
Other countries	CAN$10.00	CAN$20.00

All prices exclude sales taxes.

The printed version of this publication can be ordered by

- Phone (Canada and United States) 1-800-267-6677
- Fax (Canada and United States) 1-877-287-4369
- E-mail infostats@statcan.ca
- Mail Statistics Canada
  Finance Division
  R.H. Coats Bldg., 6th Floor
  100 Tunney's Pasture Driveway
  Ottawa (Ontario) K1A 0T6
- In person from authorised agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable, courteous, and fair manner. To this end, the Agency has developed standards of service that its employees observe in serving its clients. 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.ca under About us > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

December 2006

Published by authority of the Minister responsible for Statistics Canada

December 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Catalogue no. 12-001-XPB
ISSN: 0714-0045

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

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 cooperation and goodwill.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in 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.

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. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (smj@statcan.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the Journal.

Subscription Rates

The price of printed versions of *Survey Methodology* (Catalogue No. 12-001-XPB) is CDN $58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN $12 ($6 × 2 issues); Other Countries, CDN $30 ($15 × 2 issues). Subscription order should be sent to Statistics Canada, Dissemination Division, Circulation Management, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 or by dialling 1 800 700-1033, by fax 1 800 889-9734 or by E-mail: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec. Electronic versions are available on Statistics Canada's web site: www.statcan.ca.

Survey Methodology

A journal Published by Statistics Canada

Volume 32, Number 2, December 2006

Contents

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.
∞

Le papier utilisé dans la présente publication répond aux exigences minimales de l'"American National Standard for Information Sciences" – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.
∞

Survey Methodology, December 2006
Vol. 32, No. 2, pp. 121-122
Statistics Canada, Catalogue No. 12-001

In This Issue

This issue of *Survey Methodology* opens with the sixth paper in the annual invited paper series in honour of Joseph Waksberg. It is with sadness that we note the passing of Joseph Waksberg in January of 2005. A short biography of Joseph Waksberg was given in the June 2001 issue of the journal, along with the first paper in the series. For more information about the life and work of Joseph Waksberg, see the Statistical Science article (Vol. 15, No 3) "A Conversation with Joseph Waksberg," by David Morganstein and David Marker available at http://projecteuclid.org/Dienst/UI/1.0/Home. I would like to thank the members of the selection committee – David Bellhouse, chair, Gordon Brackstone, Sharon Lohr and Wayne Fuller – for having selected Alastair Scott as the author of this year's Waksberg Award paper.

In his paper entitled "Population-Based Case Control Studies", Scott discusses the analysis of case control studies in which the controls are obtained from a complex sample survey. Using the example of logistic regression, he shows how the survey weighted estimates can be quite inefficient because of the relatively small weight given to the cases. Drawing on an analogy with maximum likelihood estimation, he then proposes a simple, much more efficient alternative that is, however, biased for the intercept term. Efficiency and robustness properties are illustrated through examples. Finally, he briefly discusses the problem of case-control family studies.

Kott considers the use of weight calibration to correct for nonresponse and coverage errors. He gives a general description of calibration estimation, and extends Estavao and Särndal's functional form approach to general calibration. He then discusses properties of this calibration method to correct for unit nonresponse and coverage errors under a quasi-randomization model. He concludes with an empirical example and discussion of some issues.

Reiter, Raghunathan and Kinney investigate through a simulation study the effect of ignoring sampling design variables when building imputation models in a multiple imputation context. They show that potential biases can be reduced by controlling for these design variables in the imputation model, either through a fixed-effect or mixed-effect model. They conclude that a useful prescription for imputers is to include as predictors all variables that are related to the variables being imputed, particularly sampling design variables, so as to make the usual assumption of ignorable non-response satisfied.

The article by Funaoka, Saigo, Sitter and Toida investigates the use of bootstrap variance estimators in stratified multi-stage sampling where the sampling fractions are large. They propose a Bernoulli-type bootstrap that provides consistent bootstrap variance estimates when simple random sampling without replacement is used at each stage. The proposed method is simple to implement and can be extended to any number of stages without much complication. The method is illustrated through a limited simulation study and using data from the 1997 Japanese National Survey of Prices.

In the Kozak and Verma paper, the geometric approach to stratification proposed by Gunning and Horgan (2004) is compared with two optimization approaches; the Lavallée-Hidiroglou algorithm (Lavallée and Hidiroglou 1988) and an optimization algorithm proposed by Kozak (2004). Using five artificial populations of various sizes, the three methods are compared under two scenarios; comparison of the resulting CV under a fixed sample size and comparison of the resulting sample sizes under a fixed level of precision.

Deville and Lavallée present general theoretical foundations for the weight share method in indirect sampling. They define the important concept of a link matrix in indirect sampling, which specifies how the elements of the sampled population are linked to the target population and gives weights to these links that permit unbiased estimation. They discuss important properties of the link matrix, and derive necessary and sufficient conditions for an optimal link matrix to exist. The theory is illustrated with some interesting examples.

Deville and Maumy-Bertrand study the determination of a sampling design and an estimation method for a tourist survey. The main issue that this type of survey has to address is the absence of a sampling frame that can be used to directly reach tourists. To get around this problem, authors suggest to sample services for tourists. This is thus a situation of indirect sampling for which the generalized weight-share method is used to obtain estimates of parameters of interest. Some extensions to the method become necessary. The authors focus more specifically on one of them and describe it in greater detail.

Félix-Medina and Monjardin consider a variant of link-tracing sampling. They use a Bayesian approach to construct estimators of population size, however in order to make inferences about the population size that are robust to erroneous specification of the assumed model, the authors make inferences under the frequentist design-based approach. Based on the results of the simulation study, the proposed estimators perform better than the maximum likelihood estimators that are currently used.

The paper by Dorfman, Lent, Leaver, and Wegman presents a comparison of the Consumer Price Index design methodologies of the United Kingdom and the United States employing the same "scanner" data. They conclude that in the population studied, the UK approach, which involves tighter stratification and, more importantly, more restrictive judgment sampling within strata than the probability sampling of the US approach, does better in estimating a target superlative index. This is shown to be the case, whichever low level price index estimator (the ratio of averages, the geometric mean, or the average of ratios) is employed.

In their paper, Thomas, Raghunathan, Schenker, Katzoff and Johnson use multiple imputation to analyze data with missing values caused by a matrix sampling design. In matrix sampling, only a subset of questions is administered to each respondent in order to reduce respondent burden. The authors develop a method for creating matrix sampling forms, each form containing a subset of questions to be administered to randomly selected respondents. The method is designed so that each form includes questions that are predictive of the excluded questions in order to recover some of the information about the latter. The proposed method and multiple imputation are evaluated using data from the National Health and Nutrition Examination Survey.

Harold Mantel, Deputy Editor

Survey Methodology, December 2006
Vol. 32, No. 2, pp. 123-132
Statistics Canada, Catalogue No. 12-001

Population-Based Case Control Studies

Alastair Scott [1]

Abstract

We discuss methods for the analysis of case-control studies in which the controls are drawn using a complex sample survey. The most straightforward method is the standard survey approach based on weighted versions of population estimating equations. We also look at more efficient methods and compare their robustness to model mis-specification in simple cases. Case-control family studies, where the within-cluster structure is of interest in its own right, are also discussed briefly.

Key Words: Case-control studies; Response-selective sampling; Retrospective sampling; Weighting.

1. Introduction

The case-control study, in which separate samples are drawn from 'cases' (people with a disease of interest, say) and from 'controls' (people without the disease), is one of the most common designs in health research. In fact, Breslow (1996) has described such studies as "the backbone of epidemiology". We shall concentrate on biostatistical applications, but the basic design is an efficient sampling strategy whenever cases are rare and examples are common in many other fields as well (business, social science, ecology, market research, for example). In particular, there has been a parallel development of much of the theory in the econometric literature on choice-based sampling (see Manski and McFadden 1981, Cosslett 1981 for example).

There are two fundamentally different types of case-control study: (set-)matched studies, in which each case is matched with one or more controls, and unmatched studies, in which the case and control samples are drawn independently, although there may be loose "frequency matching", with the control sample allocated across strata defined by basic demographic variables in such a way that the distribution of these variables in the control sample is similar to their expected distribution in the case sample. We are only concerned with unmatched studies here and, more specifically, only with the restricted class of population-based studies in which the controls (and occasionally the cases as well) are selected using standard survey sampling techniques.

An excellent introduction to the strengths and potential pitfalls of case-control sampling is given by Breslow (1996, 2004). One of the most important and difficult challenges confronting anyone designing such a study is to ensure that controls really are drawn from the same population, using the same protocols, as the cases. In the words of Miettinen (1985), cases and controls "should be representative of the same base experience". Failure to ensure this adequately in some early examples led to case-control sampling being regarded with some suspicion by many researchers. A comprehensive discussion on the principles that should govern the selection of controls is given in Wacholder, McLaughlin, Silverman and Mandel (1991). Since the essence of survey sampling lies in methods for drawing representative samples from a target population, it became natural at some stage to think about using survey methods for obtaining controls. Increasingly over the last 25 years or so, the controls (and occasionally the cases as well) are being drawn using complex stratified multi-stage designs. A good history of this development can be found in Chapter 9 of Korn and Graubard (1999).

The analysis of such studies is a particularly appropriate topic for this paper since Joe Waksberg himself was one of the principal drivers behind the adoption of survey methods (and random digit dialing, in particular) for obtaining controls (see, for example, Waksberg 1998 and DiGaetano and Waksberg 2002).

2. Examples

We start with two examples to illustrate the sort of problem that we want to handle. The first example is typical of the large scale studies conducted by the National Cancer Institute whose personnel have been responsible for much of the development of the area. Joe Waksberg, along with his colleagues at Westat, had a strong influence on the sampling methods used for these studies (see Hartge, Brinton, Rosenthal, Cahill, Hoover and Waksberg 1984, who also gives a description of a number of other similar studies) so it is a natural place to start.

Example 1.

In 1977–78, the National Cancer Institute and the US Environmental Protection Agency conducted a population-based case-control study to examine the effects of ultraviolet radiation on non-melanoma skin cancer over a one-year period (Hartge, Brinton, Rosenthal, Cahill, Hoover and

1. Alastair Scott, Department of Statistics, University of Auckland, Auckland 1, New Zealand. E-mail: a.scott@auckland.ac.nz.

Waksberg 1984, Fears and Gail 2000). The study was conducted at eight geographic locations with varying solar ultraviolet intensities. Samples of non-melanoma skin cancer patients aged 20 to 74 and samples of general population controls from each region were interviewed by telephone to obtain information on risk factors. At each location, a simple random sample of 450 patients and an additional sample of 50 patients in the 20 – 49 age group were selected for contact. For the controls, 500 households were sampled at each location using Mitofsky-Waksberg random-digit dialing (Waksberg 1978). An attempt was made to interview all adults aged 65 – 74 as well as a randomly selected individual of each sex aged 20 to 64. In addition, a second Mitofsky-Waksberg sample of between 500 to 2,100 households was taken and information gathered on all adults aged 65 to 74. This resulted in samples of approximately 3,000 cases and 8,000 controls, with the sampling rate for cases being roughly 300 times the rate for controls, depending on age.

The second example is important to me personally since it first introduced Chris Wild and myself to the area.

Example 2.

The Auckland Meningitis Study was commissioned by the NZ Ministry of Health and Health Research Council to study risk factors for meningitis in young children which was reaching epidemic proportions in Auckland at that time (see Baker, McNicholas, Garrett, Jones, Stewart, Koberstein and Lennon 2000). The target population was all children under the age of nine in the Auckland region in 1997 – 2000.

All cases of meningitis in the target age group over the three year duration of the study were included in the study, resulting in about 250 cases. A similar number of controls was drawn from the remaining children in the study population using a complex multi-stage design. At the first stage of sampling, 300 census mesh blocks (each containing roughly 70 households) were drawn with probabilities proportional to the number of houses in the block. At the second stage, a systematic sample of 20 households was selected from each chosen mesh block and children from these households were selected for the study with varying probabilities that depended on age and ethnicity and were chosen to match the expected frequencies among the cases. Selection probabilities are shown in the table below: (PI means Pacific Islander) Cluster sample sizes varied from one to six and a total of approximately 250 controls was achieved. This corresponds to a sampling fraction of about 1 in 400 on average, so that cases are sampled at a rate that is 400 times that for controls here.

These two studies are fairly typical of the sort of study that we want to discuss. They also illustrate the two main sampling methods used, namely random digit dialing and area sampling. A lively discussion of the relative merits of these two strategies are given in Brogan, Denniston, Liff, Flagg, Coates and Brinton (2001) and DiGaetano and Waksberg (2002).

Table 1
Selection Probabilities

AGE	MAORI	PACIFIC ISLANDER	OTHER
≤ 1 year	0.29	0.70	0.10
≤ 3 years	0.15	0.50	0.07
≤ 5 years	0.15	0.31	0.04
≤ 8 years	0.15	0.17	0.04

3. General Set-Up

Suppose that we have a binary response variable, Y, with $Y = 1$ denoting a case and $Y = 0$ denoting a control, and a vector of potential explanatory variables, $\mathbf{x}$. We assume that the value of Y is known for all N units in some target population but that at least some components of $\mathbf{x}$ are unknown. We stratify the population into cases and controls, draw a sample from each stratum based on the variables that we know for all units, and measure the values of the missing covariates for the sampled units (in practice, the control sample is often drawn from the whole population, rather than the units with $Y = 0$. If the proportion of cases is small, the difference will be negligible. Otherwise it is simple to adapt the results below to this variant – for a rigorous development, see Lee, Scott and Wild 2006). Typically, we then want to use the sample data to fit a binary regression model for the marginal probability of a being a case as a function of the covariates. The model used is almost always logistic with

$$\text{logit}\,\{P(Y = 1 \mid \mathbf{x})\} = \log\left(\frac{P(Y = 1 \mid \mathbf{x})}{P(Y = 0 \mid \mathbf{x})}\right) = \beta_0 + \mathbf{x}^T\boldsymbol{\beta}_1 \qquad (1)$$

say, where β_0 and $\boldsymbol{\beta}_1$ are unknown parameters, and we shall assume model (1) throughout the paper. Extensions to more general regression models are straightforward in principle (see Scott and Wild 2001b) but the resulting expressions are somewhat clumsier than those for the logistic model.

How should we go about fitting the model (1) given sample data? Efficient methods are straightforward with simple or stratified random sampling, but we are interested in more complex sampling procedures here. Very often the complex sampling is simply ignored. Potentially, this could lead to all the usual problems that arise from ignoring sampling design structure. Varying selection probabilities can distort the mean structure and estimates produced by standard programs may be inconsistent. Intra-cluster

correlation can reduce the effective sample size so that routinely-produced standard errors are too small, confidence intervals are too short, p – values too low, and so on. A simple strategy that has been adopted by some researchers to minimize the effect is to keep the numbers of subjects in each cluster small (see Graubard, Fears and Gail 1989, for example). This reduces the design effect and hence the impact of clustering, but it can be a very expensive remedy. We look at some possible ways of coping with standard, more cost-effective, sampling schemes in the next few sections.

4. Survey Weighted Approach

One obvious possibility is to use the standard weighted estimating equation approach embodied in most modern packages for analyzing survey data (see Binder 1983). Suppose first that we had data from the whole finite population. If we assume this finite population is drawn from a superpopulation in which the conditional logistic model **(**1**)** holds, then we could estimate $\boldsymbol{\beta}$ by solving the whole-population or census estimating equations

$$\mathbf{S}(\boldsymbol{\beta}) = \sum_{1}^{N} \mathbf{x}_i (y_i - p_1(\mathbf{x}_i; \boldsymbol{\beta})) = 0, \qquad (2)$$

where $p_1(\mathbf{x}; \boldsymbol{\beta}) = e^{\beta_0 + \mathbf{x}^T\boldsymbol{\beta}_1} / (1 + e^{\beta_0 + \mathbf{x}^T\boldsymbol{\beta}_1})$. (These are the likelihood equations if population units are assumed to be sampled independently from a superpopulation but the resulting estimators are consistent under much more realistic population structures as long as model (1) holds marginally – see Rao, Scott and Skinner 1998 for more discussion.)

Now, for any fixed value of $\boldsymbol{\beta}$, $\mathbf{S}(\boldsymbol{\beta})$ in equation (2) is just a vector of population totals. This means that we can estimate it from the sample, say by

$$\hat{\mathbf{S}}(\boldsymbol{\beta}) = \sum_{\text{sample}} w_i \mathbf{x}_i (y_i - p_1(\mathbf{x}_i; \boldsymbol{\beta})), \qquad (3)$$

where w_i is the inverse of the selection probability, perhaps adjusted for non-response and post-stratification. Setting $\hat{\mathbf{S}}(\boldsymbol{\beta})$ equal to $\mathbf{0}$ gives us our estimator, $\hat{\boldsymbol{\beta}}$. We could use linearization or the jackknife directly on $\hat{\boldsymbol{\beta}}$ to get standard errors. Alternatively, we can expand $\hat{\mathbf{S}}(\hat{\boldsymbol{\beta}})$ about the true value, $\boldsymbol{\beta}$, and obtain as our estimated covariance matrix the "sandwich" estimator

$$\hat{\text{Cov}}\{\hat{\boldsymbol{\beta}}\} \approx \mathbf{J}(\hat{\boldsymbol{\beta}})^{-1} \hat{\text{Cov}}\{\hat{S}(\hat{\boldsymbol{\beta}})\} \mathbf{J}(\hat{\boldsymbol{\beta}})^{-1}, \qquad (4)$$

where $\mathbf{J}(\boldsymbol{\beta}) = -\partial\hat{\mathbf{S}}/\partial\boldsymbol{\beta}^T = \sum_{\text{sample}} w_i\, p_1(\mathbf{x}_i;\boldsymbol{\beta})\, p_0(\mathbf{x}_i;\boldsymbol{\beta})\, \mathbf{x}_i \mathbf{x}_i^T$ with $p_0 = 1 - p_1$. Since $\hat{\mathbf{S}}(\boldsymbol{\beta})$ is a vector of totals, $\hat{\text{Cov}}\{\hat{\mathbf{S}}(\boldsymbol{\beta})\}$ should be available as a matter of course for any standard design. Most major statistical packages (for example, SAS (PROC SURVEYLOGISTIC), SPSS (CSLOGISTIC), STATA (SVY:LOGIT), SUDAAN (LOGISTIC)) can handle logistic regression with complex sampling and weighting routinely these days. Thus producing weighted estimates and making associated inferences is reasonably straightforward.

Strictly speaking, the selection probabilities will themselves often be random variables in our model-based framework, based on a finite population that we assume is generated from the model. We can account for this by using the results in Rao (1973), but the correction is of order $1/N$ and can be ignored in most large studies.

The downside of weighting in general is that it tends to be inefficient when the weights are highly variable. **(**A rule-of-thumb sometimes suggested is that $w_{\max} / w_{\min}$ should be no more than 10.**)** In case-control studies, the variation in weights is about as extreme as it can get. For instance, the ratio of $w_{\max}$ to $w_{\min}$ is approximately 300:1 in Example 1 and approximately 1,000:1 in Example 2. Even more extreme ratios are not uncommon. No experienced survey sampler would be surprised to find that weighting is not very efficient under these circumstances.

Can we do something more efficient? The answer is certainly "Yes" in some special cases. Fully efficient likelihood methods have been developed in situations where both cases and controls are drawn using simple or stratified random sampling and these can be very much more efficient than weighted methods. We review these results in the next section.

5. Review: Simple Case

We start with the very simplest case where cases and controls are selected by simple random sampling and we have no population information about any of the covariates at the design stage. Here fully efficient semi-parametric maximum-likelihood procedures are well-developed. Moreover, these methods are very simple to implement using standard software (Prentice and Pyke 1979). (The methods are *semi-parametric* because the full likelihood depends on the unknown distribution of the covariates and we do not want to model this in general.)

It turns out that all we have to do is fit model (1) using a standard logistic regression program without any weighting at all. More specifically, solving the unweighted equation

$$\sum_{\text{sample}} \mathbf{x}_i (y_i - p_1(\mathbf{x}_i; \boldsymbol{\beta})) = \mathbf{0}, \qquad (5)$$

produces efficient estimates of all the coefficients except the intercept. Perhaps more importantly, all the standard errors and resulting inferences that we get from the standard program are also valid, again with the exception of anything

involving the intercept. It is simple enough to correct inferences involving the intercept provided that we know the ratio of the two sampling fractions but we are often only interested in the other coefficients anyway.

The results extend directly to stratified random sampling, provided that separate intercepts for each stratum are included in the model. Again efficient semi-parametric estimators of all coefficients except the stratum intercepts can be obtained simply by running the data through an ordinary (unweighted) logistic regression program. Again, the estimated standard errors and associated inferences are also valid. As with simple random sampling, we can correct the results for the stratum intercepts provided that we know the stratum sampling fractions but, again, these are usually of minor interest.

Thus in these simple situations, maximum likelihood estimates are simpler to compute than the weighted estimates, as well as being more efficient. How much more efficient are they? This depends on the number of covariates, the magnitude of their coefficients and the ratio of the sampling fractions, but the difference is often substantial. (The weighted estimates are about 50% efficient in Example 2 of the introduction, for example, and less than 20% efficient in the brain cancer example we look at in Section 8. Lawless, Kalbfleisch and Wild 1999 discuss situations where the efficiency is even lower than this.)

Finally, we note that the maximum likelihood estimates have yet another advantage over weighted estimates: they tend to have much better small sample performance, especially in situations where the efficiency of the weighted estimates is low. Essentially, weighting results in a reduction in the effective sample size and it is this effective sample size that governs when the asymptotic theory starts to give a good approximation. (See Scott and Wild 2001a for more details.) Clearly we can pay a very heavy price for a rigid adherence to population weights.

6. More Complex Sampling

In both the examples in Section 2, the controls were obtained from a complex multi-stage survey rather than a simple random sample. As we noted in the introduction, this is increasingly common in large scale case-control studies. (Occasionally, as in Example 1, the cases are also selected using a complex sampling scheme.) It is possible to derive semi-parametric efficient estimators for stratified multistage sampling, assuming that primary sampling units are selected independently within strata (which is the assumption that all the computer packages are making with the survey-weighted approach anyway), but this requires us to build multivariate models for the vector of responses within a primary sampling unit. Details can be found in Neuhaus, Scott and Wild (2002, 2006). Unless we are interested in the within-cluster structure in its own right (as in the family case-control studies considered in Section 9, for example), this requires far too much effort for it to be practicable, certainly for routine analysis.

Can we do something simpler without losing too much efficiency? Weighted estimates are always available, of course. However, they are just as inefficient with complex designs as they are in the simple case considered in the previous section. It turns out that we can do considerably better without too much extra complication.

Return for a moment to the situation of the previous section where we have a simple random sample of size n_1 from the case stratum and an independent simple random sample of size n_0 from the control stratum. Here all units in Stratum ℓ have weight $w_i \propto W_\ell / n_\ell$, where W_ℓ denotes the proportion of the population in the stratum, for $\ell = 0,1$. If we divide throughout by N and set $p_0(\mathbf{x};\boldsymbol{\beta}) = 1 - p_1(\mathbf{x};\boldsymbol{\beta})$, then we can re-write equation (3) for the weighted estimator in the form

$$W_1 \frac{\sum_{\text{cases}} \mathbf{x}_i \, p_0(\mathbf{x}_i;\boldsymbol{\beta})}{n_1} - W_0 \frac{\sum_{\text{controls}} \mathbf{x}_i \, p_1(\mathbf{x}_i;\boldsymbol{\beta})}{n_0} = \mathbf{0}. \quad (6)$$

Similarly, we can write equation (5) for the efficient maximum likelihood estimator in the form

$$\omega_1 \frac{\sum_{\text{cases}} \mathbf{x}_i p_0(\mathbf{x}_i;\boldsymbol{\beta})}{n_1} - \omega_0 \frac{\sum_{\text{controls}} \mathbf{x}_i p_1(\mathbf{x}_i;\boldsymbol{\beta})}{n_0} = \mathbf{0}, \quad (7)$$

where $\omega_\ell = n_\ell / (n_0 + n_1)$, for $\ell = 0,1$. Both these are special cases of the general set of estimating equations

$$\lambda_1 \frac{\sum_{\text{cases}} \mathbf{x}_i p_0(\mathbf{x}_i;\boldsymbol{\beta})}{n_1} - \lambda_0 \frac{\sum_{\text{controls}} \mathbf{x}_i p_1(\mathbf{x}_i;\boldsymbol{\beta})}{n_0} = \mathbf{0}. \quad (8)$$

As $n_0, n_1 \to \infty$, under mild conditions on the way that the finite population is generated from the superpopulation the solution of (8) converges almost surely to the solution $\boldsymbol{\beta}^*$ of

$$\lambda_1 E_1\{\mathbf{X} \, p_0(\mathbf{X};\boldsymbol{\beta}^*)\} - \lambda_0 E_0\{\mathbf{X} \, p_1(\mathbf{X};\boldsymbol{\beta}^*)\} = \mathbf{0}, \quad (9)$$

where $E_\ell\{\cdot\}$ denotes the conditional expectation given that $Y = \ell$ for $\ell = 0,1$. If model (1) is true, then equation (8) has solution $\boldsymbol{\beta}_1^* = \boldsymbol{\beta}_1$ and $\beta_0^* = \beta_0 + b_\lambda$ with $b_\lambda = \log(\lambda_1 W_0 / \lambda_0 W_1)$ for any positive λ_0, λ_1 (see Scott and Wild 1986 for details of the proof). Thus the solution to equation (8) produces consistent estimators of all the regression coefficients apart from the constant term for any $\lambda_\ell > 0$ ($\ell = 0,1$). As in the simple case, it is easy to correct the inferences about the constant term, provided that we know the proportion of cases in the population.

Now turn to more complex sampling schemes. Since the left hand side of equation (9) just involves two subpopulation means, we can still estimate these means for any standard survey design. This suggests an estimator, $\hat{\boldsymbol{\beta}}_\lambda$ say, for general sampling schemes satisfying

$$\hat{\mathbf{S}}_\lambda(\boldsymbol{\beta}) = \lambda_1 \hat{\boldsymbol{\mu}}_1(\boldsymbol{\beta}) - \lambda_0 \hat{\boldsymbol{\mu}}_0(\boldsymbol{\beta}) = \mathbf{0}, \qquad (10)$$

where $\hat{\boldsymbol{\mu}}_\ell(\boldsymbol{\beta})$ is the sample estimator of the subpopulation mean $E_\ell\{\mathbf{X}(1 - p_\ell(\mathbf{X}; \boldsymbol{\beta}))\} (\ell = 0,1)$. The covariance matrix of $\hat{\boldsymbol{\beta}}_\lambda$ can then be obtained by standard linearization arguments. This leads to an estimated ('sandwich') covariance matrix

$$\hat{\mathrm{Cov}}\{\hat{\boldsymbol{\beta}}_\lambda\} \approx \mathbf{J}_\lambda(\hat{\boldsymbol{\beta}}_\lambda)^{-1} \hat{\mathrm{Cov}}\{\hat{\mathbf{S}}_\lambda(\hat{\boldsymbol{\beta}}_\lambda)\} \mathbf{J}_\lambda(\hat{\boldsymbol{\beta}}_\lambda)^{-1}, \quad (11)$$

with $\mathbf{J}_\lambda(\boldsymbol{\beta}) = (-\partial \hat{\mathbf{S}}_\lambda(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}^T)$ and $\hat{\mathrm{Cov}}\{\hat{\mathbf{S}}_\lambda(\boldsymbol{\beta})\} = \lambda_1^2 \hat{\mathrm{Cov}}\{\hat{\boldsymbol{\mu}}_1(\boldsymbol{\beta})\} + \lambda_0^2 \hat{\mathrm{Cov}}\{\hat{\boldsymbol{\mu}}_0(\boldsymbol{\beta})\}$. Here, $\hat{\mathrm{Cov}}\{\hat{\boldsymbol{\mu}}_\ell(\boldsymbol{\beta})\}$ denotes the usual survey estimate which should be available routinely for any standard survey design since $\hat{\boldsymbol{\mu}}_\ell(\boldsymbol{\beta})$ is just an estimated mean.

All of this can also be carried out straightforwardly in any package that can handle logistic regression for complex survey designs simply by specifying the appropriate vector of weights. More specifically, suppose that

$$\hat{\boldsymbol{\mu}}_\ell(\boldsymbol{\beta}) = \frac{\sum_{i \in S_\ell} w_i \mathbf{x}_i (y_i - p_1(\mathbf{x}_i; \boldsymbol{\beta}))}{\sum_{i \in S_\ell} w_i}, \qquad (12)$$

where S_1 denotes the case subpopulation (*i.e.*, the set of all units with $Y = 1$) and S_0 denotes the control subpopulation (the set of all units with $Y = 0$). Then the estimating equation (9) can be written in the form

$$\hat{\mathbf{S}}_\lambda(\boldsymbol{\beta}) = \sum_{\text{sample}} w_i^* \mathbf{x}_i (y_i - p_1(\mathbf{x}_i; \boldsymbol{\beta})) = \mathbf{0}, \qquad (13)$$

with $w_i^* \propto \lambda_\ell w_i / \sum_{i \in S_\ell} w_i$ for units in $S_\ell (\ell = 0,1)$. In other words, we simply have to scale the case weights and control weights separately so that the sum of the case weights is proportional to λ_1 and the sum of the control weights is proportional to λ_0 and put them, along with the usual specification of the design structure (strata, primary sampling units), into our program of choice. Note that the choice of proportionality constant does not affect the result.

We still have to decide on good values for λ_1 and λ_0. We can often get substantial gains using sample weights ($\lambda_\ell = n_\ell / n$) compared with using population weights ($\lambda_\ell = W_\ell$). Scott and Wild (2002) report efficiency gains of 50% or more in Example 2 and in simulations based on that population. The gains became larger as the strength of the relationship increased, and as the effect of clustering increased. Moreover the coverage of confidence intervals was closer to the nominal value for sample weighting in the simulations.

Using sample weights is the most efficient possible strategy when we have simple random samples of cases and controls but for more complex schemes using the sample weights will no longer be fully efficient. We might expect weights based on some form of equivalent sample sizes to perform better. This does indeed produce some gain in efficiency in some limited simulations reported in Scott and Wild (2001a). However, the gains are relatively small, at least when the control sample design effect is less than 2, since $\mathrm{Cov}\{\hat{\boldsymbol{\beta}}_\lambda\}$ is very flat as a function of $\boldsymbol{\lambda}$ near its minimum. Considerations of robustness that we discuss in Section 8 are possibly more important in the choice of $\boldsymbol{\lambda}$.

The gains from sample weighting may depend very much on the particular problem under examination. Korn and Graubard (1999, page 327) comment that, in their experience, the sample weighting strategy rarely produces big gains in efficiency. Obviously more work, both empirical and theoretical, is needed here. In any event, it would seem prudent to fit the model using both sample weights and population weights routinely. If the coefficient estimates are similar, then we can make a judgement based on the estimated standard errors. However, significant differences in the coefficient estimates indicate that the model has been mis-specified. If we are unable to fix up the deficiencies in the model, then we need to think very carefully about just what it is that we are trying to estimate. We look at this again in Section 8.

7. Stratified Sampling

The compromise suggested in the previous section (*i.e.*, use standard survey weighting within the subpopulations defined by case/control status but combine the subpopulations using sample proportions) seems to work reasonably well in practice but it is all completely *ad hoc*. Could we do better with a more systematic approach?

In the special case of stratified random sampling, where independent case-control samples are taken within each stratum, fully efficient procedures are well-developed and easy to implement. In particular, if our model includes a separate intercept for each stratum, then ordinary unweighted logistic regression (with a simple adjustment for the stratum intercepts if they are wanted) is the efficient semi-parametric maximum likelihood procedure (Prentice and Pyke 1979). It is reasonably straightforward to extend this to more general stratified designs. Our model is now

$$\mathrm{logit}\{P(Y = 1 \mid \mathbf{x}, \text{ Stratum h})\} = \beta_{0h} + \mathbf{x}^T \boldsymbol{\beta}_1, \qquad (14)$$

and the stratified equivalent of the estimating equation (7) is

$$\sum_h \left(\lambda_{1h} \frac{\sum_{\text{cases}} \mathbf{x}_i p_{0h}(\mathbf{x}_i; \boldsymbol{\beta})}{n_{1h}} - \lambda_{0h} \frac{\sum_{\text{controls}} \mathbf{x}_i p_{1h}(\mathbf{x}_i; \boldsymbol{\beta})}{n_{0h}} \right) = \mathbf{0}. \quad (15)$$

As $n_{0h}, n_{1h} \to \infty$, the solution of (7) converges almost surely to the solution of

$$\sum_h (\lambda_{1h} E_{1h}\{\mathbf{X} p_{0h}(\mathbf{X}; \boldsymbol{\beta})\} - \lambda_{0h} E_0\{\mathbf{X} p_{1h}(\mathbf{X}; \boldsymbol{\beta})\}) = \mathbf{0}, \quad (16)$$

with the obvious extension of the notation from the unstratified case. If model (13) is true, then equation (8) has solution $\boldsymbol{\beta}_1^* = \boldsymbol{\beta}_1$ and $\beta_{0h}^* = \beta_{0h} + b_{\lambda h}$ with $b_{\lambda h} = \log(\lambda_{1h} W_{0h} / \lambda_{0h} W_{1h})$. Since equation (14) only involves stratum means, we can estimate them easily using the data coming from any reasonable survey design, for example by

$$\hat{\boldsymbol{\mu}}_{\ell h}(\boldsymbol{\beta}) = \frac{\sum_{i \in S_{\ell h}} w_{ih} \mathbf{x}_{ih} (y_{ih} - p_1(\mathbf{x}_{ih}; \boldsymbol{\beta}))}{\sum_{i \in S_{\ell h}} w_{ih}}.$$

Substituting these estimators in place of the sample means in equation (14) leads to the estimating equation

$$\hat{\mathbf{S}}_\lambda(\boldsymbol{\beta}) = \sum_h \sum_{i \in S_h} w_{ih}^* \mathbf{x}_i (y_i - p_{1h}(\mathbf{x}_i; \boldsymbol{\beta})) = \mathbf{0}, \quad (17)$$

with $w_{ih}^* \propto \lambda_{\ell h} w_{ih} / \sum_{i \in S_{\ell h}} w_{ih}$ for units in $S_{\ell h} (\ell = 0,1; h = 1, ..., H)$. This can be fitted in any standard survey program by including these weights and the appropriate design information. Note that we need to be careful about how we include the so-called 'strata' in the design specification. If primary sampling units are nested within the 'strata', as with the geographical locations in Example 1, there is no problem and the strata should be included in the standard way. However, if the primary sampling units cut across the 'strata', as with age in Example 1 and age and ethnicity in Example 2, then these are not strata in the usual survey sampling sense. They should not be included in the design specifications but simply handled through the weights.

Sometimes we want to model the contribution of the stratum variables using some smooth parametric curve rather than including them through dummy variables. For example, we might well want to include a linear function of age in our model in both Examples 1 and 2. The survey weighted method and the compromise weighting suggested in Section 6 both apply directly and no new theory is needed. More efficient methods are not nearly so simple, however. Fully efficient methods have been developed in the case where simple random samples of cases and controls are drawn within each of the strata (see Scott and Wild 1997, and Breslow and Holubkov 1997) but the resulting estimating equations are not linear combinations of stratum means and there is no obvious way of generalizing them to more complex sampling schemes. There is a slightly less efficient way that does extend easily, however. If we modify model (14) by including $b_{\lambda h} = \log(\lambda_{1h} W_{0h} / \lambda_{0h} W_{1h})$ as an offset, *i.e.*, we set

$$\text{logit}\{P^*(Y = 1 \mid \mathbf{x}, \text{Stratum h})\} = b_{\lambda h} + \beta_{0h} + \mathbf{x}^T \beta_1, \quad (18)$$

then equation (15) produces consistent, fully efficient, estimates of all the coefficients including $\beta_{0h} (h = 1, ..., H)$. Including the same offsets in models where there is no β_{0h} term and the $\mathbf{x}$ vector includes functions of the stratifying variable produces consistent estimators of all the coefficients with typically high (although not full) efficiency (see Fears and Brown 1986, and Breslow and Cain 1988). This generalizes to arbitrary designs immediately. We just use equation (16) with p_{1h} replaced by p_{1h}^* defined by setting $\text{logit}(p_{1h}^*) = b_{\lambda h} + \mathbf{x}^T \beta$. Then any survey program that caters for offsets can be used to fit the model and provide estimated standard error, *etc*.

How much extra efficiency do we get in this case? We have carried out a number of simulations, some of which are reported in Scott and Wild (2002). Most of the scenarios are based on the meningitis study in Example 2 and we set the ratio of the largest to smallest stratum sampling fraction in the control sample at about 10:1. Without any clustering, the gain in efficiency from using the offset method (which is full maximum likelihood in this case) compared to the *ad hoc* procedure was never more than 10%. The relative efficiencies stayed about the same as clustering that cut across strata was introduced. When clustering nested within strata was introduced, the gains disappeared progressively as the design effect increased and the *ad hoc* procedure actually became more efficient than the offset method when the design effect reached about 1.5.

As we stated earlier, it is possible to produce fully efficient semi-parametric estimators if we are willing to model the dependence structure within primary sampling units. We have begun to carry out some simulation. The early results suggest that the extra work involved in the modeling will almost never be worth the effort if we are only interested in the parameters of the marginal model (1). Our tentative conclusion is that, the *ad hoc* partially weighted procedures (with sample weights) are simple to use and work well enough for most practical purposes in the range covered by our experience but this is another area where more empirical work is needed yet. We note, however, that there are some problems, like the case-control family design discussed in Section 9, where the within-cluster behavior is of interest in its own right. These require more sophisticated methods.

8. Robustness

There must be a catch somewhere. What if the model is not correct? What price do we pay for efficiency then?

By its construction, the population-weighted estimator is always estimating the linear logistic approximation that we would get if we had data from the whole population. By contrast, what the more efficient sample-weighted estimator is estimating depends on the particular sample sizes used. Some people would regard this alone as a strong enough reason for using the population weighted estimator and I suspect that very few people would regard it as completely satisfactory to have the target of their inference depend on the arbitrary choice of sample size.

Our general estimator $\hat{\boldsymbol{\beta}}_\lambda$ satisfying (10) converges to the solution of equation (9), $\mathbf{B}_\gamma$ say, with $\gamma = \lambda_0/(\lambda_0 + \lambda_1)$, which depends on the true model and distribution of the covariates, as well as on γ. In Scott and Wild (2002), we looked at what happens to $\mathbf{B}_\gamma$ under mild deviations from the assumed model. (We are interested in small deviations since large ones should be picked up by routine model-checking procedures and the model then improved.) For simplicity, suppose that we fit a linear model with a single explanatory variable for the log odds ratio but that the true model is quadratic, say

$$\text{logit}\{P(Y = 1 \mid x)\} = \beta_0 + \beta_1 x + \delta x^2 \qquad (19)$$

with δ small.

Obviously, the actual slope on the logit scale, $\beta_1 + 2\delta x$, changes as we move along the curve. For any $0 < \gamma < 1$, $\mathbf{B}_{\gamma 1}$ is equal to the actual slope at some point along the curve. Denote this value by $x = x_\gamma$. Let x_0 be the expected value of x in the control population and let x_1 the expected value of x in the case population. We shall assume that $\beta_1 > 0$ so that $x_0 < x_1$. It turns out that x_γ always lies between x_0 and x_1 .and that x_γ increases as γ increases from 0 to 1. Recall that survey weighting corresponds to $\gamma = W_0$ and sample weighting to $\gamma = \omega_0 = n_0/n$. Typically, W_0 is much larger than ω_0 so that survey weighting gives an estimate of the slope at larger values of x, where the probability of a case is higher, while the slope estimated from sample weighting is closer to the average value of x in the population. Figure 1, adapted from Scott and Wild (2002), illustrates the position in two scenarios, one with positive curvature and one with negative, based roughly on Example 2. The value of δ is chosen so that it would be detected with a standard likelihood ratio test about 50% of the time if we took simple random samples of $n_0 = n_1 = 200$ from the population.

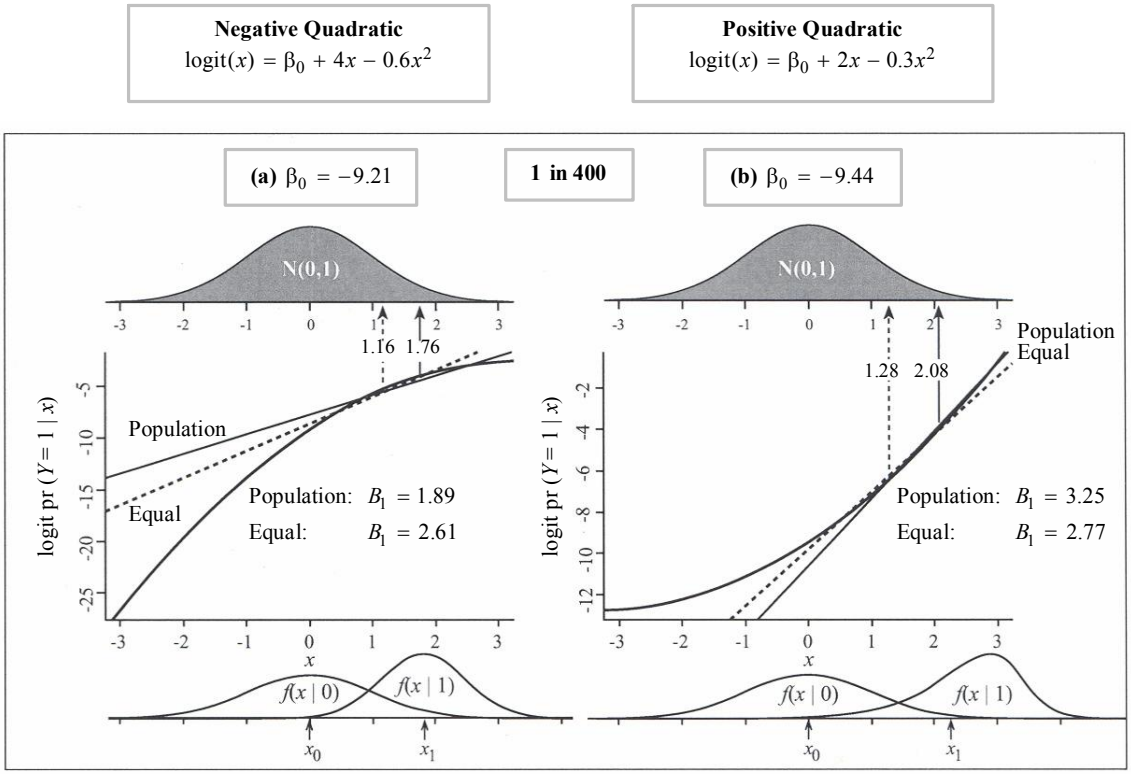


Figure 1. Comparison of population and equal weights.

In both scenarios, the value of β_0 is set so that the proportion of cases in the population is 1 in 400, *i.e.*, so that $W_0 = 0.9975$. The overall density of x is shown at the top of the graph and the conditional densities for cases and controls are shown at the bottom. Values of x_γ and $\mathbf{B}_{\gamma 1}$ are shown for $\gamma = W_0$ (labeled "population") and $\gamma = 0.5$ (labeled "equal"). The latter value corresponds to sample weighting if we draw equal numbers of cases and controls. Clearly, survey weighting is estimating the appropriate slope for values of x further out in the upper tail of the distribution (*i.e.*, for individuals at higher risk) than equal weighting in both scenarios.

Note that if we took simple random samples of $n_0 = n_1 = 200$ from the population in Figure 1 (a), the relative efficiency of survey weighting is only about 16%, and the small sample bias is 0.24. In this case, even if we take the population value as our target, the survey weighting leads to a larger mean squared error than sample weighting.

More results are given in Scott and Wild (2002) where we also look at the effect of omitted covariates. This turns out to have a similar, but somewhat smaller, effect to omitting a quadratic term.

Which is the right value of γ to use? That clearly depends on what we want to use the resulting model for. If our primary interest is in using the model for estimating odds ratios at values of x where the probability of a case is higher, and the sample is large enough so that variance and small sample bias are less important, we might use population weights. For smaller sample sizes, or if we are interested in values of x closer to the population mean, sample weights would be better. A value intermediate between population weighting and sample weighting might sometimes be a sensible compromise. For example trimming the weights to 10:1 (*i.e.*, setting $\gamma \approx 0.91$) in the example, instead of 1:1 (sample weighting) or 400:1 (population weighting), leads to an efficiency of 70% and a small sample bias of 0.04. The corresponding values for population weighting were 16% and 0.24. The value of $x_{0.91}$ lies almost exactly half way between $x_{0.5}$ and $x_{0.9975}$.

9. Case-Control Family Studies

If we are primarily interested in the parameters of the marginal model (1), then the methods that we have discussed in previous sections are simple to implement and reasonably efficient. Fully efficient methods require building parametric models for the within-cluster dependence and the extra effort that this would entail is rarely worthwhile. However, there are situations where the dependence structure is of interest in its own right. In particular, it has become increasingly common for genetic epidemiologists to augment data from a standard case-control study with response and covariate information from family members, in an attempt to gain information on the role of genetics and environment. This can be regarded as a stratified cluster sample, with families as clusters, and the intra-cluster structure is of the primary focus of attention here. The following example is fairly typical.

Example 3.

Wrensch, Lee, Miike, Newman, Barger, Davis, Wiencke and Neuhaus (1997) conducted a population-based case-control study of glioma, the most common type of malignant brain tumor, in the San Francisco Bay Area. They collected information on all cases of glioma that were diagnosed in a specified time interval and on a comparable sample of controls obtained through random digit dialing. They also collected brain tumor status and covariate information from family members of the participants in the original case-control sample. There were 476 brain cancer case families and 462 control families in the study.

We could use the methods that we have been discussing to fit a marginal model for the probability of becoming a glioma victim but a major interest of the researchers was the estimation of within-family characteristics. One way of approaching this would be to fit a mixed logistic model with one or more random family effects.

Note that, strictly speaking, the original sampling scheme in Example 3 is not included in this case-control set-up. The stratification here is related to the response variable but not completely determined by it. Stratum 1 contains the 476 families with a case diagnosed in a particular small time interval while Stratum 2 contains the remaining 1,942,490 families, some of which contain brain cancer victims.

In Neuhaus *et al*. (2006) we develop efficient semi-parametric methods for stratified multi-stage sampling in situations where the stratification depends on the response, possibly in an unspecified way that has to be modeled, and observations within a primary sampling unit are related through some parametric model. The estimates require the solution of $p+1$ estimating equations, where p is the dimension of the parameter vector. The covariance matrix can also be estimated in a straightforward way using an analogue of the inverse observed information matrix. The whole procedure can be implemented using any reasonably general maximization routine but this still requires some computing expertise.

We could also fit the same models using survey weighted estimators, which has the big advantage of requiring no specialist software. In our example, case families would have weight 1 and control families would have weight $1{,}942{,}490/462 \approx 4{,}200$. With such a huge disparity, we might expect the weighted estimates to be very inefficient indeed. Unfortunately it turned out to be almost impossible to fit an interesting model for which the weighted estimates

converged. One problem is that the weighted estimates are based almost entirely on the control sample and there is very little information about family effects in the control families. (Another problem is that we did not have information on age for family members and any model without age was grossly mis-specified!) For this reason, we had to resort to simulation which is far from complete at this stage. It seems, however, that the efficiency of weighted estimates is less than 10% of the efficient semiparametric estimates here. More details are given in Neuhaus *et al.* (2002, 2006).

Although our simulations are at a very early stage, it is possible to draw a few tentative conclusions. The main one is that within-family quantities are very poorly estimated, even using fully-efficient procedures. Case-control family designs, where the information on family members is obtained as an add-on to a standard case-control design, simply do not contain enough information to estimate the parameters of interest to genetic epidemiologists unless the associations are extremely (even unrealistically) strong. (I should note that not all genetic epidemiologists would agree with this.) More efficient variants are possible, however. For example, if we can identify families containing more than one case, then it is possible to get much greater efficiency by heavily over-sampling such families. In essence, we would be taking the family as the sampling unit, defining a 'case family' as one containing multiple individual cases and then taking a case-control sample of families. This is an important area where a lot of work still needs to be done.

10. Conclusion

The population-based case-control study is one of those subjects where practice has forged ahead of theory. As far as I know, the only book that discusses the topic in any depth is Korn and Graubard (1999, Chapter 9). One aspect that has received a reasonable amount of theoretical attention in the literature is stratification. Efficient procedures for incorporating stratifying variables in the analysis have been developed by Scott and Wild (1997), Breslow and Holubkov (1997), and Lawless *et al.* (1999), among others, when the variables can take only a finite set of values. Breslow and Chatterjee (1999) have considered how best to use such information at the design stage. The extension of all this (both analysis and design) to situations where we have information on continuous variables such as age for all members of the population is an area that still needs work. Much less has been written on the effect of clustering, even though multi-stage sampling is in common use. Exceptions are Graubard *et al.* (1989), Fears and Gail (2000) and Scott and Wild (2001a). Perhaps this paper might stimulate more work on an important topic. In particular, since the essence of the problem boils down to estimating two population means (see equation (8)), it should be possible to transfer a lot of the expertise about efficient survey design across to this problem.

Acknowledgements

I would like to thank the referees and Barry Graubard and Graham Kalton, whose thoughtful discussion of an early version of this paper helped my understanding of the subject considerably. Finally, I would to give special thanks to my long term collaborators Chris Wild, with whom almost all the work underlying this paper was done, and Jon Rao, with whom I learnt essentially everything that I know about the analysis of survey data.

References

Baker, M., McNicholas, A., Garrett, N., Jones, N., Stewart, J., Koberstein, V. and Lennon, D. (2000). Household crowding: A major risk factor for epidemic meningococcal disease in Auckland children. *Pediatric Infectious Disease Journal*, 19, 983-990

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Breslow, N.E. (1996). Statistics in epidemiology: the case-control study. *Journal of the American Statistical Association*, 91, 14-28.

Breslow, N.E. (2004). Case-control studies. In *Handbook of Epidemiology*. (Eds. W. Aherns and I. Pigeot). New York: Springer. 287-319.

Breslow, N.E., and Cain, K.C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, 75, 11-20.

Breslow, N.E., and Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumor prognosis. *Applied Statistics*, 48, 457-468.

Breslow, N.E., and Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase outcome-dependent sampling. *Journal of the Royal Statistical Society*, B, 59, 447-461.

Brogan, D.J., Denniston, M.M., Liff, J.M., Flagg, E.W., Coates, R.J. and Brinton, L.A. (2001). Comparison of telephone sampling and area sampling: Response rates and within-household coverage. *American Journal of Epidemiology*, 153, 1119-1127.

Cosslett, S.R. (1981). Maximum likelihood estimation for choice-based samples. *Econometrica*, 49, 1289-1316.

DiGaetano, R., and Waksberg, J. (2002). Trade-offs in the development of a sample design for case-control studies. *American Journal of Epidemiology*, 155, 771-775.

Fears, T.R., and Brown, C.C. (1986). Logistic regression models for retrospective case-control studies using complex sampling procedures. *Biometrics*, 42, 955-960.

Fears, T.R., and Gail, M.H. (2000). Analysis of a two-stage case-control study with cluster sampling of controls: Application to nonmelanoma skin cancer. *Biometrics*, 56, 190-198.

Graubard, B.I., Fears, T.R. and Gail, M.H. (1989). Effects of cluster sampling on epidemiologic analysis in population-based case-control sampling. *Biometrics*, 45, 1053-1071.

Hartge, P., Brinton, L.A., Rosenthal, J.F., Cahill, J.I., Hoover, R.N. and Waksberg, J. (1984). Random digit dialing in selecting a population-based control group. *American Journal of Epidemiology*, 120, 825-833.

Hartge, P., Brinton, L.A., Cahill, J.I., West, D., Hauk, M., Austin, D., Silverman, D. and Hoover, R.N. (1984). Design and methods in a multi-center case-control interview study. *American Journal of Public Health*, 74, 52-56.

Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons, Inc.

Lawless, J.F., Kalbfleisch, J.D. and Wild, C.J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society*, B, 61, 413-38.

Lee, A.J., Scott, A.J. and Wild, C.J. (2006). Fitting binary regression models with case-augmented samples. *Biometrika*, 95 (to appear).

Manski, C.F., and McFadden, D. (Eds) (1981). *Structural Analysis of Discrete Data with Econometric Applications*. New York: John Wiley & Sons, Inc.

Miettinen, O.S. (1985). The case-control study: Valid selection of subjects. *American Journal of Epidemiology*, 135, 1042-1050.

Prentice, R.L., and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66, 403-411.

Neuhaus, J., Scott, A.J. and Wild, C.J. (2002). The analysis of retrospective family studies. *Biometrika*, 89, 23-37.

Neuhaus, J., Scott, A.J. and Wild, C.J. (2006). Family-specific approaches to the analysis of retrospective family data. *Biometrics*, 62, in press.

Rao, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.

Rao, J.N.K., Scott, A.J. and Skinner, C.J. (1998). Quasi-score tests with survey data. *Statistica Sinica*, 8, 1059-1070.

Scott, A.J., and Wild, C.J. (1986). Fitting logistic models under case-control or choice-based sampling. *Journal of the Royal Statistical Society*, B, 48, 170-182.

Scott, A.J., and Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 83, 57-72.

Scott, A.J., and Wild, C.J. (2001a). The analysis of clustered case-control studies. *Applied Statistics*, 50, 57-71.

Scott, A.J., and Wild, C.J. (2001b). Fitting regression models to case-control data by maximum likelihood. *Journal of Statistical Planning and Inference*, 96, 3-27.

Scott, A.J., and Wild, C.J. (2002). On the robustness of weighted methods for fitting model to case-control data by maximum likelihood. *Journal of the Royal Statistical Society*, B, 64, 207-220.

Wacholder, S., McLaughlin, J.K., Silverman, D.T. and Mandel, J.S. (1991). Selection of controls in case-control studies. I. Principles. *American Journal of Epidemiology*, 135, 1019-1028.

Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

Waksberg, J. (1998). Random digit dialing sampling for case-control studies. In *Encyclopedia of Biostatistics*. (Eds. P.Armitage and T. Colton). New York: John Wiley & Sons, Inc., 3678-3682.

Wrensch, M., Lee, M., Miike, R., Newman, B., Barger, G., Davis, R., Wiencke, J. and Neuhaus, J. (1997). Familial and personal medical history of cancer and nervous system conditions among adults with glioma and controls. *American Journal of Epidemiology*, 145, 581-93.

Survey Methodology, December 2006
Vol. 32, No. 2, pp. 133-142
Statistics Canada, Catalogue No. 12-001

Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors

Phillip S. Kott [1]

Abstract

Calibration weighting can be used to adjust for unit nonresponse and/or coverage errors under appropriate quasi-randomization models. Alternative calibration adjustments that are asymptotically identical in a purely sampling context can diverge when used in this manner. Introducing instrumental variables into calibration weighting makes it possible for nonresponse (say) to be a function of a set of characteristics other than those in the calibration vector. When the calibration adjustment has a nonlinear form, a variant of the jackknife can remove the need for iteration in variance estimation.

Key Words: Prediction model; Quasi-randomization model; Quasi-randomization consistent; Instrumental variable; Generalized raking.

1. Introduction

Calibration weighting was originally developed as a method for reducing sampling errors while retaining randomization consistency. Deville and Särndal (1992) demonstrated that many alternative forms of calibration weighting are asymptotically identical in the sampling context. This lead to a breakthrough in our understanding of common weight adjustment methods like raking that do not appear in generalized-regression (GREG) estimator format.

Folsom and Singh (2000) showed that calibration weighting can also be used to adjust for known coverage errors and/or unit nonresponse under appropriate quasi-randomization models. Their work is not in the refereed literature. The heart of this article repeats key results in Folsom and Singh including a necessary modification of the Deville-Särndal approach to model variance/randomization mean-squared-error estimation in this expanded context. An earlier, strictly linear version of calibration weighting for unit-nonresponse adjustment can be found in Fuller, Loughin and Baker (1994). See also Lundström and Särndal (1999).

A distinction is drawn between the prediction model usually underpinning calibration and the quasi-randomization model in Folsom and Singh. Unlike in Folsom and Singh, however, both properties are explored here. Furthermore, the explanatory variables in the quasi-randomization model are allowed to differ from the calibration variables. This is likewise allowed in Lundström and Särndal.

A new jackknife is proposed which is analogous to the Deville-Särndal linearization variance estimator. It employs replicate weights computed in one step even though the calibration weights themselves may be determined iteratively.

After introducing the popular notion of calibration weighting, Section 2 provides a review of the GREG special case in a purely sampling context. Section 3 describes Estevao and Särndal's (2000) extension of calibration weighting in its linear form to include instrumental variables. Section 4 expands Deville and Särndal's treatment of calibration weighting to include the possibility of instrumental variables. Section 5 reviews variance/mean squared error estimation, proposing a new jackknife for certain designs. Section 6 describes how calibration weighting can be used to adjust for nonresponse. In this context, alternative functional forms of calibration weighting need no longer be asymptotically identical. Section 7 discusses quasi-randomization models for coverage errors, that is, frame under- or over-coverage. Section 8 contains a small empirical example supporting the new jackknife. Section 9 provides a discussion of alternative approaches and areas for future research.

2. Calibration Weighting and the GREG Estimator

Suppose we knew the selection probability, π_k, for each sample element k in the sample S. We can estimate any population total, $T_y = \sum_U y_k$, where U denotes the population, with the expansion estimator $t_{y\_E} = \sum_S y_k / \pi_k = \sum_U y_k I_k / \pi_k$, where $I_k = 1$ when $k \in S$ and 0 otherwise. Treating the I_k as random variables, it is easy to see that t_{y_E} is an unbiased estimator for T_y. Properties arising when the I_k are treated as random variables are called *randomization-based*. We can also write $t_{y\_E} = \sum_U a_k y_k = \sum_S a_k y_k$, where $a_k = I_k / \pi_k$ is the *sampling weight* of element k.

Deville and Särndal (1992) coined the term "calibration estimator" to describe an estimator of the form $t_{y\_\mathrm{CAL}} = \sum_S w_k y_k$, where $\sum_S w_k \mathbf{x}_k = \sum_U \mathbf{x}_k = T_{\mathbf{x}}$ for some

1. Phillip S. Kott, National Agricultural Statistics Service, 3251 Old Lee Highway, Fairfax, VA 22030, U.S.A. E-mail: pkott@nass.usda.gov.

row vector of auxiliary variables, $\mathbf{x}_k = (x_{1k}, \ldots, x_{Pk})$, about which $T_{\mathbf{x}}$ is known. Since there is generally a continuum of sets $\{w_k | k \in S\}$ that satisfy the *calibration equation*:

$$\sum_{k \in S} w_k \mathbf{x}_k = T_{\mathbf{x}}, \tag{1}$$

Deville and Särndal required that the difference between the set of weights, $\{w_k | k \in S\}$, satisfying equation (1) and $\{a_k | k \in S\}$ minimize some loss function.

An alternative approach to survey sampling treats the y_k as random variables satisfying the linear prediction model:

$$y_k = \mathbf{x}_k \boldsymbol{\beta} + \varepsilon_k, \tag{2}$$

where $E(\varepsilon_k | \{\mathbf{x}_g, I_g | g \in U\}) = 0$ for all $k \in U$. By conditioning this expectation on the I_g, we are assuming the sampling mechanism can be ignored. This is a crucial, and sometimes unreasonable, aspect of the (prediction) *model-based* framework.

It is easy to see that t_{y_CAL} is an unbiased estimator for T_y under the model in the sense that $E_\varepsilon(t_{y\_\mathrm{CAL}} - T_y) = 0$ (suppressing the conditioning for notational convenience); the subscript ε refers to treating the ε_k as random variables (and the I_k as fixed constants).

For our purposes, the general(ized) regression or GREG estimator has the form:

$$t_{y\_\mathrm{GREG}} = t_{y\_E} + \left(T_{\mathbf{x}} - \sum_{k \in S} a_k \mathbf{x}_k\right)\left(\sum_{k \in S} c_k a_k \mathbf{x}_k' \mathbf{x}_k\right)^{-1} \sum_{k \in S} c_k a_k \mathbf{x}_k' y_k, \tag{3}$$

where c_k is an arbitrary constant which may or may not be a function of $\mathbf{x}_k$, and $\lim_{N \to \infty} \sum_U c_k \mathbf{x}_k' \mathbf{x}_k / N = \boldsymbol{\Lambda}$ is a positive definite matrix, where N is the size of U. This last condition means that $\sum_S c_k a_k \mathbf{x}_k' \mathbf{x}_k$ will usually be invertible in practice. We will assume that it is always invertible for convenience.

The GREG estimator in equation (3) can be rewritten in calibration form as $t_{y\_\mathrm{GREG}} = \sum_S w_k y_k$, where

$$w_k = a_k + \left(T_{\mathbf{x}} - \sum_{j \in S} a_j \mathbf{x}_j\right)\left(\sum_{j \in S} c_j a_j \mathbf{x}_j' \mathbf{x}_j\right)^{-1} c_k a_k \mathbf{x}_k'.$$

Strictly speaking, the w_k are functions of the realized sample, S, and the $c_k a_k$, but we suppress that in the notation for convenience. Observe that the calibration weights can be expressed as

$$w_k = a_k (1 + c_k \mathbf{x}_k \mathbf{q}), \tag{4}$$

where $\mathbf{q} = [(\sum_S a_j c_j \mathbf{x}_j' \mathbf{x}_j)^{-1}]'(T_{\mathbf{x}} - \sum_S a_j \mathbf{x}_j)'$ is a column vector, since $\mathbf{x}_k \mathbf{q} = \mathbf{q}' \mathbf{x}_k'$.

Let us assume that reasonable regularity conditions hold (see, for example, Kott 2004a for a more thorough treatment) and the sample plan is such that $t_{y\_E} - T_y = O_P(N/\sqrt{n})$, where n is the expected size of S (the actual size can be random), $\sum_S a_k \mathbf{x}_k - T_{\mathbf{x}} = \mathbf{O}_P(N/\sqrt{n})$, and $\sum_S a_k c_k \mathbf{x}_k' \mathbf{f}_k - \sum_U c_k \mathbf{x}_k' \mathbf{f}_k = \mathbf{O}_P(N/\sqrt{n})$, where $\mathbf{f}_k$ can be $\mathbf{x}_k$ or y_k. Let $e_k = y_k - \mathbf{x}_k (\sum_U c_i \mathbf{x}_i' \mathbf{x}_i)^{-1} \sum_U c_i \mathbf{x}_i' y_i$, so that $\sum_U c_i \mathbf{x}_i' e_i = 0$, and $\sum_S a_k c_k \mathbf{x}_k' e_k = \mathbf{O}_P(N/\sqrt{n})$. We can express the error of t_{y_GREG} as

$$\begin{aligned}
&t_{y_\mathrm{GREG}} - T_y \\
&= \sum_{k \in S} w_k y_k - \sum_{k \in U} y_k \\
&= \sum_{k \in S} w_k e_k - \sum_{k \in U} e_k \quad \left(\text{since } \sum_{k \in S} w_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k\right) \\
&= \sum_{k \in S} a_k e_k + \left(T_{\mathbf{x}} - \sum_{k \in S} a_k \mathbf{x}_k\right)\left(\sum_{k \in S} a_k c_k \mathbf{x}_k' \mathbf{x}_k\right)^{-1} \sum_{k \in S} a_k c_k \mathbf{x}_k' e_k \\
&\qquad - \sum_{k \in U} e_k \\
&= \sum_{k \in S} a_k e_k - \sum_{k \in U} e_k + O_P(N/n).
\end{aligned} \tag{5}$$

It is now not hard to see that the GREG estimator is randomization consistent; that is, $p\lim_{n \to \infty} [(t_{y\_\mathrm{GREG}} - T_y)/N] = 0$. Moreover, both the relative randomization bias and relative randomization mean squared error of the GREG estimator are order $1/n$. Since mean squared error $= \mathrm{bias}^2 +$ variance, we can conclude that the randomization bias of the GREG estimator is usually an asymptotically insignificant contributor to its mean squared error.

3. Redefining Calibration Weights

In their original definition of calibration weights, Deville and Särndal (1992) required that the set of calibration weights, $\{w_k | k \in S\}$ minimize some distance function between the members of the set and the original sampling weights, the a_k, subject to satisfying the calibration equation. As a result, the calibration estimator, $t_{y\_\mathrm{CAL}} = \sum_S w_k y_k$, was both unbiased under the model in equation (2) and usually randomization consistent.

Estevao and Särndal (2002) suggested removing the requirement that the calibration weights minimize a distance function. Instead, they essentially proposed that the w_k need only satisfy the calibration equation and be of the "functional form:"

$$w_k = a_k (1 + \mathbf{h}_k \mathbf{q}), \tag{6}$$

where $\mathbf{h}_k$ is a row vector with the same dimension as $\mathbf{x}_k$ such that $\sum_S a_k \mathbf{h}_k' \mathbf{x}_k$ is invertible, and $\mathbf{q}$ is a column vector of that same dimension. Equation (6) is a mild generalization of (4) where $\mathbf{h}_k$ effectively replaces $c_k \mathbf{x}_k$.

It is not hard to see that $\mathbf{q} = [(\sum_S a_j \mathbf{h}_j' \mathbf{x}_j)^{-1}]'(T_{\mathbf{x}} - \sum_S a_j \mathbf{x}_j)'$. Moreover, under mild conditions we assume to hold, $t_{y\_\mathrm{CAL}} = \sum_S w_k y_k = \sum_S a_k y_k + (T_{\mathbf{x}} - \sum_S a_j \mathbf{x}_j)(\sum_S a_j \mathbf{h}_j' \mathbf{x}_j)^{-1} \sum_S a_k \mathbf{h}_k' y_k$ is randomization consistent

whenever t_{y_E} is. It is unbiased under the linear prediction model in equation (2) when $E(\varepsilon_k|\{\mathbf{x}_g, \mathbf{h}_g|g\in S\}, \{I_g|g\in U\}) = 0$ for all $k \in U$.

This suggests another alternative definition of calibration weights: a set of weights, $\{w_k|k \in S\}$, such that,

i. the w_k satisfy the calibration equation for $\{\mathbf{x}_k|k \in U\}$ and,

ii. $t_{y\_\mathrm{CAL}} = \sum_S w_k y_k$ is randomization consistent whenever t_{y_E} is under mild conditions.

That is the definition we will use. This broadened definition of calibration weighting will prove very helpful when using calibration to adjust for nonresponse or coverage errors.

It follows from our new definition that Estevao and Särdnal's functional-form calibration is indeed a form a calibration weighting. Borrowing from econometric theory, the components of $\mathbf{h}_k$ that are not linear combinations of components of $\mathbf{x}_k$ are called "instrumental variables."

4. Possibly Nonlinear Calibration

Building on ideas in Deville and Särndal (1992), we can generalize the linear form for the calibration weights in equation (6) to

$$w_{k\_\mathrm{GEN}} = a_k f(\mathbf{h}_k \mathbf{q}^*), \qquad (7)$$

where f is a monotonic, twice-differentiable function with $f(0) = 1, f'(0) = 1$ ($f'(0)$ is the first derivative of f evaluated at 0), and $\mathbf{q}^*$ is chosen so that the calibration equation holds. Unlike the calibration-weight equation above, the calibration equation itself, $\sum_S w_k \mathbf{x}_k = T_\mathbf{x}$, remains linear. Note that since $f(0) = 1, f'(0) = 1$, $f(\mathbf{h}_k \mathbf{q}^*) \approx 1 + \mathbf{h}_k \mathbf{q}^*$.

Strictly speaking, there should be an additional symbol on w_{k_GEN} (and later on w_{k_LIN}) to denote the particular choice of $\mathbf{h}_k$. It has been dropped for convenience.

A solution, $\mathbf{q}^*$, to equation (7) can often be reached iteratively. One can start with $\mathbf{q}^{(0)} = \mathbf{0}$; that is, $\sum_S w_k^{(0)} y_k$, where $w_k^{(0)} = a_k f(0)$. For $r = 1, 2, \ldots$, one then sets $\mathbf{q}^{(r)} = \mathbf{q}^{(r-1)} + \{[\sum_s f'(\mathbf{h}_k \mathbf{q}^{(r-1)}) a_k \mathbf{x}_k' \mathbf{h}_k]^{-1}\}'(T_\mathbf{x} - \sum_S w_k^{(r-1)} \mathbf{x}_k)'$, and $w_k^{(r)} = a_k f(\mathbf{h}_k \mathbf{q}^{(r)})$. Iteration stops at r^* when $T_\mathbf{x} = \sum_S w_k^{(r^*)} \mathbf{x}_k$ for all practical purposes. One should be aware, however, that *there may not be a set of weights that can be expressed in the form of equation* (7) *while satisfying the calibration equation*.

Note that $\mathbf{q}^{(1)}$ above equals the $\mathbf{q}$ in $w_{k\_\mathrm{LIN}} = a_k (1 + \mathbf{h}_k \mathbf{q})$. A Taylor expansion around zero reveals $f(\mathbf{h}_k \mathbf{q}^{(1)}) = 1 + \mathbf{h}_k \mathbf{q}^{(1)} + O_P(1/n)$ under mild conditions, so $\sum_S w_k^{(1)} y_k = \sum_S w_{k\_\mathrm{LIN}} y_k + O_P(N/n) = T_y [1 + O_P(1/n)]$. Furthermore, it is not difficult to see that $w_{k\_\mathrm{GEN}} = w_{k\_\mathrm{LIN}}[1 + O_P(1/n)]$, an equality that proves helpful in variance estimation.

The most common example in practice of a nonlinear f is $f(\mathbf{h}_k \mathbf{q}) = \exp(\mathbf{x}_k \mathbf{q})$, where the values of each of the components of $\mathbf{x}_k$, denoted $x_{1k}, \ldots, x_{Pk}$, are either 0 or 1. That is effectively the form of Deming and Stephan's (1940) raking weights computed via iterative proportional fitting. Many have observed that the iterative routine described above can be used even when the components of $\mathbf{x}_k$ are not binary as they are in Deming and Stephan. Note that the *generalized raking* calibration weights that result are always nonnegative.

5. Variance Estimation

Särndal, Swensson, and Wretman (1989) proposed this *plug-in* model variance/randomization mean-squared-error estimator for t_{y_GREG} under an arbitrary sampling plan:

$$v_{\mathrm{SSW}} = \sum_{k\in S} \sum_{j\in S} [(\pi_{kj} - \pi_k \pi_j)/\pi_{kj}](w_k r_k)(w_j r_j). \qquad (8)$$

The term derives from r_k being "plugged into" $v_{\mathbf{ssw}}$ in place of the unknown $e_k = y_k - \mathbf{x}_k (\sum_U \mathbf{h}_i' \mathbf{x}_i)^{-1} \sum_U \mathbf{h}_i' y_i$ for randomization-mean-squared-error estimation.

Paralleling arguments in Deville and Särndal (1992), $v_{\mathbf{ssw}}$ also applies more generally to t_{y_CAL} with calibration weights defined by equation (7) with

$$r_k = y_k - \mathbf{x}_k \left(\sum_{j\in S} a_j \mathbf{h}_j' \mathbf{x}_j \right)^{-1} \sum_{j\in S} a_j \mathbf{h}_j' \mathbf{x}_j. \qquad (9)$$

This is because $w_{k\_\mathrm{GEN}} = w_{k\_\mathrm{LIN}}[1 + O_P(1/n)]$, so $\sum_S w_{k\_\mathrm{GEN}} e_k = \sum_S w_{k\_\mathrm{LIN}} e_k + O_P(N/n) = \sum_S a_k e_k + O_P(N/n)$. The last step uses reasoning exhibited in equation (5) with $\mathbf{h}_j$ serving in place of the $c_j \mathbf{x}_j$.

In their article, Deville and Särndal effectively replace the a_j in equation (9) with $w_j = a_j f(\mathbf{h}_j \mathbf{q}^*)$. A different version is given in Demanti and Rao (2004), where the a_j in the equation are replaced by $a_j f'(\mathbf{h}_j \mathbf{q}^*)$. This author noted in a comment accompanying the latter that all three versions of the r_k are asymptotically identical since $f(0) = f'(0) = 1$ and $\mathbf{q}^*$ is asymptotically $\mathbf{0}$. These asymptotic identities may no longer hold when calibration weighting is used to adjust for nonresponse as we shall see in the following section.

Developing asymptotic properties for $v_{\mathbf{ssw}}$ under stratified simple random sampling is a simple matter. In this context, $v_{\mathbf{ssw}}$ collapses to

$$v_{\mathrm{ST1}} = \sum_{\alpha=1}^{A} (n_\alpha/[n_\alpha - 1]) \sum_{k\in S_\alpha} (1 - n_\alpha/N_\alpha) \times \left(w_k r_k - \sum_{j\in S_\alpha} w_j r_j / n_\alpha \right)^2,$$

where S_α denotes the sample of n_α units in stratum $\alpha(\alpha = 1, \ldots, A)$, and U_α the stratum population containing N_α elements.

For a multi-stage sample it makes sense to allow the possibility that ε_k and ε_j in the prediction model are correlated when k and j are elements in the same PSU, but not otherwise. When finite-population correction can be ignored, the model variance of a calibration estimator is approximately $V_m = \sum_{i \in S'} E_\varepsilon[(\sum_{k \in S(i)} w_k \varepsilon_k)^2]$ under mild conditions, where $S(i)$ is the set of sampled elements in PSU i, and S' is the set of PSUs selected in the first stage of sampling.

The following variance estimator, not strictly equal to $v_{\mathbf{SSW}}$, often has good randomization and model-based properties (when the first-stage selection probabilities are all small):

$$v_{\mathrm{ST2}} = \sum_{\alpha=1}^{A} (n_\alpha / [n_\alpha - 1]) \times \left\{ \sum_{j \in S_\alpha} - \left(\sum_{k \in S_{\alpha j}} w_k r_k \right)^2 \frac{\left(\sum_{j \in S_\alpha} \sum_{k \in S_{\alpha j}} w_k r_k \right)^2}{n_\alpha} \right\}, \quad (10)$$

where α denotes a first-stage stratum of PSU's, $n_{1\alpha}$ the number of sampled PSU's in stratum α, S_α the set of sampled PSU's in α, and $S_{\alpha j}$ the set of subsampled elements from PSU j of stratum α. There can be many stages of sampling involved.

It is not hard to show that v_{ST2} is asymptotically indistinguishable from the jackknife variance estimator:

$$v_J = \sum_{\alpha=1}^{A} ([n_\alpha - 1]/n_\alpha) \left\{ \sum_{j \in S_\alpha} (t_{y\_\mathrm{CAL}(\alpha j)} - t_{y\_\mathrm{CAL}})^2 \right\}, \quad (11)$$

where $t_{y\_\mathrm{CAL}(\alpha j)} = \sum_{k \in S} w_{k(\alpha j)} y_k$, and the *jackknife replicate calibration weights* are

$$w_{k(\alpha j)} = w_k a_{k(\alpha j)} / a_k + \left(\sum_{m \in U} \mathbf{x}_m - \sum_{m \in S} w_m [a_{m(\alpha j)} / a_m] \mathbf{x}_m \right) \times \left(\sum_{m \in S} a_{m(\alpha j)} \mathbf{h}'_m \mathbf{x}_m \right)^{-1} a_{k(\alpha j)} \mathbf{h}'_k, \quad (12)$$

where $a_{k(\alpha j)} = 0$ when k is in PSU j of stratum α, $a_{k(\alpha j)} = a_k$ when k is not in stratum α at all, and $a_{k(\alpha j)} = (n_\alpha / [n_\alpha - 1])\, a_k$ otherwise. The $w_{k(\alpha j)}$ are constrained so that $\sum_{k \in S} w_{k(\alpha j)} \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k$ for all αj.

Let $S(\alpha+)$ be the set of *elements* in stratum α (not PSU's like S_α), and $S(\alpha j)$ the set of elements in PSU j of stratum α. Under mild conditions we assume to hold,

$$\sum_U \mathbf{x}_m - \sum_S w_m [a_{m(\alpha j)} / a_m] \mathbf{x}_m = (n_\alpha / [n_\alpha -]) \left(\sum_{S(\alpha j)} w_k \mathbf{x}_k - \sum_{S(\alpha+)} w_k \mathbf{x}_k / n_\alpha \right) = \mathbf{O}_P(N/n),$$

$$\sum_S a_{m(\alpha j)} \mathbf{h}'_m \mathbf{x}_m = \mathbf{O}_P(N),$$

and $$\sum_S a_{m(\alpha j)} \mathbf{h}'_m e_m = \mathbf{O}_P(N/\sqrt{n}).$$

As a result,

$$t_{y\_\mathrm{CAL}} - t_{y\_\mathrm{CAL}} = \sum_s w_{k(\alpha j)} e_k - \sum_s w_k e_k = (n_\alpha / [n_\alpha - 1]) \left(\sum_{s(\alpha+)} w_k e_k / n_\alpha - \sum_{S(\alpha j)} w_k e_k \right) + O_P(N / n^{3/2}),$$

and $v_J = v_{\mathrm{ST2}}[1 + O_P(1/\sqrt{n})]$ when $p\lim_{n \to \infty} (n v_{\mathrm{ST2}} / N^2) > 0$.

The replicate weights defined in equation (12) do not require iteration even when the calibration weights are themselves produced that way. This is a great computation convenience. It not only saves computer time, it avoids the possibility that at iterative solution for the w_k may exist while one for the replicate weights does not.

6. Unit Nonresponse

6.1 Quasi-randomization and Prediction Modeling

In this section we explore handling unit (whole-element) nonresponse as an additional phase of Poisson sampling. That is the essence of a *quasi-randomization* model. Each element k in the original sample, now denoted F, is assumed to have a probability of response, p_k. The probability of elements k and j jointly responding is $p_k p_j$, and whether element k would respond (given a vector of covariates) is independent of whether it is chosen for the original sample.

It is often possible to construct a set of weights so that the calibration estimator is randomization consistent under the quasi-randomization model. We are interested here in a particular way of constructing those weights. To this end, we assume that the quasi-randomization model is correct. Each element has attached to it a row vector of auxiliary variables, $\mathbf{x}_k$, for which $T_{\mathbf{x}} = \sum_U \mathbf{x}_j$ is known. Finally, each p_k is assumed to have the form:

$$p_k = 1 / f(\mathbf{h}_k \boldsymbol{\phi}), \quad (13)$$

where $\boldsymbol{\phi}$ is an unknown column vector, $\mathbf{h}_k$ is a row vector with the same dimension as $\mathbf{x}_k$, and $\sum_S a_k \mathbf{h}'_k \mathbf{x}_k / N$, where S now denotes the "subsample" of respondents, is invertible both for the realized population size, N, and in the probability limit.

The function $f(\cdot)$ in equation (13) is assumed to be monotonic and twice differentiable. Its functional form is known, but the value of the governing parameter, $\boldsymbol{\phi}$, is not. When plugged into the calibration-weight equation,

$w_k = a_k f(\mathbf{h}_k \mathbf{q})$, so that the calibration equation itself, $\sum_s w_k \mathbf{x}_k = T_\mathbf{x}$ holds, $f(\mathbf{h}_k \mathbf{q})$ implicitly estimates the inverse of the element response probabilities. Unlike when calibration is used to correct for $\sum_s a_k \mathbf{x}$ differing from $T_\mathbf{x}$ due purely to sampling error, $f(0)$ and $f'(0)$ do not need to be 1 nor does $\mathbf{h}_k \boldsymbol{\phi}$ need to be zero.

The most obvious choice for $\mathbf{h}_k$ when postulating the response model in equation (13) is $\mathbf{x}_k$ itself. In a common example of calibration weighting for nonresponse, the components of $\mathbf{x}_k$ are indicator variables: $x_{gk} = 1$ when k is in group g and zero otherwise. When the groups are mutually exclusive, calibration weighting is the same thing as reweighting within post-stratification classes. See, for example, Särndal, Swensson and Wretman (1992, page 585). The prediction model usually underpinning calibration (the prefix "prediction" is needed to distinguish this model from the quasi-randomization one) assumes that every element k in group g, whether or not it would respond, has a common mean: $E_\varepsilon(y_k) = \beta_g$. The quasi-random response model is analogous: $p_k = 1/\phi_g$. The two models are conceptually very different, however.

When the groups are not mutually exclusive, raking is one method of determining calibration weights. There are others depending on the exact form of the assumed response function $f(\cdot)$. The prediction model remains linear, $E_\varepsilon(y_k) = \mathbf{x}_k \boldsymbol{\beta}$, while the response model that leads to raking, $p_k = \exp\{-\mathbf{x}_k \boldsymbol{\phi}\}$, does not. Berry, Flatt, and Pierce (1996) provides an example of using raking to adjust for nonresponse.

In many applications of calibration weighting the components of $\mathbf{x}_k$ are continuous or semi-continuous rather than dichotomous. In an annual crop survey, for example, let x_{1k} be the quantity of corn harvested in the previous census of agriculture by farm k, x_{2k} be the farm's harvested wheat, x_{3k} its harvested potatoes, and so forth. The annual crop survey has an assumed prediction model for farm k's planted corn acres, y_{1k}, of the form: $y_{1k} = \mathbf{x}_k \beta_{1k} + \varepsilon_{1k}$. The subscript, 1, is corn-specific. There are other survey values of interest, like planted wheat acres, and potentially assumed prediction models for each.

The quasi-random response model for the crop survey depends on assumptions about $f(\cdot)$ and $\mathbf{h}_k$ in equation (13) with $\mathbf{h}_k$ possibly equal to $\mathbf{x}_k$. Unlike the prediction model, the same assumed quasi-randomization model applies for all survey variables.

Promising choices for $f(\cdot)$ are $\exp(\cdot)$ and $1 + \exp(\cdot)$, the latter corresponding to a response probabilities being fit by a logistic function of $\mathbf{h}_k \boldsymbol{\phi}$. It may also be reasonable to assume $h_{gk} = x_{gk}^\lambda$ for $\lambda < 1$. In particular, setting $\lambda = 0$ means that the probability of farm k responding to the annual crop survey depends only on whether the farm had corn, wheat, or potatoes on the previous census of agriculture rather than on how much of those crops it had.

In the crop-survey example, the components of $\mathbf{x}_k$ from the previous census were the best predictors available for the corresponding annual survey values *before* sampling. Whether farm k responds to the survey, however, is more likely a function of the farm's current planted corn acres, if any, than on a predetermined proxy for that value. As a result, placing survey values in $\mathbf{h}_k$ rather than corresponding census values is tempting. There is a theoretical problem with this procedure as we shall see.

Given an $f(\cdot)$, the iterative method described in Section 4 will often be able to uncover a row vector $\mathbf{q}$ such that $T_\mathbf{x} = \sum_s a_k f(\mathbf{h}_k \mathbf{q}) \mathbf{x}_k$. When that happens, estimating T_y with $t_{y\_\mathrm{CAL}} = \sum_s w_k y_k$, where $w_k = a_k f(\mathbf{h}_k \mathbf{q})$, will have good properties under the linear prediction model: $y_k = \mathbf{x}_k \boldsymbol{\beta} + \varepsilon_k$, where $E(\varepsilon_k | \{\mathbf{x}_g, \mathbf{h}_g, I_g | g \in U\}) = 0$ for all $k \in U$, $I_k = 1$ if element k is both in the original sample and responds, 0 otherwise.

Prediction-model unbiasedness is simply a result of the weights satisfying the calibration equation. Note, however, that if components of $\mathbf{h}_k$ come from the survey rather than $\mathbf{x}_k$, the prediction-model assumption that $E(\varepsilon_k | \mathbf{h}_k) = 0$ can be problematic. At the extreme, consider the case where one such component is y_k itself. Usually, $E(\varepsilon_k | y_k)$ is not 0. In the crop-survey example described earlier, y_k can be the annual corn acres planted on farm k. Putting this value in $\mathbf{h}_k$ makes the associated calibration estimator for corn prediction-model biased.

When the prediction model is correct (treating $E(\varepsilon_k | \{\mathbf{x}_g, \mathbf{h}_g, I_g | g \in U\}) = 0$ as an integral part of the model), however, calibration weighting based on any choice of $f(\cdot)$ will produce estimators with good prediction-model-based properties. These estimators will also have good quasi-randomization properties when the response model in equation (13) is correct for that choice of $f(\cdot)$. In some sense, one model provides protection against the failure of the other. See Kott (1994).

As noted, the prediction model is more likely to hold when $\mathbf{h}_g = \mathbf{x}_g$. Even then, sometimes the ε_k in the model in equation (2) satisfy $E(\varepsilon_k | \{\mathbf{x}_g | g \in U\}) = 0$, but not $E(\varepsilon_k | \{\mathbf{x}_g I_g | g \in U\}) = 0$; that is to say, the sampling mechanism – including response – is not ignorable with respect to the prediction model.

We can factor I_k into $I_{k1} I_{k2}$, where $I_{k1} = 1$ if and only if k is in the original sample, and $I_{k2} = 1$ if and only if k would respond if sampled. The interested reader can confirm that calibration weighting provides some protection against bias if the prediction model in equation (2) holds when $E(\varepsilon_k | \{\mathbf{x}_g, \mathbf{h}_g, I_{g2} | g \in U\}) = 0$; that is when the response mechanism is ignorable with respect to the

prediction model but not necessarily the original sampling mechanism.

6.2 Quasi-randomization Mean Squared Error Estimation

Whether or not t_{y_CAL} can reasonably be called prediction-model unbiased has no effect on its quasi-randomization-based properties. Note that $\mathbf{h}_k\boldsymbol{\phi}$ are $\mathbf{h}_k\mathbf{q}$ are scalar values not vectors. Since $T_{\mathbf{x}} = \sum_s a_k f(\mathbf{h}_k\mathbf{q})\mathbf{x}_i$, our assumptions and the mean value theorem $(f(\mathbf{h}_k\boldsymbol{\phi}) = f(\mathbf{h}_k\mathbf{q}) + f'(\theta_k)(\mathbf{h}_k\boldsymbol{\phi} - \mathbf{h}_k\mathbf{q}))$ reveal

$$T_{\mathbf{x}} - \sum_{k\in S} a_k f(\mathbf{h}_k\boldsymbol{\phi})\mathbf{x}_k = \sum_{k\in S} a_k[f'(\theta_k)\mathbf{h}_k(\mathbf{q}-\boldsymbol{\phi})]\mathbf{x}_k = \mathbf{O}_P(N/\sqrt{n})$$

for some scalar θ_k between each $\mathbf{h}_k\mathbf{q}$ and $\mathbf{h}_k\boldsymbol{\phi}$. From this we see that if $\sum_s a_j f'(\mathbf{h}_j\boldsymbol{\phi})_j\mathbf{h}'_j\mathbf{x}_j/N$ is invertible both for the realized N and at the probability limit (recall that f is monotonic so f' is never zero), then

$$\begin{aligned}\mathbf{q}-\boldsymbol{\phi} &= \left\{\left[\sum_{j\in S} a_j f'(\mathbf{h}_j\mathbf{q})_j\mathbf{h}'_j\mathbf{x}_j\right]^{-1}\right\}'\left[T_{\mathbf{x}} - \sum_{i\in S} a_i f(\mathbf{h}_i\boldsymbol{\phi})\mathbf{x}_i\right]' \\ &= \mathbf{O}_P(1/\sqrt{n}) \\ &= \left\{\left[\sum a_j f'(\mathbf{h}_j\boldsymbol{\phi})_j\mathbf{h}'_j\mathbf{x}_j\right]^{-1}\right\}'\left[T_{\mathbf{x}} - \sum a_i f(\mathbf{h}_i\boldsymbol{\phi})\mathbf{x}_i\right]' \\ &\quad + \mathbf{O}_P(1/n).\end{aligned}$$

The estimator t_{y_CAL} has an error of

$$t_{y\_\mathrm{CAL}} - T_y = \sum_{k\in S} a_k f(\mathbf{h}_k\mathbf{q})y_k - \sum_{k\in U} y_k = \sum a_k f(\mathbf{h}_k\mathbf{q})e_k - \sum e_k,$$

where

$$e_k = y_k - \mathbf{x}_k\left(\sum\nolimits_U f'(\mathbf{h}_j\boldsymbol{\phi})p_j\mathbf{h}'_j\mathbf{x}_j\right)^{-1}\sum\nolimits_U f'(\mathbf{h}_j\boldsymbol{\phi})p_j\mathbf{h}'_j y_j,$$

and $p_j = 1/f(\mathbf{h}_j\boldsymbol{\phi})$. The e_k are again unknown. They have been design so that $\sum_S a_k f'(\mathbf{h}_k\boldsymbol{\phi})\mathbf{h}'_k e_k = \mathbf{O}_P(N/\sqrt{n})$. Continuing:

$$\begin{aligned}&t_{y\_\mathrm{CAL}} - T_y \\ &= \sum_{k\in S} a_k f(\mathbf{h}_k\boldsymbol{\phi})e_k - \sum_{k\in U} e_k + \sum_{k\in S} a_k\{f(\mathbf{h}_k\mathbf{q}) - f(\mathbf{h}_k\boldsymbol{\phi})\}e_k \\ &= \sum a_k f(\mathbf{h}_k\boldsymbol{\phi})e_k - \sum e_k + \sum a_k f'(\mathbf{h}_k\boldsymbol{\phi})\mathbf{h}_k(\mathbf{q}-\boldsymbol{\phi})e_k \\ &\quad + O_P(N/n) \\ &= \sum a_k f(\mathbf{h}_k\boldsymbol{\phi})e_k - \sum e_k + (\mathbf{q}-\boldsymbol{\phi})'\sum a_k f'(\mathbf{h}_k\boldsymbol{\phi})\mathbf{h}'_k e_k \\ &\quad + O_P(N/n) \\ &= \sum a_k f(\mathbf{h}_k\boldsymbol{\phi})e_k - \sum e_k + O_P(N/n). \qquad (14)\end{aligned}$$

Thus, t_{y_CAL} is quasi-randomization consistent under mild conditions whenever $t = \sum_S a_k f(\mathbf{h}_k\boldsymbol{\phi})y_k$ is.

To estimate the quasi-randomization mean squared error of t_{y_CAL} (*i.e.*, the estimator's randomization mean squared error under the response model), we first note that the probability that elements k and $j, k \neq j$, are both in the respondent subsample is $\pi^*_{kj} = \pi_{kj}p_k p_j$. Let $\pi^*_k = \pi_k p_k$, and recall that $a_k = 1/\pi_k$ and $1/p_k = f(\mathbf{h}_k\boldsymbol{\phi})$. From equation (14), we see that the quasi-randomization mean squared error of t_{y_CAL} is approximately

$$\begin{aligned}&E_1[(t_{y\_\mathrm{CAL}} - T_y)^2] \\ &\approx \sum_{k\in U}\sum_{j\in U}(\pi^*_{kj} - \pi^*_k\pi^*_j)(e_k/\pi^*_k)(e_j/\pi^*_j) \\ &= \sum_{k\in U}(1-\pi^*_k)e_k^2/\pi^*_k \\ &\quad + \sum_{k\in U}\sum_{\substack{j\in U\\ k\neq j}}(\pi_{kj} - \pi_k\pi_j)(e_k/\pi_k)(e_j/\pi_j). \qquad (15)\end{aligned}$$

If the original sample is Poisson, then $v_m = \sum_s (w_k^2 - w_k)r_k^2$ with

$$r_k = y_k - \mathbf{x}_k\left[\sum_{j\in S} a_j f'(\mathbf{h}_j\mathbf{q})\mathbf{h}'_j\mathbf{x}_j\right]^{-1}\sum_{j\in S} a_j f'(\mathbf{h}_j\mathbf{q})\mathbf{h}'_j y_j, \quad (16)$$

serves as both a reasonable estimator for prediction-model variance and quasi-randomization mean squared error under mild conditions, since $w_k \approx 1/\pi^*_k$ and $r_k \approx e_k$. A close relative of the non-intuitive sample residual in equation (16) can be found in Folsom and Singh (2000). See Kott (2004a) for a further discussion of v_m in a purely sampling context.

For a general design, we can get close to a good variance/mean-squared-error estimator with

$$\begin{aligned}v_{\mathrm{com}} &= \sum_{k\in S}(w_k^2 - w_k)r_k^2 \\ &\quad + \sum_{k\in S}\sum_{\substack{j\in S\\ k\neq j}}[(\pi_{kj} - \pi_k\pi_j)/\pi_{kj}](w_k r_k)(w_j r_j). \qquad (17)\end{aligned}$$

The right hand side of equation (17) estimates the right hand side of equation (15) with r_k replacing e_k. Note that $\sum_U (1-\pi^*_k)e_k^2/\pi^*_k$ in equation (15) is estimated by $\sum_S (w_k^2 - w_k)r_k^2$ rather than $\sum_S w_k^2(1-\pi^*_k)r_k^2$, which would make v_{com} more consistent with v_{SSW} in equation (8). This substitution results in a variance estimator with good prediction-model-based properties when the ε_k are uncorrelated, and $\sigma_k^2 = \mathbf{x}_k\boldsymbol{\zeta}$, for some $\boldsymbol{\zeta}$. It can be made even in the absence of nonresponse.

When the actual sample is multistage, and the first stage selection probabilities are ignorably small, v_{ST2} in equation (10) can be used as the variance/mean-squared-error estimator with r_k defined once more by equation (16).

When f is linear, $f'(\theta) = 1$, and the r_k in equation (16) are computed as if there were no nonresponse. The same holds true for the variance/mean-squared-error estimator v_{ST2}. Unfortunately, this f corresponds to an awkward-looking response-probability function: $p_k = 1/\mathbf{h}_k\boldsymbol{\phi}$. Fuller, Loughin and Baker (1994) made these observations for the case where $\mathbf{h}_k = c_k\mathbf{x}_k$.

The jackknife, v_J, in equation (11) can be computed with these jackknife replicate weights:

$$w_{k(\alpha j)} = w_k a_{k(\alpha j)}/a_k + \left(\sum_{m\in U}\mathbf{x}_m - \sum_{m\in S} w_m[a_{m(\alpha j)}/a_m]\mathbf{x}_m\right) \times \left(\sum_{m\in S} a_{m(\alpha j)} f'(\mathbf{h}_m\mathbf{q})\mathbf{h}'_m\mathbf{x}_m\right)^{-1} a_{k(\alpha j)} f'(\mathbf{h}_k\mathbf{q})\mathbf{h}'_k, \qquad (18)$$

an obvious generalization of the jackknife replicate weights in equation (12). Again when $f'(\theta)=1, v_J$ can be computed as if there were no nonresponse.

7. Coverage Modeling

Folsom and Singh (2000) pointed out that the treatment of nonresponse through calibration weighting can also be used to adjust for undercoverage. In the context, the quasi-random phase as sampling occurs conceptually before the actual sample is drawn. The population associated with the sampling frame is assumed to be a Poisson sample from a hypothetical complete population for which the vector $T_{\mathbf{x}}$ must be known. The frame population becomes *F*, while the hypothetical complete population is *U*. The probability that element $k\in U$ is in *F* is assumed to be modeled correctly by equation (13). If the first (from *U* to *F*) and second (from *F* to *S*) phases of sampling are independent, then all the theory developed for using calibration weighting to handle nonresponse carries over to handling undercoverage.

It should be noted that coverage adjustment through calibration is a extension of the well-known practice of coverage adjustment through post-stratification often used with telephone surveys. As with the post-stratification special case, one needs quantities for the calibration targets for *U* that can be assumed to be free of error or to have very little mean squared error compared to the calibration estimators themselves.

Folsom and Singh noted that overcoverage (duplication) or a combination of under and overcoverage can be handled with their methodology. The definition of p_k in equation (13) becomes the expected number of times *k* is in the frame, which can now exceed 1 due to potential duplication.

Folsom and Singh further suggested that $f(\cdot)$ have the flexible form:

$$f(\mathbf{x}_k\boldsymbol{\phi}) = \frac{U(C-L)\exp(\mathbf{x}_k\boldsymbol{\phi}) + L(U-C)}{(U-C)+(C-L)\exp(\mathbf{x}_k\boldsymbol{\phi})}, \qquad (19)$$

where $L\ge 0, 1<U\le\infty$, and $L<C\le U$ are predetermined constants. They call this the "General Exponential Model" or "GEM." Observe that when $C=1, U=\infty$, and $L=0$, $p_k = 1/f(\mathbf{x}_k\boldsymbol{\phi}) = \exp(-\mathbf{x}_k\boldsymbol{\phi})$. Similarly, when $C=2$, $U=\infty$, and $L=1$, $p_k = [1+\exp(\mathbf{x}_k\boldsymbol{\phi})]^{-1}$; that is to say, the probability of coverage (or response) is logistic. The values *L* and *U* serve as bounds on the *calibration adjustment*, $f(\cdot)$, while $C=f(0)$ is effectively its center.

The authors made the calibration adjustment in GEM even more flexible by postulating three classes of sampling units, each with its own set of *U*, *C*, and *L* values. They proposed its use both for coverage-error and unit-non-response adjustment

8. A Small Empirical Example

Since the jackknife replicate weights expressed in equation (18) are new, it is prudent to investigate whether they actually work with real data. To this end, the author took the MU281 data from Särndal, Swensson and Wretman (1992) and replicated it 20 times (so N = 5,620). Using stratified simple random sampling, 16 units were selected from each of the eight unequally-sized strata. The variable RMT85 served as y_k and P75 as x_k in $\mathbf{x}_k = (1, x_k)$. Each of the 128 sampled units was given a probability of being in the respondent subsample, *S*, which decreased with the size of x_k; in particular, $p_k = \exp(-0.35\, x_k/M_x)$, where M_x was the population mean of the x_k. In 1,600 simulations, the size of the *S* ranged from 78 to 110, with an average of approximately 93.8.

The total T_y was estimated two ways, with $t_{y\_\mathrm{LIN}} = \sum_S a_k(1+\mathbf{x}_k\mathbf{q})y_k$ and with $t_{y\_\mathrm{EXP}} = \sum_S a_k \exp(\mathbf{x}_k\mathbf{q}^{(\exp)})y_k$, where $\mathbf{q}$ and $\mathbf{q}^{(\exp)}$ were respectively selected so that the calibration equation held. The former was a GREG estimator, while the latter was a generalized raking estimator. Both estimators were unbiased under the implied prediction model $(y_k = \mathbf{x}_k\boldsymbol{\beta}+\varepsilon_k)$, but only t_{y_EXP} was randomization consistent under the correct response model. The GREG implicitly assumed $p_k = 1/(\phi_0^{(\mathrm{LIN})} + \phi_1^{(\mathrm{LIN})}\mathbf{x}_k)$ for unknown $\phi_0^{(\mathrm{LIN})}$ and $\phi_1^{(\mathrm{LIN})}$.

The small size of the sample relative to the population in each stratum allowed the ignoring of finite population correction in variance/mean-squared-error estimation (called "variance estimation" from now on). Variances were estimated using, *i*, the linearization estimator, v_{ST2}, in equation (10) with r_k defined by equation (16), and, *ii*, the proposed jackknife, v_J, in equation (11) with replicate weights defined by equation (18). To make the jackknife computations easier, the 16 samples in each stratum were randomly assigned to one of four clusters, so that only 32 jackknife replicates had to be computed.

For comparison purposes, a better version of the linearization variance estimator, labeled $v_{\mathrm{ST2}(e)}$, was also computed with r_k replaced by $e_k = y_k - \mathbf{x}_k(\sum_U f'(\mathbf{x}_j\boldsymbol{\phi})p_j\mathbf{x}'_j\mathbf{x}_j)^{-1}\sum_U f'(\mathbf{x}_j\boldsymbol{\phi})p_j\mathbf{x}'_j y_j$, where $\boldsymbol{\phi}$ and p_j were known. In practice, e_k is rarely known, but computing $v_{\mathrm{ST2}(e)}$ is useful here for comparison.

One should note that computations of r_k and e_k were slightly different depending on whether the variance estimator for t_{y_LIN} or for t_{y_EXP} was of interest. For t_{y_LIN}, $f'(\mathbf{x}_j \boldsymbol{\phi}) = f'(\mathbf{x}_j \mathbf{q}) = 1$; for t_{y_EXP}, $f'(\mathbf{x}_j \mathbf{q}^{(\mathrm{exp})}) = \exp(\mathbf{x}_j \mathbf{q}^{(\mathrm{exp})})$, and $f'(\mathbf{x}_j \boldsymbol{\phi}) = 1/p_j$.

Table 1 displays the empirical means (the mean over the 1,600 simulations) of the two estimators for T_y normalized so that $T_y = 100$. Although both are close to unbiased, t_{y_LIN} is significantly different from 100 at the 0.05 level; t_{y_EXP} is not. This is not surprising, since only the latter is based on the correct response model.

The variance estimators and empirical mean squared errors of each estimator were normalized so that the empirical means of the respective $v_{\mathrm{ST2}(e)}$'s were 100. Neither $v_{\mathrm{ST2}(e)}$ had an empirical mean significantly different from the empirical mean squared error (EMSE) of the associated estimator. This was a bit disappointing. It seems that although t_{y_LIN} had a significant empirical bias, this bias was such a small component of the estimator's mean squared error, that the difference between its EMSE and the empirical mean of $v_{\mathrm{ST2}(e)}$ was not significant.

The $v_{\mathrm{ST2}(e)}$ were chosen as benchmarks for the table rather than the empirical mean squared errors because each $v_{\mathrm{ST2}(e)}$ had roughly half the empirical standard error of the corresponding EMSE (which itself was the average of 1,600 squared differences) and correlated more strongly with the variance estimators. The $t-$values for this part of the table were also computed with respect to the $v_{\mathrm{ST2}(e)}$.

The two linearization variance estimators had surprisingly large downward biases. Apparently, there was a tendency for unusually large w_{k_LIN} and w_{k_EXP} to cause associated r_k to be appreciably smaller than e_k in absolute terms. The problems associated with unusually large w_{k_LIN} and w_{k_EXP} seem to be more muted with the jackknives.

To speed up the asymptotics of the linearization variance estimators (*i.e.*, reduce the difference between r_k and e_k), an *ad-hoc* adjustment of v_{ST2} was computed by replacing each r_k with $r_{k(\mathrm{adjusted})} = r_k / \omega_k$, where $\omega_k^2 = 1 - \mathbf{x}_k (\sum_S a_j f'(\mathbf{x}_j \mathbf{q})\, \mathbf{x}'_j \mathbf{x}_j)^{-1} a_k f'(\mathbf{x}_k \mathbf{q})\, \mathbf{x}'_k = 1 + O_P(1/n)$. Observe that under the prediction model with the ε_k uncorrelated and $E(\varepsilon_k^2) = \sigma_k^2$, $E(r_{k(\mathrm{adjusted})}^2) \approx \sigma_k^2$. The near equality is exact when all the $a_j f'(\mathbf{x}_j \mathbf{q})$ and σ_j, respectively, are equal.

The adjusted v_{ST2} for both t_{y_LIN} and t_{y_EXP} remained biased downward, while the v_J were biased upward by a slightly smaller amount. Although these biases were significant, they were reasonably small (from 4.5 to 11.2%) and suggest that the variance estimators may have indeed been asymptotically unbiased as theoretically demonstrated in previous sections.

Using $v_{\mathrm{ST2}(e)}$ as an efficient proxy for EMSE, the empirical mean squared error of t_{y_EXP}, which incorporated the correct response model, was more than 13% larger than that of the t_{y_LIN}, which did not. One should not generalize broadly based on one data set involving only two calibration variables, however. See Crouse and Kott (2004) for a different set of results.

Table 1
Empirical Means of Estimators Based on 1,600 Simulations*

Empirical mean (standard error)		$t-$value (two-sided significance)	
The Estimators for $T_y (T_y = 100)$			
t_{y_LIN}	99.84 (0.06)	−2.79 (0.02)	difference from
t_{y_EXP}	100.04 (0.06)	0.58 (0.56)	T_y
Variance Estimators for $t_{y\_\mathrm{LIN}} (E_{\mathrm{EMP}}(v_{\mathrm{ST2}(e)}) = 100)$			
v_{ST2}	83.59 (1.53)	−19.96 (< 0.0001)	difference from
$v_{\mathrm{ST2(adjusted)}}$	95.53 (1.80)	−6.09 (< 0.0001)	$v_{\mathrm{ST2}(e)}$
v_J	104.69 (2.28)	3.60 (0.0003)	
EMSE	99.35 –	−0.18 (0.85)	
Variance Estimators for $t_{y\_\mathrm{EXP}} (E_{\mathrm{EMP}}(v_{\mathrm{ST2}(e)}) = 100)$			
v_{ST2}	73.12 (1.54)	−18.22 (< 0.0001)	difference from
$v_{\mathrm{ST2(adjusted)}}$	88.79 (1.98)	−8.57 (< 0.0001)	$v_{\mathrm{ST2}(e)}$
v_J	107.00 (2.73)	4.09 (< 0.0001)	
EMSE	101.21 –	0.33 (0.74)	
Other Statistics			
relvar $(v_{\mathrm{ST2}(e)[\mathrm{LIN}]})$	0.051 –	–	
relvar $(v_{\mathrm{ST2}(e)[\mathrm{EXP}]})$	0.059 –	–	
$\dfrac{(v_{\mathrm{ST2}(e)[\mathrm{LIN}]} - v_{\mathrm{ST2}(e)[\mathrm{EXP}]})}{(E_{\mathrm{EMP}}(v_{\mathrm{ST2}(e)[\mathrm{EXP}]})}$	−0.1340 (0.010)	−13.87 (< 0.0001)	

* In four additional simulations, convergence was not reached in 10 iterations for t_{y_EXP}. They were excluded from the analysis.

Whether or not one is better off incorporating the correct response model in the calibration estimator, if one does so, then the variance estimators discussed in the previous section, perhaps with the linearization estimator adjusted as suggested in this section, appear to be serviceable.

A second set of 1,600 simulations (not displayed) were done using the same population and stratified sampling design but with each sampled element given a 70% chance of being in the respondent sample (the average respondent sample size was roughly 89.8). In this set of simulations, both estimators for T_y are randomization consistent under the response model. Consequently, it is not surprising, that the empirical means of t_{y_LIN} and t_{y_EXP} were virtually identical (within 0.01% of each other) as were their empirical mean squared errors (within 1% of each other). The empirical means of each pair of variance estimators (*e.g.*, $\mathrm{var}_{\mathrm{ST2}}$ for t_{y_LIN} and t_{y_EXP}) were likewise very close (within 1% of each other). The relative bias of the adjusted v_{ST2} (compared to $\mathrm{var}_{\mathrm{ST2}(e)}$) was -1.3% when estimating the variance of t_{y_LIN} and -2.2% when estimating the variance of t_{y_EXP}. The relative biases of the unadjusted linearization variances were -9.0% and -10.3%, respectively. The relative bias of both jackknives was 3.6%.

9. Discussion

9.1 Estimating a Response Model Explicitly

When faced with unit nonresponse, many have attempted to estimate the element probabilities of response, $p_k = 1/f(\mathbf{h}_k\boldsymbol{\phi})$, directly. This method requires one to have information on $\mathbf{h}_k$ for every element in the sample whether it responded to the survey or not, but $\mathbf{h}_k$ need not have the same dimension as $\mathbf{x}_k$. The direct-adjustment method is generally not available for handling coverage errors.

Fuller (2002) noted that there can be an extra term in the quasi-randomization mean squared error of $t_{y\_\mathrm{GREG}} = \sum_S a_k^* y_k + (T_{\mathbf{x}} - \sum_S a_j^* \mathbf{x}_j)(\sum_S c_j a_j^* \mathbf{x}_j' \mathbf{x}_j)^{-1} \sum_S c_k a_k^* \mathbf{x}_k' \mathbf{x}_k$, where S is the respondent subsample, $a_k^* = a_k[1 + f(\mathbf{h}_k\mathbf{q})]$, and $\mathbf{q}$ is a consistent direct estimator for the quasi-randomization model parameter, $\boldsymbol{\phi}$. This does not imply that direct estimation of the response model based on a given $f(\cdot)$ and $\mathbf{h}_k$ is less efficient than analogous calibration when $\mathbf{h}_k$ has the same dimension of $\mathbf{x}_k$. See Kim (2004) for a suggestion otherwise. Nevertheless, the convenience of incorporating nonresponse adjustment into calibration is appealing when variance estimates need to be produced.

A reasonable compromise is to choose the form of $f(\cdot)$ and $\mathbf{h}_k$ by modeling the response behavior of the entire sample and then estimating the parameter of $f(\cdot)$ implicitly through calibration. This compromise also overcomes a striking weakness of using calibration weighting to adjust for nonresponse (as well as for coverage errors). The choices for $f(\cdot)$ and $\mathbf{h}_k$ are motivated primarily by plausibility and convenience and not by a statistical analysis of the data.

9.2 Response Homogeneity Groups

To control the magnitude of the weight adjustment due to nonresponse, Little (1986) recommended that one estimate $\mathbf{q}$ explicitly and then divide the sample into C mutually exclusive groups based on the sizes of the fitted $f(\mathbf{h}_k\mathbf{q})$ values. One then computes the adjusted weight for each element k in group c as with post-stratification: $w_{k\_\mathrm{ADJ}} = (\sum_{F(c)} w_g / \sum_{S(c)} w_g) w_k$, where $F_{(c)}$ is that part of the original sample in group c, $S(c)$ is the subsample of $F(c)$ that respond, and w_k is the sampling weight assigned to element k after sampling but before quasi-random subsampling. This approach assumes that each element in a group has (roughly) the same probability of response, hence the term "response homogeneity group."

An alternative way of incorporating fitted $f(\mathbf{h}_k\mathbf{q})$ values into the estimation based on methodology developed in the text follows. Divide the fitted values into P groups based in their sizes, where P is again the dimension of $\mathbf{x}_k$, and let $\mathbf{d}_k$ be a row vector of indicator variables for the P cells. By setting each $w_k = a_k[1 + (T_{\mathbf{x}} - \sum_S a_j \mathbf{x}_j) \times (\sum_S a_j \mathbf{d}_j' \mathbf{x}_j)^{-1} \mathbf{d}_k']$, one computes a set of weights for the respondent subsample that, unlike $\{w_{k\_\mathrm{ADJ}}\}$ above, satisfies the calibration equation for the respondent sample. Because of the nature of $\mathbf{d}_k$, this linear method returns the same set of calibration weights as fitting $w_k = a_k \exp(\mathbf{d}_k\mathbf{f})$ would – if both produce a set of weights. Note that since calibration weights can be negative with the linear method, it may be able to find a set that the generalized raking method cannot. The linear method effectively scales the a_k –value for every element in the same group by a fixed amount. Thus, it may not produce surprisingly small or surprisingly large weights when the dimension of $\mathbf{x}_k$ is small compared to the sample size.

9.3 Breaking Up Sample and Nonresponse Calibration

In the previous section we noted that it is possible for components of $\mathbf{h}_k$ in equation (13), the quasi-random response model, to be unknown before enumeration. When such an $\mathbf{h}_k$ is used in calibration, it might no longer to reasonable to assert that the resulting t_{y_CAL} is prediction-model unbiased. This is particularly troublesome when nonresponse is modest compared to the sample size. An intriguing idea is to calibrate in two phases. The first phase, sample calibration, adjusts for the difference between $T_{\mathbf{x}}$ and $\sum_F a_k \mathbf{x}_k$, and would not include any components in $\mathbf{h}_k$ unavailable at the time of sampling. The second phase,

nonresponse calibration, adjusts for the difference between $\sum_F a_k \mathbf{x}_k$ and $\sum_S a_k \mathbf{x}_k$ and would include component variables only available after the respondent subsample is enumerated.

A more thorough analysis of this idea must wait for another time.

9.4 Work at NASS

The National Agricultural Statistics Service (NASS) used variants of the Fuller *et al*. (1994) approach for handling undercoverage in the 2002 Census of Agriculture (see Fetter and Kott 2003) and for adjusting an agricultural economics survey with large nonresponse to match totals from more reliable surveys (see Crouse and Kott 2004). In this approach, $f(\cdot)$ has the form:

$$f(\mathbf{x}_k\boldsymbol{\phi}) = \begin{cases} L & \text{when} \quad \mathbf{x}_k\boldsymbol{\phi} < L \\ \mathbf{x}_k\boldsymbol{\phi} & \text{when} \quad \mathrm{L} \le \mathbf{x}_k\boldsymbol{\phi} \le U \\ U & \text{when} \quad \mathbf{x}_k\boldsymbol{\phi} > U, \end{cases} \qquad (20)$$

which truncates linear calibration at pre-specified values, L and U, to control the size of the weight adjustment. Note that when $f(\cdot) = U$ or L, $f'(\cdot) = 0$. Unlike the calibration adjustment in equation (19), $f(\cdot)$ in equation (20) is not twice differentiable at L or U. This does not cause a problem in practice.

The agency's original justification for calibration in these contexts was based on prediction-modeling. Equation (20) is simple to implement and appears to produce weights within an acceptable range more often than readily available alternatives.

NASS is investigating the following questions: How sensitive is t_{y_CAL} to the choice of $f(\cdot)$ in practice? Would a different choice for $f(\cdot)$ result in less bias, and if so, would the reduction in absolute bias translate into a lower mean squared error? What would be the effect of replacing some component of the vector of calibration variables with a better predictor of nonresponse/undercoverage?

References

Berry, C.C., Flatt, S.W. and Pierce, J.P. (1996). Correcting unit nonresponse via nonresponse modeling and raking in the California Tobacco Survey. *Journal of Official Statistics*, 12, 349-363.

Crouse, C., and Kott, P.S. (2004). Evaluation alternative calibration schemes for an economic survey with large nonresponse. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Deming, W.E., and Stephan, F.F. (1940). On a least squares adjustment of a sample frequency table when the expected marginal total are known. *Annals of Mathematical Statistics*, 11, 427-444.

Demnati, A., and Rao, J.N.K. (2004). Linearization variance estimation for survey data. *Survey Methodology*, 30, 17-26.

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

Estevao, V.M., and Särndal, C.-E. (2000). A functional form approach to calibration. *Journal of Official Statistics*, 16, 379-399.

Fetter, M.J., and Kott., P.S. (2003). Developing a coverage adjustment strategy for the 2002 Census of Agriculture. Presented at 2003 Federal Committee on Statistical Methodology Research Conference, http://www.fcsm.gov/03papers/fetter_kott.pdf.

Folsom., R.E., and Singh, A.C. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 598-603.

Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28, 5-23.

Fuller, W.A., Loughin, M.M. and Baker, H.D. (1994). Regression weighting for the 1987-88 National Food Consumption Survey. *Survey Methodology*, 20, 75-85.

Kim, J.K. (2004). Efficient nonresponse weighting adjustment using estimated response probability. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Kott, P.S. (1990). The design consistent regression estimator and its conditional variance. *Journal of Statistical Planning and Inference*, 24, 287-296.

Kott, P.S. (1994). A note on handling nonresponse in surveys. *Journal of the American Statistical Association*, 89, 693-696.

Kott, P.S. (2004a). Randomization-assisted model-based survey sampling. *Journal of Statistical Planning and Inference*, 48, 263-277.

Kott, P.S. (2004b). Comment. *Survey Methodology*, 30, 27-28.

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.

Lundström, S., and Särndal, C.-E. (1999). Calibration as a standard method for the treatment of nonresponse. *Journal of Official Statistics*, 15, 305-327.

Särndal, C.-E., Swensson, B. and Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of a finite population total. *Biometrika*, 76, 527-537.

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

Survey Methodology, December 2006
Vol. 32, No. 2, pp. 143-149
Statistics Canada, Catalogue No. 12-001

The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data

Jerome P. Reiter, Trivellore E. Raghunathan and Satkartar K. Kinney [1]

Abstract

The theory of multiple imputation for missing data requires that imputations be made conditional on the sampling design. However, most standard software packages for performing model-based multiple imputation assume simple random samples, leading many practitioners not to account for complex sample design features, such as stratification and clustering, in their imputations. Theory predicts that analyses of such multiply-imputed data sets can yield biased estimates from the design-based perspective. In this article, we illustrate through simulation that (i) the bias can be severe when the design features are related to the survey variables of interest, and (ii) the bias can be reduced by controlling for the design features in the imputation models. The simulations also illustrate that conditioning on irrelevant design features in the imputation models can yield conservative inferences, provided that the models include other relevant predictors. These results suggest a prescription for imputers: the safest course of action is to include design variables in the specification of imputation models. Using real data, we demonstrate a simple approach for incorporating complex design features that can be used with some of the standard software packages for creating multiple imputations.

Key Words: Complex sampling design; Multiple imputation; Nonresponse; Surveys.

1. Introduction

Typically in large surveys, less than 100% of the sampled units respond fully to the survey. Some units do not respond at all, and others respond only to certain items. One approach to handle such nonresponse is multiple imputation of missing data (Rubin 1987). It has been used in, for example, the Fatality Analysis Reporting System (Heitjan and Little 1991), the Consumer Expenditures Survey (Raghunathan and Paulin 1998), the National Health and Nutrition Examination Survey (Schafer, Ezzati-Rice, Johnson, Khare, Little and Rubin 1998), the Survey of Consumer Finances (Kennickell 1998) and the National Health Interview Survey (Schenker, Raghunathan, Chiu, Makuc, Zhang and Cohen 2005). Multiple imputation also has been suggested to protect confidentiality of public-release data (Rubin 1993; Little 1993; Raghunathan, Reiter and Rubin 2003; Reiter 2003, 2004, 2005). See Rubin (1996) and Barnard and Meng (1999) for a review of other applications.

Multiple imputation, in theory, conditions on the sampling design when deriving methods for obtaining inferences from multiply-imputed datasets (Rubin 1987). However, imputers seldom account for complex sampling design features, such as stratification and clustering, when using available software packages to construct imputation models. They instead use multivariate normal or general location models (*e.g.*, the software NORM written by Joe Schafer), or use sequential regression models (Raghunathan, Lepkowshi, van Hoewyk and Solenberger 2001). These methods can be modified to incorporate design features, but this is infrequently done.

This paper has two objectives. First, we illustrate the bias that can arise when imputers fail to account for complex design features in imputation models. To do so, we simulate multiple imputation in two-stage, stratified-cluster samples. The simulations indicate these biases can be severe, even when using design-based estimators in multiply-imputed data sets with moderate amounts of missing data. Second, we suggest two simple approaches to account for design features in imputation models. The first approach, which is relatively easy to implement, includes dummy variables for stratum or cluster effects in the imputation models. The second approach, which is computationally more complex than the first, uses hierarchical models where (i) the effects of clustering are incorporated using random effects, and (ii) the effects of stratification are incorporated using fixed effects. The simulations show that accounting for the design in these ways can reduce the bias. They also illustrate that controlling for design features that are unrelated to the survey variables can result in inefficient, but conservative, inferences relative to those from models that do not condition on such features, provided that the models include the predictors required to make the missing at random assumption (Rubin 1976) plausible. We demonstrate the first approach to incorporating the design features by imputing missing data from the National Health and Nutrition Examination Survey using a sequential regression approach.

1. Jerome P. Reiter and Satkartar K. Kinney, Institute of Statistics and Decision Sciences, Box 90251, Duke University, Durham, NC 27708, U.S.A.; Trivellore E. Raghunathan, Department of Biostatistics and Institute for Social Research, University of Michigan, Ann Arbor, MI 48106, U.S.A.

2. Inferences from Multiply-Imputed Data Sets

To describe construction of and inferences from multiply-imputed data sets, we use the notation of Rubin (1987). For a finite population of size *N*, let $I_j = 1$ if unit *j* is selected in the original survey, and $I_j = 0$ otherwise, where $j = 1, 2, \ldots, N$. Let $I = (I_1, \ldots, I_N)$. Let *n* be the size of the sample obtained using a complex design. To simplify notation, assume only one variable in the survey is subject to nonresponse. Let $R_j = 1$ if unit *j* responds to the original survey, and $R_j = 0$ otherwise. Let $R = (R_1, \ldots, R_N)$. The notation can be extended to handle multivariate item nonresponse, but such complication is not necessary for our purposes.

Let *Y* be the $N \times p$ matrix of survey data for all units in the population. Let $Y_{\text{inc}} = (Y_{\text{obs}}, Y_{\text{mis}})$ be the $n \times p$ matrix of survey data for units with $I_j = 1$; Y_{obs} is the portion of Y_{inc} that is observed, and Y_{mis} is the portion of Y_{inc} that is missing due to nonresponse. Let *Z* be the $N \times d$ matrix of design variables for all *N* units in the population, *e.g.*, stratum or cluster indicators or size measures. We assume that such design information is known at least approximately, for example from census records or the sampling frames.

Values for Y_{mis} are usually constructed from draws from some approximation to the Bayesian posterior predictive distribution of $(Y_{\text{mis}} | Z, Y_{\text{obs}}, I, R)$. These draws are repeated independently $l = 1, \ldots, M$ times to obtain *M* completed data sets, $D^{(l)} = (Z, Y_{\text{obs}}, Y_{\text{mis}}^{(l)}, I, R)$.

From these multiply-imputed data sets, some user of the data seeks inferences about some estimand $Q = Q(Z, Y)$. For example, *Q* could be a population mean or a population regression coefficient. In each imputed data set $D^{(l)}$, the analyst estimates *Q* with some estimator *q* and the variance of *q* with some estimator *u*. We assume that the analyst specifies *q* and *u* by acting as if each $D^{(l)}$ was in fact collected data from a random sample of (Z, Y) based on the original sampling design *I*, *i.e.*, *q* and *u* are complete-data estimators.

For $l = 1, \ldots, M$, let $q^{(l)}$ and $u^{(l)}$ be respectively the values of *q* and *u* in data set $D^{(l)}$. Under assumptions described in (Rubin 1987), the analyst can obtain valid inferences for scalar *Q* by combining the $q^{(l)}$ and $u^{(l)}$. Specifically, the following quantities are needed for inferences:

$$\bar{q}_M = \sum_{l=1}^{M} q^{(l)} / M \tag{1}$$

$$b_M = \sum_{l=1}^{M} (q^{(l)} - \bar{q}_M)^2 / (M - 1) \tag{2}$$

$$\bar{u}_M = \sum_{l=1}^{M} u^{(l)} / M. \tag{3}$$

The analyst then can use $\bar{q}_M$ to estimate *Q* and $T_M = (1 + \frac{1}{M}) b_M + \bar{u}_M$ to estimate the variance of $\bar{q}_M$. When *n* and *M* are large, inferences for scalar *Q* can be based on normal distributions, so that a $(1-\alpha)\%$ confidence interval for *Q* is $\bar{q}_M \pm z(\alpha/2)\sqrt{T_M}$. For moderate *M*, inferences can be based on *t*–distributions with degrees of freedom $v_M = (M-1)(1 + r_M^{-1})^2$, where $r_M = (1 + M^{-1}) b_M / \bar{u}_M$, so that a $(1-\alpha)\%$ confidence interval for *Q* is $\bar{q}_M \pm t_{vM}(\alpha/2)\sqrt{T_M}$. Refinements of these basic combining rules have been proposed by several authors, including Li, Raghunathan and Rubin (1991a), Li, Meng and Rubin (1991b), Raghunathan and Siscovick (1996), and Barnard and Rubin (1999).

3. Illustrative Simulations

In this section, we use simulations to illustrate the biases/inefficiencies associated with incorporating design features in imputation models. We simulate three target populations of $N = 100,000$ units that are stratified and clustered within strata. In the first population, *Y* depends on both stratum and cluster effects. In the second population, *Y* depends on strata but not on cluster effects. In the third population, *Y* is unrelated to the stratum and cluster indicators. The first population is used to demonstrate the importance of including all relevant design variables, and the second and third populations are used to examine the effect of including irrelevant design variables. The simulated populations are stylized to illustrate the importance of modeling the survey design; hence, the magnitudes of biases/inefficiencies may not be generalizable to other settings.

Each population is divided into five equally-sized strata comprised of $N_h = 200$ clusters, for $h = 1, \ldots, 5$. Each cluster *c* in stratum *h* is comprised of N_{hc} units. In each stratum, ten clusters have $N_{hc} = 300$, twenty clusters have $N_{hc} = 200$, sixty clusters have $N_{hc} = 100$, sixty clusters have $N_{hc} = 75$, and fifty clusters have $N_{hc} = 50$. Cluster sizes are varied to magnify design effects when taking multi-stage cluster samples. For each target population, there are two survey variables, *X* and *Y*. In all three populations, for simplicity we generate each X_{hcj}, where *j* indexes a unit within stratum and cluster *hc*, from $X_{hcj} \sim N(0, 10^2)$. To generate *Y*, we use different methods for each population, as shall be described in subsequent sections.

We randomly sample units from each population using multi-stage cluster sampling. First, we take a simple random sample of $n_1 = 40$ clusters from stratum 1, $n_2 = 20$ clusters from stratum 2, $n_3 = 30$ clusters from stratum 3, $n_4 = 10$ clusters from stratum 4, and $n_5 = 15$ clusters from stratum 5. The cluster sample sizes differ across strata to magnify

design effects relative to equal sampling. We then take a simple random sample of twenty units from each sampled cluster. Hence, there are 2,300 units with $I_{hcj} = 1$.

The estimands of interest in each population are $Q = \bar{Y}$, the population mean of Y, and the coefficients for the population regression of Y on X. The complete-data estimator of $\bar{Y}$ is the usual, unbiased design-based estimator,

$$q = \frac{1}{100,000}\left(\sum_{h=1}^{5} \frac{200}{n_h}\sum_{c\in h} \hat{y}_{hc}\right),$$

where $\hat{y}_{hc} = N_{hc}\bar{y}_{hc}$ is the estimated total in cluster *hc*. The complete-data estimator of the variance of q is,

$$u = \frac{1}{100,000^2}\left(\begin{array}{l}\sum_{h=1}^{5} 200^2\left(1-\frac{n_h}{200}\right)s_h^2/n_h \\ +\sum_{h=1}^{5} \frac{200}{n_h}\sum_{c\in h} N_{hc}^2\left(1-\frac{20}{N_{hc}}\right)s_{hc}^2/20\end{array}\right),$$

where s_h^2 is the sample variance of the $\hat{y}_{hc}$ and s_{hc}^2 is the sample variance of Y within cluster *hc*. The estimators of the coefficients in the regression of Y on X are the usual approximately unbiased, design-based estimators, which are computed using the "survey" routines (Lumley 2004) in the software package R. These routines estimate variances using Taylor series linearizations. These estimators are used for all multiply-imputed data sets in all simulations.

For each sample, we let X be fully observed, and let Y be missing for about 30% of the sampled units.

Each unit's binary response variable, R_{hcj}, is drawn from a Bernoulli distribution:

$$\Pr(R_{hcj} = 1) = \frac{\exp(-0.847 - 0.1X_{hcj})}{1+\exp(-0.847 - 0.1X_{hcj})} \qquad (4)$$

Here, $R_{hcj} = 1$ means that the unit's value of Y is missing. Equation 4 implies that Y_{mis} is missing at random (Rubin 1976). We can ignore the missing data mechanism provided that imputations for missing data are conditional on X. We purposefully do not allow missingness to depend on stratum or cluster membership to illustrate that bias can arise from failing to account for the survey design even when the ignorable missing data mechanism does not depend on the sampling design. Of course, if the sampling design is related to missingness, as it is in many real datasets, one must condition on the sampling design to make the missing data mechanism ignorable.

We examine three strategies to impute Y_{mis} that make different use of the design information. These strategies are summarized in Table 1. The first strategy, labeled SRS, completely disregards the sampling design. The second strategy, FX, incorporates the stratification and the clustering by using fixed effects for each cluster within stratum. The third strategy, HM, uses normal random effects models that incorporate the stratification and clustering. For SRS, one model is fit to the entire data set. For FX and HM, models are fit separately in each stratum. All three strategies regress on X because it is part of the missing data mechanism; not conditioning on X would violate ignorability and cause bias.

Table 1
Imputation Strategies

Label	Imputation model for missing Y_{hcj}
SRS	$N(\beta_0 + \beta_1 X_{hcj}, \sigma^2)$
FX	$N(\beta_{0h} + \beta_{1h} X_{hcj} + \omega_{hc}, \sigma_h^2)$
HM	$N(\beta_{0h} + \beta_{1h} X_{hcl} + \omega_{hc}, \sigma_h^2)$, $\omega_{hc} \sim N(0, \tau^2)$

All imputations are draws from the appropriate Bayesian posterior predictive distributions. First, we draw parameters of the imputation models from their posterior distributions given the components of the observed data, $(Z, X, Y_{\text{obs}}, I, R)$, that are included in the models. Second, we draw values of the missing data from the distributions given in Table 1. Diffuse priors are used for all parameters. For strategy HM, we draw values of the parameters using a Gibbs sampler (Gelfand and Smith 1990). We run the sampler for a burn-in period to get approximate convergence, then we use every tenth draw for imputations. Finally, we use $M = 5$ independently drawn imputations in each data set for each strategy.

3.1 Simulation A: Illustration of Disregarding Relevant Design Features

In this simulation, we generate a population in which the distributions of Y differ across strata and clusters. We call this "Population 1". Specifically, for unit j in stratum h and cluster c, we construct the population value of Y_{hcj} from

$$Y_{hcj} = 10X_{hcj} + \beta_{0h} + \omega_{hc} + \in_{hcj} \qquad (5)$$

where β_{0h} is a scalar constant for stratum h, the ω_{hc} is a scalar constant for cluster *hc*, and $\in_{hcj}$ is a random error term drawn from $N(0, 200^2)$. The values of the stratum effects are $\beta_{01} = 500, \beta_{02} = -250, \beta_{03} = 0, \beta_{04} = 250,$ and $\beta_{05} = -500$. The values of the ω_{hc} are obtained by drawing five sets of $N_h = 200$ values from independent $N(0, 70^2)$. The stratum and cluster effects are widely dispersed to magnify design effects relative to simple random sampling, which in turn magnifies the effects of disregarding the design in imputations. We then sample from this population using the stratified cluster sampling scheme outlined previously. We create the missing data indicator R using equation 4.

Table 2 shows the results of 1,000 replications of the three imputation strategies outlined in Table 1. The additional row labeled "Complete data" shows the results using the data for all sampled units, *i.e.*, assuming no units with $I_{hcj} = 1$ have $R_{hcj} = 0$. The column labeled "95% CI cov." contains the percentage of the 1,000 simulated confidence intervals that contain the population parameter. The column labeled "Pt. Est." contains the averages of the 1,000 point estimates of Q. The column labeled "Var" contains the variances of the 1,000 point estimates of Q. The column labeled "Est. Var" contains the averages across the 1,000 replications of the estimated variances of the point estimates. The columns labeled "Var(Est.Var)" and "MSE(Est.Var)" give the variance and mean squared error of the 1,000 estimated variances.

Imputations based on method SRS lead to severely biased estimates and very poor confidence interval coverage in this population. These problems exist even though there is not much missing information and despite the fact that we use design-unbiased estimators for inferences. Both FX and HM have point estimates that approximately match the complete-data point estimates, and both have coverage rates that approximately match the rates for the complete data inferences. FX and HM have similar profiles because the fixed effect models and the hierarchical models produce similar estimates of the parameters in equation 5.

When estimating the population mean, the variance associated with FX or HM is only slightly larger than the variance associated with the complete-data estimator. This is because of the large cluster effects, which makes the within-imputation variance a dominant factor relative to the between-imputation variance. That is, the fraction of missing information due to missing data is relatively small when compared to the effect of clustering.

3.2 Simulation B: Illustration of including irrelevant predictors

Modeling the design features is essential when the features are related to the survey variables of interest. How does modeling irrelevant design features affect inferences? In this section, we present the results of two simulation studies that explore this question.

First, we generate "Population 2" in which the distribution of Y differs across strata but does not depend on the clusters. To do so, we use the same generation method as in Equation 5, setting the ω_{hc} equal to zero. The $\in_{hcj}$ are drawn from $N(0, 100^2)$. We sample from Population 2 and generate missing data using the schemes outlined previously. The results for 1,000 replications are displayed in Table 3.

SRS continues to have severe bias and poor confidence interval coverage because it ignores the stratification. For FX and HM, the averages of their point estimates are within simulation error of the average of the point estimates for the complete data. Additionally, their confidence interval coverage rates approximately match the coverage rate for the complete-data intervals. This indicates that FX and HM are reasonable for these populations, even though the irrelevant cluster features are included in their imputation models.

We next generate "Population" 3 in which the distribution of Y is independent of the strata and cluster membership indicators. Specifically, to generate Y, we subtract the β_{0h} from the values of Y generated in Population 2. We then sample from Population 3 using the stratified cluster sampling scheme and create missing data using the methods outlined previously. The results for 1,000 replications are displayed in Table 4.

Table 2
Performance of Imputation Procedures when the Design Features are Related to the Survey Variable of Interest. The Population Mean Equals 3.2 and the Population Regression Coefficients Equal 3.0 and 10.1

	Method	95% CI cov.	Pt. Est.	Var	Est. Var	Var(Est. Var)	MSE (Est. Var)
Mean Y	Complete data	94.2	2.0	544.91	527.31	31,626.19	31,936.07
	SRS	38.0	45.8	327.79	360.74	11,927.97	13,013.35
	FX	94.8	2.4	554.09	579.92	37,474.82	38,141.70
	HM	94.5	2.3	551.02	553.16	34,056.39	34,060.99
Intercept	Complete data	93.0	2.4	529.51	499.73	18,543.13	19,430.21
	SRS	39.5	46.8	340.09	365.50	9,351.15	9,996.99
	FX	94.5	2.8	539.19	551.68	21,529.16	21,685.33
	HM	93.9	2.7	536.82	524.82	19,256.24	19,400.11
Slope	Complete data	93.3	10.1	1.24	1.15	0.14	0.15
	SRS	64.8	7.6	2.10	2.20	0.55	0.56
	FX	94.5	10.1	1.45	1.44	0.18	0.18
	HM	95.7	10.1	1.53	1.65	0.29	0.30

Table 3
Performance of Imputation Procedures when the Population has Stratum Effects but no Cluster Effects. The Population Mean Equals 0.34 and the Population Regression Coefficients Equal 0.14 and 10.13

	Method	95% CI cov.	Pt. Est.	Var	Est. Var	Var(Est. Var)	MSE (Est. Var)
Mean Y	Complete data	93.6	0.37	468.97	461.88	29,301.77	29,352.04
	SRS	31.1	42.90	259.46	303.46	10,228.40	12,164.74
	FX	93.7	0.32	473.86	474.21	30,408.95	30,409.07
	HM	93.4	0.34	476.03	465.53	29,406.61	29,516.85
Intercept	Complete data	93.0	0.72	451.46	432.74	14,955.20	15,305.73
	SRS	31.5	43.10	275.22	311.36	8,134.04	9,440.57
	FX	93.2	0.66	456.08	444.88	15,539.21	15,664.64
	HM	92.3	0.68	457.48	436.25	14,941.00	15,391.75
Slope	Complete data	93.1	10.09	0.99	0.91	0.09	0.10
	SRS	59.0	7.72	1.67	1.77	0.35	0.36
	FX	93.4	10.10	1.03	0.98	0.10	0.10
	HM	93.3	10.10	1.03	0.96	0.10	0.10

Table 4
Performance of Imputation Procedures when the Design Variables are Completely Unrelated to the Survey Variable of Interest. The Population Mean Equals 0.34 and the Population Regression Coefficients Equal 0.14 and 10.04

	Method	95% CI cov.	Pt. Est.	Var	Est. Var	Var(Est. Var)	MSE (Est. Var)
Mean Y	Complete data	94.7	0.35	14.61	14.73	32.65	32.66
	SRS	95.7	0.12	16.45	19.22	40.65	48.31
	FX	97.8	0.40	19.64	28.29	97.66	172.38
	HM	95.1	0.26	18.77	19.16	47.29	47.44
Intercept	Complete data	93.7	0.12	7.13	7.20	5.31	5.32
	SRS	96.8	–0.10	8.97	11.72	13.59	21.10
	FX	98.6	0.17	12.23	20.62	39.84	110.24
	HM	96.2	0.03	10.45	11.61	15.09	16.45
Slope	Complete data	94.5	10.04	0.07	0.07	0.001	0.001
	SRS	96.4	10.07	0.10	0.13	0.002	0.003
	FX	96.4	10.04	0.12	0.15	0.003	0.004
	HM	95.2	10.05	0.11	0.12	0.002	0.002

SRS finally produces point estimates whose averages are within simulation error of the complete data average point estimate. This is because the imputations in SRS reflect the population structure reasonably well. This suggests that disregarding the design in imputation models may provide acceptable inferences when the design variables are only weakly correlated with the survey outcomes. As in the previous simulations, FX and HM continue to have average point estimates within simulation error of the complete-data average point estimate. When comparing the three imputation strategies, we see that FX and HM are inefficient relative to SRS. This is because the imputation models for FX and HM estimate parameters that equal approximately zero in the population, whereas SRS sets them equal to zero. HM has smaller variance than FX does, because the hierarchical imputation model smoothes the estimated cluster effects towards zero.

For FX, the percentage of confidence intervals that cover Q is larger than the percentages for the complete-data intervals and HM intervals. This is because the estimated variance for FX tends to be larger than its actual variance. This apparent upward bias in T_M also exists for SRS, resulting in a larger coverage percentage than those for the complete-data and HM.

4. Real Data Example

We next examine the effect of accounting for stratification and clustering when imputing missing data in a genuine dataset. The data are taken from the public use file for the 1999–2002 National Health and Nutrition Examination Surveys. Individuals are grouped in 56 clusters divided among 28 strata. Many variables have 5% to 10% missing data.

We imputed missing data using two strategies: one ignoring design variables (like SRS) and one incorporating the design variables using fixed effects for cluster indicators (like FX). In the imputation model, we included 27 dummy variables to represent 28 strata and one dummy variable within-each stratum to represent the two clusters nested within each stratum. That is, a total of 55 dummy variables

were included as predictors. We used a stepwise variable selection procedure to identify statistically significant interactions between these dummy variables and survey variables, and we included these interactions as predictors in the imputation model as well. The values were imputed using the sequential regression approach implemented in the software package IVEWARE (www.isr.umich.edu/src/smp/ive). We generate $M = 10$ data sets for each strategy.

We consider three estimands. The first is the population percentage of people who have ever had their blood cholesterol level checked (BPQ060). This variable has about 15% missing values. The second and third are the population regression coefficients in a logistic regression of BPQ060 on family poverty income ratio (INDFMPIR), a continuous variable that has about 12% missing values. These estimands are estimated using design-based methods computed with the "survey" routines in the software package R.

Table 5 displays the results for both imputation strategies. The two sets of estimates for all analyses are very similar. In this case, incorporating the design variables into the imputation model hardly impacts the results. This is due in part to the small fractions of missing information and the relative unimportance of stratum and cluster effects. However, there is minimal penalty for including the design features in the imputation model. In light of the results of the simulations in section 3, we would incorporate the design features in this imputation model.

Table 5
Comparison of Real Data Results when Design Features are Included in Imputation Model and when Design Features are Ignored

	Pt. Est.	S.E.	95% CI
Mean BPQ060			
design	0.319	0.010	(0.299, 0.339)
no design	0.319	0.011	(0.296, 0.341)
Intercept: Logistic Regression			
design	0.362	0.054	(0.256, 0.467)
no design	0.352	0.052	(0.251, 0.454)
Slope: Logistic Regression			
design	–0.409	0.020	(–0.449, –0.369)
no design	–0.407	0.019	(–0.444, –0.371)

5. Concluding Remarks

The simulation studies, though limited, suggest disregarding the sampling design in multiple imputation can be a risky practice. When the design variables are related to the survey variables, as in our Simulation A, failing to include the design variables can lead to severe bias. On the other hand, including irrelevant design variables, as in our Simulation B and the NHANES example, leads at worst to inefficient and conservative inferences when the imputation models are otherwise properly specified.

Including dummy variables for cluster effects greatly reduced the bias relative to disregarding the design completely. However, blindly including dummy variables is not an automatic solution. When the regression slopes or variances differ across clusters, using FX or HM may result in biased estimates, since important design features are disregarded. Imputers suspecting such relationships should include appropriate interactions with the dummy variables for the design features, as we did in the NHANES example. In some surveys the design may be so complicated that it is impractical to include dummy variables for every cluster. In these cases, imputers can simplify the model for the design variables, for example collapsing cluster categories or including proxy variables (*e.g.*, cluster size) that are related to the outcome of interest.

The simulations suggest that there can be payoffs to using hierarchical models for imputation of missing data relative to using fixed effects models, particularly when cluster effects are similar. However, hierarchical models are more difficult to fit than fixed effect models. For example, it is daunting to fit hierarchical models in complex designs when data are missing for several continuous and categorical variables. It may be possible to fit sequential hierarchical models in a spirit similar to the sequential regression imputations of Raghunathan *et al.* (2001). This is an area for future research. A further disadvantage of hierarchical models is that they are easier to mis-specify than fixed effects models. For example, if the cluster effects follow a non-normal distribution, the hierarchical normal model used in this paper could provide implausible imputations.

With multiple imputation, the key to success is specifying an imputation model that reasonably describes the conditional distribution of the missing values given the observed values. Design features frequently are related to survey variables, so that including them in the imputation models reduces the risks of model mis-specification. We believe that in many cases the potential biases resulting from excluding important design variables, or other variables related to the missing data mechanism, outweigh the potential inefficiencies from estimating small coefficients. This reinforces the general advice provided by many on multiple imputation: include all variables that are related to the missing data in imputation models to make the missing data mechanism ignorable (*e.g.*, Meng 1994; Little and Raghunathan 1997; Schafer 1997, and Collins, Schafer and Kam 2001).

Acknowledgements

This research was funded by the National Science Foundation grant ITR-0427889. The authors thank the associate editor and reviewers for their comments and suggestions.

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.

Barnard, J., and Rubin, D.B. (1999). Small-sample degrees of freedom with multiple-imputation. *Biometrika*, 86, 948-955.

Collins, L.M., Schafer, J.L. and Kam, C.K. (2001). A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychological Methods*, 6, 330-351.

Gelfand, A.E., and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

Heitjan, D.F., and Little, R.J.A. (1991). Multiple imputation for the Fatal Accident Reporting System. *Applied Statistics*, 40, 13-29.

Kennickell, A.B. (1998). Multiple imputation in survey of consumer finances. In *Proceedings of the Section on Business and Economic Statistics*, American Statistical Association, 11-20.

Li, K.H., Raghunathan, T.E. and Rubin, D.B. (1991a). Large-sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association*, 86, 1065-1073.

Li, K.H., R.T.E., Meng, X.L. and Rubin, D.B. (1991b). Significance levels from repeated $p-$values with multiply-imputed data. *Statistica Sinica*, 1, 65-92.

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

Little, R.J.A., and Raghunathan, T.E. (1997). Should imputation of missing data condition on all observed variables? In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 617-622.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9, 8.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (disc: P558-573). *Statistical Science*, 9, 538-558.

Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J. and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology*, 27, 85-96.

Raghunathan, T.E., and Paulin, G.S. (1998). Multiple imputation of income in the Consumer Expenditure Survey: Evaluation of statistical inference. In *Proceedings of the Section on Business and Economic Statistics*, American Statistical Association, 1-10.

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., and Siscovick, D.S. (1996). A multiple-imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. *Applied Statistics*, 45, 335-352.

Reiter, J.P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29, 181-189.

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

Reiter, J.P. (2005). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society*, Series A, 168, 185-205.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-590.

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

Rubin, D.B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9, 462-468.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

Schafer, J.L., Ezzati-Rice, T.M., Johnson, W., Khare, M., Little, R.J.A. and Rubin, D.B. (1998). The NHANES III multiple imputation project. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 28-37.

Schenker, N., Raghunathan, T.E., Chiu, P.-L., Makuc, D.M., Zhang, G. and Cohen, A.J. Multiple imputation of missing income data in the National Health Interview Survey. *Journal of the American Statistical Association*, forthcoming.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Survey Methodology, December 2006
Vol. 32, No. 2, pp. 151-156
Statistics Canada, Catalogue No. 12-001

Bernoulli Bootstrap for Stratified Multistage Sampling

Fumio Funaoka, Hiroshi Saigo, Randy R. Sitter and Tsutom Toida [1]

Abstract

In this article, we propose a Bernoulli-type bootstrap method that can easily handle multi-stage stratified designs where sampling fractions are large, provided simple random sampling without replacement is used at each stage. The method provides a set of replicate weights which yield consistent variance estimates for both smooth and non-smooth estimators. The method's strength is in its simplicity. It can easily be extended to any number of stages without much complication. The main idea is to either keep or replace a sampling unit at each stage with preassigned probabilities, to construct the bootstrap sample. A limited simulation study is presented to evaluate performance and, as an illustration, we apply the method to the 1997 Japanese National Survey of Prices.

Key Words: Complex survey; Linearization; Quantiles; Resampling; Stratification.

1. Introduction

Many large scale surveys are conducted using a stratified multi-stage sampling design. Variance estimation in this type of design can be analytically involved or even impossible. In addition, for publicly released data sets the specific forms of estimators the end-user may wish to obtain variance estimates for are unknown. As a result, resampling methods are often carried out to obtain a set of replicate weights that can be supplied with the data set and used for the purpose of variance estimation for a broad class of possible estimators. The bootstrap is particularly useful since it can handle both smooth and nonsmooth sample statistics under multistage designs. A concise summary of several bootstrap methods for finite population sampling is found in Shao and Tu (1995, pages 232–282) (see also, Gross 1980; Bickel and Freedman 1984; McCarthy and Snowden 1985; Rao and Wu 1988; Kovar, Rao and Wu 1988; Sitter 1992a, b; Booth, Butler and Hall 1994; Shao and Sitter 1996).

If the first-stage sampling fraction is small, there are various bootstrap methods available that treat the first-stage sampling as having been with-replacement for the purposes of variance estimation. In the case where the first-stage sampling fraction is not negligible, there are fewer results available. For bootstrapping in two-stage sampling with simple random sampling (SRS) at each stage see Sitter (1992a, 1992b) and with unequal probabilities Rao and Wu (1988). However, if the first-stage sampling fractions are not negligible no simple bootstrap procedure is available for three or more stages of sampling. In this paper, we propose a new bootstrap method which easily accommodates such cases when the sampling is simple random sampling (SRS) at each stage. We call it a Bernoulli bootstrap (BBE) because of its resemblance to Bernoulli sampling. The National Survey of Prices (NSP) in Japan is used for illustration.

The paper is organized as follows. Section 2 introduces notation for three-stage stratified sampling. Section 3 describes two types of BBE. Section 4 investigates properties of the methods via simulation. Section 5 describes the sampling design of the 1997 NSP and illustrates the use of BBE on the NSP data. Concluding remarks are made in section 6.

2. Stratified Three-Stage Sampling

In stratified random sampling, the finite population, consisting of N primary sampling units (PSU's), is partitioned into H nonoverlapping *strata* of $N_1, N_2, \ldots, N_H$ PSU's, respectively; thus, $\sum_{h=1}^{H} N_h = N$. A simple random sample without replacement (SRSWOR) of PSU's is taken independently from each stratum. The sample sizes within each stratum are denoted by $n_1, n_2, \ldots, n_H$, and the total PSU sample size is $n = \sum_{h=1}^{H} n_h$. At the second stage, a sample of m_{hi} secondary sampling units (SSU's) are selected from PSU i of size M_{hi} within stratum h by SRSWOR. At the third stage, a sample of l_{hij} ultimate sampling units (USU's) are selected from SSU ij of size L_{hij} within stratum h by SRSWOR. A vector of measurements of some unit characteristics is represented as $\mathbf{y}_{hijk} = (y_{1hijk}, y_{2hijk}, y_{\tau hijk})^T$, where the subscripts $hijk$ refer to the stratum label, PSU label, SSU label and USU label, respectively. The population parameter of interest $\theta = \theta(\mathrm{S})$, where $\mathrm{S} = \{\mathbf{y}_{hijk} : h = 1, 2, \ldots, H; i = 1, 2, \ldots, N_h;$

1. F. Funaoka, Professor, Faculty of Economics, Shinshu University, 3-1-1 Asahi, Matsumoto, Nagano, 390-8621, Japan; H. Saigo, Professor, School of Political Science and Economics, Waseda University, 1-6-1 Nishiwaseda Shinjuku, Tokyo, 169-8050, Japan; R.R. Sitter, Professor, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, B.C., V5A 1S6, Canada; T. Toida, Associate Professor, Faculty of Social and Information Studies, Gunma University, 2-4 Aramakicho, Maebashi, Gunma 371-8510, Japan.

$j = 1, \ldots, M_{hi}; k = 1, \ldots, L_{hij}\}$, is usually estimated by $\hat{\theta} = \hat{\theta}(s)$, where $s = \{\mathbf{y}_{hijk} : h = 1, 2, \ldots, H; i = 1, 2, \ldots, n_h; j = 1, \ldots, m_{hi}; k = 1, \ldots, l_{hij}\}$. The population total vector is denoted $\mathbf{Y} = (Y_1, \ldots, Y_\tau)^T$. In this case, its unbiased estimate is

$$\hat{\mathbf{Y}} = \sum_{h=1}^{H} \hat{\mathbf{Y}}_\mathbf{h} = \sum_{h=1}^{H} (N_h / n_h) \sum_{i=1}^{n_h} \hat{\mathbf{Y}}_\mathbf{hi},$$

where $\hat{\mathbf{Y}}_\mathbf{hi} = (M_{hi} / m_{hi}) \sum_{j=1}^{m_{hi}} \hat{\mathbf{Y}}_\mathbf{hij}$ and $\hat{\mathbf{Y}}_\mathbf{hij} = (L_{hij} / l_{hij}) \sum_{k=1}^{l_{hij}} \mathbf{y}_\mathbf{hijk}$. This may be written as $\hat{\mathbf{Y}} = \sum_{hijk} w_{hij} \mathbf{y}_\mathbf{hijk}$, where $w_{hij} = (N_h / n_h)(M_{hi} / m_{hi})(L_{hij} / l_{hij})$.

For $\tau = 1$, an unbiased estimate of $\mathrm{Var}(\hat{Y})$ is $v(\hat{Y}) = \sum_{h=1}^{H} v(\hat{Y}_h)$, where

$$v(\hat{Y}_h) = \frac{N_h^2 (1 - f_{1h}) s_h^2}{n_h} + \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}^2 (1 - f_{2hi}) s_{hi}^2}{m_{hi}} + \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} \frac{L_{hij}^2 (1 - f_{3hij}) s_{hij}^2}{l_{hij}}$$

with $\bar{\hat{Y}}_{h\cdot} = n_h^{-1} \sum_i \hat{Y}_{hi}$, $\bar{\hat{Y}}_{hi\cdot} = m_{hi}^{-1} \sum_j \hat{Y}_{hij}$, $\bar{y}_{hij\cdot} = l_{hij}^{-1} \sum_k y_{hijk}$, $f_{1h} = n_h / N_h$, $f_{2hi} = m_{hi} / M_{hi}$, $f_{3hij} = l_{hij} / L_{hij}$, $s_h^2 = \sum_i (\hat{Y}_{hi} - \bar{\hat{Y}}_{h\cdot})^2 / (n_h - 1)$, $s_{hi}^2 = \sum_j (\hat{Y}_{hij} - \bar{\hat{Y}}_{hi\cdot})^2 / (m_{hi} - 1)$, and $s_{hij}^2 = \sum_k (y_{hijk} - \bar{y}_{hij\cdot})^2 / (l_{hij} - 1)$ (Särndal, Swensson and Wretman 1992, pages 148–149).

3. Proposed Bernoulli Bootstrap

To handle the multi-stage aspect of the sampling within stratum, we propose a multi-stage bootstrap. To simplify ideas, we first introduce a simple version that has some limitations in applicability. We will then subsequently describe a more general form that avoids these difficulties.

A Short Cut BBE

Step I. For each sample PSU, *hi*, within stratum $h, h = 1, \ldots, H$, we: (a) keep it in the bootstrap sample with probability

$$p_h = \sqrt{1 - \frac{(1 - f_{1h})}{(1 - n_h^{-1})}}; \qquad (3.1)$$

or (b) replace it with one selected randomly from the n_h PSU's. If (a) is the case, go to Step II.

Step II. For each SSU *hij* in PSU *hi* of stratum *h* kept at Step I, we: (c) keep it in a bootstrap sample with probability

$$q_{hi} = \sqrt{1 - \frac{f_{1h}}{p_h^{-1}} \frac{(1 - f_{2hi})}{(1 - m_{hi}^{-1})}}; \qquad (3.2)$$

or (d) replace it with one selected randomly from the m_{hi} SSU's in PSU *hi* of stratum *h*. If (c) is the case, go to Step III.

Step III. For each USU *hijk* in SSU *hij* in PSU *hi* of stratum *h*, we: (e) keep it in the bootstrap sample with probability

$$r_{hij} = \sqrt{1 - \frac{f_{1h}}{p_h^{-1}} \frac{f_{2hi}}{q_{hi}^{-1}} \frac{(1 - f_{3hij})}{(1 - l_{hij}^{-1})}}; \qquad (3.3)$$

or (f) replace it with one randomly selected from the l_{hij} USU's in SSU *hij* in PSU *hi* of stratum *h*.

If we let K_{hij}^* denote the number of times unit *hijk* appears in the bootstrap resample, then the bootstrap estimate of the total is $\hat{\mathbf{Y}}^* = \sum_{hijk} w_{hij}^* \mathbf{y}_\mathbf{hijk}$, where $w_{hij}^* = K_{hij}^* w_{hij}$, and the bootstrap estimate of $V(\hat{\theta})$ is $v_B(\hat{\theta}) = V_*(\hat{\theta}^*)$, where $\hat{\theta}^* = \theta(\hat{\mathbf{Y}}^*)$ and V_* represents the variance under the resampling procedure. Typically, the bootstrap estimate of variance is obtained by Monte Carlo simulation. That is, repeat Steps I–III a large number of times, *B*, to get $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ and use

$$v_B(\hat{\theta}) = \sum_{b=1}^{B} (\hat{\theta}_b^* - \bar{\theta}_{(\cdot)}^*)^2 / B,$$

where $\bar{\theta}_{(\cdot)}^* = \sum_{b=1}^{B} \hat{\theta}^* / B$. In most cases one can replace $\bar{\theta}_{(\cdot)}^*$ by $\hat{\theta}$. This allows the survey methodologist to create a set of replicate weights w_{hij}^* for each bootstrap resample and release these with the data released to the public.

Obviously, the short cut BBE is feasible only when $p_h, q_{hi}, r_{hij} \in [0, 1] \; \forall \; h, i, j$. For instance, it is necessary that $f_{1h} \geq n_h^{-1}$. To handle arbitrary $n_h, m_{hi}, l_{hij} \geq 2$, we may modify each step and change p_h, q_{hi}, r_{hij} accordingly:

A General BBE

Step I′. Choose $(n_h - 1)$ PSU's by SRS with replacement from n_h PSU's in the sample, $h = 1, \ldots, H$. Denote the candidate set by $\{\mathrm{P\tilde{S}U}_{hi} : i = 1, 2, \ldots, n_h - 1\}$. For each PSU *i* in the sample in stratum *h*, we: (a) keep it in the bootstrap sample with probability

$$p_h = 1 - \frac{1}{2} \frac{(1 - f_{1h})}{(1 - n_h^{-1})}; \qquad (3.4)$$

or (b) replace it with one selected randomly from $\{\mathrm{P\tilde{S}U}_{hi} : i = 1, 2, \ldots, n_h - 1\}$. If (a) is the case, go to Step II′.

Step II′. For *hi* kept at Step I′, choose $(m_{hi} - 1)$ SSU's by SRS with replacement from m_{hi} SSU's in PSU *hi*. Denote the candidate set by $\{\mathrm{S\tilde{S}U}_{hij} : j = 1, 2, \ldots, m_{hi} - 1\}$. For each SSU

hij in PSU *hi* kept at Step **I′**, we: (c) keep it in the bootstrap sample with probability

$$q_{hi} = 1 - \frac{1}{2}\frac{f_{1h}}{p_h^{-1}}\frac{(1-f_{2hi})}{(1-m_{hi}^{-1})}; \qquad (3.5)$$

or (d) replace it with one selected randomly from $\{\tilde{\text{SSU}}_{hij}: j = 1, 2, \ldots, m_{hi} - 1\}$. If (c) is the case, go to Step **III′**.

Step III′. For *hij* kept at Step **II′**, choose $l_{hij} - 1$ USU's by SRS with replacement from l_{hij} USU's in SSU *hij* in PSU *hi*. Denote the candidate set by $\{\tilde{\text{USU}}_{hijk}: k = 1, 2, \ldots, l_{hij} - 1\}$. For each USU *hijk* in SSU *hij* in PSU *hi*, we: (e) keep in the bootstrap sample with probability

$$r_{hij} = 1 - \frac{1}{2}\frac{f_{1h}}{p_h^{-1}}\frac{f_{2hi}}{q_{hi}^{-1}}\frac{(1-f_{3hij})}{(1-l_{hij}^{-1})}; \qquad (3.6)$$

or (f) replace it with one randomly selected from $\{\tilde{\text{USU}}_{hijk}: k = 1, 2, \ldots, l_{hij} - 1\}$.

It is readily seen that $p_h, q_{hi}, r_{hij} \in [0, 1] \; \forall \; n_h, m_{hi}, l_{hij} \geq 2$.

The reason for randomly selecting a candidate set in the general BBE can be explained as follows. To fix the idea, consider single-stratum one-stage SRSWOR. Let $\bar{y}^*$ be a bootstrap sample mean under the short cut BBE with some arbitrary $p \in [0, 1]$. Then, it can be shown that $V_*(\bar{y}^*) = n^{-1}(1-n^{-1})s^2(1-p^2)$, where $s^2 = \sum_i (y_i - \bar{y})^2/(n-1)$. Note that $V_*(\bar{y}^*)$ is monotone decreasing with respect to p in $[0, 1]$. So, $\min_{p\in[0,1]} V_*(\bar{y}^*) = 0$ and $\max_{p\in[0,1]} V_*(\bar{y}^*) = n^{-1}(1-n^{-1})s^2$. If $f_1 < n^{-1}$, then $\max_p V_*(\bar{y}^*) < v(\bar{y})$. The key idea of the general BBE is that we can make $\max_p V_*(\bar{y}^*)$ greater than $v(\bar{y})$ by putting extra variation into unit replacement through randomly selecting a candidate set.

It can be shown that both the short cut BBE and the general BBE provide consistent variance estimation for smooth functions of estimated population totals. Moreover, under appropriate regularity conditions for the population distribution function, they also provide consistent variance estimation for sample quantiles. In addition, both BBE methods use resample sizes equal to the original sample sizes. This can be a desirable property when we deal with imputed survey data (see Saigo, Shao and Sitter 2001).

It is not difficult to extend the BBE approach to designs with more than three stages. For example, for a four stage stratified design, a USU at the fourth stage within stratum *h* is kept with probability

$$\sqrt{1 - p_h^{-1} f_{1h} q_{hi}^{-1} f_{2hi} r_{hij}^{-1} f_{3hij} (1 - g_{hijk}^{-1})^{-1} (1 - f_{4hijk})}$$

or replaced in the short cut BBE, where g_{hijk} is the fourth stage sample size and f_{4hijk} is the fourth stage sampling fraction. Further extensions are analogous.

The general BBE randomizes a candidate set in order to merely fix infeasibility of the short cut BBE. This idea has similarities to the approximately Bayesian bootstrap of Rubin and Schenker (1986).

A disadvantage of the general BBE versus the short cut BBE is that the former requires, on the average, $\sum_h \{(n_h - 1) + p_h \sum_i (m_{hi} - 1) + p_h \sum_i q_{hi} \sum_j (l_{hij} - 1)\}$ more random number generations than the latter, where p_h, q_{hi}, and r_{hij} are given by (3.4), (3.5), and (3.6), respectively. This may be time-consuming when the sample sizes and/or the number of strata are large. To reduce random number generations in the general BBE, one can create a candidate set by randomly deleting one unit from the original sample and use

$$p_h = (n_h + 1/2) - \sqrt{(n_h + 1/2)^2 - n_h(1 + f_{1h})}, \qquad (3.7)$$

$$q_{hi} = (m_{hi} + 1/2) - \sqrt{(m_{hi} + 1/2)^2 - f_{h1} p_h^{-1} m_{hi}(1 + f_{2hi})}, \qquad (3.8)$$

$$r_{hij} = (l_{hij} + 1/2) - \sqrt{(l_{hij} + 1/2)^2 - f_{h1} p_h^{-1} f_{2hi} q_{hi}^{-1} l_{hij}(1 + f_{3hij})}, \qquad (3.9)$$

instead. It can be shown that $p_h, q_{hi}, r_{hij} \in [0, 1]$. The proof for this modified version of the general BBE is similar.

4. A Simulation Study

In this section, we perform limited simulations to examine the BBE for ratio estimation and quantile estimation. For simplicity, we consider two-stage SRSWOR and restrict to a single stratum.

4.1 General Description of Simulation

A single-stratum finite population is generated by the following procedure and fixed over all simulation runs to observe design-based properties of the BBE. First, the average of the auxiliary variables in cluster *i* is generated by $\mu_i \sim N(\mu, \sigma^2)$ for $i = 1, 2, \ldots, N$. Then, the auxiliary variable x_{ik} of unit *k* in cluster *i* is generated by

$$x_{ik} = \mu_i + \varepsilon_{ik} (k = 1, 2, \ldots, M_i; i = 1, 2, \ldots, N), \qquad (4.1)$$

where $\varepsilon_{ik} \sim N(0, (1-\rho)\sigma^2/\rho)$. The target variable y_{ik} of unit *k* in cluster *i* is obtained by

$$y_{ik} = a + bx_{ik} + e_{ik} (k = 1, 2, \ldots, M_i; i = 1, 2, \ldots, N), \qquad (4.2)$$

where $e_{ik} \sim N(0, \sigma^2/4)$. The parameter values used are $\mu = 100, \sigma = 10, \rho = 0.1(0.3), a = 0$, and $b = 1$, and two-stage SRSWOR is used throughout the simulation study.

4.2 Ratio Estimation

Let $N = 50, n = 15, M_i = 20$ and $m_i = 3,$ for $i = 1, \ldots, n.$ Consider the ratio estimator of the population total, Y,

$$\hat{Y}_R = \hat{R}\, X,$$

where $X = \sum_{i=1}^{N} \sum_{k=1}^{M_i} x_{ik}$ is the population total of the x's $\hat{R} = \hat{Y}/\hat{X}$, $\hat{Y} = \sum_{h=1}^{H} \hat{Y}_h = \sum_{h=1}^{H} (N_h/n_h) \sum_{i=1}^{n_h} \hat{Y}_{hi}$, $\hat{X} = \sum_{h=1}^{H} \hat{X}_h = \sum_{h=1}^{H} (N_h/n_h) \sum_{i=1}^{n_h} \hat{X}_{hi}$, $\hat{Y}_{hi} = (M_{hi}/m_{hi}) \sum_{k=1}^{m_{hi}} \hat{Y}_{hik}$ and $\hat{X}_{hi} = (M_{hi}/m_{hi}) \sum_{k=1}^{m_{hi}} \hat{X}_{hik}$.

For the purpose of comparison, we consider a number of alternate variance estimators that are available in this simple context:

1) The conventional variance estimator is denoted

$$v_0(\hat{Y}_R) = N^2 \frac{1-f_1}{n} \frac{\sum_i (\hat{Y}_i - \hat{R}\hat{X}_i)^2}{n-1} + \frac{N}{n} \sum_i \frac{M_i^2 (1-f_{2i}) s_{d'2i}^2}{m_i}, \qquad (4.3)$$

where $f_1 = n/N$, $f_{2i} = m_i/M_i$ and

$$s_{d'2i}^2 = \sum_j (y_{ij} - \hat{R}x_{ij})^2 / (m_i - 1).$$

2) The delete 1 PSU at a time jackknife corrected for the first-stage sampling fraction is sometimes used, even though it is not entirely correct,

$$v_{cj}(\hat{Y}_R) = (1-f_1) \frac{n-1}{n} \sum_i (\hat{Y}_{R(i)} - \hat{Y}_{R(\cdot)})^2, \qquad (4.4)$$

where $\hat{Y}_{R(i)}$ is the estimator recalculated with the i^{th} PSU removed and $\hat{Y}_{R(\cdot)} = \sum_i \hat{Y}_{R(i)}/n.$

3) An externally weighted jackknife (see Folsom, Bayless and Shah 1971) can be derived that corrects for both stages of sampling as

$$v_{ewj}(\hat{Y}_R) = (1-f_1) \frac{n-1}{n} \sum_i (\hat{Y}_{R(i)} - \hat{Y}_{R(\cdot)})^2 + f_1 \sum_i (1-f_{2i}) \frac{m_i - 1}{m_i} \sum_j (\hat{Y}_{R(ij)} - \hat{Y}_{R(\cdot\cdot)})^2, \qquad (4.5)$$

where $\hat{Y}_{R(i)}$ is the i^{th} jackknife pseudo value by deleting PSU i, $\hat{Y}_{R(ij)}$ is the ij^{th} jackknife pseudo value by deleting unit j in PSU i, $\hat{Y}_{R(\cdot)} = \sum_i \hat{Y}_{R(i)}/n$, and $\hat{Y}_{R(\cdot\cdot)} = \sum_j \hat{Y}_{R(ij)}/m_i$.

4) A model-assisted variance estimator is also available (see Särndal, Swensson and Wretman (1992), equation (8.10.6)),

$$v_{ma}(\hat{Y}_R) = (X/\hat{X})^2 v_0(\hat{Y}_R). \qquad (4.6)$$

We use $B = 100$ bootstrap resamples in each of $S = 1{,}000$ simulation runs. The true MSE's are approximated by 10,000 simulation runs and we use Monte Carlo estimates of percent relative bias and coefficient of variation of the various variance estimators as measures of their relative performance, as well as, empirical coverage probabilities of 90% confidence intervals.

We see in Table 1 that $v_{\text{BBE}}, v_0, v_{ewj}$ and v_{ma} perform comparably and well, except that the CV of the resampling methods are a bit higher than the non-resampling methods, as is typical. The delete 1 PSU at a time jackknife performs poorly.

To investigate the conditional properties, we ordered the 1,000 simulation runs on $X/\hat{X}$ and divided the runs into 20 equally sized groups. For each group the average of each variance estimator is calculated. Figure 1 plots these grouped averages for each variance estimator (excluding v_{cj} since it has large negative bias) versus the grouped average $X/\hat{X}$, for $\rho = 0.3$. The true MSE is included in the plot, as well. This is a similar plot to that used by Royall and Cumberland (1981a, 1981b). One can see that v_{BBE} tracks the true MSE much like v_{ewj} and v_{ma}, whereas v_0 does not. Thus, the BBE seems to have a desirable conditional property.

Table 1
Comparison of Variance Estimators for $\hat{Y}_r$

ρ		% Bias	CV	Coverage (90%)
0.1	v_0	−1.70	0.28	89.2
	v_{BBE}	−0.62	0.33	88.9
	v_{ewj}	−0.33	0.30	89.4
	v_{cj}	−26.55	0.39	80.5
	v_{ma}	−0.39	0.30	89.4
0.3	v_0	−0.67	0.28	86.6
	v_{BBE}	−1.63	0.33	86.5
	v_{ewj}	−0.74	0.29	86.5
	v_{cj}	−26.85	0.39	80.2
	v_{ma}	−0.87	0.29	86.4

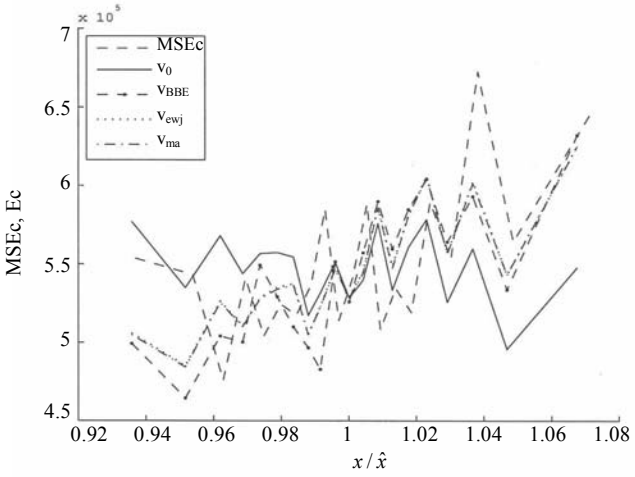


Figure 1. MSEc and Ec(v) for the ratio estimation.

4.3 Quantile Estimation

For quantile estimation, we set $N = 100, n = 30,$ $M_i = 100$ and $m_i = 10,$ for $i = 1, \ldots, n.$ We use $B = 500$ bootstrap resamples in each of $S = 5,000$ simulation runs. The true MSE's are approximated by 50,000 simulation runs. Only the results for v_{BBE} and v_{ewj} when $\rho = 0.1$ are summarized in Table 2 because those when $\rho = 0.3$ are similar. We see that the BBE method performs quite well, with a slight upward bias, while the externally weighted jackknife method has serious bias because of its inconsistency in variance estimation for quantiles.

Table 2
Performance of v_{BBE} and v_{ewj} for the 0.10, 0.25, 0.50, 0.75 and 0.90 quantiles

	v_{BBE}			v_{ewj}		
Quantile	%Bias	CV	Coverage (90%)	%Bias	CV	Coverage (90%)
0.10	8.40	0.51	87.7	51.87	1.93	81.3
0.25	6.21	0.42	88.2	21.19	1.28	83.3
0.50	2.53	0.37	87.4	14.27	1.00	83.0
0.75	6.23	0.42	87.8	28.07	1.33	83.4
0.90	6.32	0.50	88.0	54.47	2.05	80.3

5. Application to the 1997 National Survey of Prices in Japan

The objective of the NSP is to analyze price formations of major consumers' goods, such as food, clothes and home appliances. To this end, quantile estimation plays a central role, and many quantile estimates based on several post-stratifications are included in the NSP reports.

The stratified multistage sampling used in NSP 1997 is summarized as follows:

Stratification. Municipalities form the PSU's and are stratified into 537 strata, first according to prefectures and economic sphere that each municipality forms and then further by their population sizes.

First Stage Sampling. These PSU's are selected via SRSWOR independently within each stratum. An overview of the first-stage sampling fractions is given in Table 3.

Second Stage Sampling. In a selected municipality, all the large scale outlets are enumerated. In other words, single stage cluster sampling is employed for large scale outlets. For small scale outlets, on the other hand, a sampled municipality is divided into survey areas (SSU's) each consisting of about 100 outlets. Systematic sampling is used to sample survey areas. The sampling fractions at the second stage are between 0.1 and 1.0.

Third Stage Sampling. In each selected survey area, 40 outlets (USU's) are chosen by ordered systematic sampling with respect to the types of outlets and the annual sales reported in the 1994 Census of Commerce.

Strictly speaking, there is no valid variance formula for the NSP data because it contains systematic sampling. For estimating variance, however, it is assumed that systematic sampling can be approximated by SRSWOR. Even under this simplified condition, there is no closed variance formula for sample quantiles. In fact, no variance estimates are associated with estimated price quantiles in the NSP report, while the average prices are reported with their variance estimates.

In this section, we apply the short cut BBE to the NSP data, assuming that systematic sampling can be approximated by SRSWOR. Some strata have only one PSU. In addition, $f_{1h} < n_h^{-1}$ in some strata. Such strata are grouped into adjacent strata so that p_h given by (3.1) is in $[0, 1]$. After grouping, there are more than 280 strata. The effect of reforming strata is assumed to be negligible.

Table 3
The First Stage Sampling Fractions in NSP 1997

Area Category	Population Size	# of PSU's	Sampling Fraction	Sample Size
Cities	$\geq 100{,}000$	221	1/1	221
Cities	$50{,}000 - 99{,}999$	220	2/3	179
Cities	$< 50{,}000$	224	1/3	80
Towns and villages	$\geq 40{,}000$	32	1/5	4
Towns and villages	$< 40{,}000$	2,536	1/15	187

After reforming strata, the short cut BBE is employed in those strata composed by cities. On the other hand, the with-replacement bootstrap (Shao and Tu 1995, page 247) using resample size $(n_h - 1)$ is used in those composed of towns and villages, where the first stage sampling fractions are small. The quantile estimates and their standard errors for selected commodities in small-sized outlets are shown in Table 4. Note that the prices of a given commodity are discrete. However, we apply the bootstrap as if prices of commodities are continuous. This approximation should be acceptable for many commodities, but not for very inexpensive ones, since in such a case, a large percentage of observations concentrate on a specific price and the estimated standard error can be 0.

Table 4
Sample Quantiles (Standard Errors) of Selected Commodities for Small Outlets in NSP

Commodity	p	0.10	0.25	0.5	0.75	0.90
Rice (5kg)[a]	Sample quantile	239.4	255.2	278.3	299.1	315.0
(10 yens)	(standard error)	(0.24)	(0.53)	(0.21)	(0.02)	(0.61)
Instant Coffee (1 bottle)[b]	Sample quantile	714	788	859	893	914
(yen)	(standard error)	(0.13)	(0.40)	(0.00)	(2.68)	(1.43)
Beer (24 cans)[c]	Sample quantile	467.3	500.0	536.8	549.4	549.4
(10 yens)	(standard error)	(1.01)	(0.64)	(0.82)	(0.00)	(0.00)
PC[d]	Sample quantile	248.8	260.4	299.3	346.5	375.9
(1,000 yens)	(standard error)	(2.03)	(0.35)	(3.25)	(7.17)	(1.48)

The specified brands [a]Koshihikari; [b]Nescafe Gold Blend, 100g; [c]Sapporo (Nama) Black Label, 350ml; [d]NEC PC9821 NW133/D14.

6. Conclusions

The bootstrap is useful for estimating variances in complex surveys, particularly when quantile estimation is important. We have proposed two Bernoulli-type bootstrap methods that can easily handle multi-stage stratified SRSWOR designs where sampling fractions are large: the short cut BBE and the general BBE. In both methods, a sampling unit at a given stage is either kept or replaced with preassigned probabilities to construct a bootstrap sample. The general BBE has an advantage in that it can handle any combination of sample sizes ≥ 2 although it requires more random number generations than the short cut BBE. As an illustration, we applied the short cut BBE to Japanese 1997 National Survey of Prices data.

Acknowledgements

The second author was supported by the Japan Statistical Association. The third author was supported by a grant from the Natural Science and Engineering Research Council of Canada. The authors thank the Statistics Bureau, Ministry of Public Management, Home Affairs, Posts and Telecommunications, and the Ministry of Economy, Trade, and Industry, Japan, for providing the 1997 NSP data.

References

Bickel, P.J., and Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12, 470-482.

Booth, J.G., Butler, R.W. and Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89, 1282-1289.

Folsom, R.E., Bayless, D.L. and Shah, B.V. (1971). Jackknifing for variance components in complex sample survey designs. *Proceedings of the Social Statistics Section*, American Statistical Association, 36-39.

Gross, S. (1980). Median estimation in sample surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 181-184.

Kovar, J.G., Rao, J.N.K. and Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statististics*, 16, Supplement, 25-45.

McCarthy, P.J., and Snowden, C.B. (1985). The bootstrap and finite population sampling. *Vital and Health Statistics*, Serie 2, 95, Public Health Service Publication, 85-1369, Washington, DC: U.S. Government Printing Office.

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.

Royall, R.M., and Cumberland, W.G. (1981a). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-77.

Royall, R.M., and Cumberland, W.G. (1981b). The finite population linear regression estimator: An empirical study. *Journal of the American Statistical Association*, 76, 924-930.

Rubin, D.B., and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statististical Association*, 81, 366-374.

Saigo, H., Shao, J. and Sitter, R.R. (2001). A repeated half-sample bootstrap and balanced repeated replication for randomly imputed data. *Survey Methodology*, 27, 189-196.

Särndal, C.-E., Swenson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Shao, J., and Sitter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278-1288.

Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer-Verlag.

Sitter, R.R. (1992a). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87, 755-765.

Sitter, R.R. (1992b). Comparing three bootstrap methods for survey data. *Canadian Journal of Statististics*, 20, 135-154.

Survey Methodology, December 2006
Vol. 32, No. 2, pp. 157-163
Statistics Canada, Catalogue No. 12-001

Geometric Versus Optimization Approach to Stratification: A Comparison of Efficiency

Marcin Kozak and Med Ram Verma [1]

Abstract

In this paper, the geometric, optimization-based, and Lavallée and Hidiroglou (LH) approaches to stratification are compared. The geometric stratification method is an approximation, whereas the other two approaches, which employ numerical methods to perform stratification, may be seen as optimal stratification methods. The algorithm of the geometric stratification is very simple compared to the two other approaches, but it does not take into account the construction of a take-all stratum, which is usually constructed when a positively skewed population is stratified. In the optimization-based stratification, one may consider any form of optimization function and its constraints. In a comparative numerical study based on five positively skewed artificial populations, the optimization approach was more efficient in each of the cases studied compared to the geometric stratification. In addition, the geometric and optimization approaches are compared with the LH algorithm. In this comparison, the geometric stratification approach was found to be less efficient than the LH algorithm, whereas efficiency of the optimization approach was similar to the efficiency of the LH algorithm. Nevertheless, strata boundaries evaluated via the geometric stratification may be seen as efficient starting points for the optimization approach.

Key Words: Optimum stratification; Geometric Stratification; Numerical Optimization; Lavallée-Hidiroglou algorithm.

1. Introduction

Gunning and Horgan (2004) proposed a stratification algorithm based on a geometric progression. For the sake of simplicity, we will call this technique the "geometric approach to stratification," "geometric stratification," or just "geometric approach." The geometric stratification aims to equalize values of the coefficient of variation of a stratification variable within strata, based on the assumption that the variable is uniformly distributed within each stratum. Gunning and Horgan (2004) showed that their algorithm is much easier to implement and more efficient than the classical cumulative root frequency method (Dalenius and Hodges 1959) as well as the Lavallée and Hidiroglou (LH) algorithm (Lavallée and Hidiroglou 1988). Horgan (2006) compared the geometric stratification with the Dalenius and Hodges' (1959), Ekman's (1959), and Lavallée and Hidiroglou (1988) procedures; again, in their study the geometric stratification occurred to be the most efficient among the procedures compared. Gunning, Horgan and Yancey (2004) applied this method to stratify accounting populations.

Like the cumulative square root frequency method, the geometric approach is an approximate stratification technique, and hence the stratification points it provides may be quite far from optimum stratification points. On the other hand, there exist approaches, especially for univariate stratification, that lead to near-optimum stratification points. These approaches are based on the use of self-implemented algorithms or numerical optimization methods to provide strata boundaries (*e.g.*, Lavallée and Hidiroglou 1988; Lednicki and Wieczorkowski 2003; Kozak 2004). Such methods, however, usually require initial strata boundaries to start an optimization process; approximate stratification methods can be employed to find such initial points. Of course, initial strata boundaries should be of high quality, as their low quality may cause the optimization to provide a local minimum (Rivest 2002).

Many surveys deal with positively skewed study variables. If this is the case, it is important to take into account this attribute when stratifying a population. Many researchers have attempted to create stratification methods that would construct a so-called "take-all" stratum (*e.g.*, Glasser 1962; Hidiroglou 1986), from which all the elements are selected in the sample with probability 1. In stratified sampling, this is the best manner of dealing with positively skewed variables. Such methods are usually more efficient (certainly, only if a population is positively skewed) than stratification methods in which a take-all stratum is not constructed. A take-all stratum is not constructed in the geometric stratification (Gunning and Horgan 2004).

The aim of this paper is to compare the efficiency of the geometric stratification, proposed by Gunning and Horgan (2004), and two optimization approaches to stratification (Lavallée and Hidiroglou 1988; Lednicki and

1. Marcin Kozak, Department of Biometry, Warsaw Agricultural University, Nowoursynowska 159, 02-776 Warsaw, Poland. E-mail: marcin.kozak@omega.sggw.waw.pl; Med Ram Verma, Division of Agricultural Economics & Statistics, ICAR Research Complex for N.E.H. Region, Umroi Road, Umiam (Barapani) Meghalaya, India, Pin 793 103. E-mail: mrverma19@yahoo.co.in.

Wieczorkowski 2003; Kozak 2004), which is based on the use of numerical optimization methods.

2. Stratification Approaches Compared

Suppose we aim to stratify an $N-$element positively skewed population, U, based on an $N-$vector $\mathbf{x} = (x_1, \ldots, x_N)^T$ of values, known at the outset (*i.e.*, prior to the study), of a stratification variable X.

In this paper, we consider two stratification problems. In the first problem, L strata are to be constructed subject to a given sample size n. Suppose we are looking for an $(L+1)$ –vector of strata boundaries $\mathbf{k} = (k_0, \ldots, k_L)^T$ $(k_0 < k_1 < \ldots < k_L, k_0$ being the minimum and k_L the maximum value of $X)$ that minimizes the variance of an estimator of the population mean of X under stratified sampling with simple random sampling without replacement within strata (*STSI*) sampling combined with a take-all stratum approach. (Note that we treat the stratification variable as identical to the corresponding survey variable.) The variance of $\bar{x}_{\text{st}}$ is given by

$$V(\bar{x}_{\text{st}}) = \sum_{h=1}^{L-1} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h},$$

$$\bar{x}_{\text{st}} = \sum_{h=1}^{L} \frac{N_h}{N} \bar{x}_h, \bar{x}_h = \frac{1}{n_h} \sum_{k=1}^{n_h} x_{kh} (h = 1, \ldots, L), \qquad (1)$$

where n_h is the sample size from the h^{th} stratum, N_h is the size of the h^{th} stratum, S_h^2 is the population variance of X restricted to the h^{th} stratum, $\bar{x}_{\text{st}}$ is the estimator of the population mean of X under *STSI* sampling, $\bar{x}_h$ is the estimator of the population mean of X in the h^{th} stratum under simple random sampling without replacement (*SI*) sampling, and x_{kh} is the value of X for the k^{th} sample element of the h^{th} stratum and $h = 1, \ldots, L$.

The optimum sample allocation, which is in our problem obtained by minimizing the variance (1) subject to a given sample size n, is given by the following Neyman-optimum formula adjusted to a take-all stratum approach (Lednicki and Wieczorkowski 2003):

$$n_h = (n - N_L) \frac{N_h S_h}{\sum_{h=1}^{L-1} N_h S_h}, h = 1, \ldots, L-1. \qquad (2)$$

The geometric approach to stratification aims to equalize values of the coefficient of variation of X within the L strata. It simply consists of applying the following formula based on a geometric progression (Gunning and Horgan 2004)

$$k_h = ar^h, h = 0, \ldots, L, \qquad (3)$$

where $a = \min(X), k_L = \max(X),$ and $r = (k_L / k_0)^{1/L}$. The formula (3) is based on the assumption that X is uniformly distributed within each stratum.

The optimization approach applied to this particular stratification problem is based on the numerical optimization of the following problem: Minimize

$$f(\mathbf{k}) = V(\bar{x}_{\text{st}}), \qquad (4)$$

where $V(\bar{x}_{\text{st}})$ is the variance (1) under the optimum allocation (2), subject to constraints

$$N_h \geq 2 \text{ and } 2 \leq n_h \leq N_h \text{ for } h = 1, \ldots, L-1, \qquad (5)$$

and

$$\sum_{h=1}^{L-1} n_h = n - N_L. \qquad (6)$$

Sometimes, when one wants to obtain more or less equal levels of precision of estimation in each stratum, a power allocation may be applied (Bankier 1988; Rivest 2002; Lednicki and Wieczorkowski 2003):

$$n_h = \frac{(n - N_L)(N_h \bar{x}_h)^P}{\sum_{h=1}^{L-1} (N_h \bar{x}_h)^p}, p \in (0,1]; h = 1, \ldots, L-1. \qquad (7)$$

The optimization approach is more difficult to apply than the geometric stratification approach due in large part to the fact that the algorithm for the geometric approach is significantly more simplistic than for the optimization approach. An optimization method has to be chosen from among various available methods. Lednicki and Wieczorkowski (2003) used the simplex method of Nelder and Mead (1965); however, more efficient methods, which often require self-implemented algorithms (*e.g.*, Kozak 2004), can be applied, too.

Note that the geometric stratification does not take into account the formulae for the variance (1), the sample allocation (2), and the constraints (5). It may happen that one of the constraints (5) is not fulfilled. For these reasons, the geometric stratification is an approximate stratification procedure.

In this study, the algorithm proposed by Kozak (2004) was applied to stratify several populations. It is a random search algorithm adjusted to the problem of stratification. It is a simple algorithm; in each step, a stratum boundary is randomly selected and randomly changed. If the new set of strata boundaries is better than the previous one, the new one replaces the previous one. In the Appendix, the algorithm based on the paper by Kozak (2004) is given in detail.

The second problem considered in the paper is construction of strata that minimize a sample size from a population with respect to a given level of precision of estimation (the precision of estimation being given by the variance of an estimator of the population mean or total). The Lavallée-Hidiroglou (LH) algorithm (Lavallée and Hidiroglou 1988) can be seen as a particular optimization method to solve this particular stratification problem; it does not, however, work in other problems, *e.g.*, in the one considered earlier. For details of the algorithm, see the paper by Lavallée and Hidiroglou (1988). Besides the LH algorithm, the geometric stratification and random search method were applied to construct the strata.

The R language and environment (R Development Core Team 2005) was used to perform all the computation work in the present study.

3. Numerical Comparison of Efficiency of the Approaches in Stratification Under Fixed Sample Size

In this section, we compare two stratification approaches, the geometric stratification (geom) and optimization approach (optim), applied to a problem of searching for the strata boundaries that minimize the variance of the considered estimator with respect to a fixed sample size. In order to perform the comparison, five artificial populations of various sizes (from 2,000 to 10,000) were generated. Their summary statistics are presented in Table 1; the histograms of the stratification variables in the populations are given in Figure 1. In each case, the stratification variable was positively skewed (the skewness ranged between 1.40 for the 1st population to 5.02 for the 5th population). As it is usually the case in real populations, values of the stratification variables were integers. The sample size, n_i, from the i^{th} population was $n_i = f N_i$, where $f = 0.15$ is an assumed sample fraction and N_i is the size of the i^{th} population.

Table 1
Summary Statistics for Studied Artificial Populations

Population	Size	Range	Skewness	Mean	Variance
1	4,000	3–72	1.40	16.11	45.8
2	4,000	243–28,578	2.66	2,823.95	4.8×10^6
3	2,000	6–2,793	3.55	224.12	6.0×10^4
4	10,000	62–74,398	4.20	3,616.41	2.1×10^7
5	2,000	259–186,685	5.02	9,265.36	1.1×10^8

First, each population was stratified using the geometric stratification method into 4, 5, 6, and 7 strata. Then, the optimization approach was applied; as initial parameters in the optimization approach, the strata boundaries determined via the geometric stratification were used.

Like Gunning and Horgan (2004), to compare the efficiency of the two approaches, the relative efficiency was calculated via the formula:

$$\text{eff}_{\text{geom, optim}} = \frac{V_{\text{geom}}(\bar{x}_{\text{st}})}{V_{\text{optim}}(\bar{x}_{\text{st}})}, \tag{8}$$

where $V_{\text{geom}}(\bar{x}_{\text{st}})$ and $V_{\text{optim}}(\bar{x}_{\text{st}})$ are the variances (1) under the geometric and optimization approach, respectively. In addition, we calculated the coefficients of variation of the estimator of the population mean under both approaches:

$$\text{cv}_{\text{geom}} = \frac{\sqrt{V_{\text{geom}}(\bar{x}_{\text{st}})}}{\bar{x}_{\text{st}}};\ \text{cv}_{\text{optim}} = \frac{\sqrt{V_{\text{optim}}(\bar{x}_{\text{st}})}}{\bar{x}_{\text{st}}}. \tag{9}$$

Table 2 contains the values of the relative efficiencies (8) and the coefficients of variation (9) for each combination studied (population × number of strata).

Table 2
Coefficients of Variation of the Estimator of the Population Mean Under the Geometric Stratification (CV_{geom}) and Optimization Approach (CV_{optim}), and Efficiencies of the Geometric Stratification Relative to the Optimization Approach ($eff_{geom,optim}$)

Number of strata *L*	CV_{geom}	CV_{optim}	$eff_{geom,optim}$
Population 1			
4	0.0086	0.0056	1.53
5	0.0070	0.0042	1.66
6	0.0057	0.0034	1.66
7	0.0051	0.0029	1.75
Population 2			
4	0.0116	0.0084	1.37
5	0.0095	0.0065	1.47
6	0.0085	0.0051	1.66
7	0.0073	0.0042	1.72
Population 3			
4	0.0235	0.0133	1.76
5	0.0174	0.0100	1.74
6	0.0146	0.0081	1.80
7	0.0129	0.0067	1.91
Population 4			
4	0.0104	0.0063	1.64
5	0.0089	0.0047	1.88
6	0.0073	0.0038	1.93
7	0.0064	0.0032	2.00
Population 5			
4	0.0235	0.0134	1.76
5	0.0185	0.0100	1.86
6	0.0161	0.0080	2.00
7	0.0134	0.0074	1.82

In each case, the optimization approach was more efficient than the geometric stratification. The efficiency was smaller than 1.5 for only two combinations; in the rest of combinations, it ranged between 1.5 and 2. Usually, the more strata constructed the greater the gain in efficiency.

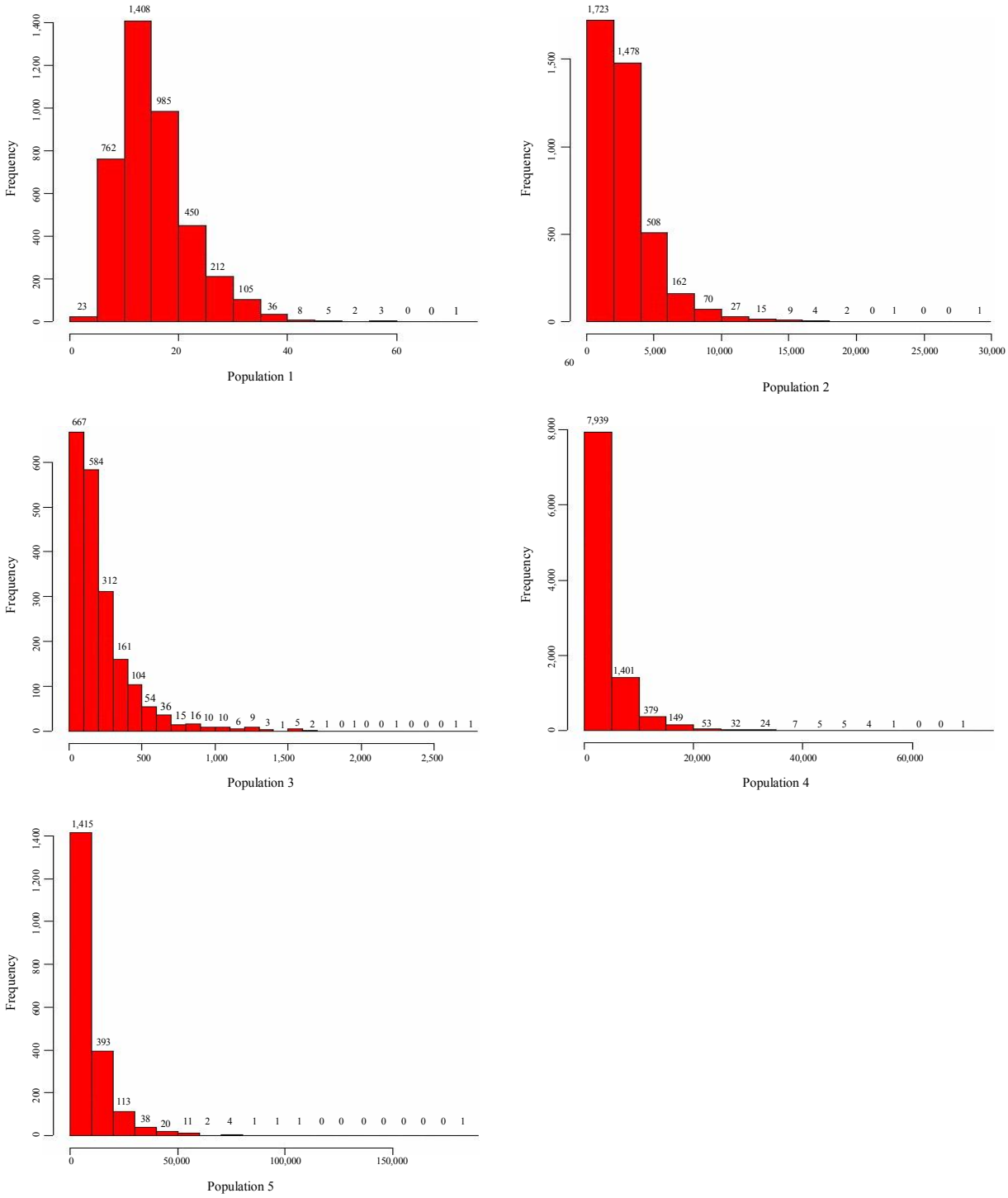


Figure 1. Histograms of stratification variable in studied artificial populations.

4. Numerical Comparison of Efficiency of the Stratification Approaches Under Fixed Level of Precision of Estimation

Gunning and Horgan (2004) and Horgan (2006) compared the geometric stratification with the Lavallée and Hidiroglou (Lavallée and Hidiroglou 1988) algorithm and found that the former was usually more efficient. In this section, we compare the three stratification approaches: the geometric stratification, the LH algorithm, and the optimization approach via a random search method. In this study, the same five populations as in the previous section were used (see Table 1 and Figure 1).

The relative efficiencies of two approaches were evaluated as

$$\text{eff}_{i,j} = \frac{n_i(\text{cv})}{n_j(\text{cv})}, \qquad (10)$$

where i and j are the indices of the stratification approaches (i, j = geom, optim, LH), and $n_i(\text{cv})$ and $n_j(\text{cv})$ are the minimum sample sizes required to obtain a desired level of precision (cv) under the i^{th} and j^{th} approaches, respectively.

Using the three approaches, each population was stratified into $L = 4, \ldots, 7$ strata; the required level of precision was 0.01 in each case. Minimum sample sizes required for this level of precision and relative efficiencies (10) are given in Table 3.

Table 3
Minimum Sample Sizes Required to Obtain a Value Equal to 0.01 for the Coefficient of Variation of the Estimator of the Population Mean, Under the Geometric Stratification (n_{geom}), Optimization Approach (n_{optim}), and LH Algorithm (n_{LH}); and Efficiencies of the Geometric Stratification Relative to the Optimization Approach ($\text{eff}_{\text{geom, optim}}$), the Geometric Stratification Relative to the LH Algorithm ($\text{eff}_{\text{geom, LH}}$), and LH Algorithm Relative to the Optimization Approach ($\text{eff}_{\text{LH, optim}}$)

Number of strata L	n_{geom}	n_{optim}	n_{LH}	$\text{eff}_{\text{geom, optim}}$	$\text{eff}_{\text{geom, LH}}$	$\text{eff}_{\text{LH, optim}}$
Population 1						
4	805	496	496	1.63	1.63	1.00
5	613	344	344	1.78	1.78	1.00
6	460	252	252	1.83	1.83	1.00
7	357	192	192	1.86	1.86	1.00
Population 2						
4	483	248	259	1.94	1.86	1.04
5	329	154	163	2.14	2.02	1.06
6	224	113	117	1.98	1.92	1.03
7	180	83	83	2.17	2.17	1.00
Population 3						
4	782	410	411	1.91	1.90	1.00
5	601	303	304	1.98	1.98	1.00
6	495	242	241	2.04	2.05	1.00
7	422	195	195	2.11	2.16	1.00
Population 4						
4	839	409	409	2.05	2.05	1.00
5	650	301	301	2.15	2.15	1.00
6	552	240	242	2.30	2.28	1.01
7	–[1]	200	200	–	–	1.00
Population 5						
4	1,768	894	894	1.98	1.98	1.00
5	1,274	628	628	2.03	2.03	1.00
6	949	459	459	2.07	2.07	1.00
7	758	355	355	2.13	2.13	1.00

[1] There were numerical problems with obtaining stratum boundaries (sample sizes from some strata were bigger than the sizes of these strata).

From the results it follows that the optimization approach was more efficient than the geometric stratification; this outcome was obtained for each population and number of strata. The relative efficiency was always greater than 1.6. Moreover, an interesting conclusion follows from the comparison of the efficiency of the geometric and LH stratifications. As already mentioned, Gunning and Horgan (2004) and Horgan (2006) found the geometric stratification more efficient than the LH algorithm. On the contrary, in our study, the LH algorithm was always more efficient than the geometric stratification. This situation occurred also for other generated populations of various sizes and skewness (results not included in this paper). Nevertheless, we do not state that the LH algorithm is always more efficient than the geometric stratification. It may happen that the geometric stratification will be better, as Gunning and Horgan (2004) and Horgan (2006) obtained in their studies.

From the comparison of the LH algorithm and the optimization approach it follows that both approaches provides stratification points leading to similar sample sizes. In some cases, the LH stratification was slightly better and in some other cases slightly worse than the optimization approach. Nevertheless, these differences do not mean that we could indicate either of these two approaches as more efficient. In fact, these two approaches have the same aim (in this particular stratification problem) and they just differ in the algorithm to achieve this aim. In summary, on the basis of our results we conclude that, in general, the LH stratification and optimization approach are more efficient than the geometric stratification.

5. Conclusions

The stratification technique based on a geometric progression proposed by Gunning and Horgan (2004) has a significant advantage; namely, its algorithm is very simple to implement compared to the cumulative square root of frequency method of Dalenius and Hodges (1959) and to other stratification methods. It is, however, an approximate stratification procedure, so the stratification points it provides may lead to poor precision of estimation (or a large sample size required to achieve a required level of precision). Furthermore, it is likely that some of the strata constructed will not fulfill the constraints (5); *e.g.*, some strata may be empty (so they would not comprise any population element) or/and sample sizes from some strata may be smaller than two or greater than their population sizes.

In our study, the optimization approach (via the LH and random search algorithms) was more efficient than the geometric stratification for each population studied and number of strata constructed. Nevertheless, the strata boundaries provided by the geometric stratification can be seen as efficient initial parameters required in the optimization approach; they should not be considered, however, as the optimal or efficient strata boundaries. Furthermore, our results conclusively show that the geometric stratification is less efficient than the stratification

presented by Lavallée and Hidiroglou (1988), which is the result opposite to the one obtained by Gunning and Horgan (2004) and Horgan (2006). This problem needs further studies on real skewed populations; investigations on artificial populations univocally show that the LH algorithm and the optimization approach are more efficient than the geometric stratification.

At first look, one could be surprised that the gain in efficiency after applying the LH and optimization approaches compared to the geometric stratification increases after increasing the number of strata. This can be easily explained. The aim of the geometric stratification is to equalize cvs of the stratification variable within the strata. Therefore, this is not the same aim as the aim of stratification, which is to optimize the efficiency of estimation or to minimize a sample size. Furthermore, there is no certainty that under the optimum stratification the distribution of the stratification/survey variable is uniform within the strata. These two sets of strata boundaries (*i.e.*, provided by the geometric and optimization approaches) are not necessarily the same; in fact, they are likely different.

Note that we applied the random search method as the algorithm of the optimization approach to stratification. In fact, Lavallée and Hidiroglou's (1988) algorithm is a representative of optimization approaches, too. When the aim of stratification is to minimize a sample size required to achieve a desired level of precision, the two approaches will likely provide similar results, as they did in our study. Nevertheless, the random search algorithm may be applied to any stratification problem (*i.e.*, any optimization function and its constraints), contrary to the LH algorithm, which is applicable only when a sample size is minimized with respect to a given level of precision. It is to be noted that the random search algorithm, as a global optimization method, provides random results.

Our aim, however, was not to promote any of these two algorithms by showing that they are more efficient than the geometric stratification. In addition, we applied Nelder and Mead's (1965) simplex method to stratify the populations (results not presented in the paper); its results were very similar to those of the LH and random search method algorithms. Each of these methods has some drawbacks. For instance, numerical difficulties may occur while using the LH algorithm (Slanta and Krenzke 1996); the random search method provides random results (Kozak 2004); Nelder and Mead's (1965) method may be inefficient under large number of strata and large populations (Kozak 2004); and, in fact, none of the methods has been proven to provide optimum stratification points. Therefore, there is still a need of constructing a stratification algorithm that would be optimum irrespective of the situation (*e.g.*, of a population size or variable's skewness) and that would provide results that are not random. Our main aim was to prove that the geometric stratification is not optimum, although the stratification points it provides may be useful as initial parameters in other approaches to stratification.

Acknowledgements

The authors are very indebted to the referees and the Associate Editor of *Survey Methodology* for their valuable comments, which helped to improve the first version of this paper.

Appendix

The algorithm given below was proposed by Kozak (2004); we have adapted some of its details to the general stratification problem. In the algorithm, we do not refer to the particular problem of stratification (*i.e.*, we do not define the optimization function and its constraints), since the algorithm works for both problems presented in the paper as well as for other stratification problems. Where required, we refer to "optimization function" (which may be either the variance of an estimator considered or a sample size from a population) and "constraints" (which, depending on the optimization function, may be the constraints (5) and (6), or the constraints (5) combined with the constraint on the level of precision of estimation); certainly, other forms of the optimization function and its constraints may be considered as well.

Let us define a vector **a** as follows. It takes values on the interval $(1, N)$, N being the population size. Provided that a population is sorted by the values of a stratification variable X, two elements a_{h-1} and a_h of the vector **a** define the stratum h in such a way that this stratum consists of the elements with the index I (which gives the order of an element in the population sorted) that $a_{h-1} < I \le a_h$, $h = 1, \ldots, L$, $a_0 = 0$, $a_L = N$. The algorithm is as follows.

1. Sort the population by the values of the stratification variable.
2. Choose an initial vector **a**, *i.e.*, the vector of initial strata boundaries. You may use random integers that satisfy the constraints, but practice shows that better results may be achieved by using approximate strata boundaries obtained via some approximate stratification methods. Calculate the value of the optimization function. Check the constraints; if they are not fulfilled, the initial points have to be changed.

3. For $r = 0, 1, \ldots, R$ repeat the following step:

   a. Generate point $\mathbf{a}'$ by drawing one stratum boundary a_i and changing it as follows

$$a_i' = a_i + j,$$
$$a_k' = a_k \quad \text{for } k = 1, \ldots, L-1, k \neq i, \qquad (11)$$

   where j is the random integer, $j \in \langle -p; -1 \rangle \cup \langle 1; p \rangle$, p being a given integer chosen based on the population size (the larger the population, the larger the p value); usually, it should be between 3 and 5.

   b. Calculate the value of the optimization function.

   c. If the constraints are satisfied and the value of the optimization function under the vector $\mathbf{a}'$ is smaller than the value under the vector $\mathbf{a}$, accept the new vector, *i.e.*, $\mathbf{a}_{r+1} = \mathbf{a}'$ (where $\mathbf{a}_{r+1}$ is the vector of strata boundaries in a next iteration); otherwise do not accept the vector, *i.e.*, $\mathbf{a}_{r+1} = \mathbf{a}$.

4. Finish the algorithm if the stopping rule is fulfilled, *e.g.*, if $r = R$, where R is given number of steps, or if in the last m (for instance, 50) steps the value of the optimization function did not change. Finally, calculate the vector $\mathbf{k}$ (the vector of final strata boundaries) on the basis of the values of the vector $\mathbf{a}$.

References

Bankier, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician*, 42, 174-177.

Dalenius, T., and Hodges, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54, 88-101.

Ekman, G. (1959). An approximation useful in univariate stratification. *Annals of Mathematical Statistics*, 30, 219-229.

Glasser, G.J. (1962). On the complete coverage of large units in a statistical study. *Review of the International Statistical Institute*, 30, 28-32.

Gunning, P., and Horgan, J.M. (2004). A simple algorithm for stratifying skewed populations. *Survey Methodology*, 30, 159-166.

Gunning, P., Horgan, J.M. and Yancey, W. (2004). Geometric stratification of accounting data. *J. de Contaduria y Administration*, 214, septiembre-diciembre.

Hidiroglou, M.A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40, 27-31.

Horgan, J.M. (2006). Stratification of skewed populations: A review. *International Statistical Review*, 74(1): 67-76.

Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6(5), 797-806.

Lavallée, P., and Hidiroglou, M. (1988). On the stratification of skewed populations. *Survey Methodology*, 14, 33-43.

Lednicki, B., and Wieczorkowski, R. (2003). Optimal stratification and sample allocation between subpopulations and strata. *Statistics in Transition*, 6, 287-306.

Nelder, J.A., and Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, 7, 308-313.

R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; URL http://www.R-project.org.

Rivest, L.-P. (2002). A generalization of Lavallée and Hidiroglou algorithm for stratification in business surveys. *Survey Methodology*, 28, 191-198 (http://www.mat.ulaval.ca/pages/lpr/).

Slanta, J., and Krenzke, T. (1996). Applying the Lavallée and Hidiroglou method to obtain stratification boundaries for the Census Bureau's annual Capital Expenditure Survey. *Survey Methodology*, 22, 65-75.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Survey Methodology, December 2006
Vol. 32, No. 2, pp. 165-176
Statistics Canada, Catalogue No. 12-001

Indirect Sampling: The Foundations of the Generalized Weight Share Method

Jean-Claude Deville and Pierre Lavallée [1]

Abstract

To select a survey sample, it happens that one does not have a frame containing the desired collection units, but rather another frame of units linked in a certain way to the list of collection units. It can then be considered to select a sample from the available frame in order to produce an estimate for the desired target population by using the links existing between the two. This can be designated by *Indirect Sampling*.

Estimation for the target population surveyed by Indirect Sampling can constitute a big challenge, in particular if the links between the units of the two are not one-to-one. The problem comes especially from the difficulty to associate a selection probability, or an estimation weight, to the surveyed units of the target population. In order to solve this type of estimation problem, the Generalized Weight Share Method (GWSM) has been developed by Lavallée (1995) and Lavallée (2002). The GWSM provides an estimation weight for every surveyed unit of the target population.

This paper first describes Indirect Sampling, which constitutes the foundations of the GWSM. Second, an overview of the GWSM is given where we formulate the GWSM in a theoretical framework using matrix notation. Third, we present some properties of the GWSM such as unbiasedness and transitivity. Fourth, we consider the special case where the links between the two populations are expressed by indicator variables. Fifth, some special typical linkages are studied to assess their impact on the GWSM. Finally, we consider the problem of optimality. We obtain optimal weights in a weak sense (for specific values of the variable of interest), and conditions for which these weights are also optimal in a strong sense and independent of the variable of interest.

Key Words: Indirect Sampling; Generalized Weight Share Method; Unbiasedness; Optimal Weights.

1. Introduction

To select the samples needed for social or economic surveys, it is useful to have sampling frames, *i.e.*, lists of units intended to provide a way to reach desired target populations. Unfortunately, it happens that one does not have a list containing the desired collection units, but rather another list of units linked in a certain way to the list of collection units. One can speak therefore of two populations U^A and U^B linked to each other, where one wants to produce an estimate for U^B. Unfortunately, a sampling frame is only available for U^A. It can then be considered to select a sample s^A from U^A in order to produce an estimate for U^B by using the correspondence existing between the two populations. This can be designated by *Indirect Sampling*.

Estimation for a target population U^B surveyed by Indirect Sampling can constitute a big challenge, in particular if the links between the units of the two populations are not one-to-one. The problem comes especially from the difficulty to associate a selection probability, or an estimation weight, to the surveyed units of the target population. In order to solve this type of estimation problem, the Generalized Weight Share Method (GWSM) has been developed by Lavallée (1995) and Lavallée (2002), and presented also in Lavallée and Caron (2001). The GWSM provides an estimation weight for every surveyed unit of the target population U^B. Basically, this estimation weight corresponds to a weighted average of the survey weights of the units of the sample s^A. The GWSM is an extension of the Weight Share Method described by Ernst (1989) in the context of longitudinal household surveys.

The purposes of this paper are to describe Indirect Sampling – the foundations underlying the GWSM – and to obtain optimal weights from the GWSM that provide unbiased estimates with minimum variance. First, we will describe Indirect Sampling together with the GWSM in a theoretical framework that will use, for instance, matrix notation. The use of matrix notation for the GWSM has previously been presented by Deville (1998). Second, we will use this theoretical framework to state some general properties associated with the GWSM that include unbiasedness and transitivity. Transitivity is to go from the population U^A to a target population U^C, through an intermediate population U^B. Third, we will show the correspondence between the matrix formulation and the one that has been described in Lavallée (1995), Lavallée (2002), and Lavallée and Caron (2001). Fourth, we will study the effect of various typical link matrices between U^A and U^B on the precision of the estimates obtained from the GWSM. Finally, we will assess the problem of optimality. We will obtain optimal weights in a weak sense (for specific values

1. Jean-Claude Deville, Laboratoire de Statistique d'Enquête (ENSAI/CREST), Campus de Ker Lann, rue Blaise Pascal, 35170 Bruz, FRANCE. E-mail: deville@ensai.fr; Pierre Lavallée, Statistics Canada, Ottawa, Ontario, K1A 0T6, CANADA. E-mail: pierre.lavallee@statcan.ca.

of the variable of interest), and conditions under which these weights are also optimal in a strong sense and independent of the variable of interest.

2. Indirect Sampling

As mentioned in the introduction, with Indirect Sampling, we select a sample s^A from a population U^A in order to produce an estimate for a target population U^B. For that, we use the correspondence existing between the two populations. For example, assume that we want to produce estimates for a population of children (collection units) while we only have a sampling frame of parents. The target population U^B is the one of the children, but we need to select a sample of parents before being able to interview the children. This is illustrated in Figure 1.

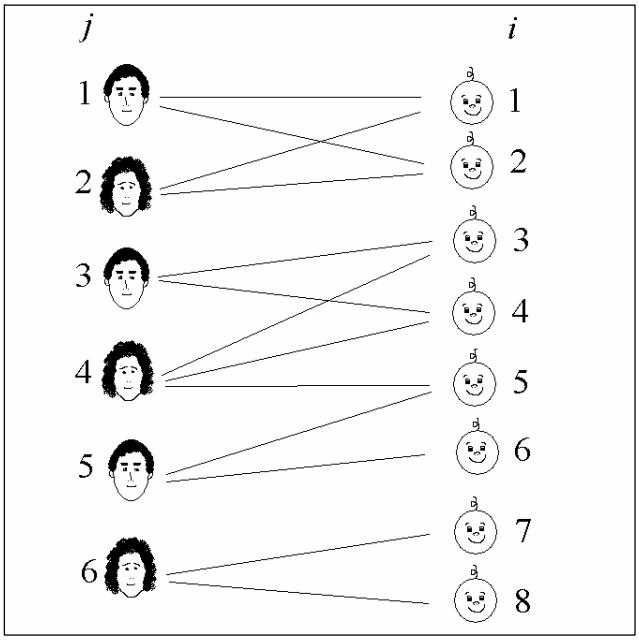


Figure 1. Population U^A of parents and population U^B of children with the links between the two.

Let the population U^A contain N^A units, where each unit is labeled by the letter j. Similarly, let the target population U^B contain N^B units, where each unit is labeled by the letter i. The correspondence between the two populations U^A and U^B can be represented by a *link matrix* $\mathbf{\Theta}_{AB} = [\theta_{ji}^{AB}]$ of size $N^A \times N^B$ where each element $\theta_{ji}^{AB} \geq 0$. That is, unit j of U^A is related to unit i of U^B provided that $\theta_{ji}^{AB} > 0$, otherwise the two units are not related to each other. For the above example, the link matrix is given by

$$\mathbf{\Theta}_{AB} = \begin{bmatrix} \theta_{11}^{AB} & \theta_{12}^{AB} & 0 & 0 & 0 & 0 & 0 & 0 \\ \theta_{21}^{AB} & \theta_{22}^{AB} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \theta_{33}^{AB} & \theta_{34}^{AB} & 0 & 0 & 0 & 0 \\ 0 & 0 & \theta_{43}^{AB} & \theta_{44}^{AB} & \theta_{45}^{AB} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \theta_{55}^{AB} & \theta_{56}^{AB} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \theta_{67}^{AB} & \theta_{68}^{AB} \end{bmatrix}.$$

Obtaining the link matrix *link matrix* $\mathbf{\Theta}_{AB} = [\theta_{ji}^{AB}]$ is a critical issue in Indirect Sampling. For the case where two units $j \in U^A$ and $i \in U^B$ are not linked, we simply set $\theta_{ji}^{AB} = 0$. When there is a link between two units j and i, the choice of $\theta_{ji}^{AB} > 0$ is important. As we will see, it influences the precision of the estimates issued from Indirect Sampling. Now, in several applications, the values of θ_{ji}^{AB} for the linked units are simply set to 1. Of course, the values of θ_{ji}^{AB} for the linked units can be chosen to be different from 1. Lavallée and Caron (2001) discussed the use of the linkage weights obtained from a record linkage process between U^A and U^B for assigning values to the θ_{ji}^{AB}. The linkage weights are proportional to the probability of two units $j \in U^A$ and $i \in U^B$ being linked. Since the choice of $\theta_{ji}^{AB} > 0$ for two linked units j and i can affect the precision of the estimates, it is natural to seek for those θ_{ji}^{AB} that will minimize the variance of the estimates. This optimization problem is considered in section 6 of the paper.

With Indirect Sampling, we select the sample s^A of n^A units from U^A using some sampling design. Let π_j^A be the selection probability of unit j. We assume $\pi_j^A > 0$ for all $j \in U^A$. For each unit j selected in s^A, we identify the units i of U^B that have a non-zero correspondence, *i.e.*, with $\theta_{ji}^{AB} > 0$. Let Ω^B be the set of the n^B units of U^B identified by the units $j \in s^A$, *i.e.*, $\Omega^B = \{i \in U^B \mid \exists\, j \in s^A$ and $\theta_{ji}^{AB} > 0\}$. For each unit i of the set Ω^B, we measure a variable of interest y_i from the target population U^B. Let $\mathbf{Y} = \{y_1, \ldots, y_{N^B}\}'$ be the column vector of that variable of interest. In a practical view point, it is important to mention that although the sample size n^A is usually determined in advance, the number of units n^B is difficult to control because it depends on the selected sample s^A and the link matrix $\mathbf{\Theta}_{AB}$. As a consequence, it turns out to be difficult in general to establish a budget for measuring the variable of interest y_i. Fortunately, in most applications (*e.g.*, the parents-children case above), the number of links that start from a given unit j of s^A is somewhat predictable (for example, a parent typically has one, two, or three children), which helps to assess how many units i of U^B will finally be measured.

We assume that for any unit j of s^A, the correspondences for $i = 1, \ldots, N^B$ can be obtained. That is, we can identify all the links between the two populations by direct interview or by some administrative source for any sampled

unit j. Also, for any identified unit i of U^B, we assume that the links for $j = 1, ..., N^A$ can be obtained (as mentioned by Lavallée (2002), there are cases where this last constraint can be difficult to satisfy in practice. Referring to the example of parents and children, it might not be easy for a very young child, selected through his mother, to mention back his father, when the two parents are divorced. In order to simplify the discussion, such a problem of identification of links will be assumed to be negligible). Therefore, the values of the links need not to be known between the entire populations U^A and U^B. In fact, we need to know the links (and consequently the values of θ_{ji}^{AB}) only for the lines j of $\mathbf{\Theta}_{AB}$ where $j \in s^A$, and also for columns i of $\mathbf{\Theta}_{AB}$ where $i \in \Omega^B$.

Suppose that we are interested in estimating the total Y^B of the target population U^B where $Y^B = \sum_{i=1}^{N^B} y_i$. We can also write $Y^B = \mathbf{1}_B' \mathbf{Y}$ where $\mathbf{1}_B$ is the column vector of 1's of size N^B (note that we use for simplification the notation $\mathbf{1}_B$ instead of $\mathbf{1}_{N^B}$). Now let $\theta_{+i}^{AB} = \sum_{j=1}^{N^A} \theta_{ji}^{AB}$ and let $\tilde{\theta}_{ji}^{AB} = \theta_{ji}^{AB} / \theta_{+i}^{AB}$. We have $\mathbf{1}_A' \mathbf{\Theta}_{AB} = \{\theta_{+1}^{AB}, ..., \theta_{+N^B}^{AB}\}$. We then define the *standardized link matrix* $\tilde{\mathbf{\Theta}}_{AB} = \mathbf{\Theta}_{AB} [\text{diag}(\mathbf{1}_A' \mathbf{\Theta}_{AB})]^{-1}$, where $\text{diag}(\mathbf{v})$ is the square matrix obtained by putting the elements of the row-vector (or column-vector) $\mathbf{v}$ in the diagonal, and 0 elsewhere. Note that in order for the matrix $\tilde{\mathbf{\Theta}}_{AB}$ to be well defined, we must have $[\text{diag}(\mathbf{1}_A' \mathbf{\Theta}_{AB})]^{-1}$ to exist, which is the case if and only if $\theta_{+i}^{AB} > 0$ for all $i = 1, ..., N^B$. For the parents-children example, this means that every child must be linked to at least a parent.

Result 1:

The link matrix $\tilde{\mathbf{\Theta}}_{AB}$ is a standardized link matrix if and only if

$$\tilde{\mathbf{\Theta}}_{AB}' \mathbf{1}_A = \mathbf{1}_B. \tag{2.1}$$

The proof of Result 1 follows directly from the definition of a standadized link matrix. Using Result 1, we directly obtain Result 2 that can also be found in Deville (1998):

Result 2:

$$\begin{aligned} Y^B &= \mathbf{1}_B' \mathbf{Y} \\ &= \mathbf{1}_A' \tilde{\mathbf{\Theta}}_{AB} \mathbf{Y} = \sum_{j=1}^{N^A} \sum_{i=1}^{N^B} \frac{\theta_{ji}^{AB}}{\theta_{+i}^{AB}} y_i. \end{aligned} \tag{2.2}$$

Let us define the column vector $\mathbf{Z} = \tilde{\mathbf{\Theta}}_{AB} \mathbf{Y}$ of size N^A. Considering each line of $\mathbf{Z}$, the variable $z_j = \sum_{i=1}^{N^B} \tilde{\theta}_{ji}^{AB} y_i$ is defined for each unit j of the population U^A and measured for each unit $j \in s^A$.

For estimating Y^B, we want to use the values of y_i measured from set Ω^B. For this, we will use an estimator of the form:

$$\hat{Y}^B = \sum_{i=1}^{N^B} w_i y_i \tag{2.3}$$

where w_i is the estimation weight of the unit i of Ω^B, with $w_i = 0$ for $i \notin \Omega^B$. Let $\mathbf{W}' = \{w_1, ..., w_{N^B}\}$. The estimator (2.3) can be rewritten as

$$\hat{Y}^B = \mathbf{W}'\mathbf{Y}. \tag{2.4}$$

Usually, to get an unbiased estimate of Y^B, one can simply use as the weight the inverse of the selection probability π_i^B of unit i. As mentioned by Lavallée (1995) and Lavallée (2002), with Indirect Sampling, this probability can however be difficult, or even impossible, to obtain. It is then proposed to use the GWSM, which is defined as follows.

Let $\boldsymbol{\pi}^A = \{\pi_1^A, ..., \pi_{N^A}^A\}'$ and let $\mathbf{\Pi}_A = \text{diag}(\boldsymbol{\pi}^A)$ be the diagonal matrix of size $N^A \times N^A$ containing the selection probabilities used for the selection of sample s^A. Accordingly, let $\mathbf{t}^A = \{t_1^A, ..., t_{N^A}^A\}'$ where $t_j^A = 1$ if $j \in s^A$, and 0 otherwise. Let $\mathbf{T}_A = \text{diag}(\mathbf{t}^A)$ be the diagonal matrix of size $N^A \times N^A$ containing the indicator variables t_j^A. Starting from $Y^B = \mathbf{1}_A' \tilde{\mathbf{\Theta}}_{AB} \mathbf{Y} = \mathbf{1}_A' \mathbf{Z}$, we can directly form the following Horvitz-Thompson estimator in terms of the vector $\mathbf{Z}$:

$$\hat{Y}^B = \mathbf{1}_A' \mathbf{T}_A \mathbf{\Pi}_A^{-1} \mathbf{Z}. \tag{2.5}$$

Using the fact that $\mathbf{Z} = \tilde{\mathbf{\Theta}}_{AB} \mathbf{Y}$, we have $\hat{Y}^B = \mathbf{1}_A' \mathbf{T}_A \mathbf{\Pi}_A^{-1} \tilde{\mathbf{\Theta}}_{AB} \mathbf{Y}$ and therefore we can define the column vector $\mathbf{W}$ of weights:

$$\mathbf{W} = \tilde{\mathbf{\Theta}}_{AB}' \mathbf{T}_A \mathbf{\Pi}_A^{-1} \mathbf{1}_A. \tag{2.6}$$

The vector $\mathbf{W}$ is of size N^B and for each $i = 1, ..., N^B$, we have $w_i = \sum_{j=1}^{N^A} t_j^A \tilde{\theta}_{ji}^{AB} / \pi_j^A$. The weights w_i of that vector are said to be obtained from the GWSM, as described by Lavallée (2002).

3. Properties of the GWSM

3.1 Unbiasedness

As mentioned by Ernst (1989), to get an unbiased estimator, we only need to have $E(\mathbf{W}) = \mathbf{1}_B$. By construction, because the estimator (2.5) is a Horvitz-Thompson estimator, this condition is directly satisfied and therefore, the GWSM produces unbiased estimates.

From this discussion, we can in addition obtain the following result:

Result 3:

The vector of weights $\mathbf{W}$ given by (2.6) provides unbiased estimates if and only if the matrix $\tilde{\mathbf{\Theta}}_{AB}$ is a standardized link matrix.

Proof:

Starting from (2.6), we have

$$E(\mathbf{W}) = \tilde{\mathbf{\Theta}}'_{AB}\mathbf{1}_A. \tag{3.1}$$

Using Result 1, we directly get $E(\mathbf{W}) = \mathbf{1}_B$ and therefore we have unbiased estimates. Now, assume that $E(\mathbf{W}) = \mathbf{1}_B$. From (3.1), we must have $\tilde{\mathbf{\Theta}}'_{AB}\mathbf{1}_A = \mathbf{1}_B$ and therefore, $\tilde{\mathbf{\Theta}}_{AB}$ is a standardized link matrix.

3.2 Variance

Because the estimator (2.5) is a Horvitz-Thompson estimator, we directly obtain the following result:

Result 4:

The variance of $\hat{Y}^B$ is given by

$$\begin{aligned}\mathrm{Var}(\hat{Y}^B) &= \mathbf{Z}'\mathbf{\Delta}_A\mathbf{Z} \\ &= \mathbf{Y}'\mathbf{\Delta}_B\mathbf{Y}\end{aligned} \tag{3.2}$$

where $\mathbf{\Delta}_A = [(\pi^A_{jj'} - \pi^A_j\pi^A_{j'})/\pi^A_j\pi^A_{j'}]_{N^A\times N^A}$ is a non-negative definite matrix of size $N^A \times N^A$ and where $\pi^A_{jj'}$ is the joint selection probability of units j and j' from U^A, and where $\Delta_B = \tilde{\mathbf{\Theta}}'_{AB}\mathbf{\Delta}_A\tilde{\mathbf{\Theta}}_{AB}$.

For a proof of the variance of the Horvitz-Thompson estimator, see Särndal, Swensson and Wretman (1992).

3.3 Transitivity

Let us suppose that we are interested in producing estimates for a target population U^C that can only be reached through the population U^B. We assume that the target population U^C contains N^C units, where each unit is labeled by the letter k. The correspondence between the two populations U^B and U^C can be represented by the link matrix $\mathbf{\Theta}_{BC} = [\theta^{BC}_{ik}]$ of size $N^B \times N^C$ where each element $\theta^{BC}_{ik} \geq 0$. That is, unit i of U^B is related to unit k of U^C provided that $\theta^{BC}_{ik} > 0$, otherwise the two units are not related to each other.

We can now use Indirect Sampling by *transitivity*. For this, we select a sample s^A from the population U^A and first identify the set Ω^B of U^B. From this set Ω^B, we then identify the units of U^C that are associated in order to form the set $\Omega^C = \{k \in U^C \mid \exists\, i \in \Omega^B \text{ and } \theta^{BC}_{ik} > 0\}$ of units to be measured from the target population U^C. An important question is to see if the GWSM, when applied in the context of Indirect Sampling by transitivity, is also transitive. That is, is applying the GWSM from U^A to U^B, and then from U^B to U^C, is equivalent to directly applying the GWSM from U^A to U^C?

First, consider using Indirect Sampling from U^A directly to the target population U^C. By going from the population U^A to U^B, and then to U^C, this can relate to having the link matrix $\mathbf{\Theta}_{AC} = [\theta^{AC}_{jk}]$ of size $N^A \times N^C$ defined as $\mathbf{\Theta}_{AC} = \mathbf{\Theta}_{AB}\mathbf{\Theta}_{BC}$. For each unit j selected in s^A, we identify the units k of U^C that have a non-zero correspondence, *i.e.*, with $\theta^{AC}_{jk} > 0$, to obtain the set $\vec{\Omega}^C = \{k \in U^C \mid \exists\, j \in s^A \text{ and } \theta^{AC}_{jk} > 0\}$. We measure the variable of interest y_k from the target population U^C. Applying the GWSM, we obtain from (2.6) the following weights:

$$\vec{\mathbf{W}}_C = \tilde{\mathbf{\Theta}}'_{AC}\mathbf{T}_A\mathbf{\Pi}^{-1}_A\mathbf{1}_A \tag{3.3}$$

where $\tilde{\mathbf{\Theta}}_{AC} = \mathbf{\Theta}_{AC}\,[\mathrm{diag}(\mathbf{1}'_A\mathbf{\Theta}_{AC})]^{-1}$.

Let us now consider using Indirect Sampling in two steps. For each unit j selected in s^A, we identify the units i of U^B that have a non-zero correspondence, *i.e.*, with $\theta^{AB}_{ji} > 0$. As before, we have $\Omega^B = \{i \in U^B \mid \exists\, j \in s^A \text{ and } \theta^{AB}_{ji} > 0\}$. For each unit i of the set Ω^B, we then identify the units k of U^C that have a non-zero correspondence, *i.e.*, with $\theta^{BC}_{ik} > 0$. We then have the set $\Omega^C = \{k \in U^C \mid \exists\, i \in \Omega^B \text{ and } \theta^{BC}_{ik} > 0\}$. From (2.6), we have the column vector $\mathbf{W}_B$ of weights associated to the units of population U^B:

$$\mathbf{W}_B = \tilde{\mathbf{\Theta}}'_{AB}\mathbf{T}_A\mathbf{\Pi}^{-1}_A\mathbf{1}_A. \tag{3.4}$$

For each unit i of the set Ω^B, we then have a non-zero weight w^B_i. Now, the set Ω^B can be seen as a sample of units that are used in an Indirect Sampling process to identify the set Ω^C. By similarity with Indirect Sampling from the sample s^A to the target population U^B, applying the GWSM in the context of Indirect Sampling from the set Ω^B to the target population U^C produces the following weights:

$$\mathbf{W}_C = \tilde{\mathbf{\Theta}}'_{BC}\,\mathbf{T}_B\,\mathrm{diag}(\mathbf{W}_B)\mathbf{1}_B \tag{3.5}$$

where $\tilde{\mathbf{\Theta}}_{BC} = \mathbf{\Theta}_{BC}[\,\mathrm{diag}(\mathbf{1}'_B\mathbf{\Theta}_{BC})]^{-1}$ and $\mathbf{T}_B = \mathrm{diag}(\mathbf{t}_B)$ with $\mathbf{t}_B = (t^B_1, \ldots, t^B_{N^B})'$ and $t^B_i = 1$ if $i \in \Omega^B$, and 0 otherwise. Because the weights $w^B_i = 0$ for $i \notin \Omega^B$, we have $\mathbf{T}_B\,\mathrm{diag}(\mathbf{W}_B) = \mathrm{diag}(\mathbf{W}_B)$. Therefore, we obtain

$$\mathbf{W}_C = \tilde{\mathbf{\Theta}}'_{BC}\,\mathrm{diag}(\mathbf{W}_B)\mathbf{1}_B. \tag{3.6}$$

Replacing $\mathbf{W}_B$ by (3.4) in equation (3.6), we get

$$\begin{aligned}\mathbf{W}_C &= \tilde{\mathbf{\Theta}}'_{BC}\,\mathrm{diag}(\tilde{\mathbf{\Theta}}'_{AB}\,\mathbf{T}_A\,\mathbf{\Pi}^{-1}_A\mathbf{1}_A)\mathbf{1}_B \\ &= \tilde{\mathbf{\Theta}}'_{BC}\,\tilde{\mathbf{\Theta}}'_{AB}\,\mathbf{T}_A\,\mathbf{\Pi}^{-1}_A\mathbf{1}_A.\end{aligned} \tag{3.7}$$

Since $\tilde{\mathbf{\Theta}}'_{BC}\tilde{\mathbf{\Theta}}'_{AB}\mathbf{1}_A = \tilde{\mathbf{\Theta}}'_{BC}\mathbf{1}_B = \mathbf{1}_C$, from Result 1, the matrix $\tilde{\mathbf{\Theta}}_{AB}\tilde{\mathbf{\Theta}}_{BC}$ is a standardized link matrix. Because of this, the GWSM is therefore transitive, at least in some sense. That is, the weights $\mathbf{W}_C$ can be obtained in a single step by using the standardized link matrix $\tilde{\mathbf{\Theta}}_{AB}\tilde{\mathbf{\Theta}}_{BC}$ into the GWSM. Now, for the GWSM to be perfectly transitive, the weights $\mathbf{W}_C$ provided (3.7) would need to be exactly the same as the weights $\vec{\mathbf{W}}_C$ provided by (3.3). By comparing equations (3.3) and (3.7), we obtain the following result:

Result 5:

Applying the GWSM from U^A to U^B, and then from U^B to U^C, is transitive if and only if

$$\tilde{\boldsymbol{\Theta}}_{AC} = \tilde{\boldsymbol{\Theta}}_{AB}\tilde{\boldsymbol{\Theta}}_{BC}. \qquad (3.8)$$

Unfortunately, condition (3.8) does not hold in general. In fact, it is relatively easy to construct examples where $\tilde{\boldsymbol{\Theta}}_{AC} \neq \tilde{\boldsymbol{\Theta}}_{AB}\tilde{\boldsymbol{\Theta}}_{BC}$.

4. A Structural Property of the GWSM

In the present section, we stress the fact that with Indirect Sampling, the sampling process depends only on the links between the two populations U^A and U^B. The values of the θ_{ji}^{AB} themselves, apart from being zero or not, do not interfere in the sampling process. On the other hand, the values of the θ_{ji}^{AB} **do** have a role in the weights, and therefore the estimator, issued from the GWSM. We extend this idea in the following paragraphs.

Indirect Sampling associates to each sample s^A in U^A a sample Ω^B in U^B, namely $\Omega^B = \{i \in U^B \mid \exists\, j \in s^A$ and $\theta_{ji}^{AB} > 0\}$. Thus, a function $f : s^A \to \Omega^B$ that maps the sample s^A to the sample Ω^B is uniquely determined by the set of couples (j, i) with $\theta_{ji}^{AB} > 0$. Let $l_{ji}^{AB} = 1$ if $\theta_{ji}^{AB} > 0$, and 0 otherwise. These are the elements of the incidence matrix of the graph linking U^A to U^B.

Suppose we are given a function ϕ from the set of subsets of U^A into the set of subsets of U^B. Like f, suppose that ϕ satisfies the "Union Property": $\phi(s_1^A \cup s_2^A) = \phi(s_1^A) \cup \phi(s_2^A)$, where s_1^A and s_2^A are two subsets of U^A.

Result 6:

The function ϕ is determined unequivocally by *a zero-one link matrix*.

Proof:

This can be shown as follows: Take $s_j^A = \{j\}$ for some unit j in U^A. Then, $\phi(s_j^A)$ is a set in U^B. Let $l_{ji}^{AB} = 1$ if unit i of U^B belongs to $\phi(s_j^A)$, and 0 otherwise. By the Union Property, $\phi(s^A) = \bigcup_{j \in s^A} \phi(s_j^A)$ and the set of l_{ji}^{AB} defines the *zero-one link matrix* $\mathbf{L}_{AB} = [l_{ji}^{AB}]$ of size $N^A \times N^B$, which precisely defines the function ϕ.

This provides us an equivalence relation between link matrices, associated with a deeper property. Let p^A be a sampling design on U^A (*i.e.*, a probability distribution on the set of subsets of U^A). The function f induces a sampling design on U^B by $p^B(\Omega^B) = \sum_{s^A : \Omega^B = f(s^A)} p^A(s^A)$. As the design is induced by f, it does not depend on the particular link matrix $\boldsymbol{\Theta}_{AB}$ defining the function, but is rather a characteristic of the equivalence class through the zero-one link matrix $\mathbf{L}_{AB}$. As a consequence, the Horvitz-Thompson estimator in U^B depends only on this class. It is therefore of some interest to choose in this class a matrix $\boldsymbol{\Theta}_{AB}$ having, in some sense, an optimal characteristic (see section 6).

5. Special Link matrices

As it can be seen from the previous sections, the link matrix $\boldsymbol{\Theta}_{AB}$ drives the form of the estimator (2.4) obtained from the GWSM. In this section, we present some special link matrices $\boldsymbol{\Theta}_{AB}$ that correspond to extreme cases. Although not all such cases are likely to be seen it practice, they illustrates the effect of the link matrix on the estimator (2.4).

5.1 Identity Matrix

Assume that the link matrix $\boldsymbol{\Theta}_{AB}$ is given by the identity matrix $\mathbf{I}$. In practice, this means that the population U^A and the target population U^B have a one-to-one relationship. Of course, this implies that $N^A = N^B = N$ and that the identity matrix $\mathbf{I}$ is of size $N \times N$.

As a first result, we have $\tilde{\boldsymbol{\Theta}}_{AB} = \mathbf{I}$. As a consequence, the vector of weights (2.6) is given by $\mathbf{W}' = (t_1^A / \pi_1^A, ..., t_{N^A}^A / \pi_{N^A}^A)$ and we also have $\mathbf{Z} = \tilde{\boldsymbol{\Theta}}_{AB}\mathbf{Y} = \mathbf{Y}$. Therefore, the estimator $\hat{Y}^B$ given by (2.5) turns out to be nothing else than the Horvitz-Thompson estimator $\hat{Y}^B = \mathbf{1}'_A \mathbf{T}_A \boldsymbol{\Pi}_A^{-1} \mathbf{Y}$.

5.2 One for All (Within Clusters)

Consider the case where the population U^B is divided into Γ clusters where each cluster γ is of size N_γ^B. Theses clusters are such that each cluster γ from U^B is associated to exactly one unit j of U^A. Because of this, we can use the letter γ for both the units j from U^A and the clusters from U^B. Note also that $\Gamma = N^A$.

This situation corresponds to a link matrix $\boldsymbol{\Theta}_{AB}$ being block diagonal where each submatrix contains only one line. Let the row vector $\mathbf{1}'_{B\gamma}$ be of size N_γ^B and containing only 1's. The link matrix $\boldsymbol{\Theta}_{AB}$ is then defined as

$$\boldsymbol{\Theta}_{AB} = \begin{bmatrix} \mathbf{1}'_{B1} & \mathbf{0} & \cdots & \mathbf{0} \\ & \ddots & & \ddots \\ \mathbf{0} & & \mathbf{1}'_{B\gamma} & \mathbf{0} \\ \vdots & \ddots & & \ddots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{1}'_{B\Gamma} \end{bmatrix}. \qquad (5.1)$$

We can also write $\boldsymbol{\Theta}_{AB} = \text{diag}(\{\mathbf{1}'_{B1}, ..., \mathbf{1}'_{B\Gamma}\})$. Using this, we have $\text{diag}(\mathbf{1}'_A \boldsymbol{\Theta}_{AB}) = \text{diag}(\mathbf{1}'_A \text{diag}(\{\mathbf{1}'_{B1}, ..., \mathbf{1}'_{B\Gamma}\})) = \text{diag}(\{\mathbf{1}'_{B1}, ..., \mathbf{1}'_{B\Gamma}\})$ and hence $\tilde{\boldsymbol{\Theta}}_{AB} = \boldsymbol{\Theta}_{AB}$. From equation (2.6), we obtain the column vector of weights $\mathbf{W}' = (t_1^A / \pi_1^A \mathbf{1}'_{B1}, ..., t_\Gamma^A / \pi_\Gamma^A \mathbf{1}'_{B\Gamma})$. As we can see, the elements of the column vector $\mathbf{W}$ have the values $t_\gamma^A / \pi_\gamma^A$ repeated within each cluster γ of U^B. From (2.4), we obtain

$$\hat{Y}^B = \sum_{\gamma=1}^{\Gamma} \frac{t_\gamma^A}{\pi_\gamma^A} Y_\gamma^B \tag{5.2}$$

where $Y_\gamma^B = \sum_{i=1}^{N_\gamma^B} y_i$.

5.3 All for One (Within Clusters)

Consider the case where the population U^A is divided into Γ clusters where each cluster γ is of size N_γ^A. Theses clusters are such that each cluster γ from U^A is associated to exactly one unit i of U^B. Because of this, we can use the letter γ for both the clusters from U^A and the units i from U^B. Note also that $\Gamma = N^B$.

This situation corresponds to a link matrix $\mathbf{\Theta}_{AB}$ being block diagonal where each submatrix contains only one column. Let the column vector $\mathbf{1}_{A\gamma}$ be of size N_γ^A and containing only 1's. The link matrix $\mathbf{\Theta}_{AB}$ is then defined as

$$\mathbf{\Theta}_{AB} = \begin{bmatrix} \mathbf{1}_{A1} & \mathbf{0} & \cdots & \mathbf{0} \\ & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{1}_{A\gamma} & & \mathbf{0} \\ \vdots & \ddots & \ddots & \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{1}_{A\Gamma} \end{bmatrix}. \tag{5.3}$$

We can also write $\mathbf{\Theta}_{AB} = \text{diag}(\{\mathbf{1}_{A1}, ..., \mathbf{1}_{A\Gamma}\})$. Using this, we have $\tilde{\mathbf{\Theta}}_{AB} = \text{diag}(\{1/N_1^A \mathbf{1}_{A1}, ..., 1/N_\Gamma^A \mathbf{1}_{A\Gamma}\})$. From equation (2.6), we obtain the column vector of weights $\mathbf{W}' = (1/N_1^A \sum_{j=1}^{N_1^A} t_j^A/\pi_j^A, ..., 1/N_\Gamma^A \sum_{j=1}^{N_\Gamma^A} t_j^A/\pi_j^A)$. Thus, the elements γ (or i) of the column vector $\mathbf{W}$ have the averaged values $\sum_{j=1}^{N_\gamma^A} t_j^A/\pi_j^A N_{\gamma,\gamma}^A \gamma = 1, ..., \Gamma$. From (2.4), we obtain $\hat{Y}^B = \sum_{\gamma=1}^{\Gamma} y_\gamma / N_\gamma^A \sum_{j=1}^{N_\gamma^A} t_j^A/\pi_j^A$.

5.4 Inefficient Sampling

Suppose that some rows of the link matrix $\mathbf{\Theta}_{AB}$ contain only zeros. This means that some units of the population U^A are not associated to any unit of the target population U^B. Then, if such units are selected in the sample s^A, this will lead to the identification of no unit from U^B. This can be seen as inefficient in a sampling point of view. In a more formal way, assume that each of the first N^{1A} rows of the link matrix $\mathbf{\Theta}_{AB}$ contains at least one $\theta_{ji} > 0$, and that they form the submatrix $\mathbf{\Theta}_1$. Assume that the other N^{0A} rows of $\mathbf{\Theta}_{AB}$ have $\theta_{ji} = 0$ for $i = 1, ..., N^B$. We therefore have

$$\mathbf{\Theta}_{AB} = \begin{bmatrix} \mathbf{\Theta}_1 \\ \mathbf{0} \end{bmatrix}.$$

As a first result, we obtain

$$\tilde{\mathbf{\Theta}}_{AB} = \begin{bmatrix} \mathbf{\Theta}_1 [\text{diag}(\mathbf{1}_{1A}' \mathbf{\Theta}_1)]^{-1} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{\Theta}}_1 \\ \mathbf{0} \end{bmatrix} \tag{5.4}$$

where $\mathbf{1}_{1A}$ is the column vector of 1's of size N^{1A}. From equation (2.6), we obtain the column vector of weights $\mathbf{W} = [\tilde{\mathbf{\Theta}}_1' \ \mathbf{0}'] \mathbf{T}_A \mathbf{\Pi}_A^{-1} \mathbf{1}_A$. Let $\mathbf{\Pi}_{1A} = \text{diag}(\{\pi_1^A, ..., \pi_{N^{1A}}^A\})$ be the diagonal matrix of size $N^{1A} \times N^{1A}$ and accordingly, let $\mathbf{T}_{1A} = \text{diag}(\{t_1^A, ..., t_{N^{1A}}^A\})$ be the diagonal matrix of size $N^{1A} \times N^{1A}$. We then get

$$\begin{aligned} \mathbf{W} &= [\tilde{\mathbf{\Theta}}_1' \ \mathbf{0}'] \mathbf{T}_A \mathbf{\Pi}_A^{-1} \mathbf{1}_A \\ &= \tilde{\mathbf{\Theta}}_1' \mathbf{T}_{1A} \mathbf{\Pi}_{1A}^{-1} \mathbf{1}_{1A}. \end{aligned} \tag{5.5}$$

As we can see from (5.5), the weights only depend on the probabilities of selection π_j^A of the units of U^A that have at least one $\theta_{ji} > 0$ for $i = 1, ..., N^B$. From (2.4), we finally obtain $\hat{Y}^B = \mathbf{1}_{1A}' \mathbf{T}_{1A} \mathbf{\Pi}_{1A}^{-1} \tilde{\mathbf{\Theta}}_1 \mathbf{Y}$.

5.5 Biased Estimator

Suppose that some columns of the link matrix $\mathbf{\Theta}_{AB}$ contain only zeros. This means that some units of the population U^B are not associated to any unit of the target population U^A. Recall that in order for the matrix $\tilde{\mathbf{\Theta}}_{AB}$ to be well defined, we must have $\text{diag}(\mathbf{1}_A' \mathbf{\Theta}_{AB})^{-1}$ to exist. As we will see, the present case does not satisfy this condition. This results in a biased estimator for the total Y^B.

In a more formal way, assume that each of the first N^{1B} columns of the link matrix $\mathbf{\Theta}_{AB}$ contains at least one $\theta_{ji} > 0$, and let them form the submatrix $\mathbf{\Theta}_1$, different from the one of the previous section. Assume that the other N^{0B} columns of $\mathbf{\Theta}_{AB}$ have $\theta_{ji} = 0$ for $j = 1, ..., N^A$. We therefore have $\mathbf{\Theta}_{AB} = [\mathbf{\Theta}_1, \mathbf{0}]$.

From this definition, we directly have

$$\begin{aligned} [\text{diag}(\mathbf{1}_A' \mathbf{\Theta}_{AB})]^{-1} &= [\text{diag}([\mathbf{1}_A' \mathbf{\Theta}_1, \mathbf{1}_A' \mathbf{0}])]^{-1} \\ &= \begin{bmatrix} \text{diag}(\mathbf{1}_A' \mathbf{\Theta}_1) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}^{-1}. \end{aligned} \tag{5.6}$$

Since this matrix is singular, $[\text{diag}(\mathbf{1}_A' \mathbf{\Theta}_{AB})]^{-1}$ does not exist. As a solution to this problem, it could be possible to use a *generalized inverse*. Recall that for a given square matrix $\mathbf{A}$, the matrix $\mathbf{A}^-$ is a generalized inverse of $\mathbf{A}$ provided that $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$ (Searle 1971). One possible generalized inverse of (5.6) is

$$[\text{diag}(\mathbf{1}_A' \mathbf{\Theta}_{AB})]^- = \begin{bmatrix} [\text{diag}(\mathbf{1}_A' \mathbf{\Theta}_1)]^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{5.7}$$

With this generalized inverse, we have the following standardized link matrix $\tilde{\mathbf{\Theta}}_- = \mathbf{\Theta}_{AB} [\text{diag}(\mathbf{1}_A' \mathbf{\Theta}_{AB})]^- = [\tilde{\mathbf{\Theta}}_1, \mathbf{0}]$. Starting from equation (2.6), we can obtain the column vector $\mathbf{W}_-$ of weights:

$$\mathbf{W}_- = \begin{bmatrix} \tilde{\mathbf{\Theta}}_1' \mathbf{T}_A \mathbf{\Pi}_A^{-1} \mathbf{1}_A \\ \mathbf{0}' \end{bmatrix}. \tag{5.8}$$

As we can see from (5.8), the weights are null for the units i of the target population U^B that $\mathbf{\Theta}_{AB}$ have $\theta_{ji} = 0$ for $j = 1, ..., N^A$. From (2.4) and using $\mathbf{W}_-$ instead of $\mathbf{W}$, we obtain $\hat{Y}_-^B = \mathbf{1}_A' \mathbf{T}_A \mathbf{\Pi}_A^{-1} \tilde{\mathbf{\Theta}}_1 \mathbf{Y}_1$ where $\mathbf{Y}_1 = \{y_1, ..., y_{N^{1B}}\}'$ is the subvector constructed from the N^{1B} first elements of $\mathbf{Y}$. Since in general $E(\hat{Y}_-^B) = \mathbf{1}_A' \tilde{\mathbf{\Theta}}_1 \mathbf{Y}_1 \neq \mathbf{1}_A' \mathbf{Y} = Y^B$, this estimator is biased for the total Y^B.

6. Optimality

Optimality is an important aspect of the GWSM. As it has been shown in Result 3, the estimator $\hat{Y}^B$ obtained by the GWSM will provide unbiased estimates provided that the matrix $\tilde{\mathbf{\Theta}}_{AB}$ is a standardized link matrix. Now, given that the variance (3.2) of this estimator depends on this matrix, there should be at least one matrix $\tilde{\mathbf{\Theta}}_{AB,\text{opt}}$ such that the variance of the estimator $\hat{Y}^B$ will be minimum. That is, for the θ_{ji}^{AB} that are greater than 0, we are interested in finding the values that these θ_{ji}^{AB} should have to obtain the most precise estimator $\hat{Y}^B$.

This optimality problem was first assessed by Kalton and Brick (1995). They obtained results based on the simplified situation where $N^A = 2$ and with s^A obtained through equal probability sampling. Their conclusions suggested the use of $\theta_{ji}^{AB,\text{opt}} = 1$ when $\theta_{ji}^{AB} > 0$, and $\theta_{ji}^{AB,\text{opt}} = 0$ when $\theta_{ji}^{AB} = 0$. Lavallée (2002) and Lavallée and Caron (2001) obtained results along the same lines by the use of simulations. In the present section, we present new results on the optimality of the GWSM.

6.1 Factorization

Factorization is the reverse problem of transitivity. It consists in finding a population U^G and standardized link matrices $\tilde{\mathbf{\Theta}}_{AG}$ and $\tilde{\mathbf{\Theta}}_{GB}$ such that $\tilde{\mathbf{\Theta}}_{AB} = \tilde{\mathbf{\Theta}}_{AG} \tilde{\mathbf{\Theta}}_{GB}$. This leads to an important simplification in searching for an optimal standardized link matrix $\tilde{\mathbf{\Theta}}_{AB,\text{opt}}$.

The population U^G can be taken as being one of clusters, the factorization being achieved in the context of "one for all (within clusters)" (from U^A to U^G) and "all for one (within clusters)" (from U^G to U^B), as presented in sections 5.2 and 5.3. This can be described in a very general way as follows. Consider a population U^G containing as many units as there are links starting from the units j of U^A. The population size N^G is then given by the number of θ_{ji}^{AB} of $\mathbf{\Theta}_{AB}$ that are greater than 0. Each unit g of U^G can be seen as the extremity of an "arrow" starting from some unit j of U^A. From this graph, there is only one link matrix $\mathbf{\Theta}_{AG}$ of size $N^A \times N^G$ keeping unbiasedness, namely $\mathbf{\Theta}_{AG} = [\theta_{jg}^{AG}]$ where $\theta_{jg}^{AG} = 1$ if there is a link (or an "arrow") leaving unit j of U^A to unit g from U^G, and $\theta_{jg}^{AG} = 0$ otherwise. Note that by construction, each unit g from U^G is linked to at most one unit j from U^A and therefore $\tilde{\mathbf{\Theta}}_{AG} = \mathbf{\Theta}_{AG}$. This corresponds to the "one to all within clusters" situation presented in section 5.2. Indirect Sampling from U^A to U^G is in fact standard Cluster Sampling and leading the GWSM to the usual Horvitz-Thompson estimator (see Lavallée 2002). For the parent-children example, the result of this factorization would be given by Figure 2.

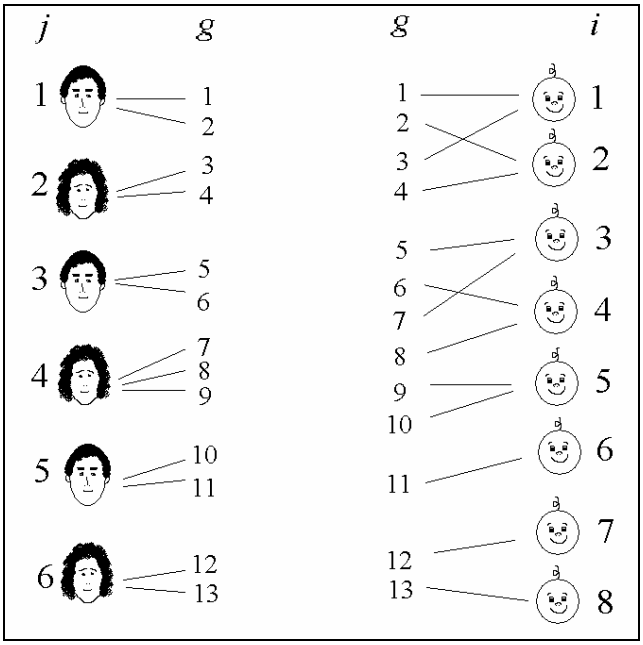


Figure 2. Result of the factorization of the parents-children populations.

Considering the graph from U^G to U^B, we can construct the link matrix $\mathbf{\Theta}_{GB}$ of size $N^G \times N^B$ as follows. Because of the definition of the population U^G, each unit g of U^G is linked to exactly one unit i of U^B. Note that Indirect Sampling in this context can be seen as sampling clusters (*i.e.*, the units i of U^B) from their elements (*i.e.*, the units g of U^G). It can also be seen as the "all to one within clusters" presented in section 5.3. Let $\tilde{\mathbf{\Theta}}_{GB} = \mathbf{\Theta}_{GB}[\text{diag}(\mathbf{1}_G' \mathbf{\Theta}_{GB})]^{-1}$ be the standardized link matrix obtained from $\mathbf{\Theta}_{GB}$. We have $\text{diag}(\mathbf{1}_G' \mathbf{\Theta}_{GB}) = \text{diag}(\mathbf{1}_A' \mathbf{\Theta}_{AB})$, and therefore $\tilde{\mathbf{\Theta}}_{GB} = \mathbf{\Theta}_{GB}[\text{diag}(\mathbf{1}_A' \mathbf{\Theta}_{AB})]^{-1}$.

Now,

$$\begin{aligned} \tilde{\mathbf{\Theta}}_{AG} \tilde{\mathbf{\Theta}}_{GB} &= \mathbf{\Theta}_{AG} \tilde{\mathbf{\Theta}}_{GB} \\ &= \mathbf{\Theta}_{AG} \mathbf{\Theta}_{GB} [\text{diag}(\mathbf{1}_A' \mathbf{\Theta}_{AB})]^{-1} \\ &= \mathbf{\Theta}_{AB} [\text{diag}(\mathbf{1}_A' \mathbf{\Theta}_{AB})]^{-1} \\ &= \tilde{\mathbf{\Theta}}_{AB}. \end{aligned} \tag{6.1}$$

Therefore, using this construction, the standardized link matrix $\tilde{\mathbf{\Theta}}_{AB}$ from U^A to U^B can always be factorized into the two matrices $\tilde{\mathbf{\Theta}}_{AG}$ and $\tilde{\mathbf{\Theta}}_{GB}$.

6.2 Strong Optimality: Statement of the Problem

As mentioned before, the optimality problem that we consider here is to minimize the variance (3.2) with respect to the standardized link matrix $\tilde{\mathbf{\Theta}}_{AB}$. Now, using the factorization presented in section 6.1, we have

$$\begin{aligned}\mathrm{Var}(\hat{Y}^B) &= \mathbf{Y}'\tilde{\mathbf{\Theta}}'_{AB}\mathbf{\Delta}_A\tilde{\mathbf{\Theta}}_{AB}\mathbf{Y} \\ &= \mathbf{Y}'\tilde{\mathbf{\Theta}}'_{GB}\tilde{\mathbf{\Theta}}'_{AG}\mathbf{\Delta}_A\tilde{\mathbf{\Theta}}_{AG}\tilde{\mathbf{\Theta}}_{GB}\mathbf{Y} \\ &= \mathbf{Y}'\tilde{\mathbf{\Theta}}'_{GB}\mathbf{\Delta}_G\tilde{\mathbf{\Theta}}_{GB}\mathbf{Y}\end{aligned} \qquad (6.2)$$

where $\mathbf{\Delta}_G = \tilde{\mathbf{\Theta}}'_{AG}\mathbf{\Delta}_A\tilde{\mathbf{\Theta}}_{AG}$.

For any standardized link matrix $\tilde{\mathbf{\Theta}}_{AB}$, the factorization presented in section 6.1 always produces the same first factor $\tilde{\mathbf{\Theta}}_{AG}$. Therefore, if we seek for some optimal matrix $\tilde{\mathbf{\Theta}}_{AB,\mathrm{opt}}$ that minimizes the variance (3.2), it is sufficient to optimize the second factor $\tilde{\mathbf{\Theta}}_{GB}$. We would also like the optimal matrix $\tilde{\mathbf{\Theta}}_{AB,\mathrm{opt}}$ to produce unbiased estimates.

Let U_i^G be the subpopulation of U^G containing the N_i^G links to the unit i of U^B. Note that the subpopulations U_i^G are disjoint. Thus, without loss of generality, we can order the links from U^A to U^B so that, for every i, the links to unit i in U^B are indexed consecutively. Now, let $\tilde{\tilde{\boldsymbol{\theta}}}_{GB,i}$ be the i^{th} column vector of the matrix $\tilde{\mathbf{\Theta}}_{GB}$, $i = 1, \ldots, N^B$. By construction, the vector $\tilde{\tilde{\boldsymbol{\theta}}}_{GB,i}$ contains non null elements only for the N_i^G links to the unit i of U^B. Hence, letting $\tilde{\boldsymbol{\theta}}_{GB,i}$ be a column vector of size N_i^G containing the non null elements of $\tilde{\tilde{\boldsymbol{\theta}}}_{GB,i}$, we have

$$\tilde{\tilde{\boldsymbol{\theta}}}_{GB,i} = \begin{bmatrix} \mathbf{0} \\ \tilde{\boldsymbol{\theta}}_{GB,i} \\ \mathbf{0} \end{bmatrix}.$$

Similarly, let $\dot{\mathbf{1}}_{G,i}$ be the column vector of size N^G containing 1's for N_i^G elements, and 0's elsewhere. Letting $\mathbf{1}_{G,i}$ be a column vector of size N_i^G containing 1's, we have

$$\dot{\mathbf{1}}_{G,i} = \begin{bmatrix} \mathbf{0} \\ \mathbf{1}_{G,i} \\ \mathbf{0} \end{bmatrix}.$$

Now, for the GWSM from U^G to U^B to be unbiased, we need to have $\tilde{\boldsymbol{\theta}}'_{GB,i}\mathbf{1}_{G,i} = 1$ for all i, or equivalently $\tilde{\tilde{\boldsymbol{\theta}}}'_{GB,i}\dot{\mathbf{1}}_{G,i} = 1$. All this together leads to the following optimization problem:

Find a matrix $\tilde{\mathbf{\Theta}}_{GB,\mathrm{opt}} = \{\tilde{\tilde{\boldsymbol{\theta}}}_{GB,\mathrm{opt},1}, \ldots, \tilde{\tilde{\boldsymbol{\theta}}}_{GB,\mathrm{opt},N^B}\}$ satisfying $\tilde{\tilde{\boldsymbol{\theta}}}'_{GB,\mathrm{opt},i}\dot{\mathbf{1}}_{G,i} = 1$ for all $i = 1, \ldots, N^B$, and minimizing the quadratic form $\mathrm{Var}(\hat{Y}^B) = \mathbf{Y}'\tilde{\mathbf{\Theta}}'_{GB}\mathbf{\Delta}_G\tilde{\mathbf{\Theta}}_{GB}\mathbf{Y}$.

This problem turns out to be nothing else than the minimization of a positive quadratic form under linear constraints. This is a relatively standard and simple problem to solve. It is well known that a solution always exists and is unique if the form (6.2) is positive definite, or if the null subspace of $\tilde{\mathbf{\Theta}}_{GB}$ is not included in the null-space of $\mathbf{\Delta}_G$.

The above optimization problem can be rewritten in a different form. Let $\mathbf{\Delta}_{G,ii'}$ be the submatrix of $\mathbf{\Delta}_G$ corresponding to the elements in positions g and g' if g has a link with unit i and g' has a link with unit i'. These matrices constitute a partition of $\mathbf{\Delta}_G$. Note that the matrices $\mathbf{\Delta}_{G,ii}$ are symmetric, positive definite, and $\mathbf{\Delta}'_{G,ii'} = \mathbf{\Delta}_{G,i'i}$. With these notations, the optimization problem can be written as:

Minimize

$$\sum_{i=1}^{N^B}\sum_{i'=1}^{N^B} y_i y_{i'} \tilde{\boldsymbol{\theta}}'_{GB,i}\mathbf{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,i'} \qquad (6.3)$$

under the constraints $\tilde{\boldsymbol{\theta}}'_{GB,i}\mathbf{1}_{G,i} = 1$ for all $i = 1, \ldots, N^B$.

Minimization is achieved for vectors $\tilde{\boldsymbol{\theta}}_{GB,\mathrm{opt},i}$ satisfying

$$y_i \sum_{i'=1}^{N^B} \mathbf{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\mathrm{opt},i'} y_{i'} = \lambda_i \mathbf{1}_{G,i} \qquad (6.4)$$

for all $i = 1, \ldots, N^B$ and where λ_i are the Lagrange multipliers entering into the constrained minimization of (6.3). As we can see from (6.4), the optimal choice $\tilde{\boldsymbol{\theta}}_{GB,\mathrm{opt},i}$ (and therefore $\tilde{\mathbf{\Theta}}_{GB,\mathrm{opt}}$) will depend in general explicitly on the vector $\mathbf{Y}$, which is not useful in practice. Observe that the set of λ_i depends also of the variable $\mathbf{Y}$. This will appear more explicitly in section 6.3. This is the reason why we will seek, instead of a strong optimization, for a weaker form of optimality that will lead to the existence of an "optimal" solution $\tilde{\mathbf{\Theta}}_{GB,\mathrm{opt}}$ (and $\tilde{\mathbf{\Theta}}_{AB,\mathrm{opt}}$) not depending on $\mathbf{Y}$.

6.3 Weak Optimality

Equations (6.4) must be valid for any vector $\mathbf{Y}$. In particular, a necessary condition is to hold for a particular variable of interest, such as $y_i = 1$ for a unit i of U^B and $y_{i'} = 0$ for all other units i' of U^B ($i' \neq i$). This leads to the necessary conditions (one for each of those particular variables) $\mathbf{\Delta}_{G,ii}\tilde{\boldsymbol{\theta}}_{GB,\mathrm{opt},i} = \lambda_i \mathbf{1}_{G,i}$. Assuming that $\mathbf{\Delta}_{G,ii}$ is invertible, we then have $\tilde{\boldsymbol{\theta}}_{GB,\mathrm{opt},i} = \lambda_i \mathbf{\Delta}_{G,ii}^{-1}\mathbf{1}_{G,i}$. It can be shown that this is also a sufficient condition. Now, because $\tilde{\boldsymbol{\theta}}'_{GB,\mathrm{opt},i}\mathbf{1}_{G,i} = 1$, we have $\lambda_i = 1/\mathbf{1}'_{G,i}\mathbf{\Delta}_{G,ii}^{-1}\mathbf{1}_{G,i}$. Therefore, a necessary and sufficient condition for equation (6.4) to be satisfied is when

$$\tilde{\boldsymbol{\theta}}_{GB,\mathrm{opt},i} = \frac{\mathbf{\Delta}_{G,ii}^{-1}\mathbf{1}_{G,i}}{\mathbf{1}'_{G,i}\mathbf{\Delta}_{G,ii}^{-1}\mathbf{1}_{G,i}}. \qquad (6.5)$$

This result corresponds to weak optimization in the following sense. The weight w_i given by (2.6) satisfies $E(w_i) = 1$ and moreover $E(w_i \mid i \in \Omega^B) = 1/\pi_i^B$ where π_i^B is the inclusion probability of unit i in Ω^B, which is generally difficult or even impossible to compute in practice. Now, note that the Horvitz-Thompson estimator is characterized by $\mathrm{Var}(w_i \mid i \in \Omega^B) = 0$. The weak optimization

obtained here consists in minimizing $\mathrm{Var}(w_i \mid i \in \Omega^B)$ over all possible standardized link matrices $\tilde{\boldsymbol{\Theta}}_{GB}$, or equivalently $\tilde{\boldsymbol{\Theta}}_{AB}$. This variance is strictly positive for the cases where unit i of U^B is in position to receive more than a unique weight for different sample s^A. Moreover, using (6.3), the multiplier λ_i appears to be the variance of the weight w_i and is, therefore, always strictly positive (except, a case that we exclude, when unit i is selected with a weight equal to one).

6.4 Strong Optimality Independent of Y

Weak optimality is a necessary condition for strong optimality independent of the vector $\mathbf{Y}$ of a variable of interest. It provides the necessary form of the vectors $\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i}$ in (6.4). To get sufficient conditions for strong optimality independent of $\mathbf{Y}$, we go back to the equations (6.4). These equations need to be satisfied for all vectors $\mathbf{Y}$ and they must therefore be satisfied for a particular variable of interest such as $y_i = 1$ for a unit i of U^B, $y_{i'} = 1$ for another unit i' of U^B, and $y_{i''} = 0$ for all other units i'' of U^B $(i'' \neq i' \neq i)$. In that case, to satisfy equations (6.4), it is necessary to have the following relations for any i and i':

$$\boldsymbol{\Delta}_{G,ii}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i} + \boldsymbol{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'} = \lambda_i^{ii'}\mathbf{1}_{G,i} \qquad (6.6)$$

$$\boldsymbol{\Delta}_{G,i'i'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'} + \boldsymbol{\Delta}_{G,i'i}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i} = \lambda_{i'}^{ii'}\mathbf{1}_{G,i'}.$$

As we must necessarily have weak optimality, we have $\boldsymbol{\Delta}_{G,ii}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i} = \lambda_i \mathbf{1}_{G,i}$. Considering the first line of (6.6), we then get

$$\begin{aligned}\boldsymbol{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'} &= (\lambda_i^{ii'} - \lambda_i)\mathbf{1}_{G,i} \\ &= \Phi_{ii'}\mathbf{1}_{G,i}. \qquad (6.7)\end{aligned}$$

Multiplying both sides of (6.7) by $\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i}$, we obtain

$$\begin{aligned}\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i}\boldsymbol{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'} &= \Phi_{ii'}\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i}\mathbf{1}_{G,i} \\ &= \Phi_{ii'}\end{aligned}$$

since $\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i}\mathbf{1}_{G,i} = 1$. Let $\boldsymbol{\Phi}$ be the matrix with elements $\Phi_{ii'}$ off the diagonal and $\Phi_{ii} = \lambda_i$ on the diagonal. Using again (6.2), it can be shown that the optimal variance (whenever it exists) has the expression $\mathbf{Y}'\boldsymbol{\Phi}\mathbf{Y}$.

Let us show that this set of conditions is also sufficient. Assume that (6.7) holds. Note that for $i = i'$, condition (6.7) is nothing else than (6.5) which gives the necessary values for the $\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i}$. It is now straightforward to verify that (6.4) holds whatever the value of $\mathbf{Y}$ and that we have obtained the strong optimality. Now, the values of λ_i depend on $\mathbf{Y}$, as well as the variance $\mathrm{Var}(\hat{Y}^B)$, but we have that equations (6.4) always have the same solution (6.5) that does not depend on $\mathbf{Y}$. We therefore have the following result:

Result 7:

The conditions $\boldsymbol{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'} = \boldsymbol{\Phi}_{ii'}\mathbf{1}_{G,i}$ are necessary and sufficient for the existence of a standardized link matrix $\tilde{\boldsymbol{\Theta}}_{GB,\,\mathrm{opt}}$, or equivalently $\tilde{\boldsymbol{\Theta}}_{AB,\,\mathrm{opt}}$, that achieves strong optimality independent of the vector $\mathbf{Y}$ of the variable of interest. The values in the columns of this strong optimal matrix are given by (6.5), which are the vectors $\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i}$ obtained from weak optimality.

It should be noted that since $\boldsymbol{\Delta}_{G,ii}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i} = \lambda_i \mathbf{1}_{G,i}$ (6.7) can be written in an equivalent way as

$$\Phi_{ii'}^{**}\,\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i} = \boldsymbol{\Delta}_{G,ii}^{-1}\boldsymbol{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'} \qquad (6.8a)$$

or

$$\Phi_{ii'}^{*}\,\mathbf{1}_{G,i} = \boldsymbol{\Delta}_{G,ii'}\boldsymbol{\Delta}_{G,i'i'}^{-1}\mathbf{1}_{G,i'} \qquad (6.8b)$$

where $\Phi_{ii'}^{**} = (\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i}\boldsymbol{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'})(\mathbf{1}'_{G,i}\boldsymbol{\Delta}_{G,ii}^{-1}\mathbf{1}_{G,i})$ and $\Phi_{ii'}^{*} = (\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i}\boldsymbol{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'})(\mathbf{1}'_{G,i'}\boldsymbol{\Delta}_{G,i'i'}^{-1}\mathbf{1}_{G,i'})$. In some situations, these can proved to be easier to use that the expresssion (6.7) stated in Result 7.

6.5 Two Examples

We now present two examples that illustrate the preceding theory on weak optimality and strong optimality independent of $\mathbf{Y}$.

Example 1: Poisson Sampling

Let us suppose that the sample s^A is selected using Bernoulli or Poisson Sampling. In that case, the $N^A \times N^A$ matrix $\boldsymbol{\Delta}_A$ is given by $\boldsymbol{\Delta}_A = \mathrm{diag}(1/\pi_j^A - 1)$. Considering the factorization of section 6.1, we have $\boldsymbol{\Delta}_G = \tilde{\boldsymbol{\Theta}}'_{AG}\boldsymbol{\Delta}_A\tilde{\boldsymbol{\Theta}}_{AG} = \tilde{\boldsymbol{\Theta}}'_{AG}[\mathrm{diag}(1/\pi_j^A - 1)]\tilde{\boldsymbol{\Theta}}_{AG} = [\mathrm{diag}((1/\pi_j^A - 1)\mathbf{1}_{A,jj})]$ where $\mathbf{1}_{A,jj}$ is a square matrix of size N_j^A, with N_j^A being the number of links (or "arrows") starting from unit j of U^A. From $\boldsymbol{\Delta}_G$, we extract the submatrices $\boldsymbol{\Delta}_{G,ii}$ that are, in the present case, diagonal. Each submatrix $\boldsymbol{\Delta}_{G,ii}$ is given by $\Delta_{G,ii} = \mathrm{diag}(1/\pi_g^A - 1)$, which is of size N_i^G. Note that each value $(1/\pi_g^A - 1)$ simply corresponds to a unit j of U^A that has previously been linked to the unit g of U^G, which is in turn linked to the unit i of U^B. Now, from (6.5), we directly obtain the optimal values $\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i}$ that minimize $\mathrm{Var}(\hat{Y}^B)$, in the weak sense. These values are given by the vectors

$$\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i} = \left\{\frac{\pi_1^A}{(1-\pi_1^A)\tau_i^G}, \ldots, \frac{\pi_{N_i^G}^A}{(1-\pi_{N_i^G}^A)\tau_i^G}\right\}$$

where

$$\tau_i^G = \sum_{g=1}^{N_i^G} \pi_g^A /(1-\pi_g^A),\ i = 1,\ ...,\ N^B.$$

The $\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i}$ are used to construct the vectors $\tilde{\tilde{\boldsymbol{\theta}}}'_{GB,\,\mathrm{opt},\,i}$, and then the matrix $\tilde{\boldsymbol{\Theta}}_{GB,\,\mathrm{opt}} = \{\tilde{\tilde{\boldsymbol{\theta}}}_{GB,\,\mathrm{opt},\,1},\ ...,\ \tilde{\tilde{\boldsymbol{\theta}}}_{GB,\,\mathrm{opt},\,N^B}\}$. Finally, after computing the optimal matrix $\tilde{\boldsymbol{\Theta}}_{AB,\,\mathrm{opt}} = \boldsymbol{\Theta}_{AG}\tilde{\boldsymbol{\Theta}}_{GB,\,\mathrm{opt}}$, we obtain the optimal weights $\mathbf{W}_{\mathrm{opt}}$ using (2.6).

It should be noted that if the inclusion probabilities π_j^A are equal, we get

$$\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i} = \left\{\frac{1}{N_i^G},\ ...,\ \frac{1}{N_i^G}\right\} = \frac{1}{N_i^G}\mathbf{1}_{GB,\,i},$$

where N_i^G is nothing else than the number of units of U^A linked to unit i of U^B. In other words, in the context of Bernoulli Sampling (*i.e.*, Poisson Sampling with equal probabilities), to minimize the variance $\mathrm{Var}(\hat{Y}^B)$, the choice of the values $\theta^{AB}_{\mathrm{opt},\,ji}$ should be given by 1 if there is a link between unit j of U^A and i of U^B, and 0 otherwise. This corresponds to the results obtained by Kalton and Brick (1995), Lavallée (2002), and Lavallée and Caron (2001).

Using Result 7, we now verify if conditions (6.7), (6.8a) or (6.8b) are satisfied for the optimal matrix $\tilde{\boldsymbol{\Theta}}_{AB,\,\mathrm{opt}}$ that we obtained through weak optimization. If it is the case, this matrix also provides strong optimality independent of the variable of interest y_i. First, we have

$$\Delta_{G,\,ii}^{-1} = \mathrm{diag}\left(\frac{\pi_g^A}{1-\pi_g^A}\right).$$

Also, each submatrix $\Delta_{G,\,ii'}$ of size $N_i^G \times N_{i'}^G$ has somewhat a diagonal structure, but "padded" with zeros. That is, a typical element of $\Delta_{G,ii'}$ is given by $(1/\pi_g^A - 1)$ on a part of the diagonal if both i and i' are linked to the same unit j of U^A (that is linked to unit g of U^G coming from the same j of U^A), and 0 otherwise. Because of this, if two units i and i' are not linked to the same units of U^A, then $\Delta_{G,\,ii'}$ is a matrix of zeros, and then the conditions (6.7), (6.8a) and (6.8b) are automatically satisfied. Referring to Figure 1, children $i = 2$ and $i' = 3$ of U^B are not related to the same parents j of U^A. If the selection of the parents is done using Poisson or Bernoulli Sampling, the 2×2 matrix $\boldsymbol{\Delta}_{G,\,23}$ will then contain only zeros, *i.e.*,

$$\boldsymbol{\Delta}_{G,23} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Because if this, the relations (6.7), (6.8a) or (6.8b) will be satisfied with $\Phi_{23} = 0$, expressing the fact that the weights of i and i' are not correlated.

If two units i and i' are linked to the same unit j of U^A, then, using (6.7), the column vector $\boldsymbol{\Delta}_{G,\,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'}$ contains the scalar $(\tau_{i'}^G)^{-1} = [\sum_{g=1}^{N_{i'}^G} \pi_g^A /(1-\pi_g^A)]^{-1}$ for its first N_i^B components, and 0 for the remaining $N_{i'}^B - N_i^B$ ones (assuming $N_{i'}^B \geq N_i^B$). Because the quantity $\boldsymbol{\Delta}_{G,\,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'}$ must be equal to $\Phi_{ii'}\mathbf{1}_{G,\,i}$ to satisfy (6.7), it must contain only the value $\Phi_{ii'}$. Since $\Phi_{ii'} = \tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i}\boldsymbol{\Delta}_{G,\,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'}$, this will occur only if the vector $\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i} = [1]$, which means that there is only one link to unit i of U^B. As we can see, this is not a condition that will be satisfied in general and therefore, it can be said that in the case of Poisson Sampling, strong optimality independent from $\mathbf{Y}$ will not occur in general.

As a conclusion, we might say that with Poisson or Bernoulli Sampling, the conditions (6.7), (6.8a) or (6.8b) will be satisfied in practice only when the units of U^A are linked to a single unit of U^B, as in the case of sampling households using a frame of individuals. In the other cases, the optimal matrix $\tilde{\boldsymbol{\Theta}}_{AB,\,\mathrm{opt}}$ obtained through weak optimality will not likely lead to strong optimization independent of $\mathbf{Y}$.

Example 2: Simple Random Sampling

Let us suppose that the sample s^A is selected using Simple Random Sampling. In that case, the $N^A \times N^A$ matrix $\boldsymbol{\Delta}_A$ is given by

$$\boldsymbol{\Delta}_A = \frac{N^A}{n^A}\frac{(N^A - n^A)}{(N^A - 1)}\left[\mathbf{I}_A - \frac{\mathbf{1}_A\mathbf{1}'_A}{N^A}\right].$$

Considering the factorization of section 6.1, we have

$$\begin{aligned}\boldsymbol{\Delta}_G &= \tilde{\boldsymbol{\Theta}}'_{AG}\boldsymbol{\Delta}_A\tilde{\boldsymbol{\Theta}}_{AG} \\ &= \frac{N^A}{n^A}\frac{(N^A - n^A)}{(N^A - 1)} \times \tilde{\boldsymbol{\Theta}}'_{AG}\left[\mathbf{I}_A - \frac{\mathbf{1}_A\mathbf{1}'_A}{N^A}\right]\tilde{\boldsymbol{\Theta}}_{AG} \\ &= \frac{N^A}{n^A}\frac{(N^A - n^A)}{(N^A - 1)} \times \left[\mathrm{diag}(\mathbf{1}_{A,jj}) - \frac{\mathbf{1}_G\mathbf{1}'_G}{N^A}\right] \qquad (6.9)\end{aligned}$$

where $\mathbf{1}_{A,jj}$ is a square matrix of size N_j^A, with N_j^A being the number of links (or "arrows") starting from unit j of U^A. From $\boldsymbol{\Delta}_G$, we extract the submatrices $\boldsymbol{\Delta}_{G,\,ii}$. Each submatrix $\boldsymbol{\Delta}_{G,\,ii}$ is given by

$$\boldsymbol{\Delta}_{G,\,ii} = \frac{N^A}{n^A}\frac{(N^A - n^A)}{(N^A - 1)} \times \left[\mathbf{I}_{G,\,i} - \frac{\mathbf{1}_{G,\,i}\mathbf{1}'_{G,\,i}}{N^A}\right],$$

which is of size N_i^G. Then, using a matrix result that can be found, amongst others, in Jazwinski (1970), we get

$$\boldsymbol{\Delta}_{G,ii}^{-1} = \frac{(N^A - 1)}{(N^A - n^A)}\frac{n^A}{N^A} \times \left[\mathbf{I}_{G,\,i} + \frac{1}{(N^A - N_i^G)}\mathbf{1}_{G,\,i}\mathbf{1}'_{G,\,i}\right].$$

Now, from (6.5), we directly obtain the optimal values

$$\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i} = \frac{1}{N_i^G}\mathbf{1}_{G,\,i}$$

that minimize $\text{Var}(\hat{Y}^B)$, in the weak sense, $i = 1, \ldots, N^B$. These values are used to construct the vectors $\tilde{\boldsymbol{\theta}}'_{GB,\text{opt},i}$, and then the matrix $\tilde{\boldsymbol{\Theta}}_{GB,\text{opt}} = \{\tilde{\boldsymbol{\theta}}_{GB,\text{opt},1}, \ldots, \tilde{\boldsymbol{\theta}}_{GB,\text{opt},N^B}\}$. Finally, after computing the optimal matrix $\tilde{\boldsymbol{\Theta}}_{AB,\text{opt}} = \boldsymbol{\Theta}_{AG}\tilde{\boldsymbol{\Theta}}_{GB,\text{opt}}$, we obtain the optimal weights $\mathbf{W}_{\text{opt}}$ using (2.6).

Again, this result is an important one because it goes directly in the direction of the results of Kalton and Brick (1995), Lavallée (2002), and Lavallée and Caron (2001). That is, with Simple Random Sampling, the optimal choice of $\theta^{AB}_{\text{opt},ji}$ should be 1 if there is a link between unit j of U^A and i of U^B, and 0 otherwise.

Using Result 7, we now verify if the conditions (6.7), (6.8a) or (6.8b) for strong optimality independent of y_i are satisfied for the optimal matrix $\tilde{\boldsymbol{\Theta}}_{AB,\text{opt}}$ that we obtain through weak optimization. First, each submatrix $\Delta_{G,ii'}$ of size $N_i^G \times N_{i'}^G$ is given by

$$\boldsymbol{\Delta}_{G,ii'} = \frac{N^A}{n^A}\frac{(N^A - n^A)}{(N^A - 1)} \times \left[\mathbf{H}_{G,ii'} - \frac{\mathbf{1}_{G,i}\mathbf{1}'_{G,i'}}{N^A}\right]$$

where $\mathbf{H}_{G,ii'}$ is a $N_i^G \times N_{i'}^G$ diagonal matrix of ones, "padded" with zeros. Exactly on the same pattern as in example 1, a typical element of $\mathbf{H}_{G,ii'}$ is given by 1 if both i and i' are linked to the same unit j of U^A (that is linked to unit g of U^G), and 0 otherwise. Therefore, we can easily see in which cases the conditions (6.7), (6.8a) or (6.8b) can be satisfied. In fact, because all components of $\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i}$ are equal, $\boldsymbol{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i'}$ is a vector proportional to the sum of the lines of $\boldsymbol{\Delta}_{G,ii'}$, *i.e.*, the sum of the lines of

$$\left[\mathbf{H}_{G,ii'} - \frac{\mathbf{1}_{G,i}\mathbf{1}'_{G,i'}}{N^A}\right].$$

But (6.7) says that this vector must have the same components. This is possible if and only if the matrix $\mathbf{H}_{G,ii'}$ contains only zeros, or if it is of dimension 1×1, which occurs when both i and i' are each linked to only one element of U^A. Therefore, as for Poisson Sampling, strong optimality independent of $\mathbf{Y}$ does not occur in general for Simple Random Sampling.

7. Conclusion

In the present paper, we discussed the use of Indirect Sampling together with the method developed to obtained estimation weights: the Generalized Weight Share Method (GWSM). We then showed the following properties of the GWSM: unbiasedness, the variance computation and transitivity. We presented after a section on the use of the GWSM when the links between the populations U^A and U^B are expressed by ones and zeros, *i.e.*, there is a link or there is not. The section after was devoted to results that are obtained with different forms of link matrices. Finally, we assessed the problem of optimality, *i.e.*, the choice of optimal values to express the links between U^A and U^B in order to minimize the variance of the estimates issued from the GWSM. We have distinguished two kind of optimization: weak and strong optimization.

Weak optimization consists in finding the values of the links to be used in order to minimize, for each unit, the variance of the weights provided by the GWSM. The solution is always uniquely defined, easy to compute and to implement in practice. Weak optimization is also a necessary condition for strong optimization. Strong optimization consists in finding the values of the links in order to minimize the variance of estimation for the total of any variable of interest y. It does not exist for all sampling designs and type of links between the populations U^A and U^B. It also depends on somewhat complicated relations.

We recommend the use of weak optimization because of its flows naturally and the fact that it is very easy to use. Moreover, if our estimation problem can be as well optimized in the strong sense, we will have achieved it through weak optimization, even if it was not demonstrated!

Acknowledgements

The authors would like to thank all the people that showed an interest in Indirect Sampling, and especially in the GWSM. They motivated the writing of this paper that goes beyond what was made previously on this subject.

References

Ernst, L. (1989). Weighting issues for longitudinal household and family estimates. In *Panel Surveys* (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley & Sons, Inc. 135-159.

Deville, J.C., (1998). Comment attraper une population en se servant d'une autre. *Insee méthodes, n°84-85-86, Actes des Journées de methodologie statistique des 17-18 mars 1998*, 63-82.

Jazminski, A.H. (1970). *Stochastic Processes and Filtering Theory*. New York: Academic Press.

Kalton, G., and Brick, J.M. (1995). Weighting Schemes for Household Panel Surveys. *Survey Methodology*, 21, 1, 33-44.

Lavallée, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21, 1, 25-32.

Lavallée, P. (2002). *Le Sondage Indirect, ou la Méthode généralisée du partage des poids*. Éditions de l'Université de Bruxelles, Brussels.

Lavallée, P., and Caron, P. (2001). Estimation using the generalised weight share method: The case of record linkage. *Survey Methodology*, 27, 2, 155-169.

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

Searle, S.R. (1971). *Linear Models*. New York: John Wiley & Sons, Inc.

Survey Methodology, December 2006
Vol. 32, No. 2, pp. 177-185
Statistics Canada, Catalogue No. 12-001

Extension of the Indirect Sampling Method and its Application to Tourism

Jean-Claude Deville and Myriam Maumy-Bertrand [1]

Abstract

A survey of tourist visits originating intra and extra-region in Brittany was needed. For concrete material reasons, "border surveys" could no longer be used. The major problem is the lack of a sampling frame that allows for direct contact with tourists. This problem was addressed by applying the *indirect sampling method*, the weighting for which is obtained using the *generalized weight share method* developed recently by Lavallée (1995), Lavallée (2002), Deville (1999) and also presented recently in Lavallée and Caron (2001). This article shows how to adapt the method to the survey. A number of extensions are required. One of the extensions, designed to estimate the total of a population from which a Bernouilli sample has been taken, will be developed.

Key Words: Generalized weight share method; Incomplete frame and multiple frames.

1. Introduction

A "border survey" of extra-region tourist visits in Brittany (those not by residents of Brittany) was conducted over the period from April to September 1997. The Observatoire Régional du Tourisme de Bretagne and the Comités Départementaux de Tourisme were interested in doing another one. Unfortunately, they no longer had the means to gather a certain mass of data at the regional or intra-regional borders because the police forces were no longer interested in collaborating on roadside surveys.

For this reason, the Observatoire Régional du Tourisme de Bretagne, with the assistance of a technical committee comprised of methodologists and field operators, decided to introduce a new survey methodology to replace the "border survey" methodology. In addition, evaluation of intra-regional tourism (of residents of Brittany vacationing in Brittany, for example) is vital to identifying development factors.

One of the major problems is the lack of a sampling frame that allows direct communication with tourists. This problem was addressed by using an approach previously used in the Asturias in Spain (Valdés, De La Ballina, Aza, Loredo, Torres, Estébanez, Domínguez and Del Valle (2001) and Torres Manzanera, Sustacha Melijosa, Menéndez Estébanez and Valdés Pelaáez (2002)), which involves sampling services intended mainly for tourists and asking them questions at the various locations of these many tourist service sites. Obviously, a tourist may use one or more of the services in the sampling frame once or several times during the survey period in question. To be able to estimate the parameters of interest with respect to tourists, it must be possible to conduct a rigorous sample of certain services and then link the set of weights of the sampled services to the set of weights of the tourists the tourists who used these services. The purpose of this article is to present a method that makes this calculation possible. This method relies mainly on the generalized weight share method (GWSM) developed by Lavallée (1995), Lavallée (2002) and Deville (1999).

2. Generalized Weight Share Method

We will briefly review the principle of the *generalized weight share method* (GWSM). For more information, see Lavallée (1995), Lavallée (2002) and Deville (1999).

We will let U^A be a finite population containing N^A units, where each unit is denoted by j and U^B is a finite population containing N^B units, where each unit is denoted by i. The correspondence between U^A and U^B can be represented by a matrix of links $\Theta_{AB} = [\theta_{ji}^{AB}]$, of size $N^A \times N^B$ where each element $\theta_{ji}^{AB} \geq 0$. In other words, the unit j of U^A is linked to unit i of U^B provided that $\theta_{ji}^{AB} > 0$; otherwise, there is no link between these two units.

In the case of the indirect survey, we select the sample s^A of n^A units from U^A based on a given sampling design. Let $\pi_j^A > 0$, be the probability of selection of the unit j. For each unit j selected in s^A, we identify the units i of U^B for which $\theta_{ji}^{AB} > 0$. Then we let s^B, be all of the n^B units of U^B identified using the units $j \in s^A$, that is,

$$s^B = \{i \in U^B ; \exists j \in s^A \text{ and } \theta_{ji}^{AB} > 0\}.$$

For each unit i of s^B, a variable of interest y_i is measured.

It is assumed that, for any unit j of s^A, it is possible to obtain the values of θ_{ji}^{AB} for $i = 1, ..., N^B$ by a direct interview or from an administrative source. For any unit i

1. Jean-Claude Deville, Laboratoire de Statistique d'Enquêtes, ENSAI/CREST, Campus de Ker-Lann, 35170 BRUZ (France). E-mail: deville@ensai.fr; Myriam Maumy-Bertrand, Laboratoire de Statistique, Université Louis Pasteur, 7, rue René Descartes 67084 STRASBOURG Cedex (France). E-mail: mmaumy@math.u-strasbg.fr.

identified of U^B (or only of s^B), it is assumed that we can obtain the values of θ_{ji}^{AB} for $j = 1, ..., N^A$. For this reason, it is not necessary to know the values of θ_{ji}^{AB} for all of the matrix of links Θ_{AB}. Indeed, we only need to know the values of θ_{ji}^{AB} for lines j of Θ_{AB}, where $j \in s^A$, and for columns i of Θ_{AB} where $i \in s^B$.

For example, if the purpose is to estimate a variable of interest Y^B of target population U^B, where

$$Y^B = \sum_{i=1}^{N^B} y_i, \qquad (2.1)$$

with y_i measured according to the aggregate U^B. We then use an estimator in the form

$$\hat{Y}^B = \sum_{i=1}^{N^B} w_i y_i, \qquad (2.1)$$

where w_i is the estimated weight of unit i of s^B, with $w_i = 0$ for $i \notin s^B$. To obtain an unbiased estimate of a variable of interest Y^B, we must use as weight w_i the inverse of the probability of selection π_i^B of unit i. As mentioned in Lavallée (1995) and Lavallée (2002), it is generally difficult, if not impossible, to obtain these probabilities. Consequently, we turn to the GWSM, where the weights are given by

$$w_i = \sum_{j \in s^A} \frac{\tilde{\theta}_{ji}^{AB}}{\pi_j^A},$$

where $\tilde{\theta}_{ji}^{AB} = \theta_{ji}^{AB} / \sum_{j=1}^{N^A} \theta_{ji}^{AB}$. Using this construction, the estimator $\hat{Y}^B$ is unbiased. Similarly, it is possible to calculate and estimate the variance of this estimator because it is the same as that of

$$\sum_{j \in s^A} \frac{z_j}{\pi_j^A},$$

with $z_j = \sum_{i=1}^{N^B} \tilde{\theta}_{ji}^{AB} y_i$.

3. Tourism Survey in an Open Environment

3.1 Survey Objectives

The principle of the survey is as follows:

"reach tourists (foreigners or French citizens whether or not they live in Brittany) through services aimed at meeting the basic or specific needs"

such as accommodation, food, leisure activities and transportation.

3.2 Population of Interest

We will let G be a *geographic field* (the four provinces of Brittany) and P be a *reference period* (in this case, it is from February 2005 to December 2005).

A tourist is defined as a person who spent at least one night in G outside his principle residence (tourist-night).

For a tourist, a *trip* is an period sej of P, the length of the cardinal of sej noted as $|sej|$, during which the tourist spends all his nights in G outside his principle residence, the nights immediately before or after the trip sej having been spent outside G (or at the principle residence.).

A tour is a group of tourists (tourist household) sharing the same trip and with the same accommodation during the trip. The term tourist household will also be used through a slight misuse of the terminology (the same tourist household can have several tours over a period, but we have no way to distinguish them).

The statistical unit i of the survey is the tour.

The sub-units of the survey are the trips, tourists and tourist-nights. A tour i consists of n_i tourists during a trip of duration $|sej|$ and thus $n_i \times |sej|$ tourist-nights. Here population U^B is therefore the aggregate of the tours in G during P. ($sej \cap P \neq \varnothing$).

3.3 Survey Sampling Design

To use the GWSM, the theoretical population U^A is formed by a "services" aggregate. In this survey, these services consist of:

- Purchases in bakeries, being the first stratum of U^A.
- Visits to a set of well known cultural, recreational or family sites. In practice, for each of them, a "mandatory pass point" has been defined. It consists of the total number of people passing by this point, which is the second stratum of U^A.
- The number of people exiting Brittany by way of the La Gravelle highway toll, which accounts for 80% of the exits by tourists from Brittany by car. This method of transport itself accounts for 80% of the trips by non-resident of Brittany. People passing this point constitute the third stratum of U^A.

In other words, the *sampling frame* is formally constructed of three strata:

1. purchases in bakeries;
2. visits to a set of sites typical of Brittany;
3. people at the La Gravelle highway toll.

In the *first stratum*, we use a three-stage sample:

- a sample of bakeries;
- a sample of survey days;
- a sample of clients in the bakery on a given day.

In the *second stratum*, we use a two-stage sample:

- a sample of survey days;
- a sample of people who pass through one of the 16 chosen sites on a given day.

Lastly, in the *third stratum*, we use a two-stage sample:

- a sample of survey days;
- a sample of people who pass through the La Gravelle highway toll on a given day.

It is acknowledged that any tourist household consumes at least one of the "services" (bakery purchases, visits to typical Brittany sites, the La Gravelle highway toll)), or at least, that very few households do not consume any of them.

Each sampling (bakery, days, "service") requires specific techniques and it would take considerable time to provide details on each of them. Nevertheless, we will provide the following key technical elements:

- bakeries are sampled using a traditional design stratified geographically (five strata: coastal area of four Brittany departments, the interior of Brittany). In each stratum, the bakeries are sampled with probabilities proportional to their "tourist potential" constructed from their business revenue, the tourist accommodation capacity, and the number of principal residences in the commune to which they belong. This was the theoretical approach, but in practice, the sample was somewhat "forced" by unforeseen circumstances (refusal of bakers, closures during certain period, for example).
- The sites are not sampled, but rather selected for their notoriety and the technical possibility of identifying a "mandatory pass point" (sometimes approximate).
- For each bakery, each site and the La Gravelle highway toll, we defined completely homogeneous "clusters of days" in each period P. A cluster was assigned randomly to each bakery, site and the La Gravelle highway toll. In practice, this means that a full-time enumerator is mobilized for several clusters.
- For each "service", tourists are sampled using the normal techniques of random selection of arrivals: pseudo-systematic sample because, while the enumerator is handing out one questionnaire, other people are going by without being counted. This means that the total number of visitors cannot be estimated directly. If a site is accessible through a ticket booth (museum or chateau, for example), the sampling relies on this means. Ultimately, the sample of users of a "service" on a given day is considered a Bernouilli sample, that is, a simple random sample if we know the size of the population (the number of visitors on a given day).

Comments 3.1. The definition of *tourist* itself is linked to accommodation and it seems natural to use a frame directly related to this service. Practice shows that this is difficult to achieve.

To begin with, there is no correct sampling frame for non-commercial accommodation (relatives, friends, secondary residence) or for seasonal furnished rentals.

In the case of hotels, campgrounds and family holiday homes, the trials runs in summer 2004 revealed the existence of catastrophic bias due to the intervention of hotel owners in the survey selection process. The hoteliers did not respect the random sample instructions and "essentially" distributed the questionnaires to their best clients. This part of the survey had to be set aside and replaced by the count through the La Gravelle highway toll, which is regularly subject to honest quality surveys by various organizations.

The questionnaires collected at the bakeries and at the Brittany tourism sites during summer 2004 apparently produced good qualitative and quantitative results regarding the various modes of accommodation.

Food consumption would undoubtedly have been captured better by questionnaires at the exit of supermarkets, but the problem there lies in the heterogeneity of these establishments and in the cutthroat competition between them; group C … agrees to the surveys in its establishments only if group I … is excluded! In contrast, the collaboration of local bakers in the survey was excellent.

Comments 3.2. By the very definition of the method used, we operate formally within the context of sampling from multiple frames. The problem has given rise to considerable literature (Hartley (1962), Lund (1968) and Hartley (1974) for a start). The GWSM applies to this problem by simply considering each sampling frame as a stratum provided that it is possible to identify for each unit sampled all of frames of which it is a part. This approach provides a rigorous and unique design-based solution to this problem. This comment is worthy of its own article, but the authors know that it is not worth the trouble: an idea that can be explained in ten lines does not need an article or a book for it to survive.

4. Parameters of Interest

Application F, which links to any service j during the reference period P in the three types of establishments of the survey coverage tour i that used this service, is defined as:

$$\begin{array}{ccc} F: \text{services} & \rightarrow & \text{tour} \\ j & \rightarrow & F(j) = i. \end{array}$$

We will let U^B, be the population of tours i of reference period P. This population of interest U^B is the image by F of the aggregate of services during reference period P

in the three types of establishments of the survey coverage. Population U^A is the image by F^{-1} of the aggregate of tours during reference period P. For all $i \in U^B$, we define $R_i(B) = \text{card}(F^{-1}(i))$, the number of antecedents of i during the survey period, that is, the number of services j used by the given tourist household i.

The *parameters of interest* can be totals, sizes or ratios. Let us assume, for example, that we are interested in the estimate of a total relative to a variable y defined on population U^B,

$$Y^B = \sum_{i \in U^B} y_i. \qquad (4.1)$$

A specific example of these totals is the size of U^B, written N^B and defined by

$$N^B = \text{card}(U^B) = \sum_{i \in U^B} 1.$$

For example, Y^B can be the number of people who practiced this activity, the total budget spent by the tourist household in Brittany, the geographic origin of the tourist households, or the number of days that the tourist household spends in Brittany. It should be noted that for many variables, the total Y^B depends on the size of the tourist household, that is, the number of people who make up this group and on the length of the trip (only those days spent in Brittany).

Now, we can write

$$Y^B = \sum_{i \in U^B} y_i = \sum_{l=1}^{3} \sum_{a_l \in A_l} \sum_{d_l \in D_l} \sum_{j \in C_{d_l}} z_j, \qquad (4.2)$$

where

$$z_j = \frac{y_i}{R_i(B)}, \quad \text{for} \quad j \in F^{-1}(i),$$

where

- A_1: the aggregate of bakeries in the survey coverage identified by index a_1
- A_2: the 16 visit locations in the survey coverage identified by index a_2
- A_3: the La Gravelle highway toll identified by index a_3
- D_l: the aggregate of survey days, identified by index d_l in an establishment a_l of A_l, for the variant of 1 to 3
- C_{d_l}: the aggregate of services in an establishment a_l of A_l of day d_l of D_l identified by index j.

5. Unbiased Estimates of a Total

In the previous paragraph, we showed that the total of interest is written as a total over the aggregate of the services in the coverage. Let us assume that we have a sample of respondent services j, to which we can link sampling weight δ_j. These weights are assumed to be unbiased because the sample of services follows the canons of a multi-stage sample, each component sample being unbiased.

To make the notations easier to read, we will not show below all stages of the sample draw based on establishment a_l. Let:

- s^B: be the aggregate of tourist household i corresponding to the aggregate of services sampled during the survey period
- s_{A_l}: be the aggregate of sampled establishments
- s_{D_l}: be the aggregate of days sampled in establishment a_l
- s_{d_l}: be the sub-sample of services j corresponding to establishment day a_l.

Since we have a set of sampling weights δ_j for the respondent services, and if we know $R_i(B)$, we can estimate the unbiased total Y^B by

$$\hat{Y}^B = \sum_{i \in s^B} w_i \, y_i \qquad (5.1)$$

where

$$w_i = \frac{\sum_{l=1}^{3} \sum_{s_{A_l}} \sum_{s_{D_l}} \sum_{s_{d_l}} \delta_j}{R_i(B)}.$$

This gives us an estimate of the population of tourist households. This formula is none other than that given by the GWSM mentioned in section 2. Note that $U^A = U^{A_1} \cup U^{A_2} \cup U^{A_3} = \bigcup_{l=1}^{3} U^{A_l}$, $\theta_{ji}^{AB} = 1$ of service j was used by tour i and then $\delta_j = 1/\pi_j^A$.

The variance can be estimated using the same principles (see Lavallée (2002)). We will not go into the details here because it is simply an application of general principles that requires somewhat onerous calculations.

Furthermore, using auxiliary information in the form of totals, whether in populations U^{A_l} or in population U^B, does not pose any particular problems for the point estimation or the estimation of the variance (see Lavallée (2002)).

Comments 5.1. The procedure we have just described for sharing weights may be considered naïve. In fact, we know how to optimize the links matrix Θ_{AB} as shown in Deville and Lavallée (2006). The application of the Brittany survey is described in Deville, Lavallée and Maumy (2005).

6. An example of a Specific Problem: Visit Points in Open Country

As has already been mentioned, developing the survey of tourism in Brittany required many complementary studies.

We have already mentioned the optimization of weight sharing. Using auxiliary data related to the various frames and to the various stages of the sampling is another task. In this section, we want to focus on estimating some of these auxiliary data, in particular for visits to tourism sites in open country.

In certain cases, we unfortunately do not know the total number of people, denoted as $T_P^{A_2}$, coming to the site on a given day. In effect, in aggregate A_2, we do not know all the services (here the number of visits) of the population. It is therefore not possible to obtain $\pi_j^{A_2}$ directly and therefore δ_j for $j \in A_2$. To overcome this problem, we estimate the number of daily visitors in order to deduct $\tilde{\pi}_j^{A_2} = n_{A_2} / \hat{T}_P^{A_2}$.

Our next step was to develop two approaches to estimating the number of daily visitors for sites accessible by vehicles only (or almost!). The first approach is based on a vehicle sampling system intended to estimate the number of visitors to the site. The second approach uses a sampling of visitors and is aimed at estimating the same quantity by interviewing individuals who give the number of people who travelled with him or her in the vehicle. These two approaches are developed in sections 7 and 8 below.

7. Constructing an Estimator of the Number of Visitors Using a Vehicle Sample

In this paragraph, we examine the approach where an enumerator counts the number of occupants in vehicles that break the line of an electronic eye, or an equivalent system has been set up to count vehicles for which the total number, written as T_V, is known with a virtually negligible measurement error.

7.1 Definition and Variance of $\hat{T}_P^{A_2}$

The total number of vehicles equals

$$T_V = \sum_{\kappa=1,\ \ldots} t_\kappa = \sum_{l \in U_v} 1, \tag{7.1}$$

where t_κ represents the number of vehicles carrying κ persons and U_V the vehicle universe.

Comments 7.1. To make the notations easier to read, we will use here and until the end this article T_P to denote $T_P^{A_2}$.

The total number of people visiting the site equals

$$T_P = \sum_{\kappa=1,\ \ldots} \kappa t_\kappa = \sum_{k \in U_P} 1, \tag{7.2}$$

where U_P denotes the universe of people. We also have the equation

$$T_P = \sum_{l \in U_v} v_l, \tag{7.3}$$

where v_l is the number of people in vehicle l.

As mentioned in the previous section, the total number of people T_P is unknown. Consequently, we must construct an estimator of T_P. If we let $\hat{T}_P$ be $\pi-$estimator based on s_V, a simple random sample of vehicles of size n and with a probability of inclusion n/T_V

$$\hat{T}_P = \frac{T_V}{n} \sum_{l \in s_V} v_l = T_V \bar{v}, \tag{7.4}$$

assuming

$$\bar{v} = \frac{1}{n}\left(\sum_{l \in s_V} v_l \right).$$

It is clear that $\hat{T}_P$ is an unbiased estimator of the total number of people T_P and that $\bar{v}$ is an unbiased estimate of the average number $\bar{V}$ of people in a vehicle.

The variance of $\hat{T}_P$ is therefore equal to

$$\begin{aligned} \mathrm{Var}[\hat{T}_P] &= T_V^2 \left(\frac{1}{n} - \frac{1}{T_V} \right) S_V^2 \\ &= \frac{1}{n} T_V^2 S_V^2 - T_V S_V^2, \end{aligned} \tag{7.5}$$

where S_V^2 denotes the corrected variance of population U_V.

7.2 Constructing an Estimator of a Variable of Interest in the Case of a Vehicle Sample

We want to estimate a variable of interest Y of population U_P written as

$$Y = \sum_{k \in U_P} y_k, \tag{7.6}$$

where y_k is the variable of interest measured in the final questionnaire. Let $\hat{Y}$ be $\pi-$estimator defined by

$$\hat{Y} = \sum_{k \in s_P} w_k^P y_k, \tag{7.7}$$

where weight w_k^P is equal to $\hat{T}_P / m$. Consequently, estimator $\hat{Y}$ can be written

$$\hat{Y} = \frac{\hat{T}_P}{m} \sum_{k \in s_P} y_k = \hat{T}_P \bar{y} \tag{7.8}$$

assuming

$$\bar{y} = \frac{1}{m}\left(\sum_{k \in s_P} y_k \right).$$

Subsequently, variables $\hat{T}_P$ and $\bar{y}$ will be assumed to be independent. The assumption is realistic, because we use two independent enumerators in the field.

7.2.1 Calculation of the Variance of the Estimator $\hat{Y}$

According to Huygens' theorem (1673), conditioning on sample s_V, we get

$$\begin{aligned} V_Y &= \mathrm{Var}[\hat{Y}] \\ &= \overline{Y}^2\,\mathrm{Var}[\hat{T}_P] + T_P^2\,\mathrm{Var}[\overline{y}] \\ &\quad + \mathrm{Var}[\hat{T}_P]\,\mathrm{Var}[\overline{y}]. \end{aligned} \tag{7.9}$$

In the present case, we liken the sample to a simple random sampling without replacement. Equation (7.9) thus becomes

$$\begin{aligned} V_Y = {} & \overline{Y}^2\left(\frac{1}{n}\,T_V^2\,S_V^2 - T_V\,S_V^2\right) \\ & + T_P^2\left(\frac{1}{m}S_Y^2 - \frac{S_Y^2}{T_P}\right) \\ & + \left(\frac{1}{n}T_V^2\,S_V^2 - T_V\,S_V^2\right)\left(\frac{1}{m}S_Y^2 - \frac{S_Y^2}{T_P}\right), \end{aligned}$$

with $S_Y^2 = 1/(T_P - 1)\,\sum_{k\in U_P}(y_k - \overline{Y})^2$. Reorganizing the terms gives

$$\begin{aligned} V_Y = {} & \left(\overline{Y}^2 - \frac{S_Y^2}{T_P}\right) T_V^2\,S_V^2\,\frac{1}{n} \\ & + (T_P^2 - T_V\,S_V^2)\,S_Y^2\,\frac{1}{m} \\ & + T_V^2\,S_V^2\,S_Y^2\,\frac{1}{nm} + \frac{T_V}{T_P}\,S_V^2\,S_Y^2 \\ & - \overline{Y}^2\,T_V\,S_V^2 - T_P\,S_Y^2. \end{aligned}$$

The next step is to determine the allocation of the sample sizes s_P and s_V that minimizes the variance of estimator $\hat{Y}$ for fixed population sizes T_P and T_V.

We must therefore minimize equation (7.10) in n, m subject to

$$C_V\,n + C_P\,m = C,$$

where C_V denotes the cost (in time for example) of the questionnaires related to vehicles, C_P the cost (in time) of the questionnaires related to people, and C the total cost.

The Lagrangian equation can be written as

$$\begin{aligned} L(n,\,m,\,\lambda) = {} & \left(\overline{Y}^2 - \frac{S_Y^2}{T_P}\right) T_V^2\,S_V^2\,\frac{1}{n} \\ & + (T_P^2 - T_V\,S_V^2)\,S_Y^2\,\frac{1}{m} \\ & + T_V^2\,S_V^2\,S_Y^2\,\frac{1}{nm} + \frac{T_V}{T_P}\,S_V^2\,S_Y^2 \\ & - \overline{Y}^2\,T_V\,S_V^2 - T_P\,S_Y^2 \\ & + \lambda(C_V\,n + C_P\,m - C). \end{aligned} \tag{7.11}$$

Taking the partial derivatives with respect to variables n, m, λ and setting them equal to zero gives

$$\begin{aligned} \frac{\partial L}{\partial n}(n,\,m,\,\lambda) = {} & \left(\overline{Y}^2 - \frac{S_Y^2}{T_P}\right) T_V^2\,S_V^2\left(-\frac{1}{n^2}\right) \\ & + T_V^2\,S_V^2\,S_Y^2\left(-\frac{1}{mn^2}\right) \\ & + \lambda\,C_V = 0, \end{aligned}$$

$$\begin{aligned} \frac{\partial L}{\partial m}(n,\,m,\,\lambda) = {} & (T_P^2 - T_V\,S_V^2)\,S_Y^2\left(-\frac{1}{m^2}\right) \\ & + T_V^2\,S_V^2\,S_Y^2\left(-\frac{1}{nm^2}\right) \\ & + \lambda\,C_P = 0, \end{aligned}$$

$$\frac{\partial L}{\partial \lambda}(n,\,m,\,\lambda) = C_V\,n + C_P\,m - C = 0.$$

After calculations, we get a third-degree equation in n that is written

$$\begin{aligned} & \lambda C_V^2\,n^3 - \lambda C_V C n^2 \\ & \quad - C_V T_V^2\,S_V^2\left(\overline{Y}^2 - \frac{S_Y^2}{T_P}\right) n \\ & \quad + T_V^2\,S_V^2\left(C\left(\overline{Y}^2 - \frac{S_Y^2}{T_P}\right) + C_P\,S_Y^2\right) = 0. \end{aligned}$$

This third-degree equation in n allows a real solution that can be determined using numeric methods.

Using the same reasoning, we get a third-degree equation in m

$$\begin{aligned} & \lambda C_P^2\,m^3 - \lambda C_P C m^2 \\ & \quad - C_P\,S_Y^2\,(T_P^2 - T_V\,S_V^2)\,m \\ & \quad + S_Y^2\,(C(T_P^2 + T_V\,S_V^2) + C_V\,T_V^2\,S_V^2) = 0. \end{aligned}$$

7.2.2 Simplified Case

To simplify the variance calculation of estimator $\hat{Y}$, we can make an approximation in equation (7.10). In effect, we can assume that term $1/nm$ is negligible before terms $1/n$ and $1/m$.

This then gives us the following transformation of equation (7.10)

$$\begin{aligned} V_Y = {} & \left(\overline{Y}^2 - \frac{S_Y^2}{T_P}\right) T_V^2\,S_V^2\,\frac{1}{n} \\ & + (T_P^2 - T_V\,S_V^2)\,S_Y^2\,\frac{1}{m} \\ & + \frac{T_V}{T_P}S_V^2\,S_Y^2 - \overline{Y}^2\,T_V\,S_V^2 \\ & - T_P\,S_Y^2. \end{aligned} \tag{7.12}$$

The next step is determining the allocation of the sample sizes s_P and s_V that minimize the variance of estimator $\hat{Y}$ for fixed population sizes T_P and T_V.

We must therefore minimize equation (7.12) in n, m subject to

$$C_V\, n + C_P\, m = C.$$

The Lagrangian equation can be written as

$$\begin{aligned} L(n,\, m,\, \lambda) = & \left(\bar{Y}^2 - \frac{S_Y^2}{T_P} \right) T_V^2\, S_V^2\, \frac{1}{n} \\ & + (T_P^2 - T_V\, S_V^2)\, S_Y^2\, \frac{1}{m} \\ & + \frac{T_V}{T_P} S_V^2\, S_Y^2 - \bar{Y}^2\, T_V\, S_V^2 \\ & - T_P\, S_Y^2 \\ & + \lambda (C_V\, n + C_P\, m - C). \end{aligned} \qquad (7.13)$$

Taking the partial derivatives with respect to variables $n,\, m,\, \lambda$ and setting them equal to zero gives

$$\begin{aligned} \frac{\partial L}{\partial n}(n,\, m,\, \lambda) &= \left(\bar{Y}^2 - \frac{S_Y^2}{T_P} \right) T_V^2\, S_V^2 \left(-\frac{1}{n^2} \right) \\ &\quad + \lambda\, C_V = 0, \\ \frac{\partial L}{\partial m}(n,\, m,\, \lambda) &= (T_P^2 - T_V\, S_V^2)\, S_Y^2 \left(-\frac{1}{m^2} \right) \\ &\quad + \lambda\, C_P = 0, \\ \frac{\partial L}{\partial \lambda}(n,\, m,\, \lambda) &= C_V\, n + C_P\, m - C = 0. \end{aligned}$$

After the calculations, we get

$$n_{\text{opt}} = \frac{C}{\left(C_V + \sqrt{C_P\, C_V\, \dfrac{T_P\, S_Y^2\, (T_P^2 - T_V\, S_V^2)}{T_V^2\, S_V^2\, (T_P\, \bar{Y}^2 - S_Y^2)}} \right)},$$

$$m_{\text{opt}} = \frac{C}{\left(C_V + \sqrt{C_P\, C_V\, \dfrac{T_V^2\, S_V^2\, (T_P\, \bar{Y}^2 - S_Y^2)}{T_P\, S_Y^2\, (T_P^2 - T_V\, S_V^2)}} \right)}.$$

8. Constructing an Estimator of the Number of Visitors Using a Sampling of Visitors

The previous method can be complicated and costly to use at certain sites. A simpler data collection method involves asking person k the number u_k of passengers in vehicle i that transported him or her. This number u_k is equal here to v_l for vehicle l that transported person k. This method has the further advantage of accurately capturing the number of passengers within the meaning of the survey (are babies counted?).

8.1 Definition of $\hat{\hat{T}}_P$

Let us go back to the following equation

$$T_P = \sum_{l \in U_V} v_l,$$

where v_l denotes the number of passengers in vehicle l. Let us also recall

$$T_P = \sum_{l \in U_P} 1.$$

The average number of passengers in a vehicle $\bar{V}$ can be expressed as

$$\bar{V} = \frac{\sum\limits_{l \in U_V} v_l}{\sum\limits_{l \in U_V} 1} = \frac{\sum\limits_{\kappa=1,\, \ldots} \kappa t_\kappa}{\sum\limits_{\kappa=1,\, \ldots} t_\kappa} = \frac{\sum\limits_{\kappa=1,\, \ldots} m_\kappa}{\sum\limits_{\kappa=1,\, \ldots} M_\kappa / \kappa}, \qquad (8.1)$$

where t_κ is the number of κ – passenger vehicles and M_κ is the number of people who came in a κ – passenger vehicle.

We can use this last relation to obtain a new version of T_P

$$T_P = T_V\, \bar{V}. \qquad (8.2)$$

Consequently, an estimator of T_P can be written as

$$\hat{\hat{T}}_P = T_V\, \hat{\bar{V}}, \qquad (8.3)$$

where the total number of vehicles T_V is perfectly known. Observing this expression, we see that, in order to know estimator $\hat{\hat{T}}_P$, all that is required is to determine the quantity $\hat{\bar{V}}$. Let us therefore introduce the following estimator of $\bar{V}$

$$\hat{\bar{V}} = \frac{\sum\limits_{\kappa \in s_P} m_\kappa}{\sum\limits_{\kappa \in s_P} m\kappa / \kappa},$$

where m_κ is the number of people in the sample travelling in a κ passenger vehicle. Estimator $\hat{\bar{V}}$ can also be written as follows:

$$\hat{\bar{V}} = \frac{\sum\limits_{k \in s_P} 1}{\sum\limits_{k \in s_P} 1/u_k}$$

or as

$$\hat{\bar{V}} = \frac{m}{\sum\limits_{k \in s_P} 1/u_k}. \qquad (8.4)$$

The last equation makes it possible to write the following equation

$$\frac{1}{\hat{\bar{V}}} = \frac{1}{m} \sum_{k \in s_P} \frac{1}{u_k}. \qquad (8.5)$$

This new quantity represents the empirical average of $1/u_k$ and $\hat{\bar{V}}$ is the harmonic average of u_k. It is also possible to calculate its variance, which is equal to

$$\mathrm{Var}\left[\frac{1}{\hat{\bar{V}}}\right] = \left(\frac{1}{m} - \frac{1}{T_P}\right) S^2_{1/u}. \qquad (8.6)$$

8.2 Calculating the Variance of Estimator $\hat{\hat{T}}_P$ Without a Vehicle Sample

Now we have to calculate the variance of estimator $\hat{\bar{V}}$ knowing (8.6). To this end, note that we can write

$$\begin{aligned}\frac{1}{\hat{\bar{V}}} &= \frac{1}{\bar{V}\left(\frac{\hat{\bar{V}}}{\bar{V}} - 1 + 1\right)} \\ &= \frac{1}{\bar{V}} \times \frac{1}{1 + \frac{\hat{\bar{V}} - \bar{V}}{\bar{V}}} \\ &= \frac{1}{\bar{V}}\left(1 - \frac{\hat{\bar{V}} - \bar{V}}{\bar{V}} + o\left(\frac{\hat{\bar{V}} - \bar{V}}{\bar{V}}\right)\right).\end{aligned}$$

Accordingly, this gives

$$\mathrm{Var}\left[\frac{1}{\hat{\bar{V}}}\right] \simeq \left(\frac{1}{\bar{V}}\right)^2 \times \frac{\mathrm{Var}[\hat{\bar{V}}]}{\bar{V}^2}.$$

Lastly, we have

$$\mathrm{Var}[\hat{\bar{V}}] \simeq \bar{V}^4 \times \mathrm{Var}\left[\frac{1}{\hat{\bar{V}}}\right],$$

or, with (8.6)

$$\mathrm{Var}[\hat{\bar{V}}] \simeq \bar{V}^4 \times \left(\frac{1}{m} - \frac{1}{T_P}\right) S^2_{1/u}. \qquad (8.7)$$

By definition, variance $S^2_{1/u}$ is equal to

$$S^2_{1/u} = \frac{1}{T_P - 1} \sum_{k \in U_P} \left(\frac{1}{u_k} - \frac{1}{\bar{V}}\right)^2. \qquad (8.8)$$

Since quantity T_P is unknown, this relation can be estimated by

$$\frac{1}{m - 1} \sum_{k \in s_P} \left(\frac{1}{u_k} - \frac{1}{\bar{V}}\right)^2. \qquad (8.9)$$

Given (8.7) and (8.9), we can easily determine the variance of estimator $\hat{\bar{V}}$ and consequently, that of estimator $\hat{T}_P$ and lastly, that of the variable of interest $\hat{Y}$.

Comments 8.1. Estimator $\hat{\hat{T}}_P$ is biased and asymptomatically unbiased.

Comment 8.2. If variables $\hat{\hat{T}}_P$ and $\bar{y}$ are not independent then we would have

$$\begin{aligned}\mathrm{Var}\left[\hat{\hat{T}}_P\, \bar{y}\right] &= \bar{Y}^2\, \mathrm{Var}\left[\hat{\hat{T}}_P\right] + T_P^2\, \mathrm{Var}[\bar{y}] \\ &\quad + \mathrm{Var}\left[\hat{\hat{T}}_P\, \bar{y}\right] \mathrm{Var}[\bar{y}] \\ &\quad + \text{terms not linked to the} \\ &\quad \text{eventual non-independance} \\ &\quad \text{of the variables } \hat{\hat{T}}_P \text{ and } \bar{y}.\end{aligned}$$

9. Numeric Illustration

A mechanical counter at a site in open country gives $T_V = 100$ vehicles. We assume that 20% of the vehicles have one person, 20% have two people, 20% have three people, 20% have four people and 20% have five people. This means there are 300 visitors to the site. The variance S^2_V is equal to two disregarding finite population corrections. The average number of passengers $\bar{V}$ is three. In effect, we have:

$$\begin{aligned}\frac{1}{V} &= \frac{1}{1} \times \frac{20}{300} + \frac{1}{2} \times \frac{40}{300} + \frac{1}{3} \times \frac{60}{300} \\ &\quad + \frac{1}{4} \times \frac{80}{300} + \frac{1}{5} \times \frac{100}{300} = \frac{1}{3}.\end{aligned}$$

which gives $\bar{V} = 3$.

Let us now calculate an estimate of $S^2_{1/u}$. After simplifications of (8.8) and assuming that T_P is large enough compared to one, we have

$$S^2_{1/u} \simeq \frac{1}{T_P} \sum_{k \in U_P} \frac{1}{u_k^2} - \left(\frac{1}{\bar{V}}\right)^2.$$

Thus, we get

$$\begin{aligned}S^2_{1/u} &= \frac{1}{30}\left(2 + 1 + \frac{2}{3} + \frac{1}{2} + \frac{2}{5}\right) - \frac{1}{3^2} \\ &= \frac{1}{30}\left(\frac{60 + 30 + 20 + 15 + 12}{30}\right) - \frac{1}{3^2} \\ &= \frac{137}{30^2} - \frac{1}{3^2} = \frac{37}{30^2}.\end{aligned}$$

Since we know $S^2_{1/u}$, we can calculate the variance of estimator $\bar{V}$. This gives

$$\mathrm{Var}[\hat{\bar{V}}] \simeq 3^4 \times \frac{37}{30^2} \times \frac{1}{m}.$$

Lastly, we can calculate the variance of estimator $\hat{\hat{T}}_P$

$$\mathrm{Var}\left[\hat{\hat{T}}_P\right] = T_V^2\ \mathrm{Var}\left[\hat{\bar{V}}\right]$$

$$\simeq 10^4 \times 3^4 \times \frac{37}{30^2} \times \frac{1}{m}.$$

The first approach gives a variance of estimator $\hat{T}_P$ equal to

$$\mathrm{Var}\left[\hat{T}_P\right] = 10^4 \times 2 \times \frac{1}{n}.$$

Thus, for estimator $\hat{\hat{T}}_P$ to have the same variance as estimator $\hat{T}_P$, size m of sample s_P must be equal to

$$m \simeq 1.66n.$$

Our initial conclusion is that the second approach makes field operations simpler and less costly in terms of personnel because it only requires one enumerator. It is more accurate than a count that does not involve direct contact to obtain the composition of the tourist household. It requires only one sample about one and a half times larger than the first approach to produce the same accuracy, which is tolerable given the resulting simplification of collection. In practice, at all sites, the second approach will be the preferred application.

Conclusion

This article presented a broad description of a new method applicable to tourism statistics. It involves capturing tourists based on the consumption of certain services on which probabilistic samples are constructed. The weight share method makes it possible to shift from statistical accuracy of the services to the accuracy of the relevant tourism statistical units: the tour, the trip, the tourist household, the tourist or the tourist-night. However, the method requires numerous adaptations and complements to the weight share. We described one of these in detail, which is the estimate of the number of visitors to a site in open country. Two methods were tested. One, which was more accurate in terms of sample size, requires a relatively extensive organization and runs the risk of unacceptable errors in measurement. At the price of collecting slightly more data, the second method is preferred.

Other studies of this nature were conducted before and during the time of the survey so that it is difficult to present the full methodology in a single article.

Acknowledgements

The authors sincerely thank the two reviewers and associate editor who all made a significant contribution to improving the readability of this paper.

References

Deville, J.-C. (1999). Les enquêtes par panel : En quoi différent-elles des autres enquêtes ? suivi de : Comment attraper une population en se servant d'une autre. *Actes des journées de méthodologie statistiques*, *INSEE Méthodes*, 84-85-86, 63-82.

Deville, J.-C., and Lavallée, P. (2006). Indirect Sampling: The Foundations of the Generalized Weight Share Method. *Survey Methodology*, 32, 2, 165-176.

Deville, J.-C., Lavallée, P. and Maumy, M. (2005). Composition, factorisation et conditions d'optimalité (faible, forte) dans la méthode de partage des poids. Application à l'enquête sur le tourisme en Bretagne. *Actes des journées de méthodologie statistiques*, *INSEE Méthodes*.

Hartley, H.O. (1962). Multiple Frame Surveys. *Proceedings of the Social Statistics Section*, American Statistical Association, 203-206.

Hartley, H.O. (1974). Multiple frame methodology and selected apllications. *Sankhyā*, Series C, 36, 99-118.

Huygens, C. (1673). Horologium Oscillatorium sive de motu pendulorum.

Lavallée, P. (1995). Cross-sectional Weighting of Longitudinal Surveys of Individuals and Households Using the Weight Share Method. *Survey Methodology*, 21, 25-32.

Lavallée, P. (2002). Le sondage indirect, ou la méthode généralisée du partage des poids. Éditions de l'Université de Bruxelles, éditions Ellipses, Bruxelles.

Lavallée, P., and Caron, P. (2001). Estimation using the generalised weight share method: The case of record linkage. *Survey Methodology*, 27, 155-169.

Lund, R.E. (1968). Estimators in multiple frame surveys. *Proceedings of the Social Statistics Section*, American Statistical Association, 282-288.

Torres Manzanera, E., Sustacha Melijosa, I., Menéndez Estébanez, J.M. and Valdés Pelaáez, L. (2002). A solution to problems and disadvantages in statistical operations of surveys of visitors at accommodation establishments and at popular visitors places. (Éd. Ákos Probáld). *Proceedings Of The Sixth International Forum On Tourisme Statistics. Hungarian Central Statistical Office*, *Budapest*.

Valdés, L., De La Ballina, J., Aza, R., Loredo, E., Torres, E., Estébanez, J.M., Domínguez, J.S. and Del Valle, E. (2001). A methodology to measure tourism expenditure and total tourism production at the regional level. In *Tourism Statistics*: *International Perspectives and Current Issues*, (Ed. Lennon, J.J.), Continuum, Grande Bretagne, 317-334.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Survey Methodology, December 2006
Vol. 32, No. 2, pp. 187-195
Statistics Canada, Catalogue No. 12-001

Combining Link-Tracing Sampling and Cluster Sampling to Estimate the Size of Hidden Populations: A Bayesian-Assisted Approach

Martín H. Félix-Medina and Pedro E. Monjardin [1]

Abstract

Félix-Medina and Thompson (2004) proposed a variant of Link-tracing sampling in which it is assumed that a portion of the population, not necessarily the major portion, is covered by a frame of disjoint sites where members of the population can be found with high probabilities. A sample of sites is selected and the people in each of the selected sites are asked to nominate other members of the population. They proposed maximum likelihood estimators of the population sizes which perform acceptably provided that for each site the probability that a member is nominated by that site, called the nomination probability, is not small. In this research we consider Félix-Medina and Thompson's variant and propose three sets of estimators of the population sizes derived under the Bayesian approach. Two of the sets of estimators were obtained using improper prior distributions of the population sizes, and the other using Poisson prior distributions. However, we use the Bayesian approach only to assist us in the construction of estimators, while inferences about the population sizes are made under the frequentist approach. We propose two types of partly design-based variance estimators and confidence intervals. One of them is obtained using a bootstrap and the other using the delta method along with the assumption of asymptotic normality. The results of a simulation study indicate that (i) when the nomination probabilities are not small each of the proposed sets of estimators performs well and very similarly to maximum likelihood estimators; (ii) when the nomination probabilities are small the set of estimators derived using Poisson prior distributions still performs acceptably and does not have the problems of bias that maximum likelihood estimators have, and (iii) the previous results do not depend on the size of the fraction of the population covered by the frame.

Key Words: Bayesian approach; Capture-recapture; Design-based approach; Finite population; Hard-to-access population; Maximum likelihood; Model-based approach; Sampling frame.

1. Introduction

Link-tracing sampling (LTS) has been found appropriate for sampling hidden and hard-to-access human populations, such as drug-user, homeless-person, or illegal-worker populations. In this sampling method, an initial sample of people from the target population is selected, and the people in the initial sample are asked to nominate other members of the population. The nominated people who are not in the initial sample are included in the sample and they might be asked to nominate other persons. This process might continue until a specified stopping rule is satisfied (for a review of LTS, see Spreen 1992, and Thompson and Frank 2000).

Although LTS allows the sampler to make valid model-based inferences about a number of population parameters, in practical applications the assumptions about the initial sample are difficult to satisfy. (See Snijders 1992, Frank and Snijders 1994, and Heckathorn 2002). For instance, Frank and Snijders (1994) developed a variant of LTS in which the initial sample is a Bernoulli sample, that is, elements in the initial sample are included independently and with equal probabilities; however, in real studies the initial recruitment is generally carried out by using records of people obtained from health centers or police stations, and this induces a selection bias known as institutional bias.

The difficulty in satisfying, in practical situations, the assumptions about the initial sample motivated Félix-Medina and Thompson (2004) to develop a variant of LTS which does not require an initial Bernoulli sample. They assume that a portion, not necessarily the major portion, of the target population is covered by a sampling frame of accessible sites where members of the population can be found with high probability (for instance bars, hospitals, blocks or parks). A simple random sample of sites is selected, and the members that belong to each site are identified. Finally, as in ordinary LTS, the people in each site are asked to nominate other members of the population.

Those authors derived maximum likelihood estimators (MLEs) of the population sizes from probability models that describe both the number of elements found in each site and the probability that a member is nominated from a site, which is called the nomination probability. They also proposed model-based and partly design-based variance estimators, that is, estimators based on both the design used to select the initial sample and the assumed models. Throughout this paper we will call this type of estimator a "design-based-like" estimator. By a simulation study, the authors showed that the MLEs of the population sizes and their design-based-like variance estimators are robust to deviations from the assumed model, but that the model-based variance estimators are not robust. In addition, they

1. Martín H. Félix-Medina and Pedro E. Monjardin, Escuela de Ciencias Físico-Matemáticas, Universidad Autónoma de Sinaloa, Ciudad Universitaria, Culiacán Sinaloa, México.

found that the MLEs tend to seriously overestimate the population size if the nomination probabilities are small.

As indicated by those authors, the problem of overestimation that appears when the nomination probabilities are small is caused by the small amount of information contained in the sample, which is not enough to obtain stable estimates of the nomination probabilities. They suggest that a possible solution to this problem is to use the Bayesian approach to construct estimators that incorporate additional information about the population parameters.

In this work we use the Bayesian approach to assist us in the construction of estimators of the population sizes, while we make inferences under a frequentist approach. Thus, in addition to deriving point estimators we construct confidence intervals. For this purpose we use the strategy proposed by Félix-Medina and Thompson (2004) to construct confidence intervals based on the normal distribution and using design-based-like variance estimators obtained by the delta method. In addition, we construct design-based-like bootstrap confidence intervals. We have named this inferential approach "Bayesian-assisted".

2. Sampling Design and Notation

The structure of the population and sampling design considered in this paper are the same as those proposed by Félix-Medina and Thompson (2004). A brief description of them follows. Let $U = \{u_1, \ldots, u_\tau\}$ be a hidden human population of unknown size τ. Let U_1 be a subset of U formed by an unknown number τ_1 of people that can be found in different accessible sites, such as bars, parks, or blocks. Two assumptions about this sampling design are that a sampling frame of N of those sites can be constructed, and that the researcher has an operational rule which allows him or her to determine whether or not a person belongs to a site in the frame and, in the affirmative case, to locate that site. Notice that the subset U_1 covered by the frame is not assumed to be the major part of U and that, as in ordinary cluster sampling, a person in the frame is assumed to belong to only one site. Let A_i be the i–th site or cluster in the frame and m_i be the number of people who belong to $A_i, i = 1, \ldots, N$; then $\tau_1 = \sum_1^N m_i$. Finally, let $U_2 = U - U_1$ be the portion of U not covered by the frame and let $\tau_2 = \tau - \tau_1$ be its size.

The sampling design is as follows. A sample $S_0 = \{A_1, \ldots, A_n\}$ of n clusters is selected from the frame by simple random sampling without replacement, and the m_i persons who belong to each $A_i \in S_0$ are identified. Note that we have used the subscripts $1, \ldots, n$ to denote the clusters in S_0; however, this does not mean that the first n clusters in the frame are necessarily the clusters in the sample. Next, the people in the sampled cluster A_i are asked to nominate members in U, but only nominees in $U - A_i$ are considered. This procedure is repeated for every cluster $A_i \in S_0$. As a convention, we will say that a person is nominated by a cluster if he or she is nominated by at least one member of that cluster. Nominations from different clusters are carried out independently, and different nomination strategies can be used in different sites. For instance, in site A_i the m_i members, as a group, could carry out the nominations; whereas in another site A_j each of the m_j members could make nominations separately. Finally, for each nominee the researcher has to register the site or sites that nominated him or her, and the section U_1 or U_2, to which the nominee belongs. Notice that this last piece of information could be obtained from the person who made the nomination or, if that is not possible, from an interview with the nominee.

The nomination of people by clusters will be indicated by the matrices $\mathbf{X}_1 = [x_{ij}^{(1)}]_{n\times\tau_1}$ and $\mathbf{X}_2 = [x_{ij}^{(2)}]_{n\times\tau_2}$, where $x_{ij}^{(1)} = 1$ if person $u_j \in U_1 - A_i$ is nominated by cluster A_i, and $x_{ij}^{(1)} = 0$ if $u_j \in A_i$ or u_j is not nominated by A_i. Similarly, $x_{ij}^{(2)} = 1$ if person $u_j \in U_2$ is nominated by cluster A_i, and $x_{ij}^{(2)} = 0$ otherwise. As noted by Félix-Medina and Thompson (2004), $\mathbf{X}_1$ and $\mathbf{X}_2$ are only known up to permutations of their columns because the people are not labelled. Therefore, inferences about τ_1 and τ_2 are based on the set of counts $\mathbf{y} = \{y_\omega\}$, where y_ω, $\omega \subseteq \mathbf{\Omega} = \{1, \ldots, n\}, \omega \neq \varnothing$, indicates the number of people in U who are nominated by every sampled cluster A_i with i in the set ω, but not otherwise. For instance, if $\omega = \{4, 7, 8\}$, y_ω would be the number of people in U who are nominated by only A_4, A_7 and A_8.

3. Estimators of the Population Sizes Based on Posterior Modes

Félix-Medina and Thompson noted the resemblance between their sampling design and that of multiple capture-recapture sampling (MCRS). This makes it possible to apply to our case some of the Bayesian models that have been proposed for analyzing MCRS. See Fienberg, Johnson and Junker (1999) for a review of Bayesian analyses of MCRS. In this work, we use a model considered by Castledine (1981) for the prior distributions of the logits of the nomination probabilities, along with some models for the prior distributions of the population sizes.

As in Félix-Medina and Thompson (2004), we will suppose that the sizes $m_1, \ldots, m_N$ of the clusters $A_1, \ldots, A_N$ are realizations of independent Poisson random variables $M_1, \ldots, M_N$ with mean λ_1. We will denote by $p_i^{(k)}$ the probability that a person in $U_k - A_i$ is nominated by the site $A_i \in S_0$. The probabilities $p_i^{(k)}$ will be called nomination probabilities. In addition, we will suppose that conditionally on the sizes $m_1, \ldots, m_n$ of the clusters in S_0, on τ_1 and τ_2,

and on the $p_i^{(k)}$'s, the variables $x_{ij}^{(k)}$ are realizations of independent Bernoulli random variables $X_{ij}^{(k)}$ with means $p_i^{(k)}$, $i = 1, \ldots, n$ and $k = 1, 2$.

Félix-Medina and Thompson (2004) used the fact that the joint conditional distribution of $(M_1, \ldots, M_n, \tau_1 - \sum_1^n M_i)$, given that $\sum_1^N m_i = \tau_1$, is a multinomial distribution with parameters τ_1 and $(1/N, \ldots, 1/N, 1 - n/N)$, and applied a procedure used by Darroch (1958) to show that the likelihood function of $\tau_1, \tau_2, \mathbf{p}_1 = \{p_i^{(1)}\}_1^n$ and $\mathbf{p}_2 = \{p_i^{(2)}\}_1^n$ is the product of the following factors:

$$f(\mathbf{m}_s|\tau_1) = \frac{\tau_1!}{(\tau_1 - m)! \prod_1^n m_i!} (1/N)^m (1 - n/N)^{\tau_1 - m}$$

$$f(\mathbf{y}^{(1-0)}|\mathbf{m}_s, \tau_1, \mathbf{p}_1) = \frac{(\tau_1 - m)!}{(\tau_1 - m - r_1)! \prod_{\omega \neq \varnothing} y_\omega^{(1-0)}!} \prod_{i=1}^n [p_i^{(1)}]^{z_i^{(1-0)}} \times [1 - p_i^{(1)}]^{\tau_1 - m - z_i^{(1-0)}}$$

$$f(\mathbf{y}^{(A_1)}, \ldots, \mathbf{y}^{(A_n)}|\mathbf{m}_s, \mathbf{p}_1) = \prod_{i=1}^n \frac{m_i!}{(m_i - w_i)! \prod_{\omega \neq \varnothing} y_\omega^{(A_i)}!} [p_i^{(1)}]^{z_i^{(0)}} \times [1 - p_i^{(1)}]^{m - m_i - z_i^{(0)}}$$

$$f(\mathbf{y}^{(2)}|\mathbf{m}_s, \tau_2, \mathbf{p}_2) = \frac{\tau_2!}{(\tau_2 - r_2)! \prod_{\omega \neq \varnothing} y_\omega^{(2)}!} \prod_{i=1}^n [p_i^{(2)}]^{z_i^{(2)}} [1 - p_i^{(2)}]^{\tau_2 - z_i^{(2)}},$$

where $\mathbf{m}_s = \{m_i\}_1^n; m = \sum_1^n m_i$ is the observed value of the random variable M that indicates the number of people in $S_0; \mathbf{y}^{(1-0)} = \{y_\omega^{(1-0)}\}_{\omega \neq \varnothing}, \mathbf{y}^{(2)} = \{y_\omega^{(2)}\}_{\omega \neq \varnothing}$, and $\mathbf{y}^{(A_i)} = \{y_\omega^{(A_i)}\}_{\omega \neq \varnothing}, A_i \in S_0$, are the sets of counts obtained from $\mathbf{y}$, that correspond to the counts of nominated people in $U_1 - S_0, U_2$ and $A_i \in S_0$, respectively; $z_i^{(0)} = \sum_{j \neq i}^n \sum_{\omega \supset i} y_\omega^{(A_j)}, z_i^{(1-0)} = \sum_{\omega \supset i} y_\omega^{(1-0)}$ and $z_i^{(2)} = \sum_{\omega \supset i} y_\omega^{(2)}$ are the observed values of the random variables $Z_i^{(0)}, Z_i^{(1-0)}$ and $Z_i^{(2)}$ that indicate the numbers of distinct people in $S_0, U_1 - S_0$ and U_2, respectively, that are nominated by A_i; and $r_1 = \sum_{\omega \neq \varnothing} y_\omega^{(1-0)}, r_2 = \sum_{\omega \neq \varnothing} y_\omega^{(2)}$ and $w_i = \sum_{\omega \neq \varnothing} y_\omega^{(A_i)}$ are the observed values of the random variables R_1, R_2 and W_i that indicate the numbers of distinct people in $U_1 - S_0, U_2$ and A_i, respectively, that are nominated by at least one of the clusters in S_0.

We will now focus on the problem of defining the prior distributions of $\tau_1, \tau_2, \mathbf{p}_1$ and $\mathbf{p}_2$. In the case of τ_1 and τ_2, we will consider the following three models for the prior distributions:

Poisson-Gamma Distributions

$\pi(\tau_1|\lambda_1) \propto (N\lambda_1)^{\tau_1}/\tau_1!$ and $\pi(\lambda_1) \propto \lambda_1^{a_1 - 1} e^{-b_1\lambda_1}$, $\pi(\tau_2|\lambda_2) \propto \lambda_2^{\tau_2}/\tau_2!$ and $\pi(\lambda_2) \propto \lambda_2^{a_2 - 1} e^{-b_2\lambda_2}$, where a_1, b_1, a_2, b_2 are known constants, and (τ_1, λ_1) and (τ_2, λ_2) are independent.

Jeffreys' Distributions

$\pi(\tau_k) \propto 1/\tau_k$, where $k = 1, 2$, and τ_1 and τ_2 are independent random variables.

Uniform Distributions

$\pi(\tau_k) \propto 1$, where $k = 1, 2$, and τ_1 and τ_2 are independent random variables.

The prior Poisson distribution of τ_1 defined in the first case is motivated by the fact that $\tau_1 = \sum_1^N M_i$, and that M_i is a Poisson variable with mean λ_1. Notice that this case allows the researcher to use information about τ_1 and τ_2 which is known prior to the observation of the sample. On the other hand, the distributions defined in the other two cases are not informative.

In the case of the nomination probabilities $p_i^{(k)}$'s, following Castledine (1981), we will suppose that the $p_i^{(k)}$'s are exchangeable and will use his two-stage normal model for the logits $\alpha_i^{(k)} = \log[p_i^{(k)}/(1 - p_i^{(k)})]$ of the $p_i^{(k)}$'s:

$$\alpha_i^{(k)}|\theta_k \sim N(\theta_k, \sigma_k^2),$$

$$\text{and } \theta_k \sim N(\mu_k, \gamma_k^2); i = 1, \ldots, n, \ k = 1, 2, \tag{1}$$

where $N(\theta_k, \sigma_k^2)$ stands for the normal distribution with mean θ_k and variance $\sigma_k^2; \sigma_k^2, \mu_k$ and γ_k^2 are known constants; and the $\alpha_i^{(k)}$'s are conditionally independent given θ_k. Under the assumption of exhangeability the $\alpha_i^{(k)}$'s are not independent, but information about any one of them is used to obtain information about any other of the $\alpha_i^{(k)}$'s. Of course, if we wanted independent priors for the $\alpha_i^{(k)}$'s, we could obtain a one-stage normal model from (1) by setting $\theta_k = \mu_k$ and $\gamma_k^2 = 0, k = 1, 2$.

Finally, we will suppose that all the random vectors (τ_k, λ_k) and (α_k, θ_k), where $\alpha_k = (\alpha_1^{(k)}, \ldots, \alpha_n^{(k)})$, $k = 1, 2$, are mutually independent.

Although we defined three types of prior distributions for τ_1 and τ_2, they can be treated in a unified way because the prior marginal distributions of τ_1 and τ_2, obtained from the Poisson-Gamma distributions, are the Negative binomial distributions:

$$\pi(\tau_1) \propto \frac{\Gamma(\tau_1 + a_1)}{\tau_1!} \left(\frac{N}{N + b_1}\right)^{\tau_1}$$

$$\text{and} \qquad \pi(\tau_2) \propto \frac{\Gamma(\tau_2 + a_2)}{\tau_2!} \left(\frac{1}{1 + b_2}\right)^{\tau_2}, \tag{2}$$

where $\Gamma(\cdot)$ denotes the Gamma function. The Jeffreys' and Uniform distributions are limiting cases of (2) obtained by making $a_k = b_k = 0, k = 1, 2$, and $a_k = 1, b_k = 0, k = 1, 2$,

respectively. Note that the Gamma distribution is not defined for these values of a_k and b_k; however, for the derivation of the estimators we can use these values in (2).

The posterior joint distribution of τ_1, τ_2, α_1, and α_2 can be expressed as

$$\pi(\tau_1, \tau_2, \alpha_1, \alpha_2 \,|\text{data})$$

$$\propto \frac{(N-n)^{\tau_1}\Gamma(\tau_1+a_1)}{(\tau_1-m-r_1)!(N+b_1)^{\tau_1}} \prod_{i=1}^{n} \frac{\exp[\alpha_i^{(1)} z_i^{(1)}]}{\left[1+\exp[\alpha_i^{(1)}]\right]^{\tau_1-m_i}}$$

$$\times \exp\left[\begin{array}{l} -\dfrac{\sum_{i=1}^{n}(\alpha_i^{(1)}-\bar{\alpha}^{(1)})^2}{2\sigma_1^2} \\ -\dfrac{(\bar{\alpha}^{(1)}-\mu_1)^2}{2v_1} \end{array}\right] \frac{\Gamma(\tau_2+a_2)}{(\tau_2-r_2)!(b_2+1)^{\tau_2}}$$

$$\times \prod_{i=1}^{n} \frac{\exp[\alpha_i^{(2)} z_i^{(2)}]}{\left[1+\exp[\alpha_i^{(2)}]\right]^{\tau_2}} \exp\left[\begin{array}{l} -\dfrac{\sum_{i=1}^{n}(\alpha_i^{(2)}-\bar{\alpha}^{(2)})^2}{2\sigma_2^2} \\ -\dfrac{(\bar{\alpha}^{(2)}-\mu_2)^2}{2v_2} \end{array}\right] \qquad (3)$$

where $z_i^{(1)} = z_i^{(0)} + z_i^{(1-0)}$ is the observed value of the random variable $Z_i^{(1)} = Z_i^{(0)} + Z_i^{(1-0)}$ that indicates the number of distinct people in U_1, either in S_0 or in $U_1 - S_0$, that are nominated by A_i; $\bar{\alpha}^{(k)}$ is the arithmetic mean of the $\alpha_i^{(k)}$; and $v_k = \gamma_k^2 + \sigma_k^2/n$, $k = 1, 2$.

Since we cannot compute the analytical integral of (3) with respect to $\alpha_i^{(1)}$ and $\alpha_i^{(2)}$, we will not try to obtain expressions for the posterior distributions of τ_1 and τ_2, but, as in Castledine (1981), we will use the mode of $\pi(\tau_1, \tau_2, \alpha_1, \alpha_2|\text{data})$ as an estimator of $(\tau_1, \tau_2, \alpha_1, \alpha_2)$. Using this strategy, we have that the proposed estimator is the solution to the system of equations:

$$\hat{\tau}_1 = \frac{M + R_1 + (1-n/N)[N(a_1-1)/(N+b_1)]\prod_{i=1}^{n}(1-\hat{p}_i^{(1)})}{1-(1-n/N)[N/(N+b_1)]\prod_{i=1}^{n}(1-\hat{p}_i^{(1)})};$$

$$\hat{p}_i^{(1)} = \frac{\exp\{\hat{\alpha}_i^{(1)}\}}{1+\exp\{\hat{\alpha}_i^{(1)}\}} = \frac{Z_i^{(1)}}{\hat{\tau}_1 - M_i} - \frac{\hat{\alpha}_i^{(1)} - \hat{\bar{\alpha}}^{(1)}}{(\hat{\tau}_1 - M_i)\sigma_1^2}$$

$$- \frac{\hat{\bar{\alpha}}^{(1)} - \mu_1}{n(\hat{\tau}_1 - M_i)v_1}; \; i = 1, \ldots, n; \qquad (4)$$

$$\hat{\tau}_2 = \frac{R_2 + [(a_2-1)/(1+b_2)]\prod_{i=1}^{n}(1-\hat{p}_i^{(2)})}{1-[1/(1+b_2)]\prod_{i=1}^{n}(1-\hat{p}_i^{(2)})};$$

$$\hat{p}_i^{(2)} = \frac{\exp\{\hat{\alpha}_i^{(2)}\}}{1+\exp\{\hat{\alpha}_i^{(2)}\}} = \frac{Z_i^{(2)}}{\hat{\tau}_2} - \frac{\hat{\alpha}_i^{(2)} - \hat{\bar{\alpha}}^{(2)}}{\hat{\tau}_2\sigma_2^2}$$

$$- \frac{\hat{\bar{\alpha}}^{(2)} - \mu_2}{n\hat{\tau}_2 v_2}; \; i = 1, \ldots, n; \qquad (5)$$

where $\hat{\bar{\alpha}}^{(k)} = \sum_1^n \hat{\alpha}_i^{(k)}/n, k = 1, 2$. From this, an estimator of τ is $\hat{\tau} = \hat{\tau}_1 + \hat{\tau}_2$.

The forms of these estimators are basically adjustments to the forms of the MLE's proposed by Félix-Medina and Thompson (2004) so that the proposed estimators incorporate the initial information about τ_k and $\alpha_i^{(k)}$, $i = 1, \ldots, n$; $k = 1, 2$. Also, as a referee has noted, the estimator $\hat{p}_i^{(k)}$ has the form of the MLE of $p_i^{(k)}$ followed by shrinkage terms, one of $\alpha_i^{(k)}$ toward the arithmetic mean $\hat{\bar{\alpha}}^{(k)}$, and another of $\hat{\bar{\alpha}}^{(k)}$ toward the prior mean μ_k.

4. Confidence Intervals for the Population Sizes

As was indicated earlier, we will use the frequentist approach to obtain design-based-like confidence intervals that are robust to deviations from the assumed Poisson distribution of the M_i's. We will consider bootstrap intervals and Wald intervals based on a normal approximation (see Agresti 2002, page 13 and Evans, Kim and O'Brien 1996 for the latter terminology).

4.1 Bootstrap Confidence Intervals

We will use a version of the bootstrap obtained by combining the bootstrap variant for finite populations proposed by Booth, Butler and Hall (1994) and the parametric bootstrap variant (see Davison and Hinkley 1997, Chapter 2).

The steps of the procedure that we propose are the following. (i) Construct an artificial population of N values of m_i's by repeating N/n times, assuming that N/n is an integer, the selected sample of n cluster sizes $m_1, \ldots, m_n$. If $N = kn + r$, where k and r are positive integers, construct the population by repeating k times the selected sample of n cluster sizes and add to this set of m_i's a simple random sample without replacement (SRSWOR) of r values of m_i's selected from the observed sample of n cluster sizes. (ii) Select a SRSWOR of size n from the population of the m_i's. Let $i_1, \ldots, i_n$ in be the indices of the m_i's in the sample. (iii) For each $i = i_1, \ldots, i_n$, draw samples of sizes $\hat{\tau}_1 - m_i$ and $\hat{\tau}_2$ from Bernoulli distributions with means $\hat{p}_i^{(1)}$ and $\hat{p}_i^{(2)}$, respectively, where $\hat{\tau}_1, \hat{\tau}_2, \hat{p}_i^{(1)}$ and $\hat{p}_i^{(2)}$ are

the estimates of $\tau_1, \tau_2, p_i^{(1)}$ and $p_i^{(2)}$ computed from the original observed sample. These samples simulate the values of the sets $\{x_{ij}^{(1)}\}$ and $\{x_{ij}^{(2)}\}$ of indicator variables. (iv) Compute estimates of τ_1, τ_2 and τ from the samples drawn in steps (ii) and (iii) using the same procedure as that used to compute the original estimates $\hat{\tau}_1, \hat{\tau}_2$ and $\hat{\tau}$. (v) Obtain the bootstrap distributions of $\hat{\tau}_1, \hat{\tau}_2$ and $\hat{\tau}$ by repeating (i)–(iv) a large number B of times, and computing the empirical distributions from the sets of B values of $\hat{\tau}_1, \hat{\tau}_2$ and $\hat{\tau}$. (vi) Construct the $100(1-\alpha)\%$ bootstrap confidence intervals for τ_1, τ_2 and τ by using either the basic or the percentile method (see Davison and Hinkley 1997, Chapter 5, for descriptions of these methods). In the basic method the interval for τ is $[2\hat{\tau} - \hat{\tau}^{(1-\alpha/2)}, 2\hat{\tau} - \hat{\tau}^{(\alpha/2)}]$, and in the percentile method it is $[\hat{\tau}^{(\alpha/2)}, \hat{\tau}^{(1-\alpha/2)}]$, where $\hat{\tau}^{(\alpha/2)}$ and $\hat{\tau}^{(1-\alpha/2)}$ are the lower and upper $\alpha/2$ points of the bootstrap distribution of the original estimate $\hat{\tau}$ of τ.

Note that this variant of the bootstrap does not use the assumed Poisson distribution of the M_i's, but it uses the sampling design employed to select the initial sample of clusters. Thus, we can consider that the resulting confidence intervals are robust to deviations from the assumed distribution of the M_i's.

If bootstrap estimates of the variances of $\hat{\tau}_1, \hat{\tau}_2$ and $\hat{\tau}$ were also desired, simple estimates could be obtained by computing the sample variances of the sets of B values of those estimators.

4.2 Wald Confidence Intervals

Though in this work we will not justify theoretically that the proposed estimators of the population sizes are asymptotically normally distributed, we will suppose that the normal distribution is a reasonable approximation to the distributions of the estimators. Thus, we will construct $100(1-\alpha)\%$ design-based-like Wald confidence intervals for the population sizes, which have the form $\hat{\tau}_k \pm z_{1-\alpha/2}\sqrt{\hat{\mathbf{V}}(\hat{\tau}_k)}$, where $z_{1-\alpha/2}$ is the upper $\alpha/2$ point of the standard normal distribution, and $\hat{\mathbf{V}}(\hat{\tau}_k)$ is a design-based-like estimator of the variance of $\hat{\tau}_k$.

To construct this type of interval, we will firstly derive design-based-like variance estimators by applying the same strategy as that used by Félix-Medina and Thompson (2004). In that strategy, the distribution of the cluster sizes is not employed, but it is replaced by the distribution of the sampling design used to select the initial sample S_0. This is carried out by means of the formula:

$$\mathbf{V}(\hat{\tau}_k) = \mathbf{V}_p[\mathbf{E}_\xi(\hat{\tau}_k|\mathbf{m}_s)] + \mathbf{E}_p[\mathbf{V}_\xi(\hat{\tau}_k|\mathbf{m}_s)], \qquad (6)$$

where $\mathbf{E}_\xi(\hat{\tau}_k|\mathbf{m}_s)$ and $\mathbf{V}_\xi(\hat{\tau}_k|\mathbf{m}_s)$ denote the conditional model-based expectation and variance operators, given that $\mathbf{M}_s = \mathbf{m}_s$; and $\mathbf{E}_p(\cdot)$ and $\mathbf{V}_p(\cdot)$ denote the design-based expectation and variance operators. Thus, the variance estimators are obtained by applying (6) to the first-order Taylor's approximations $\hat{\tau}_1^*$ and $\hat{\tau}_2^*$ of $\hat{\tau}_1$ and $\hat{\tau}_2$, respectively, about the model-based expectations of $c_s^{(1)} = (\mathbf{M}_s, \mathbf{Z}_s^{(1)}, R_1)$ and $c_s^{(2)} = (\mathbf{Z}_s^{(2)}, R_2)$, where $\mathbf{Z}_s^{(k)} = (Z_1^{(k)}, \ldots Z_n^{(k)})$, $k = 1, 2$.

Using the previously described strategy, and the fact that $Z_i^{(1)}|\mathbf{m}_s \sim \text{bin}(\tau_1 - m_i, p_i^{(1)})$ and $R_1|\mathbf{m}_s \sim \text{bin}(\tau_1 - m, 1 - Q_1)$, where $Q_1 = \prod_{i=1}^n (1 - p_i^{(1)})$, we have that an estimator of $\mathbf{V}_p[\mathbf{E}_\xi(\hat{\tau}_1^*|\mathbf{m}_s)]$ is

$$\hat{\mathbf{V}}_{11} = n(1 - n/N)\hat{K}^2 \frac{1}{n-1}\sum_{i=1}^n (m_i - \bar{m})^2, \qquad (7)$$

where $\bar{m} = n^{-1}\sum_1^n m_i$; $\hat{K} = -\hat{Q}_1/[\hat{A}_1(\hat{\tau}_1 - m - r_1)]$; $\hat{Q}_1 = \prod_{i=1}^n (1 - \hat{p}_i^{(1)})$;

$$\hat{A}_1 = \sum_{i=1}^n \frac{(\hat{p}_i^{(1)})^2}{\hat{B}_i^{(1)}} - \hat{C}_1 + \frac{1}{\hat{\tau}_1 + a_1 - 1} - \frac{1}{\hat{\tau}_1 - m - r_1};$$

$$\hat{B}_i^{(1)} = (\hat{\tau}_1 - m_i)\hat{p}_i^{(1)}(1 - \hat{p}_i^{(1)}) + \sigma_1^{-2}, \; i = 1, \ldots, n;$$

and

$$\hat{C}_1 = \frac{(v_1^{-1} - n\sigma_1^{-2})\left[n^{-1}\sum_{i=1}^n \hat{p}_i^{(1)}/\hat{B}_i^{(1)}\right]^2}{1 + n^{-1}(v_1^{-1} - n\sigma_1^{-2})n^{-1}\sum_{i=1}^n 1/\hat{B}_i^{(1)}}. \qquad (8)$$

In addition, since $\mathbf{Cov}(Z_i^{(1)}, R_1|\mathbf{m}_s) = (\tau_1 - m)Q_1 p_i^{(1)}$, an estimator of $\mathbf{E}_p[\mathbf{V}_\xi(\hat{\tau}_1^*|\mathbf{m}_s)]$ is

$$\hat{\mathbf{V}}_{12} = \hat{A}_1^{-2}\left\{\begin{array}{l} \displaystyle\sum_{i=1}^n \left(\frac{\hat{p}_i^{(1)} - \hat{D}_1}{\hat{B}_i^{(1)}}\right)^2 (\hat{\tau}_1 - m_i)\hat{p}_i^{(1)}(1 - \hat{p}_i^{(1)}) \\ \displaystyle + \frac{(\hat{\tau}_1 - m)\hat{Q}_1(1 - \hat{Q}_1)}{(\hat{\tau}_1 - m - r_1)^2} \\ \displaystyle - \frac{2(\hat{\tau}_1 - m)\hat{Q}_1}{\hat{\tau}_1 - m - r_1}\sum_{i=1}^n \left(\frac{\hat{p}_i^{(1)} - \hat{D}_1}{\hat{B}_i^{(1)}}\right)\hat{p}_i^{(1)} \end{array}\right\}, \qquad (9)$$

where

$$\hat{D}_1 = \frac{n^{-1}(v_1^{-1} - n\sigma_1^{-2})n^{-1}\sum_{i=1}^n \hat{p}_i^{(1)}/\hat{B}_i^{(1)}}{1 + n^{-1}(v_1^{-1} - n\sigma_1^{-2})n^{-1}\sum_{i=1}^n 1/\hat{B}_i^{(1)}}.$$

Therefore, a design-based-like estimator of $\mathbf{V}(\hat{\tau}_1)$ is $\hat{\mathbf{V}}(\hat{\tau}_1) = \hat{\mathbf{V}}_{11} + \hat{\mathbf{V}}_{12}$.

In the case of $\hat{\tau}_2^*$, since $Z_i^{(2)}|\mathbf{m}_s \sim \text{bin}(\tau_2, p_i^{(2)})$ and $R_2|\mathbf{m}_s \sim \text{bin}(\tau_2, 1 - Q_2)$, where $Q_2 = \prod_{i=1}^n (1 - p_i^{(2)})$, it follows that $\mathbf{E}_\xi(\hat{\tau}_2^*|\mathbf{m}_s)$ does not depend on $\mathbf{m}_s$, and consequently that $\mathbf{V}_p[\mathbf{E}_\xi(\hat{\tau}_2^*|\mathbf{m}_s)] \approx 0$. Therefore, since $\mathbf{Cov}(Z_i^{(2)}, R_2|\mathbf{m}_s) = \tau_2 Q_2 p_i^{(2)}$, an estimator of $\mathbf{V}(\hat{\tau}_2)$ is

$$\hat{\mathbf{V}}(\hat{\tau}_2) = \hat{A}_2^{-2} \left\{ \begin{array}{l} \displaystyle\sum_{i=1}^{n} \left(\frac{\hat{p}_i^{(2)} - \hat{D}_2}{\hat{B}_i^{(2)}} \right)^2 \hat{\tau}_2 \hat{p}_i^{(2)} (1 - \hat{p}_i^{(2)}) \\ + \dfrac{\hat{\tau}_2 \hat{Q}_2 (1 - \hat{Q}_2)}{(\hat{\tau}_2 - r_2)^2} \\ - \dfrac{2\hat{\tau}_2 \hat{Q}_2}{\hat{\tau}_2 - r_2} \displaystyle\sum_{i=1}^{n} \left(\frac{\hat{p}_i^{(2)} - \hat{D}_2}{\hat{B}_i^{(2)}} \right) \hat{p}_i^{(2)} \end{array} \right\} \qquad (10)$$

where $\hat{Q}_2 = \prod_{i=1}^{n} (1 - \hat{p}_i^{(2)})$,

$$\hat{A}_2 = \sum_{i=1}^{n} \frac{(\hat{p}_i^{(2)})^2}{\hat{B}_i^{(2)}} - \hat{C}_2 + \frac{1}{\hat{\tau}_2 + a_2 - 1} - \frac{1}{\hat{\tau}_2 - r_2},$$

$$\hat{B}_i^{(2)} = \hat{\tau}_2 \hat{p}_i^{(2)} (1 - \hat{p}_i^{(2)}) + \sigma_2^{-2}, \quad i = 1, \ldots, n,$$

$$\hat{C}_2 = \frac{(v_2^{-1} - n\sigma_2^{-2}) \left[n^{-1} \sum_{i=1}^{n} \hat{p}_i^{(2)} / \hat{B}_i^{(2)} \right]^2}{1 + n^{-1} (v_2^{-1} - n\sigma_2^{-2}) n^{-1} \sum_{i=1}^{n} 1 / \hat{B}_i^{(2)}},$$

and

$$\hat{D}_2 = \frac{n^{-1} (v_2^{-1} - n\sigma_2^{-2}) n^{-1} \sum_{i=1}^{n} \hat{p}_i^{(2)} / \hat{B}_i^{(2)}}{1 + n^{-1} (v_2^{-1} - n\sigma_2^{-2}) n^{-1} \sum_{i=1}^{n} 1 / \hat{B}_i^{(2)}}.$$

Finally, since the no dependency of $\mathbf{E}_{\xi}(\hat{\tau}_2^* | \mathbf{m}_s)$ on $\mathbf{m}_s$ implies that $\mathbf{Cov}(\hat{\tau}_1^*, \hat{\tau}_2^*) \approx 0$, it follows that a variance estimator of $\hat{\tau}$ is $\hat{\mathbf{V}}(\hat{\tau}) = \hat{\mathbf{V}}(\hat{\tau}_1) + \hat{\mathbf{V}}(\hat{\tau}_2)$.

5. Monte Carlo Study

We considered four populations; a description of each one is presented in Table 1. In the pair formed by Populations I and II the frame covered about 45% of the population, whereas in the pair formed by Populations III and IV the frame covered about 70% of the population. The populations of each pair were very similar, except that in one of the populations of each pair the distribution of the M_i's was Poisson, whereas in the other it was Negative Binomial. The nomination probabilities $p_i^{(k)}, i = 1, \ldots, N$, $k = 1, 2$, were generated using the model $p_i^{(k)} = 1 - \exp(-\beta_k m_i)$, where the values of β_k were set so that the following values of $\bar{p}^{(k)} = \sum_1^N p_i^{(k)} / N$ were obtained. For Populations I and II: $(\bar{p}^{(1)}, \bar{p}^{(2)}) \approx (0.05, 0.01)$ and $(\bar{p}^{(1)}, \bar{p}^{(2)}) \approx (0.01, 0.002)$. For Populations III and IV: $(\bar{p}^{(1)}, \bar{p}^{(2)}) \approx (0.05, 0.03)$ and $(\bar{p}^{(1)}, \bar{p}^{(2)}) \approx (0.01, 0.006)$. The model employed to generate the $p_i^{(k)}$'s is a model used in catch-effort methods (see Seber 1982, Chapter 7 for a description of those methods). As an associate editor has noted, this model implies that the number of people nominated by cluster A_i has expectation $(\tau_1 - m_i)(1 - \exp(-\beta_1 m_i)) + \tau_2 (1 - \exp(-\beta_2 m_i))$, and consequently the number of nominated people is approximately proportional to m_i. Notice that the assumed exchangeable model for $p_i^{(k)}$ does not entail such a relationship with m_i. Since the estimation of $p_i^{(k)}$ depends mainly on $z_i^{(k)}$, the number of people in U_k nominated by the cluster A_i, we expect the omission of this relationship not to affect the efficiency of the estimator of $p_i^{(k)}$. Darroch (1958) has shown, in the case of maximum likelihood estimation, that no significant gain is obtained by assuming the catch-effort model.

Table 1
Parameters of Simulated Populations

Population I	Population II	Population III	Population IV
$N = 250$	$N = 250$	$N = 250$	$N = 250$
M_i Poisson	M_i Neg. Binomial	M_i Poisson	M_i Neg. Binomial
$\mathbf{E}(M_i) = 7.2$	$\mathbf{E}(M_i) = 7.2$	$\mathbf{E}(M_i) = 7.2$	$\mathbf{E}(M_i) = 7.2$
$\mathbf{V}(M_i) = 7.2$	$\mathbf{V}(M_i) = 24.48$	$\mathbf{V}(M_i) = 7.2$	$\mathbf{V}(M_i) = 24.48$
$\tau_1 = 1{,}811$	$\tau_1 = 1{,}872$	$\tau_1 = 1{,}811$	$\tau_1 = 1{,}872$
$\tau_2 = 2{,}200$	$\tau_2 = 2{,}200$	$\tau_2 = 700$	$\tau_2 = 700$
$\tau = 4{,}011$	$\tau = 4{,}072$	$\tau = 2{,}511$	$\tau = 2{,}572$
$\tau_1 / \tau = 0.45$	$\tau_1 / \tau = 0.46$	$\tau_1 / \tau = 0.72$	$\tau_1 / \tau = 0.73$

For Populations I and II the values of the parameters of the prior distributions were $\sigma_k^2 = 25$, $\mu_k = -3.5$, $\gamma_k^2 = 25$, $k = 1, 2$, $a_1 = 1$, $b_1 = 0.1$, $a_2 = 7.84$, $b_2 = 0.0028$, so that $\mathbf{E}(\lambda_1) = 10$, $\mathbf{V}(\lambda_1) = 100$, $\mathbf{E}(\lambda_2) = 2{,}800$, and $\mathbf{V}(\lambda_2) = 10^6$. For Populations III and IV the values of the parameters were $\sigma_k^2 = 9$, $\mu_k = -3.5$, $\gamma_k^2 = 9$, $k = 1, 2$, $a_1 = 1$, $b_1 = 0.1$, $a_2 = 8$, $b_2 = 0.01$, so that $\mathbf{E}(\lambda_1) = 10$, $\mathbf{V}(\lambda_1) = 100$, $\mathbf{E}(\lambda_2) = 800$, and $\mathbf{V}(\lambda_2) = 80{,}000$. These values imply that the prior distributions are well dispersed over relatively large intervals that contain the parameters of interest.

The simulation experiment was carried out as follows. From each population of $N = 250$ values of m_i's, a SRSWOR of $n = 25$ values was selected. From cluster A_i in the sample, the values of $X_{ij}^{(1)}$ and $X_{ij}^{(2)}$ were generated by drawing samples of sizes $\tau_1 - m_i$ and τ_2 from Bernoulli distributions with means $p_i^{(1)}$ and $p_i^{(2)}$, respectively. These data were used to compute the following estimators of the population sizes: the set of MLEs $\hat{\tau}_1, \hat{\tau}_2$, and $\hat{\tau} = \hat{\tau}_1 + \hat{\tau}_2$ proposed by Félix-Medina and Thompson (2004); and the three sets of Bayesian estimators $\hat{\tau}_1^a, \hat{\tau}_2^a$, and $\hat{\tau}^a = \hat{\tau}_1^a + \hat{\tau}_2^a$, $a = U, J, P$, obtained by using as prior distributions the Uniform (U), Jeffreys' (J), and Poisson (P) distributions, respectively. In addition, variance estimators and confidence intervals were also computed. Bootstrap intervals were computed by the basic method, with the exception of the intervals based on the estimators $\hat{\tau}_1^P, \hat{\tau}_2^P$ and $\hat{\tau}^P$, which were computed by the percentile method. All bootstrap estimators were obtained by using 2,000 bootstrap samples. Finally, the performance of the point and

interval estimators was evaluated by using $r = 10,000$ trials of the previous procedure.

The performance of an estimator $\hat{\tau}$, say, was evaluated by its relative bias and the square root of its relative mean square error, defined as $\mathrm{r-bias} = \sum_1^r (\hat{\tau}_i - \tau)/(r\tau)$ and $\sqrt{\mathrm{r-mse}} = \sqrt{\sum_1^r (\hat{\tau}_i - \tau)^2/(r\tau^2)}$, where $\hat{\tau}_i$ was the value of $\hat{\tau}$ obtained in the i–th trial. The performance of a variance estimator was also evaluated by its relative bias and the square root of its relative mean square error, which were similarly defined to those of an estimator of the population size, but using the empirically determined variance instead of the real variance. Finally, the performance of the 95% confidence intervals was evaluated by their coverage probabilities and their average lengths.

6. Results and Discussion

Because of restrictions of space, in Tables 2 to 4 we present only a selection of the results of the numerical study. However, the next comments refer to the complete set of results.

Despite the limitations of the simulation study, we can conclude that the main factor that affects the performance of the estimators and confidence intervals is the magnitude of the $p_i^{(k)}$'s. When they are large and regardless of the distribution of the M_i's and the size of the fraction τ_1/τ covered by the frame, every one of the estimators of the τ's and design-based-like confidence intervals (Wald or bootstrap) perform satisfactorily. However, when the $p_i^{(k)}$'s are small and in spite of all the other factors, only the Bayesian estimators $\hat{\tau}_k^P$ perform acceptably. It is worth noting that when the $p_i^{(k)}$'s are small, the Bayesian estimators $\hat{\tau}_k^U$ and $\hat{\tau}_k^J$ perform better than the MLE's $\tilde{\tau}_k$; however, the performance of $\hat{\tau}_k^U$ and $\hat{\tau}_k^J$ is not good enough to make reliable inferences.

Bootstrap confidence intervals for τ_1 based on $\hat{\tau}_1^P$ did not perform as well as Wald intervals when the $p_i^{(k)}$'s were small or the M_i's were not Poisson distributed. The explanation of this result and the development of better bootstrap intervals are topics that require further research.

Finally, the best performance of the set of estimators $\hat{\tau}_k^P$ is a consequence of the greater amount of information used by them. Though we used relatively flat prior distributions for the τ_k's, the information supplied by them was enough to avoid the problems of bias and high variability observed in the other estimators. We carried out some additional simulation trials, and the results (which are not reported in the tables) indicate that, as long as the prior distributions are kept relatively flat, the estimates are not affected by the values of the parameters of the prior distributions. Obviously, misleading initial information combined with small values of the $p_i^{(k)}$'s will affect the estimates. An example of this is a prior distribution for τ_2 with a probability density function highly concentrated about a value very far from the true value of τ_2. However, we think that if the researcher has correct information, even if it is vague, it would be worthwhile using the set of estimators $\hat{\tau}_k^P$'s.

Table 2
Relative Biases and Square Roots of Relative Mean Square Errors of the Estimators of the Population Sizes

	Population I				Population II				Population III				Population IV			
$\bar{p}_1$	0.05		0.01		0.05		0.01		0.05		0.01		0.05		0.01	
$\bar{p}_2$	0.01		0.002		0.01		0.002		0.03		0.006		0.03		0.006	
	rβ	$\sqrt{r\varepsilon^2}$	rβ	$\sqrt{r\varepsilon^2}$	rβ	$\sqrt{r\varepsilon^2}$	rβ	$\sqrt{r\varepsilon^2}$	rβ	$\sqrt{r\varepsilon^2}$	rβ	$\sqrt{r\varepsilon^2}$	rβ	$\sqrt{r\varepsilon^2}$	rβ	$\sqrt{r\varepsilon^2}$
$\tilde{\tau}_1$	−0.00	0.02	−0.00	0.06	−0.00	0.02	−0.01	0.09	−0.00	0.02	−0.00	0.06	−0.00	0.02	−0.01	0.09
$\tilde{\tau}_2$	0.01	0.12	0.24[a]	0.78[a]	0.01	0.13	0.21[a]	0.76[a]	0.00	0.06	0.17[b]	0.67[b]	0.00	0.06	0.16[c]	0.63[c]
$\tilde{\tau}$	0.01	0.07	0.13[a]	0.43[a]	0.01	0.07	0.12[a]	0.42[a]	0.00	0.02	0.05[b]	0.19[b]	−0.00	0.02	0.04[c]	0.18[c]
$\hat{\tau}_1^U$	−0.00	0.02	−0.00	0.06	−0.00	0.02	−0.01	0.09	−0.00	0.02	−0.00	0.06	−0.00	0.02	−0.01	0.09
$\hat{\tau}_2^U$	0.02	0.13	0.14[a]	0.65[a]	0.01	0.12	0.14[a]	0.65[a]	0.00	0.06	0.13	0.65	0.00	0.06	0.13	0.71
$\hat{\tau}^U$	0.01	0.07	0.08[a]	0.36[a]	0.01	0.07	0.08[a]	0.36[a]	0.00	0.02	0.03	0.19	−0.00	0.02	0.03	0.20
$\hat{\tau}_1^J$	−0.00	0.02	−0.01	0.06	−0.00	0.02	−0.01	0.09	−0.00	0.02	−0.01	0.06	−0.00	0.02	−0.01	0.09
$\hat{\tau}_2^J$	−0.00	0.12	−0.14	0.48	−0.00	0.12	−0.14	0.48	−0.00	0.06	−0.04	0.37	−0.00	0.06	−0.04	0.35
$\hat{\tau}^J$	−0.00	0.07	−0.08	0.27	−0.00	0.07	−0.08	0.27	−0.00	0.02	−0.02	0.11	−0.00	0.02	−0.02	0.12
$\hat{\tau}_1^P$	−0.00	0.02	−0.01	0.06	−0.00	0.02	−0.01	0.09	−0.00	0.02	−0.01	0.06	−0.00	0.02	−0.01	0.09
$\hat{\tau}_2^P$	0.02	0.12	0.07	0.20	0.02	0.11	0.07	0.20	0.00	0.06	0.00	0.18	0.00	0.06	0.01	0.18
$\hat{\tau}^P$	0.01	0.06	0.04	0.11	0.01	0.06	0.03	0.11	0.00	0.02	−0.00	0.07	−0.00	0.02	−0.00	0.08

Notes: rβ, relative bias; $r\varepsilon^2$, relative mean square error; $\tilde{\tau}_1, \tilde{\tau}_2$ and $\tilde{\tau}$, MLEs. Superscripts *U*, *J*, and *P* of estimators $\hat{\tau}_1, \hat{\tau}_2$ and $\hat{\tau}$ indicate Bayesian estimators based on the prior Uniform, Jeffrey's and two stage Poisson-Gamma distributions, respectively. Results based on 10^4 trials. Superscripts *a*, *b* and *c* indicate results obtained by ignoring 8%, 15% and 21% of the trials. Ignored trials were those in which the corresponding estimator of τ_2 was negative or greater than 10^4.

Table 3
Coverage Probabilities and Average Lengths of 95% Confidence Intervals

	Population I								Population II							
	$\bar{p}_1 \approx 0.05, \bar{p}_2 \approx 0.01$				$\bar{p}_1 \approx 0.01, \bar{p}_2 \approx 0.002$				$\bar{p}_1 \approx 0.05, \bar{p}_2 \approx 0.01$				$\bar{p}_1 \approx 0.01, \bar{p}_2 \approx 0.002$			
	Bootstrap		Wald		Bootstrap		Wald		Bootstrap		Wald		Bootstrap		Wald	
	cp	$\bar{l}$	cp	$\bar{l}$	cp	$\bar{l}$	cp	$\bar{l}$	cp	$\bar{l}$	cp	$\bar{l}$	cp	$\bar{l}$	cp	$\bar{l}$
$\tilde{\tau}_1^M$	NC	NC	0.95	129	NC	NC	0.94	398	NC	NC	0.93	127	NC	NC	0.76	400
$\tilde{\tau}_2^M$	NC	NC	0.95	1,044	NC	NC	0.90[a]	8,181[a]	NC	NC	0.95	1,029	NC	NC	0.90[a]	7,764[a]
$\tilde{\tau}^M$	NC	NC	0.95	1,052	NC	NC	0.90[a]	8,200[a]	NC	NC	0.95	1,037	NC	NC	0.90[a]	7,784[a]
$\tilde{\tau}_1^D$	0.95	130	0.95	129	0.92	399	0.94	404	0.97	147	0.95	137	0.96	642	0.92	657
$\tilde{\tau}_2^D$	0.94	1,110	0.95	1,044	0.74	L_1	0.90[a]	8,181[a]	0.94	1,129	0.95	1,029	0.74	L_1	0.90[a]	7,764[a]
$\tilde{\tau}^D$	0.94	1,118	0.95	1,052	0.75	L_1	0.90[a]	8,201[a]	0.95	1,139	0.95	1,038	0.78	L_1	0.90[a]	7,819[a]
$\hat{\tau}_1^U$	0.94	131	0.95	129	0.92	412	0.94	403	0.97	150	0.94	137	0.97	668	0.93	657
$\hat{\tau}_2^U$	0.94	1,116	0.95	1,049	0.72	L_2	0.89[a]	6,887[a]	0.94	1,128	0.95	1,028	0.73	L_2	0.89[a]	6,738[a]
$\hat{\tau}^U$	0.94	1,124	0.95	1,057	0.73	L_2	0.90[a]	6,908[a]	0.95	1,139	0.95	1,038	0.77	L_2	0.90[a]	6,796[a]
$\hat{\tau}_1^J$	0.95	131	0.95	128	0.93	412	0.94	402	0.96	151	0.95	137	0.96	666	0.92	652
$\hat{\tau}_2^J$	0.93	1,043	0.94	998	0.58	3,122	0.71	3,142	0.93	1,057	0.93	985	0.60	3,074	0.72	3,095
$\hat{\tau}^J$	0.93	1,052	0.94	1,007	0.60	3,199	0.72	3,178	0.94	1,072	0.93	995	0.68	3,276	0.73	3,188
$\hat{\tau}_1^P$	0.94	131	0.95	129	0.91	411	0.94	402	0.89	151	0.95	137	0.86	666	0.93	654
$\hat{\tau}_2^P$	0.97	997	0.95	957	1.00	1,506	0.92	1,573	0.97	1,000	0.95	943	1.00	1,510	0.92	1,577
$\hat{\tau}^P$	0.97	1,006	0.95	966	1.00	1,575	0.94	1,624	0.97	1,011	0.95	953	1.00	1,679	0.95	1,710

Notes: cp, coverage probability; $\bar{l}$, average length. Superscripts *M* and *D* of the MLEs $\tilde{\tau}_1, \tilde{\tau}_2$ and $\tilde{\tau}$ indicate model-based and design-based confidence intervals, respectively. Bootstrap confidence intervals computed on 2,000 bootstrap samples. NC, not computed. Results based on 10^4 trials. Superscript *a* indicate results obtained by ignoring 8% of the trials. Ignored trials were those in which the corresponding estimator of τ_2 was negative or greater than 10^4. L_1 and L_2 indicate lengths greater than 10^9 and 10^4, respectively.

Table 4
Relative Biases and Square Roots of Relative Mean Square Errors of Variance Estimators

	Population I								Population II							
	$\bar{p}_1 \approx 0.05, \bar{p}_2 \approx 0.01$				$\bar{p}_1 \approx 0.01, \bar{p}_2 \approx 0.002$				$\bar{p}_1 \approx 0.05, \bar{p}_2 \approx 0.01$				$\bar{p}_1 \approx 0.01, \bar{p}_2 \approx 0.002$			
	Bootstrap		Taylor		Bootstrap		Taylor		Bootstrap		Taylor		Bootstrap		Taylor	
	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$
$\tilde{\tau}_1^M$	NC	NC	0.01	0.17	NC	NC	−0.04	0.08	NC	NC	−0.20	0.31	NC	NC	−0.64	0.65
$\tilde{\tau}_2^M$	NC	NC	0.01	0.49	NC	NC	1.9[a]	5.3[a]	NC	NC	−0.02	0.64	NC	NC	1.8[a]	5.4[a]
$\tilde{\tau}^M$	NC	NC	0.01	0.48	NC	NC	1.9[a]	5.3[a]	NC	NC	−0.02	0.64	NC	NC	1.7[a]	5.3[a]
$\tilde{\tau}_1^D$	0.03	0.19	0.01	0.17	−0.02	0.17	−0.00	0.17	0.08	0.46	−0.07	0.28	−0.05	0.40	−0.01	0.37
$\tilde{\tau}_2^D$	0.16	0.62	0.01	0.49	L_1	L_2	1.9[a]	5.3[a]	0.20	1.10	−0.02	0.64	L_2	L_2	1.7[a]	5.3[a]
$\tilde{\tau}^D$	0.15	0.61	0.01	0.48	L_1	L_2	1.9[a]	5.3[a]	0.20	1.10	−0.02	0.64	L_2	L_2	1.7[a]	5.3[a]
$\hat{\tau}_1^U$	0.02	0.20	−0.01	0.17	0.03	0.19	−0.01	0.17	0.14	0.51	−0.06	0.28	0.05	0.37	0.01	0.37
$\hat{\tau}_2^U$	0.13	0.62	−0.01	0.49	0.24	1.20	1.7[a]	4.6[a]	0.22	0.92	−0.00	0.62	0.30	1.40	1.6[a]	6.4[a]
$\hat{\tau}^U$	0.13	0.61	−0.01	0.48	0.24	1.20	1.6[a]	4.5[a]	0.23	0.91	0.01	0.61	0.30	1.40	1.6[a]	6.2[a]
$\hat{\tau}_1^J$	0.06	0.21	0.02	0.17	0.05	0.19	−0.01	0.17	0.12	0.50	−0.08	0.28	0.00	0.35	−0.04	0.36
$\hat{\tau}_2^J$	0.07	0.51	−0.03	0.44	−0.25	0.66	−0.11	1.40	0.13	0.69	−0.03	0.55	−0.25	0.74	−0.13	1.50
$\hat{\tau}^J$	0.06	0.50	−0.03	0.43	−0.25	0.66	−0.12	1.40	0.12	0.68	−0.03	0.53	−0.24	0.72	−0.15	1.40
$\hat{\tau}_1^P$	0.03	0.20	−0.01	0.17	0.03	0.18	−0.02	0.17	0.16	0.52	−0.05	0.28	0.05	0.37	0.01	0.37
$\hat{\tau}_2^P$	0.07	0.34	−0.02	0.35	−0.07	0.16	−0.03	0.12	0.10	0.42	−0.01	0.41	−0.06	0.17	−0.01	0.16
$\hat{\tau}^P$	0.06	0.34	−0.02	0.34	−0.05	0.14	−0.02	0.11	0.10	0.42	−0.01	0.41	−0.03	0.15	0.01	0.16

Notes: $r\beta$, relative bias; $r\varepsilon^2$, relative mean square error. Superscripts *M* and *D* of the MLEs $\tilde{\tau}_1, \tilde{\tau}_2$ and $\tilde{\tau}$ indicate model-based and design-based variance estimators, respectively. Bootstrap confidence intervals computed on 2,000 bootstrap samples. NC, not computed. Results based on 10^4 trials. Superscript *a* indicate results obtained by ignoring 8% of the trials. Ignored trials were those in which the corresponding estimator of τ_2 was negative or greater than 10^4. L_1 and L_2 indicate values greater than 10^2 and 10^4, respectively.

Acknowledgements

This research was supported by grant UASIN-EXB-01-01 of PROMEP and grant PAFI-UAS-2002-I-MHFM-0 of UAS. We thank Eduardo Gutierrez, the associate editor and the referees for their helpful suggestions and comments.

References

Agresti, A. (2002). *Categorical Data Analysis*. 2nd edition. New York: John Wiley & Sons, Inc.

Booth, J.G., Butler, R.W. and Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89, 1282-1289.

Castledine, B.J. (1981). A Bayesian analysis of multiple-recapture sampling for a closed population. *Biometrika*, 67, 197-210.

Darroch, J.N. (1958). The multiple-recapture census I: Estimation of a closed population. *Biometrika*, 45, 343-359.

Davison, A.C., and Hinkley, D.V. (1997). *Bootstrap Methods and their Applications*. New York: Cambridge University Press.

Evans, M.A., Kim, H.-M. and O'Brien, T.E. (1996). An application of profile-likelihood based confidence interval to capture-recapture estimators. *Journal of Agricultural*, *Biological*, *and Environmental Statistics*, 1, 131-140.

Félix-Medina, M.H., and Thompson, S.K. (2004). Combining cluster sampling and link-tracing sampling to estimate the size of hidden populations. *Journal of Official Statistics*, 20, 19-38.

Fienberg, S.E., Johnson, M.S. and Junker, B.W. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society*, A, 162, 383-405.

Frank, O., and Snijders, T.A.B. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10, 53-67.

Heckathorn, D.D. (1994). Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, 49, 11-34.

Seber, G.A.F. (1982). *The Estimation of Animal Abundance*. 2nd edition. London: Griffin.

Snijders, T.A.B. (1992). Estimation on the basis of snowball samples: How to weight? *Bulletin de Méthodologie Sociologique*, 36, 59-70.

Spreen, M. (1992). Rare populations, hidden populations, and link-tracing designs: What and why? *Bulletin de Méthodologie Sociologique*, 36, 34-58.

Thompson, S.K., and Frank, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology*, 26, 87-98.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Survey Methodology, December 2006
Vol. 32, No. 2, pp. 197-216
Statistics Canada, Catalogue No. 12-001

On Sample Survey Designs for Consumer Price Indexes

Alan H. Dorfman, Janice Lent, Sylvia G. Leaver and Edward Wegman [1]

Abstract

Survey sampling to estimate a Consumer Price Index (CPI) is quite complicated, generally requiring a combination of data from at least two surveys: one giving prices, one giving expenditure weights. Fundamentally different approaches to the sampling process–probability sampling and purposive sampling–have each been strongly advocated and are used by different countries in the collection of price data. By constructing a small "world" of purchases and prices from scanner data on cereal and then simulating various sampling and estimation techniques, we compare the results of two design and estimation approaches: the probability approach of the United States and the purposive approach of the United Kingdom. For the same amount of information collected, but given the use of different estimators, the United Kingdom's methods appear to offer better overall accuracy in targeting a population superlative consumer price index.

Key Words: Elementary index; Probability proportional to size sampling; Purposive sampling; Scanner data; Superlative index.

1. Background

From start to finish, survey sampling for the sake of a Consumer Price Index (CPI) must rank among the most complicated of sampling enterprises. The population target is hard to pin down, the appropriate domain of items debated, the definitions of the raw ingredients – prices, quantities, items – ambiguous and subject to question. The ultimate estimator – the estimator of the all-items CPI – relies on data from at least two surveys, one giving prices, and one giving "weights." Below the level of "composite items" (or "item strata") – groups of items supposed homogeneous in their price movements – there is typically no cost effective way to keep sampling weights up to date. Debate therefore goes on about the proper choice among various simple alternative estimators of price change for item categories, the "elementary aggregate indexes." The appropriate method of aggregating these price changes, using the weights, is subject also to debate.

There are two broad approaches to the sampling by which prices are collected: probability sampling and judgment sampling. The most commonly accepted approach to survey sampling in general requires injecting an element of randomness into the survey process and relying on this randomness to make inference about population characteristics of interest – probability or "design-based" sampling; see, *e.g.*, Särndal, Swensson and Wretman (1992). This approach was not always taken for granted. Early in the 20^{th} century, "purposive" or "representative" sampling was considered a viable, and possibly better, option. More recently, the prediction-based school of Royall has challenged design-based assumptions; see *e.g.*, Valliant, Dorfman and Royall (2000).

In the U.S., all CPI-related surveys are carried out using complex probability sampling techniques. Around the world, most CPI's are constructed from judgment samples, in which experts on the different item strata choose broader or narrower classes of items for which field representatives collect prices. The fundamental reason for this is the difficulty of getting all the data one needs on the plethora of items sold, and the places where they are sold, to make probability sampling feasible.

The interesting fact is that there has been very little assessment of the relative accuracy of the different approaches to sampling. Indeed it has not been clear that it is feasible to make such assessments. The underlying population price index, for even the smallest of countries, involves so many transactions on so many items in so many places as to be inaccessible. Moreover, the population of items on the market is in a constant state of flux, complicating the application of traditional population index formulas. How then can one judge the relative closeness to "truth" of different sample-based estimates? Furthermore, in most cases, not even sample information is available for a key ingredient of the population index-namely the *quantities* of items sold – so even artificially constructing a population for test purposes from sample data has not been feasible.

The relatively recent availability of *scanner* data, in the U.S. and elsewhere, presents an unprecedented opportunity for testing sampling approaches and estimators. These data include prices *and* quantities, typically on a weekly basis, of *all* the items sold in a given category within a large sample of outlets having scanner devices. Such data may be used to construct realistic populations of transactions for which the true price index is *known*. We can then use various methods to sample from this population, construct different index

1. Alan H. Dorfman, Office of Survey Methods Research, and Sylvia G. Leaver, Office of Prices and Living Conditions, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave NE, Washington, D.C., U.S.A., 20212; Janice Lent, U.S. Bureau of Transportation Statistics, 400 7^{th} Street, SW, Washington, D.C., U.S.A., 20590; Edward Wegman, Center for Computational Statistics, George Mason University, Fairfax, VA, U.S.A., 22030.

estimates of interest, and compare the results to the known population parameters. One such study, described by de Haan, Opperdoes and Schut (1999), seems to show that "cutoff sampling," the sampling of the few largest (in terms of revenue generated) items in the population, outperforms two important design-based approaches: simple random sampling (*srs*), and probability proportional to size sampling (*pps*) (where the size measure is, again, revenue).

One difficulty in any study making such comparisons is the task of maintaining a "level playing field." If one sampling method, for example, makes use of (population) information that might not actually be available in practice, while another does not, the comparison of methods is undermined. Similarly, if one method provides only one sample or very few samples, and another provides thousands, special precautions are needed in comparing the two; indeed, such a comparison might require serious qualifications. Given the complexity of the sampling and estimation methods used in price index computation, it is not surprising that these and many other difficulties complicate experiments designed to compare various methods.

Ideally, to compare the approaches, for example, of two countries, we would mimic the whole complex sampling and estimation process of each and evaluate its costs. Both processes would be allowed the same budget, and we would be able, by some predetermined and equitable measure, to evaluate each estimate's proximity to a known target index.

This paper comprises two studies, a large primary study, and a smaller secondary, follow-up study.

The main study is described in Sections 2 through 4. Section 2 describes the construction of the target population. Section 3 describes the "US" and "UK" methodologies and outlines the simulation details. No attempt is made to assess relative costs (thus falling short of the ideal), but the competing approaches are made as equal as possible in terms of the information they use. Results, which favor the UK approach, are given in Section 4.

The follow-up study, in Section 5, attempts to disentangle the effects of different components of the two approaches, in particular sampling method and elementary index formula. Section 6 gives a final summary and discussion.

Note on the target indexes. The price index literature contains myriad formulas for calculating price change between one period and another. Different indexes are compatible with different assumptions regarding the "average" consumer's buying behavior in response to price change. The "fixed market basket" indexes, the commonly employed Laspeyres and less used Paasche formulas, are compatible with the assumption that consumers continue to purchase the same items in the same quantities regardless of changes in relative prices. The Laspeyres index projects the period 1 ("base period") quantities forward to period 2 ("current period"), while the Paasche applies the period 2 quantities to period 1. The geometric index (or "Jevons" or "*geomean*"), usually weighted with base period expenditure shares, assumes that consumers adjust the quantities they purchase in such a way that the expenditure share for each item remains constant across time. The "superlative" Fisher, Törnqvist, and Walsh index formulas, which rely on quantity (or expenditure share) information for both periods, do not require these assumptions. Formulas for these indexes, with a superscript y representing the base period, $y+1$, the current period, and i the item purchased, for the indexes are given in Appendix A.

The debate on the all items target index usually comes down to choosing between the Laspeyres and one of the superlative indexes. Most countries select a Laspeyres target index, but a strong case (Diewert 1997) can be made that the proper target is a superlative index (usually the Fisher or Törnqvist), even if the formula for the estimator does not resemble one of the superlative population index formulas. Because of the form of the US elementary aggregates – geometric mean – and the fact that previous research (Dorfman, Leaver and Lent 1999) indicated that the lowest level of estimation can have a major impact, the weighted *geomean* will be included among the potential targets. Targets for a given domain are calculated based on prices and quantities of all items in the domain following the formulas in Appendix A (a single-stage aggregation of prices and quantities).

Note: These formulas are deceptively simple, requiring the notation of section 3 for their full development. Thus, in a formula such as that for the Fisher index F (which we will take as our target in the main study of sections 2 – 4) "i" represents an item i belonging to a small class c (an "ELI" or "representative item" – see section 3), where c is itself a subset of wider classes; further, the item i is sold in a particular outlet j, classified as part of a particular chain k, and located in a particular sampled geographic area, the primary sampling unit (*psu*) l. Thus, the expression for a sum $\sum_i$, in the case of the overall population index, is really shorthand for $\sum_{l=1}^{3}\sum_{k=1}^{8}\sum_{j\in(k,l)}\sum_{C=1}^{4}\sum_{h\in C}\sum_{c\in h}\sum_{i\in(j,c)}$; a similar remark holds for $\prod_i$. In short, these are sums and products over *all* items in the population. Contractions of this full expansion will give the population indexes for the different classes C, *etc*.

2. The Population for the Primary Study

The data source for the present study is a scanner data set for breakfast cereal for the years 1995 through 2000 in three separate but contiguous sections of a single large

metropolitan area. The data set was purchased from the A.C. Nielsen Corporation by the U.S. Bureau of Labor Statistics for the purpose of determining the feasibility of incorporating scanner data into the U.S. CPI; see Richardson (2000).

From these data, artificial "populations" were drawn by the method described below. Thus the study encompasses an apparently narrow world, that of cereal, within a fairly restricted geographic domain. Even this restricted world, however, allows for rather discrepant price trends over the six years. Thus, although we will not be able to generalize, in any simple fashion, to global price indexes encompassing a wide heterogeneity of products, we may be able to derive important clues on the effects of different sampling methods and the behavior of particular estimators.

The six years' worth of data available provided the opportunity for establishing fairly long-term price trends. In order to keep the data manageable and avoid the complications of seasonality, we limited ourselves to February data. For February of year y, for each item (*i.e.*, each particular combination k of brand, type, size) in a particular outlet, four weeks t of price and quantity data were combined into a single month's price and quantity, by using the sum of quantities $q_k^{\text{Feb},\,y} = \sum_{t\in \text{Feb},\,y} q_k^t$ sold during the month as the quantity, and the *unit value* $p_k^{\text{Feb},\,y} = \sum_{t\in \text{Feb},\,y} q_k^t\, p_k^t / \sum_{t\in \text{Feb},\,y} q_k^t$ as the price. Unit values computed over short periods of time (*e.g.*, a month) give perhaps the most meaningful sense of the "average" price for a particular item. The use of unit values smoothes the data and reduces it to more manageable proportions; for a discussion of circumstances under which use of unit values is appropriate or not appropriate see Balk (1999).

For the purposes of the study, the population of breakfast cereals was divided into four groups:

1. Hot Cereals (H)
2. "Sugary" cereals (S)
3. "Fruity" cereals (F)
4. "Plain" cereals (P), *i.e.*, cold cereals not falling into categories (2) and (3).

For each group, for each successive pair of years, superlative and non-superlative indexes were calculated, using item-outlet combinations available in both years. In practice, there is generally a problem with getting perfect match-ups from period to period, and finding means to deal with this by finding substitutes for original products or by other means is important; this study bypasses this particular problem.

Long range indexes ('95 to '00) were calculated both directly and by chaining annual indexes. Additionally, indexes were calculated on the "core" items, meaning those available in all six years. On a year-to-year basis, the core items represented between 53% and 61% of a given year's items available for year-to-year comparison; core expenditure was from 83% to 91% of the total expenditure on all (core and non-core) items. There were 326 core items, and a total of 848 distinct items over the course of '95 to '00.

Values of year-to-year population indexes are represented in Figures 1 through 5. Figure 1 gives the index $\hat{I}^{y,\,y+1}$ values for Sugary cereals based on all items sold in stores in both y and $y+1$, for (February of) $y =$ 1995, ..., 1999 (the "all items"). Values for five indexes are shown, including the Paasche P and, as being of academic interest, a unit value index U, the ratio of quantity weighted mean prices, averaged over all item types and outlets. Figure 2 gives results of the same calculations, but limited to "core" items. Figures 1 and 2 are almost identical, and the resemblance between indexes calculated using all items and those using just the core items held for the other cereal categories as well. Figures 3 through 5 give the results for the core-based indexes for Hot, Fruity, and Plain cereals. For any given index, the figures indicate serious differences across cereal categories. The price trends of the four major groups are quite different: H increases, S decreases sharply, F decreases modestly, and P increases modestly.

Table 1 gives long range ('95 – '00) direct indexes and chained indexes for "all items" and "core items." ("All items" for constructing an index between two given years, are all those item/outlets with positive quantities sold, both years). Again, there is very little difference between the values for "core items" and "all items," and sharp differences from one cereal category to another. The chained and direct results are close for the superlative indexes but tend to be discrepant for the *geomean*, Laspeyres, and Paasche. The chained and direct unit value indexes are close and in fact the latter would be identical to the chained based on the core items, except that, for convenience, the year to year indexes were based only on item-outlet combinations available for both years.

Except for some oscillation of position of the unit value index, we observe a clear ordering of index values by formula, the same across categories, which may be summarized as follows: (1) The superlative indexes differ relatively little from each other, a noteworthy result given the amount of variability in the item-outlet price relatives and quantities, due to "sales." (2) The traditional non-superlative indexes differ wildly from each other and the superlatives, with the *geomean*, weighted by first period expenditures, running much *higher* than the superlatives, the Laspeyres still higher, and the Paasche (not surprisingly) much lower. These results suggest that, in Cereal World, not only quantity, but expenditure share, tends to drop in period 2 on

an item having a sharp increase in price in that period. (3) The unit value index is low as well, but, except in the case of Hot cereals, is better than the tradional non-superlative indexes in approximating the superlatives. (4) In the light of later developments in this paper, and at the suggestion of a referee, two non-quantity based indexes are included in this table (although not in the figures): the *dutot* index, which is a simple ratio of average prices (RA) – see Appendix A, and an *unweighted* (that is, constant weighted) *geomean*; both are usually reserved for computing indexes at the elementary level. The results are surprising: in approximating the superlatives, they do as well as or better than the traditional, quantity based non-superlatives, about on a par with the unit value index.

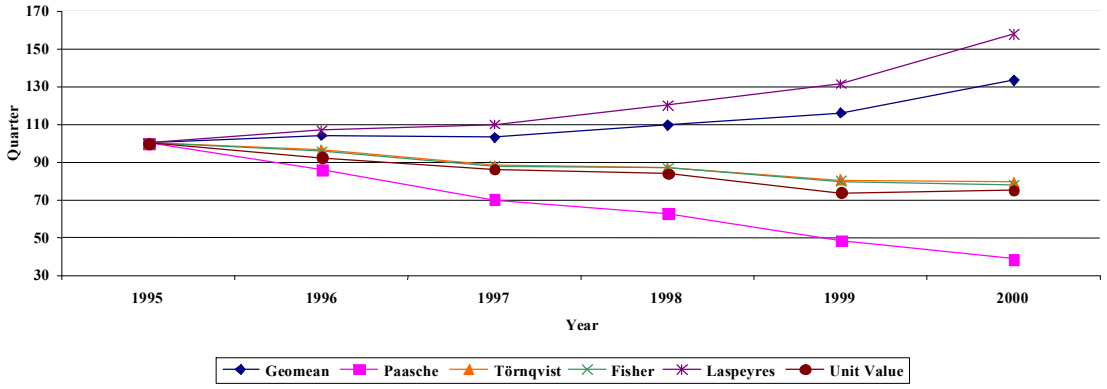


Figure 1. Annually Chained Population Target Indexes for All Sugary Cereals February-to-February Indexes, 1995 = 100.

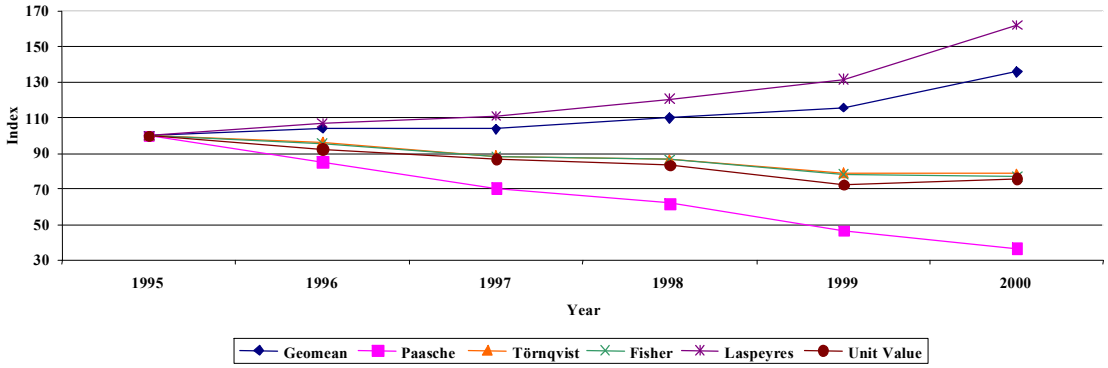


Figure 2. Annually Chained Population Target Indexes for Core Sugary Cereals February-to-February Indexes, 1995 = 100.

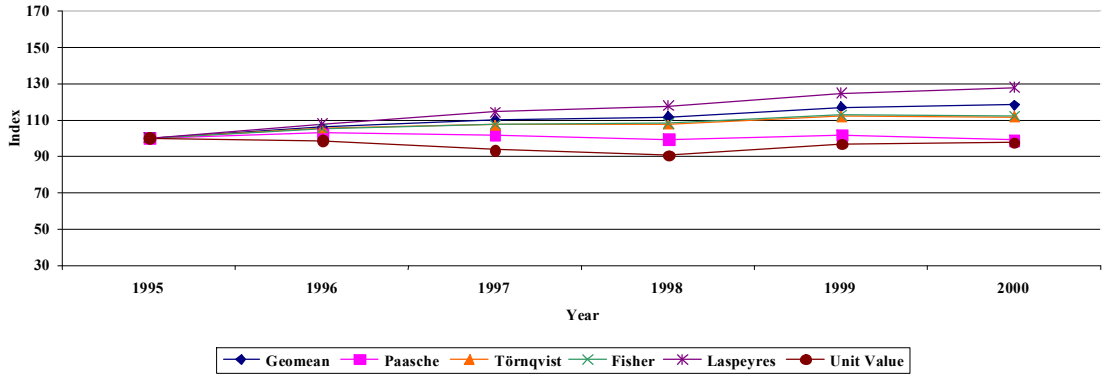


Figure 3. Annually Chained Population Target Indexes for Core Hot Cereals February-to-February Indexes, 1995 = 100.

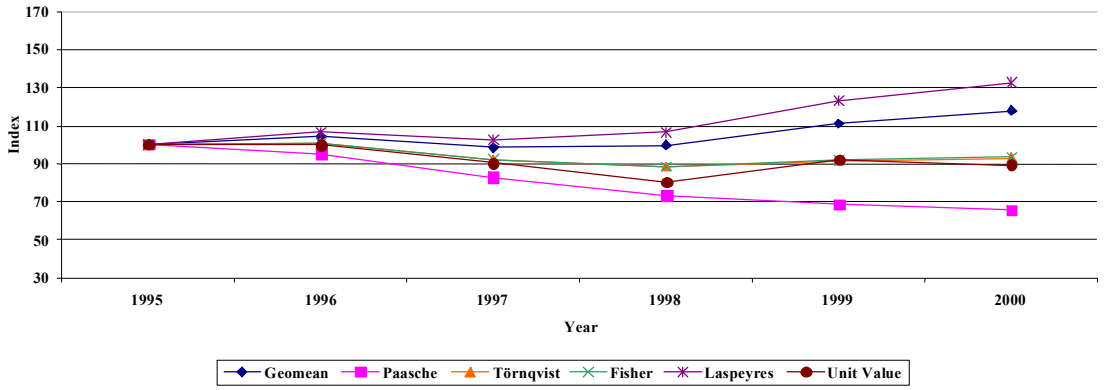


Figure 4. Annually Chained Population Target Indexes for Core Fruity Cereals February-to-February Indexes, 1995 = 100.

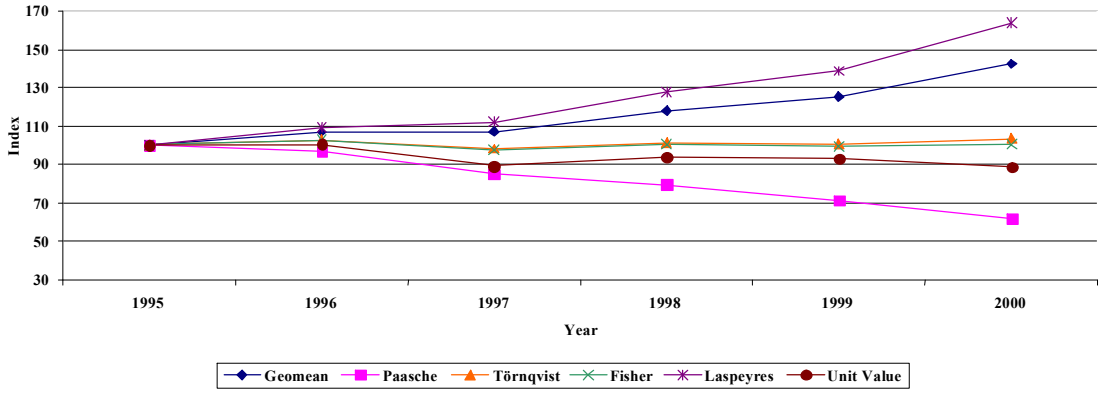


Figure 5. Annually Chained Population Target Indexes for Core Plain Cereals February-to-February Indexes, 1995 = 100.

Table 1
Direct and Chained Indexes for '95 – '00

		Geometric Mean	Paasche	Törnqvist	Fisher	Laspeyres	Unit Value
Hot	Direct	1.1176	1.0253	1.0847	1.0891	1.1569	0.9576
	Chained, All Items	1.1801	0.9874	1.1159	1.1216	1.2742	0.9453
	Chained, Core Items	1.1804	0.9865	1.1160	1.1221	1.2763	0.9759
Sugary	Direct	0.8855	0.6739	0.7913	0.7898	0.9257	0.7417
	Chained All Items	1.3341	0.3825	0.7925	0.7771	1.5786	0.7506
	Chained, Core Items	1.3591	0.3661	0.7849	0.7704	1.6212	0.7585
Fruity	Direct	0.9716	0.8676	0.9319	0.9296	0.9960	0.8932
	Chained All Items	1.2202	0.6849	0.9661	0.9696	1.3728	0.9308
	Chained, Core Items	1.1808	0.6557	0.9320	0.9328	1.3269	0.8950
Plain	Direct	1.0811	0.8641	1.0045	0.9816	1.1150	0.8554
	Chained All Items	1.3969	0.6330	1.0333	1.0053	1.5965	0.8935
	Chained, Core Items	1.4234	0.6175	1.0353	1.0054	1.6370	0.8879

Table 1
Direct and Chained Indexes for '95 – '00

		Geo-metric Mean*	Paasche	Törnqvist	Fisher	Las-peyres	Unit Value	RA	Geo-metric Mean+
Hot	Direct	1.1176	1.0253	1.0847	1.0891	1.1569	0.9576	1.1192	1.0949
	Chained, All Items	1.1801	0.9874	1.1159	1.1216	1.2742	0.9453	1.1395	1.1128
	Chained, Core Items	1.1804	0.9865	1.1160	1.1221	1.2763	0.9759	1.1374	1.1151
Sugary	Direct	0.8855	0.6739	0.7913	0.7898	0.9257	0.7417	0.8817	0.8702
	Chained All Items	1.3341	0.3825	0.7925	0.7771	1.5786	0.7506	0.9124	0.9010
	Chained, Core Items	1.3591	0.3661	0.7849	0.7704	1.6212	0.7585	0.8984	0.8894
Fruity	Direct	0.9716	0.8676	0.9319	0.9296	0.9960	0.8932	0.9815	0.9726
	Chained All Items	1.2202	0.6849	0.9661	0.9696	1.3728	0.9308	1.0263	1.0165
	Chained, Core Items	1.1808	0.6557	0.9320	0.9328	1.3269	0.8950	0.9935	0.9820
Plain	Direct	1.0811	0.8641	1.0045	0.9816	1.1150	0.8554	1.0620	1.0511
	Chained All Items	1.3969	0.6330	1.0333	1.0053	1.5965	0.8935	1.0642	1.0572
	Chained, Core Items	1.4234	0.6175	1.0353	1.0054	1.6370	0.8879	1.0653	1.0571

* Weighted by base period expenditure.
\+ Unweighted.

Based on this preliminary investigation, and for relative simplicity, we restricted our further investigations to the core data. To investigate the relative accuracy of probability and purposive sampling, as applied in practice to construct CPI's, we endeavored to approximate the sample designs used by the United States and the United Kingdom-representing probability based and judgment sampling respectively. In both cases we were fortunate to have detailed information on the complex survey processes, in the form of manuals, and contacts within the respective agencies. The basic idea was to repeatedly sample from a given population, for example the core transactions in the years '95 and '96. Each "run" was a composite of sampling and estimation activities carried out according to the methods of one country or the other. It should be borne in mind that our interest was in comparing the merits of *methodologies*, not in measuring the success of the US and UK in estimating their target population parameters.

The four "natural" population groups described above, referred to as "Major Groups" in the UK and "Expenditure Classes" in the US, were divided into finer sub-groups. In practice, sub-groups would be defined in terms of types of commodity. One justification for this, besides any intrinsic interest there might be in those commodities themselves, is that sub-groups so formed will tend to be homogeneous in their price trends. For purposes of this simulation study we therefore defined sub-groups as follows:

1) Long range price change for each of the 326 items in the core data were calculated, using unit value indexes for the items (across outlets) for '00 versus '95.
2) Noise was added to these indexes, items within a major group were sorted by their values of the perturbed index, and adjacent items were grouped together. The grouping of items with close long term indexes was to make the subgroups homogeneous, and the addition of noise was done so that the homogeneity would be realistically imperfect.

Table 2 gives the population item structure that was constructed, including the nomenclature in use in each of the two countries, the number of groups at each level of refinement, and the corresponding symbol for each class level used in this paper. The "Representative Item" is the lowest level at which an index is produced in the UK. This corresponds to the US's Entry Level Item (ELI), actually a collection of similar or related items. In the US, indexes are produced for categories one level up, *i.e.*, at the "Item Stratum" level, but these categories are further divided by the geographic areas in which the items are sold. Note that there are 2 or 3 Item Strata/Sections h in a Class/Major Group C, 3 ELI/Representative Items c per Item Stratum/Section h (except in one instance 2), and 10 or 11 items/varieties i in each ELI/Representative Item c. (*Note*: an actual UK class might be larger or smaller than the

corresponding US class; for example as a rule the ELI probably takes in more sorts of specific items than does the Representative Item. We had to force equivalence to ensure that the same amount of information was used in each approach. This adjustment will not affect our conclusions regarding the relative merits of the basic methods used in the two countries).

Table 2
Population Structure of "Cereal World": Items

UK	US	Number of Groups	Symbol
Major Group	Expenditure Class	4	C
Section	Item Strata	10	h
Representative Item	Entry Level Item (ELI)	29	c
Variety	Item	326	i

In addition to the item structure, each population of transactions has a "spatial" structure, characterizing where an item was sold. This structure is summarized in Table 3. Outlets belong to chains (*e.g.*, Safeway, Kroger), which cut across the three US geographic primary sampling units from which the cereal data were collected. (In the UK terminology, chains are called "multiples.) Outlets in a given chain share common ownership, with the exception of "Chain 8," which was a "catch-all" group consisting of outlets *not* belonging to a major chain (there may have been some "mini-chains"). In matching this "chain structure" to the classification of shops used in UK sampling, Chain 8 was considered a set of "independents" (the term used for independently-owned shops in the UK). Chain 4, which appeared to have the greatest homogeneity of pricing across outlets, was regarded as a "centrally collected multiple," the term used in the UK for groups of outlets with centrally controlled pricing. Each remaining chain was a non-centrally collected multiple. The manner of collection and estimation for each of these three types is given in the description of UK methodology below.

Thus the population consists of $N^{95} \approx 20{,}000$ records for '95 – '96 indexes, each record representing the purchase of an item i within an outlet j. Attached to each item/outlet are its PSU/Region l, its chain/shop-type k, the outlet/shop j, the item/variety i, the ELI/representative item c, the item stratum/section h, the expenditure class/major group C, and p^y, q^y, p^{y+1}, and q^{y+1}, the prices and quantities (in ounces) of the items sold in (February of) the two years in question. We used this population file (henceforth referred to simply as "the file") to simulate all phases of the US and UK operations.

3. Sampling Methodologies Simulated

The complicated sampling procedures we used to simulate the US and UK approaches are patterned on the respective practices of these two countries. These practices change over time, and have variants even at a given point in time. Our goal was not to determine which country does better, nor to encompass all variants. Rather it was to compare two distinct modes of sampling, with the range of complexity those modes entail. The interested reader can find a description of the US construction of the CPI in the *BLS Handbook of Methods* (2005), Chapter 17. For the UK's Retail Price Index (RPI), we relied on *The Retails Prices Index Technical Manual* (1998). A description of more current UK practice can be found in the *Consumer Price Indexes Technical Manual* (2005).

3.1 US Sampling Methodology

We first describe the US sampling methodology, which requires three surveys employing probability sampling: (1) a household survey, the Consumer Expenditure Survey (CEX), to estimate household allocation of expenditure to different categories of goods, (2) a second household survey, the Point of Purchase Survey (POPS) to estimate, within item groups, the relative amounts spent in different outlets, and (3) an outlets survey, through which individual items are selected and priced. In all three cases, sampling for the simulation is random with replacement (though the sampling employed in practice is considerably more complicated). The first two surveys are based on simple random samples, and the last is based on a probability proportional to size (*pps*) sample, where the size measures are a function of expenditures as estimated from the CEX and POPS. The sample for the third survey is a collection of items within outlet/ELI combinations.

Table 3
Population Structure of "Cereal World": Outlets

UK		US	#	Symbol
Region		Primary Sampling Unit	3	l
Shop type:	Independents	Chain 8		k
	Multiples: Central	Chain 4		
	Multiples: Non – central	Chains 1 – 3; 5 – 7		
Shop		Outlet	~300	j

3.1.1 CEX (Household Survey)

The goal is to estimate E_{lc}, the gross household expenditure on ELI c within PSU l. We sampled using simple random sampling with replacement (*srswr*) from the file described above, within PSU, in such a manner as to get unbiased expansion estimates

$$\hat{E}_{lc}^{95} = \frac{N_l^{95}}{n_{xl}} \sum_{j \in l \cap s(xl)} \sum_{i \in c \cap s(xl)} E_{ljci}^{95},$$

where $E_{lji}^{y} = q_{lji}^{y} p_{lji}^{y}$, N_l^{95} was the population size (number of records for *psu* l in '95 – '96), and n_{xl} was the sample size of the CEX sample $s(xl)$ in PSU l, chosen to match actual US CEX sample sizes and to achieve coefficients of variation of the estimates that approximated those achieved through the actual US CEX; the x in $s(xl)$ and n_{xl} is meant merely to differentiate the CEX from the POPS survey (which has a corresponding "p"; see below) or the prices survey. This "imitation CEX" was a simplified version of the actual survey. Our methodology tacitly assumed that all customers in a given outlet bought items in the same proportions; it did not allow for the inevitable measurement error that accompanies any actual expenditure survey, and (for '95 – '96) it was too current: real CEX data often predate by several years the outlet surveys for which they are used. Since, however, the "household data" collected were used in the corresponding UK methodology (see below) as well, the simplified version sufficed for the intended comparison of methodologies.

Higher level expenditures were estimated by simple addition. Thus, for example, the total expended across PSU's in a given ELI c is estimated by $\hat{E}_{.c}^{95} = \sum_l \hat{E}_{lc}^{95}$, *etc*. There were 500 CEX samples taken, each producing a corresponding set of expenditure estimates.

3.1.2 POPS (Household Survey)

Here the goal is to estimate the distribution of expenditures at different outlets for particular classes of goods. These classes could be ELI's or groups of ELI's; in the present study we assume they are the ELI's. The actual US TPOPS (Telephone Point of Purchase Survey) is, as its name suggests, conducted by phone, using a sample rotation scheme with a four-year cycle. We endeavored, as we did with the CEX, to match statistical properties of our procedure to the actual TPOPS, but it turned out that to match sample sizes on our file of 20,000 would have given larger than desirable sampling fractions within PSU's. We therefore cut the sample sizes in half – our "imitation POPS" should have precision about $1/\sqrt{2}$ of the actual TPOPS. Again, this modification will not affect the conclusions of this study, because we used the identical data in the UK construction. Samples $s(pl)$ of size n_{pl} were drawn by *srswr*, and estimation was by the expansion estimator:

$$\tilde{E}_{lcj}^{y} = \frac{N_l^{y}}{n_{pl}} \sum_{i \in c \cap s(pl)} E_{ljci}^{y}.$$

Since the POPS survey tends to be more up-to-date than the CEX, we allow y to be the base year of the index, '95 in '95 – '96, but '96 in '96 – '97, *etc*. There were 500 runs and sets of estimates, each to be matched with a CEX run.

3.1.3 Outlet Sampling

For each year y, selection of items from which to collect prices involves the following steps:

(a) For each PSU l, and each of the 10 item strata h, we select 2 ELIs c by probability proportional to size with replacement sampling (*ppswr*), with size measure $\hat{E}_{lc}^{95}$ derived from the CEX.

(b) For each ELI c selected, we select 8 outlets j by *ppswr*, using as size measure POPS expenditure estimates $\tilde{E}_{ljc}^{y}$. Thus altogether there are 160 ELI-outlet pairs per PSU, and 480 total, with a certain amount of repetition possible.

(c) Within outlet/ELI (j, c) we "go" (as the field representative would literally go) to the outlet and "list" all items belonging to the ELI and their corresponding first period expenditures E_{ljci}^{y}, and, with this within-outlet frame, sample 1 item by *pps*.

For each item so selected, we record the prices p_{ljhci}^{y}, $y = 1, 2$. Thus we note that all aspects of the outlet sampling are *pps* with replacement, based on estimates of expenditure from one or other of the 2 household surveys or from within the selected store. Again, we performed 500 runs, each run corresponding to a single CEX/POPS run.

3.2 US Estimation

"Elementary aggregates" $\hat{I}_{lh}^{y,\,y+1}$, index estimates at the PSU × Item stratum level, are the building blocks from which the CPI is constructed. In most CPI's around the world, the lowest level indexes are unweighted averages of one sort or another, as is the UK's RA estimator discussed below, and expenditure data are only used to aggregate these to higher levels. In the US, the elementary indexes are basically Horvitz-Thomson estimators relying explicitly or implicitly on expenditure estimates from both the CEX and POPS. In recent years, the US has for most item strata adopted a *geomean* formula (see Appendix A), so that estimates at this level take the form

$$\hat{I}_{lh}^{y,\,y+1} = \prod_{\substack{j\in l,\\ i\in c\in h\\ (i,\,j)\in s}} \left(\frac{p_{ljhci}^{y+1}}{p_{ljhci}^{y}} \right)^{s_{ljhci}},$$

where

$$s_{ljhci} = \frac{w_{ljhci}}{\sum_{\substack{j\in l,\, c\in h,\, i\in c\\ (i,\,j)\in s}} w_{ljhci}},$$

with

$$w_{ljhci} = \frac{\tilde{E}_{lc}\,\hat{E}_{lh}}{\hat{E}_{lc}}\, w_{ljhci},$$

$j \in l$, $i \in c \in h$ and $(i,\ j) \in s$. Note that the weights are not particular to the i^{th} item; we omit the time superscripts for brevity. They are not simply equal to the reciprocal of the number n_{lh} of sample items in lh, as sample unbiasedness considerations might lead one to expect (Balk 2003), because the sampling probabilities do not reflect exact base period expenditures on items; see the *BLS Handbook of Methods* (2005).

Then the elementary indexes are aggregated using estimated expenditures from the CEX according to a Laspeyres formula, for example

$$\hat{I}_{.h}^{y,\,y+1} = \frac{\sum_{l} \hat{E}_{lh}\,\hat{I}_{lh}^{y,\,y+1}}{\sum_{l} \hat{E}_{lh}}$$

to get the index for a given item stratum h, across PSU's.

3.3 UK Sampling Methodology

The UK, like the US, combines three components in its estimation methodology: (1) a household survey, the Family Expenditure Survey (FES), to get estimates of amounts spent on different item groups, (2) a shops survey, the Annual Retailing Inquiry (ARI) to get expenditure information by section and shop type, and (3) an outlet survey of shops, to select items for pricing.

3.3.1 FES (Household Survey)

The goal is to estimate expenditures $E_{..c}$ for representative items c, and $E_{l..h}$ expenditures for region/section combinations. For purposes of this study we will assume that the data for the US's CEX and the UK's FES coincide run by run, so there are, again, 500 FES data sets. Note that the UK does not aim at the more detailed estimates $E_{l..c}$ which the US targets.

3.3.2 Annual Retailing Inquiry (Shops Survey)

The goal is to get estimates of expenditures $\tilde{E}_{kh}$, by section and shop type. This is considerably broader than the outlet (shop) by ELI (representative item) that the US's POPS seeks. We use the same data, for each of 500 runs, to construct the ARI estimates that we used to construct the POPS estimates for the simulated US CPI.

3.3.3 Outlet Sampling

Selecting items from which to collect prices involves the following steps:

(a) A "judgment sample" of representative items c is selected within each section h. In the present study (only to allow for simulation), within each section, we select the two representative items with largest values of $\hat{E}_{hc}$. Note two differences from the corresponding step (a) of the US method: (i) selection is uniform across all regions l; (ii) selection is not random, and, in particular, it does not allow for duplication of representative items. (Duplication can occur in the simulated US method, due to with replacement sampling of ELI's within item strata.)

(b) The field economists select the shops in a particular locale in which to price a given representative item. Traditionally, this was *srswor,* after the field economist had constructed a frame of appropriate shops. More recently, selection has been by *pps*, where the size measure is floor space dedicated to the type of goods the representative item represents. Field economists do not draw samples of "centrally collected" items: in the case of a very large multiple, the price of an item is collected from the multiple's central office, and taken to represent the price of that item in all shops in the multiple. In the present study we proceeded as follows: for each region l and representative item c, we selected 8 shops as follows:

- 4 from non-central multiples (Chains 1, 2, 3, 5, 6, 7)
- 1 from a central multiple (Chain 4)
- 3 from independents (Chain 8)

In each case, for simplicity, we used *srs* without replacement from shops having positive expenditure for the representative item. The number of shops in the UK (8 per representative item in each region) matches the number of "outlets" in the US; there are 160 shop/representative item pairs per region, or 480 in total. Note the following differences from the U.S. methodology:

1. Information on shop type is being used for stratification (and will play a role in estimation below). This information is available in the US sample but is bypassed in favor of the *pps* methodology.
2. We are allowing the UK to have information about the presence or absence of the specific representative item *c* (equivalent to the ELI) in the list of shops before sampling, whereas the US only in effect knows of the existence of *some* ELI in the given item stratum. (This assumes a multiple ELI-to-POPS category mapping, which was typically the case until recently in US operations; the current version of ELI-to-TPOPS (telephone point of purchase survey) category mappings is 1 to 1; that is, an outlet frame is constructed for each individual ELI.)

(c) Traditionally, for each representative item *c*, within a given shop, the field economist selects that variety *i* which he/she regards as dominating its sales – a judgment sample of the most consistently purchased variety. We formalize this as follows:

1. For a given shop/representative item pair (j, c), we list all varieties *i*.
2. For each variety, we find the minimum quantity $q_i^* = \text{Min}(q_i^y, q_i^{y+1})$ over two years.
3. We sample the variety *i* with $\text{Max}\{q_i^*\}$.

This process, of course, requires more information than a field economist would have at the earlier time period (and again is not used in the US sampling described above) but may be regarded as providing a surrogate for the field economist's appraisal of the relative continuity of goods sold.

Note: It is convenient to refer to the combination of selecting an outlet by *srswor* as in (b), and an item within the shop as described in (c), as *maxminq sampling*.

3.4 UK Estimation

Elementary aggregates for the UK were calculated by a Ratio-of-Averages (RA) formula within each cross-classification cell defined by region, shop type, and representative item This is basically an unweighted estimate, given for independent shops by

$$\hat{I}_{lkhc}^{y,\,y+1} = \frac{\frac{1}{n}\sum_{i\in c,\ j\in k} p_{ljhci}^{y+1}}{\frac{1}{n}\sum_{i\in c,\ j\in k} p_{ljhci}^{y}}.$$

In the case of multiples, a weighted version of the above formula is used with expenditures by shop type, estimated from the ARI, providing relative weights of central versus non-central multiples.

A countrywide index for representative items c in the sample (aggregated over shop types k and regions l) is then calculated by a Laspeyres type estimator:

$$\hat{I}_c^{y,\,y+1} = \sum_l \sum_k \tilde{w}_{lkhc}\, \hat{I}_{lkhc}^{y,\,y+1},$$

where $c \in h$, and $\tilde{w}_{lkhc}$ is based on $\tilde{E}_{.kh}^{y}$ from the ARI and $\hat{E}_{lh}^{95}$ from the FES (using these time periods keeps information used the same between US and UK). Further aggregation (over representative items *c*) is done using $\hat{E}_{hc}^{y}$, *etc*. from the FES.

3.5 Comparison

Table 4 gives a summary comparison of the two methodologies, US and UK, that we have been considering. The predominant feature of the US method is strict probability sampling and estimation, typically *ppswr*; that of the UK is selective sampling, taking the most important item or category as judged by expenditure or quantity sold. The methods of forming elementary aggregates are different, and the weights for aggregation in the UK are estimated at a slightly coarser level at the lower stages.

Table 5 gives a summary of what might be considered the strengths and weaknesses of the US and UK methodologies. By the advantage of "brute strength," which we attribute to the UK approach, we mean the capitalizing on a combination of two factors that often play a role in pricing and price index construction. In the first place, market leaders tend to dominate the price scene; for example, if they sharply lower or raise prices, their lesser competitors selling similar goods may think it necessary or warranted to follow suit. Secondly, even if there is variation in the price trends among similar goods, the leading sellers are likely to dominate the price index by virtue of large expenditure values, that is, because of their correspondingly large weights.

Table 4
Summary Comparison of US and UK Methodologies

	US	UK
HH Exp. Survey	$\hat{E}^{95}_{lc}$	$\hat{E}^{95}_{.c}$, $\hat{E}^{95}_{lh}$
Outlet Exp./Category	HH(POPS) $\tilde{E}^{y}_{ljc}$	Shops Survey (ARI) $\tilde{E}^{y}_{.kh}$
select item categories	2 ELI's *c*/item stratum *h*/PSU *l* *ppswr* $(\hat{E}^{95}_{lc} / \hat{E}^{95}_{lh})$	2 rep. items's *c*/section *h*/Region *l* *largest* $(\hat{E}^{95}_{.c} / \hat{E}^{95}_{.h})$
select outlets	8 outlets *j*/ELI $c \times$ PSU *l* *ppswr* $(\tilde{E}^{y}_{ljc} / \tilde{E}^{y}_{l.c})$	8 outlets *j*/rep. item $c \times$ *Region* $l - srs$ within shoptype *k*, $E^{y}_{ljc} > 0$
item within outlet/category	1 item *i*/*jc pps* $(E^{y}_{ljci} / E^{y}_{ljc.})$	1 variety *i*/*jc* $\max[\mathrm{Min}(q^{y}_{ji}, q^{y+1}_{ji})]$
elementary index	$\hat{I}^{y,\,y+1}_{lh} = \prod_{\substack{j\in l,\\ i\in c\in h\\ (i,\,j)\in s}} \left(\frac{p^{y+1}_{ljhci}}{p^{y}_{ljhci}}\right)^{s_{ljhci}}$	$\hat{I}^{y,\,y+1}_{lkhc} = \frac{\frac{1}{n}\sum_{i\in c,\ j\in k} p^{y+1}_{ljhci}}{\frac{1}{n}\sum_{i\in c,\ j\in k} p^{y}_{ljhci}}$
higher aggregation	$\hat{I}^{y,\,y+1}_{.h} = \frac{\sum_{l} \hat{E}_{lh}\, \hat{I}^{y,\,y+1}_{lh}}{\sum_{l} \hat{E}_{lh}}$	$\hat{I}_{c} = \sum_{l}\sum_{k} \hat{\hat{w}}_{lkhc} \hat{I}_{lkhc}$, $\hat{\hat{w}}_{lkhc} = f(\tilde{E}_{.kh}, \hat{E}_{l.h})$

Table 5
Comparison of US, UK Approaches

	Strengths	Weaknesses
US	Gather more information More use of information Satisfies classical sampling theory Gives regional (PSU) estimates Weighted estimators at lowest level More standardized operating procedure	Possible repetition in selection Ignores stratification of shops (that is, classification into chains)
UK	Relies on "Brute Force" principle Stratification of outlets Shops survey in field Uses variety of sources	Patchwork of weights Inconsistent in Centralized pricing aggregation? Unweighted and seemingly arbitrary estimator at lowest level

4. Results of Primary Study

Indexes comparing '95 to '96 are given in Table 6, for the population (1) as a whole (the three areas combined), (2) broken down by classes/major groups, and (3) broken down even further into item strata/sections. Four indexes are given which might be taken as the targets of estimation. Recall the discussion on targets which concludes Section 1.

Table 7 gives corresponding means, variances, and mean square errors for US and UK estimates, where the mean square error is computed with respect to the Fisher indexes. We observe the following:

1) For the all-items, classes, and item strata, the US estimates appear to approximate the *geomean* G. This confirms what we have suspected from other work (Dorfman *et al*. 1999), namely that the lowest level of aggregation dominates (we used a Laspeyres formula for higher level aggregation). The fact that G lies between the Laspeyres and superlative target provides some evidence that the US switch to this method of elementary aggregation was a step in the right direction.
2) There appears to be no clear order relation of UK *Section* estimates to their corresponding targets; for example, the Section 11 index is higher than the target L, while the Section 12 index is lower

than the superlatives, *etc*. As we aggregate up to the Major Group and All Items levels, however, the estimates clearly begin to approximate the superlatives *F* or *T*. (Dalén (1998) noted a similar result in aggregating cut-off samples.)

3) If we take the Fisher as the target, *even at* the section level, the root mean square error of the UK estimator is much lower than that of the US estimator. Given the relatively restricted nature of the UK sample design, it is not surprising that the UK estimator displays lower variance, but the form of the UK estimator would not lead one to expect it to unbiasedly approximate a Fisher index. Nonetheless, our results suggest that, at least for a population of purchases such as the one used in this study, the purposive, "brute force" methods of the UK (and many other countries) work well.

Similar results were found for the succeeding pairs of years through '99 – '00. Figure 6 shows the all items year-to-year *geomean* and Fisher for five pairs of years and the means across samples of the corresponding US and UK estimators. (Note the difference in scale between Figure 6 and Figures 1 through 5). It is readily seen that the U.S. estimator tends to track the population *geomean*. The UK estimator, tracking the Fisher, tends to overestimate in the later years, although it runs much closer to the Fisher than to the population *geomean*. It should be noted that we used increasingly out-of-date expenditure data, namely the '95 data, for purposes of sampling and estimation. It is possible that outmoded expenditure data are having a greater impact on the UK estimates than on the US estimates, perhaps by leading us to oversample expensive representative items or to focus on some group of shops that are increasingly pricey.

Results for the classes ("hot," *etc*.) were very similar for the US vis-à-vis the *geomean* and are not shown. Figure 7 shows the difference between the mean year-to-year UK estimates and the Fisher, for each of the four classes. It can be seen that the tendency to overestimate in the later years affects all four classes.

Overall, the UK estimators provide better estimates of the superlative Fisher target than do the US estimators. Table 8 gives the ratio of UK root mean square error to US root mean square error, for all five pairs of years, for all items, for groups, and for sections. There are a few anomalous places, notably in the '98 – '99 indexes where, for section 2 of "hot," and consequently for the entire class "hot," the UK estimates are appreciably worse. In general, however, the UK methods provide much better estimates. This is due in part to a tighter sampling structure (mainly because purposive/cutoff sampling is much more restrictive than random selection of the set of items which can enter the sample), yielding, not surprisingly, less variance. In part though, as well, it is due to a surprising tendency of the UK estimators to target the corresponding Fisher indexes, reducing bias. Since the UK estimators do not formally resemble the Fisher index, the reasons for their tendency to approximate it merit further study. We turn to this issue in the next section.

Table 6
Potential Target '95 – '96 Indexes

Description	*geomean*	Törnqvist	Fisher	Laspeyres
All	1.053	1.002	0.997	1.079
Classes/Major Groups				
1 – Hot	1.058	1.052	1.052	1.078
2 – Sugary	1.042	0.964	0.956	1.072
3 – Fruity	1.044	1.007	1.007	1.067
4 – Plain	1.069	1.027	1.027	1.092
Item Strata/Sections				
Hot – 11	1.043	1.044	1.044	1.057
Hot – 12	1.073	1.059	1.058	1.097
Sugary – 21	1.003	0.917	0.910	1.034
Sugary – 22	1.063	0.982	0.972	1.093
Sugary – 23	1.093	1.052	1.054	1.119
Fruity – 31	0.977	0.955	0.950	0.985
Fruity – 32	1.165	1.110	1.116	1.204
Plain – 41	1.067	1.021	1.021	1.094
Plain – 42	1.030	0.996	0.996	1.050
Plain – 43	1.104	1.063	1.062	1.125

Table 7
Simulation Results for '95 – '96 Indexes

Description	Target Index	U.S. Mean	U.S. Std. Dev.	U.S. RMSE	U.K. Mean	U.K. Std. Dev.	U.K. RMSE
All	0.997	1.057	0.016	0.062	1.002	0.011	0.012
Classes/Major Groups							
1 – Hot	1.052	1.059	0.031	0.032	1.045	0.022	0.023
2 – Sugary	0.956	1.046	0.030	0.095	0.971	0.023	0.027
3 – Fruity	1.007	1.053	0.035	0.058	0.986	0.027	0.034
4 – Plain	1.027	1.072	0.025	0.051	1.025	0.016	0.016
Item Strata/Sections							
Hot – 11	1.044	1.045	0.035	0.035	1.064	0.025	0.032
Hot – 12	1.058	1.072	0.049	0.051	1.027	0.035	0.047
Sugary – 21	0.910	1.004	0.050	0.106	0.850	0.045	0.074
Sugary – 22	0.972	1.070	0.051	0.111	1.089	0.030	0.121
Sugary – 23	1.054	1.095	0.044	0.060	1.026	0.027	0.039
Fruity – 31	0.950	0.979	0.020	0.035	0.932	0.020	0.027
Fruity – 32	1.116	1.178	0.084	0.104	1.077	0.059	0.071
Plain – 41	1.021	1.069	0.050	0.070	1.060	0.030	0.049
Plain – 42	0.996	1.033	0.035	0.051	0.987	0.031	0.032
Plain – 43	1.062	1.107	0.042	0.061	1.028	0.023	0.041

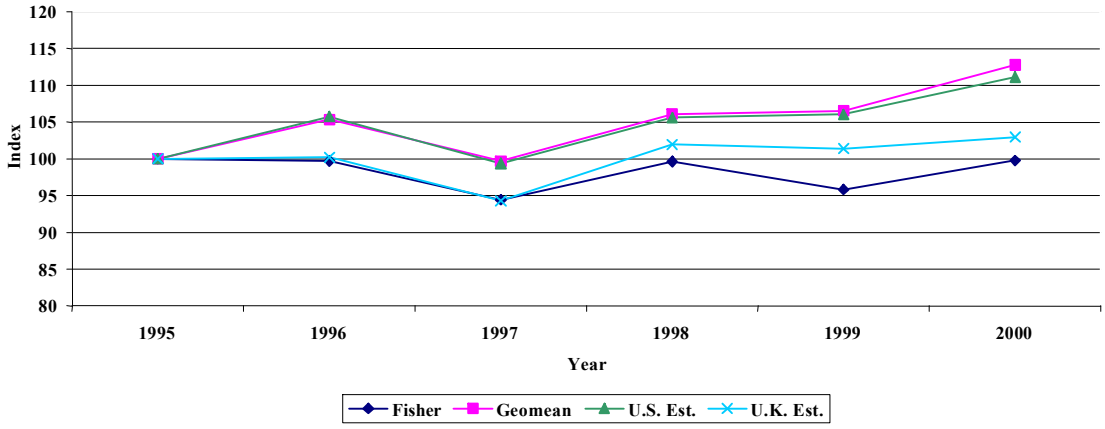


Figure 6. Index Targets and Estimates for All Cereals February-to-February Indexes and Index Estimates, 1995 = 100.

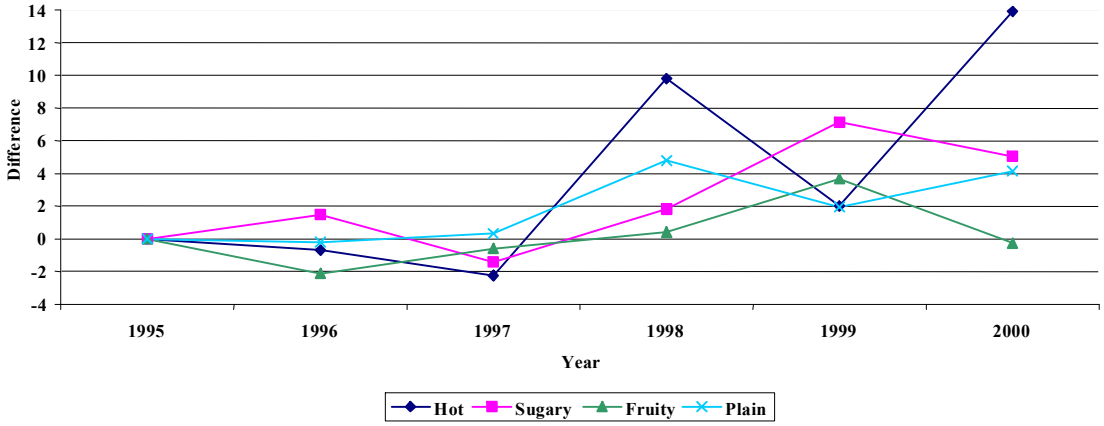


Figure 7. Differences Between U.K. Estimates and Population Fisher Indexes February-to-February Indexes and Index Estimates, 1995 = 100.

Table 8
Ratios of UK RMSE to US RMSE

Description	'95 – '96	'96 – '97	'97 – '98	'98 – '99	'99 – '00
All	0.196	0.192	0.419	0.548	0.288
Classes/Major Groups					
1 – Hot	0.713	0.517	0.483	1.437	0.589
2 – Sugary	0.286	0.336	0.314	0.522	0.282
3 – Fruity	0.595	0.508	0.308	0.501	0.405
4 – Plain	0.310	0.297	0.777	0.319	0.404
Item Strata/Sections					
Hot – 11	0.923	1.066	0.682	0.529	0.508
Hot – 12	0.920	0.850	1.169	1.860	0.842
Sugary – 21	0.702	0.392	0.421	0.595	0.330
Sugary – 22	1.092	0.426	0.380	0.341	0.365
Sugary – 23	0.650	0.455	0.448	0.925	0.851
Fruity – 31	0.778	1.059	0.637	0.581	0.618
Fruity – 32	0.683	0.809	0.314	0.457	0.356
Plain – 41	0.709	0.623	0.494	0.567	0.317
Plain – 42	0.642	0.511	1.117	1.092	1.005
Plain – 43	0.678	0.839	0.641	0.815	0.701

5. Follow-Up Study

There are four aspects in which the approaches of the UK and US differ: (1) the stratification structure, in particular, the reliance of the UK on different shops strata and, to an extent, on centralized sampling, (2) the aggregation and weighting structure, (3) the mode of sampling at different stages, and (4) the formula for elementary aggregates. This makes it difficult to disentangle the extent to which each aspect is contributing to the relative merits of US and UK index construction. In particular, as noted in the last section, it is a bit of a mystery why, especially at higher aggregations, the UK index estimator tends to target the superlative indexes.

In our follow-up study we focus on the lowest level of index construction, that is, on (3), the shop-representative item (ELI) level of sampling and on (4), the formulas for the elementary indexes. We compare the relative merits of different options, taking the within area elementary indexes as our targets. Aggregation to higher level indexes will be carried out uniformly for all alternative lower level options considered, using the true population expenditure shares. The importance of the method of construction of the elementary indexes is widely recognized; see Diewert (2004) and references; also Dorfman *et al.* (1999). The example discussed in Appendix B, with results given in Table 9, illustrates the decisive effect that the lowest level of index construction has on the index as a whole.

Thus, a likely important source of the difference in results of US and UK methodology lies in the sample estimation of the population elementary indexes. But this leaves open the question whether the differences arise because of differences in sampling method or in the formulas used in estimation, or in both. Thus we are interested in determining: (1) how judgment sampling (in this case, cutoff sampling based on *maxminq*) performs compared to probability sampling represented by *ppswr*, holding the estimator of the elementary indexes fixed, and (2) how estimators of elementary indexes compare when we keep the sampling method fixed. It will also be of some interest to determine what happens when *maxminq* sampling is based on data from the base and *previous* time period, rather than the base and current period.

5.1 Sampling Methods and Estimators at the Elementary Level

To explore these questions, we carried out further simulation studies. The data were the same Cereal Data used in the primary study (successive Februarys), but limited to the Independent Shops, *Chain 8*. This was done to make the study more manageable but also because, for the other chains, the UK elementary index estimators were more complicated than the simple *dutot*. Also, it is reasonable to expect price behavior to be most heterogeneous in this chain, so that inherent differences will be clearer. Chain 8 was the largest of the chains, comprising each year about 30% of the whole population, approximately 6,000 records.

The basic structure remained the same: 3 *psu*'s, 4 major groups/expenditure classes (hot, sugary, fruity, and plain), 10 sections/item strata, and 29 representative items/*ELI*'s. For each *ELI/representative item*, 3 outlets (one item per outlet) were selected, as opposed to 10 in the primary study above. For investigating *maxminq* based on previous time periods, the original 5 data sets, each using price and quantity data for a pair of years ('95/'96, '96/'97, *etc.*) were reduced to include only items that allowed "back matching";

that is, matching across three years to compare prices of items in outlets for '95/'96/'97, '96/'97/'98, *etc*. About 90% of the Chain 8 records allowed back matching. (In considering the results below, it is probably worth noting that the sample reduction could disproportionately impact the back matched *maxminq*). We shift our attention from the Fisher index to the superlative Walsh index, due to an astute suggestion of a referee, discussed in Appendix C.

Three estimators were used for elementary indexes: the ratio of averages (RA) (the *dutot*), the unweighted *geomean* (also known as the Jevons), and the average of ratios (*AR*). In the *pps* sampling of outlets, and then in the sampling of items within outlets, the size variable (expenditure) was assumed known (rather than being estimated, as in the main study). Besides *pps* with replacement (as in the US approach), and *maxminq*, we also investigated *pps* without replacement, on the suspicion it would be less variable than the with replacement version.

For each mode of sampling, within each *psu/ELI* combination, we took 500 samples. We calculated the mean square error of estimates with respect to a target *ELI* – level Walsh Index. Averages of *mse* across *ELI*'s were calculated for each mode of sampling/estimation, within each *psu.*

Table 10 shows the ratio of these averages to the average *mse* for the *maxminq/dutot* combination. For each estimator, for each *psu*, with one exception (*psu* 3, '99/'00), *maxminq* leads to lower *mse*, often by an appreciable margin. Sampling *pps* without replacement is second best. Holding the method of sampling fixed (comparing rows 1, 4, 7, then 2, 5, 8, *etc*. in Table 10), we note that with few exceptions, the *dutot* does better than the *geomean*, which does better than *AR*. These results suggest: (1) *maxminq* is better than *pps*(exp), and *pps*(exp) is better than *ppswr*(exp). (2) The *dutot* is more efficient than the *geomean*, and the *geomean* is more efficient than an average of ratios. There is a beneficial synergism between *maxminq* sampling and the *dutot.* Biases and variances were also studied, and the results (not shown) tended to follow the same pattern.

Table 9
Population Indexes '95 – '96, Chain 8

Description	Laspeyres	*geomean**	Fisher	Walsh	Laspeyres of Walsh Elementary
All	1.129	1.091	1.028	1.030	1.040
Classes/Major Groups					
1 – Hot	1.161	1.115	1.080	1.082	1.084
2 – Sugary	1.129	1.088	1.007	1.012	1.025
3 – Fruity	1.084	1.054	0.997	1.005	1.015
4 – Plain	1.135	1.101	1.046	1.042	1.050
Item Strata/Sections					
Hot – 11	1.157	1.117	1.088	1.089	1.090
Hot – 12	1.164	1.113	1.072	1.075	1.079
Sugary – 21	1.086	1.045	0.962	0.970	0.992
Sugary – 22	1.187	1.142	1.055	1.056	1.058
Sugary – 23	1.117	1.091	1.034	1.039	1.043
Fruity – 31	1.003	0.992	0.949	0.965	0.966
Fruity – 32	1.228	1.172	1.100	1.091	1.102
Plain – 41	1.212	1.161	1.091	1.080	1.090
Plain – 42	1.048	1.030	0.997	0.997	0.998
Plain – 43	1.136	1.107	1.048	1.046	1.056

* Weighted by base period expenditure.

Table 10
Standardized Average Relative Mean Square Error Across ELI's, Reduced Populations, Chain 8

	psu 2				*psu* 3				*psu* 4			
estimator/sampling method	'96 – '97	'97 – '98	'98 – '99	'99 – '00	'96 – '97	'97 – '98	'98 – '99	'99 – '00	'96 – '97	'97 – '98	'98 – '99	'99 – '00
dutot/*maxminq*(UK)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
dutot/*ppswor*	1.73	1.70	1.68	1.91	1.23	1.82	1.35	2.24	1.22	1.06	1.12	0.93
dutot/*ppswr*	2.13	2.10	1.91	2.14	1.42	2.10	1.46	2.67	1.45	1.23	1.36	1.07
geomean/*maxminq*	1.20	1.16	1.16	1.06	1.06	1.14	1.08	1.05	1.10	1.11	1.12	0.96
geomean/*ppswor*	2.08	1.88	1.98	2.27	1.33	1.94	1.47	2.59	1.33	1.09	1.28	0.97
geomean/*ppswr* (US)	2.49	2.29	2.18	2.53	1.58	2.23	1.58	3.09	1.59	1.30	1.52	1.12
AR/*maxminq*	1.42	1.32	1.31	1.14	1.24	1.03	1.30	1.05	1.11	1.20	1.21	1.07
AR/*ppswor*	2.81	2.35	2.49	2.85	1.70	2.31	1.77	3.43	1.57	1.30	1.42	1.17
AR/*ppswr*	3.23	2.77	2.66	3.08	2.03	2.58	1.87	3.96	1.83	1.49	1.66	1.30
dutot/*maxminq*, prior *q*'s	1.12	1.19	1.19	1.41	1.56	1.42	1.69	1.51	1.20	1.02	0.85	1.48

That the *dutot* sample index can target the Walsh population index (and hence indirectly any superlative index), when consistently largest sellers are sampled is, we suggest, the result of a very simple, "brute force" mechanism: to the extent that the Walsh can be represented by a small sample of items, it is best represented by those with the consistently largest quantities, and it is these items that the *maxminq* sampling scheme virtually always supplies. In Appendix C we discuss an alternative explanation for the good performance of the *maxminq/dutot* combination.

Average mean square errors were also calculated for the *maxminq/dutot* combination based on *previous* values of q, that is on q_i^{y-1}, q_i^y. Results are given in the last row of Table 10. There is an anticipated weakening compared to the updated *maxminq/dutot*, but the results still compare favorably to the other options. We study this further in subsection 5.2.

5.2 Effect of Lagged Quantities on *maxminq* Sampling

To put the results of Section 4 in perspective, we need to inquire what the effect is of using lagged q's in *maxminq*. The reason is simple: although at first sight, using base and current period quantities seems the obvious way to capture the UK's idea of persistent items, nonetheless, this involves using information (the current period quantities) which was not used in simulating US sampling. Perhaps this gives the UK methodology an unfair edge.

We therefore compared the US approach, viz. *ppswr* (with size variable being base period expenditure) and *geomean* at the elementary level, with the UK approach represented by *maxminq – dutot*, but now with *maxminq* based on quantities q_{y-1} and q_y. Data sets were reduced slightly to guarantee that we would have matching data for three consecutive years. Aggregation to upper level indexes used actual population expenditures for both US and UK.

Table 11 gives results for the All Cereals Indexes for chain 8, comparing biases, standard deviations, and root mean square errors with respect to the population Walsh. As expected, the results are not as good as those obtained by using current q's. Nonetheless, with respect to all three accuracy measures (bias, standard deviation, and root mean square error), the UK *maxminq/dutot* combination still does better than the US approach representing probability sampling.

For finer categories, Table 12 gives the ratios of mean square errors obtained under the UK method with lagged q's to those obtained under the US method. Although they are generally larger than those in Table 8, they still suggest that the purposive sampling approach of the UK is better.

6. Discussion

We have presented a comparison of two fundamentally different approaches to sample design and inference for a consumer price index. The inescapable conclusion is that, in the population we studied, the "UK" approach, which involves tighter stratification and, more importantly, more restrictive judgment sampling within strata than the probability sampling of the "US" approach, does better in estimating a target superlative index.

This is shown to be the case, whichever low level price index estimator (the *dutot*, or *geomean*, or the average of ratios) is employed, although the *dutot* (ratio of averages) performed best.

The UK approach does better for two reasons: (1) its tighter sampling, restrictive of items selected (for example, see Table 13 described in Appendix C), leads, not surprisingly, to lower variance, an observation made already in de Haan *et al*. (1999), and (2) the *dutot* sample indexes target the superlative indexes under dominant market sampling, which *was* surprising and called forth the investtigation described in Section 5. On the other hand, the US approach yielded an index estimator which could be described as unbiased, but it was unbiased for the (wrong) population *geometric* index weighted by first period expenditure. Thus it tended to run considerably higher than the target superlative index (whether Fisher, Walsh, or Törnqvist).

If sample sizes were allowed to increase, we could anticipate that the variances of both the US and UK would decrease, but the UK variance would remain lower. The bias of the US estimator for the superlative target would be unaffected by increased sample size, so that the relative mean square error of the UK approach would be increaseingly lower.

In practice, of course, period 2 quantities are not available at the time of sample selection (at period 1), and as part of our follow-up study we give some measure of the partial degradation that arises from using past quantities: it is not severe enough to undo the conclusion of better UK performance. Furthermore, the field economist's judgment as to the best seller might be able to invoke data more recent than a year earlier. Thus the actual effect might be somewhere between the lagged and non-lagged versions of *maxminq* which we have used. In practice, however, US field economists may often sample items within outlets based on an estimate of expenditure share that is really a smoothed average of base period *and* recent expenditure shares. This may attenuate the bias we have seen in our simulations, where only the base period expenditures were used for within-store sampling.

Table 11
Biases, Standard Deviations, and Root Mean Square Error (Each Multiplied by 1,000), in Estimating Population All Cereals Walsh Index, Chain 8, Based on Three Approaches to Sampling/Estimating Elementary Indexes*

	(a) Bias				
	'95–'96	'96–'97	'97–'98	'98–'99	'99–'00
dutot/maxminq	29	15	–13	33	2
dutot/maxminq, prior q's	–	46	32	82	36
geomean/ppswr	78	62	66	82	66
	(b) Standard Deviation				
	'95–'96	'96–'97	'97–'98	'98–'99	'99–'00
dutot/maxminq	16	13	11	14	12
dutot/maxminq, prior q's	–	14	12	15	14
geomean/ppswr	22	18	17	18	20
	(c) Root Mean Square Error				
	'95–'96	'96–'97	'97–'98	'98–'99	'99–'00
dutot/maxminq	33	20	17	36	12
dutot/maxminq, prior q's	–	48	34	83	39
geomean/ppswr	80	65	68	84	68

* At ELI/Representative Item level. To get overall index estimates, the elemenentary index estimates were aggregated using known population expenditures.

Table 12
Ratios of UK RMSE to US RMSE, Chain 8, Walsh Targets:
maxminq Using Lagged *q*'s & *dutot* Versus *ppswr*(Expenditure) & *geomean*

Description	'96–'97	'97–'98	'98–'99	'99–'00
All	0.748	0.498	0.993	0.567
Classes/Major Groups				
1–Hot	1.539	0.495	1.280	0.765
2–Sugary	0.563	0.676	0.941	0.797
3–Fruity	0.409	0.323	0.463	0.852
4–Plain	0.915	0.560	1.164	0.359
Item Strata/Sections				
Hot–11	0.748	0.607	0.660	0.657
Hot–12	1.695	0.599	1.333	0.843
Sugary–21	0.757	0.593	1.136	0.924
Sugary–22	0.370	0.776	0.751	0.671
Sugary–23	0.479	0.785	0.796	0.508
Fruity–31	0.570	0.443	0.678	1.008
Fruity–32	0.526	0.350	0.277	0.674
Plain–41	1.167	0.509	1.395	0.397
Plain–42	0.623	0.411	0.918	0.624
Plain–43	0.919	1.171	0.668	0.560

Table 13
Items Selected by *maxminq* and *pps*($\sqrt{q_y q_{y+1}}$) in 500 Samples

'95 – '96, Chain 8, *psu* 2, ELI 105											
pps	items selected	2889	2803	1564	2763	1558	2242	2344	2776	760	2850
	% of samples in which selected	43.2	32.2	10.4	5.4	3.87	1.53	1.33	0.87	0.8	0.4
maxminq	items selected	2889	2803								
	% of samples in which selected	80.87	19.13								
'95 – '96, Chain 8, *psu* 3, ELI 401											
pps	items selected	1731	2378	2866	1742	2922	2375	2528	403	871	
	% of samples in which selected	33.27	18.8	12.8	12.73	9.47	4.6	4.27	2.8	1.27	
maxminq	items selected	2378	1731	2866	1742						
	% of samples in which selected	46.27	24.47	15	14.27						
'99 – '00, Chain 8, *psu* 4, ELI 401											
pps	items selected	1731	2866	1742	2378	2922	2528	403			
	% of samples in which selected	30.07	21.93	14.3	11.07	9.53	6.8	6.27			
maxminq	items selected	1742	2866	2922	1731						
	% of samples in which selected	34.27	30.87	18	16.87						

It is generally accepted that the non-randomization approaches are intrinsically cheaper. For example, there are typically fewer outlets to visit, and price collection within outlets is less labor intensive. Thus, for a given budget we can expect the UK approach to be more efficient, compared to US probability sampling, than the present study suggests.

It would be salutary to expand this study to scanner data for products other than cereals. In particular, items with more volatile price movements would be of great interest. To some extent, the good behavior of *maxminq/dutot* may be related to the surprising closeness of the population *dutot* to the superlatives (as seen in Table 1). How typical is such closeness, and, if it is absent, will the good sampling behavior persist?

One final *caveat*. It may be a good idea in practice to inject a dose of randomness at some stage or stages of the sampling process, and in particular be a bit cautious about centralized sampling – not for statistical reasons, but to guarantee fairness and the appearance of fairness (Reinsdorf and Triplett 2005, Section II; Royall 1976).

Acknowledgements

The opinions expressed in this paper are those of the authors and do not represent US Bureau of Labor Statistics or Bureau of Transportation Statistics policy. The authors thank David Richardson and Lyuba Rozental for providing us with the cereal data and for timely assistance, Sonja Mapes and Scott Pinkerton for their work on the classification of cereals into types, and Mick Silver, Adrian Ball, and Dawn Camus for providing understanding and materials on the United Kingdom's RPI methods. The authors wish also to thank three referees and an associate editor for many insightful comments and for encouraging us to expand the study, and J. De Haan, M. Reinsdorf, and B. Moulton for their helpful suggestions. We especially wish to acknowledge the encouragement of the late M.P. Singh whose suggestions as Editor guided the final course this paper has taken.

Appendix A
Targets – Population Indexes

Laspeyres $$L = \frac{\sum_i q_i^y p_i^{y+1}}{\sum_i q_i^y p_i^y}$$

Paasche $$P = \frac{\sum_i q_i^{y+1} p_i^{y+1}}{\sum_i q_i^{y+1} p_i^y}$$

Walsh $$W = \frac{\sum_i \sqrt{q_i^y q_i^{y+1}}\, p_i^{y+1}}{\sum_i \sqrt{q_i^y q_i^{y+1}}\, p_i^y}$$

Fisher $$F = \left\{ \frac{\sum_i q_i^y p_i^{y+1}}{\sum_i q_i^y p_i^y} \frac{\sum_i q_i^{y+1} p_i^{y+1}}{\sum_i q_i^{y+1} p_i^y} \right\}^{1/2} = \sqrt{LP}$$

Törnqvist $$T = \prod_i \left(\frac{p_i^{y+1}}{p_i^y} \right)^{s_i^{y,\,y+1}},$$

where

$$s_i^{y,\,y+1} = \frac{1}{2}\left(\frac{p_i^y q_i^y}{\sum_i p_i^y q_i^y} + \frac{p_i^{y+1} q_i^{y+1}}{\sum_i p_i^{y+1} q_i^{y+1}} \right)$$

Geometric Mean $$G = \prod_i \left(\frac{p_i^{y+1}}{p_i^y} \right),$$

where

$$w_i = s_i^y = \left(\frac{p_i^y q_i^y}{\sum_i p_i^y q_i^y} \right)$$

or

$$w_i = 1/N$$

Unit Value $$U = \frac{\sum_i q_i^{y+1} p_i^{y+1} \Big/ \sum_i q_i^{y+1}}{\sum_i q_i^y p_i^y \Big/ \sum_i q_i^y}$$

dutot $$RA = \frac{\sum_i p_i^{y+1} \Big/ N}{\sum_i p_i^y \Big/ N}$$ ("Ratio of Averages")

Average of Ratios $$AR = \frac{\sum_i p_i^{y+1} \Big/ p_i^y}{N}$$

Appendix B
An Example Illustrating the Importance of Lowest Level Aggregation

We here present a simple example to illustrate the importance of the method used for constructing the elementary indexes. We compare population Walsh indexes to indexes resulting from aggregating elementary Walsh indexes according to a Laspeyres formula instead. The reason for focusing on the Walsh is given in Appendix C. The "pure" Walsh index is

$$\mathrm{W} = \frac{\sum_i \sqrt{q_i^y\, q_i^{y+1}\, p_i^{y+1}}}{\sum_i \sqrt{q_i^y\, q_i^{y+1}\, p_i^{y}}} = \sum \tilde{s}_h\, W_h^{y,\,y+1},$$

where the $W_h^{y,\,y+1}$ are the h^{th} elementary Walsh indexes and

$$\tilde{s}_h = \frac{\sum_{i\in h} \sqrt{q_i^y\, q_i^{y+1}\, p_i^{y}}}{\sum_i \sqrt{q_i^y\, q_i^{y+1}\, p_i^{y}}}$$

are proper Walsh aggegration weights. To this we compare a Laspeyres aggregation of elementary Walsh indexes ("ersatz Walsh"), $L_W^{y,\,y+1} = \sum\sum\sum s_h\, W_h^{y,\,y+1}$, where the s_h are standard base period weights.

The results are given in Table 9. We do see a perceptible difference between the actual population Walsh and the Laspeyres aggregate of elementary Walsh indexes: the latter tends to run slightly higher. However, these differences are on a par with the differences between them and the Fisher. They are minor compared to the gap between the *geomean* or Laspeyres indexes and the superlatives. This sort of result verifies that sound procedure at the lowest level is a key part of index construction.

Appendix C
The *maxminq/dutot* Combination

Why does the *maxminq/dutot* combination work so well, seeming to lead to unbiasedness for the superlative indexes?

A referee notes that *maxminq* sampling bears a strong resemblance to sampling *pps* with size variable $\sqrt{q_i^y\, q_i^{y+1}}$; for *ppswor* $(\sqrt{q^y q^{y+1}})$, the *dutot* is approximately unbiased for a Walsh target index, and so, indirectly, for any other superlative index.

Indeed, for the expectation of the numerator of the *dutot*, under this probability sampling scheme, we have

$$E_\pi\left(\sum_{i\in s} p_i^{y+1}\right) = E_\pi\left(\sum_{i'\in U} I_{i'}\, p_{i'}^{y+1}\right) = \frac{n}{\sum_{i'} \sqrt{q_{i'}^y\, q_{i'}^{y+1}}} \sum_{i'} \sqrt{q_{i'}^y\, q_{i'}^{y+1}}\; p_{i'}^{y+1},$$

where $E_\pi(\,)$ signifies expectation with respect to the sample design and $I_{i'}$ is a random indicator taking the values 1 or 0, as i' is in the sample or not. We get a similar expression for the denominator. The ratio of these two expected values is the Walsh. Therefore, apart from the usual (mild) ratio bias, which can be shown to be typically positive, the *dutot* does indeed target the Walsh, under this *pps* scheme.

We need to ask: do the two modes of sampling actually tend to have a sizeable overlap in what items get picked? For each run, for each *psu l*, *ELI c*, three items were selected either by *maxminq* or by *ppswor* $(\sqrt{q^y q^{y+1}})$ of items within *lc*. Table 13 gives the percentage of times (over 500 runs) different items make it into the sample, for some arbitrarily selected representative cases. We conclude, not entirely without surprise, that: (a) *pps* sampling leads to a wider spread of items selected, (b) the items selected by *maxminq* are a subset of those from *pps*, (c) there is a certain amount of correlation of "dominant items", that is, of those items that tend most to be selected by either method. In short, *maxminq* and *pps* $(\sqrt{q^y q^{y+1}})$ appear to be related, but loosely.

To get further insight into the relationship between the two sampling methods, we calculated bias and mean square error estimates, with respect to the Walsh population index, for the *dutot* index for each ELI, both for *maxminq* and *pps* $(\sqrt{q_y\, q_{y+1}})$ sampling. The bias and MSE estimates were based on 500 runs for each sampling method. Summary statistics were calculated across ELI's for each pair of years and each *psu*. Table 14 gives the percentage of ELI's for which the *dutot* elementary indexes are positively biased for each mode of sampling. As anticipated, *pps* sampling tends to result in positive bias; we find that *maxminq* is equally biased positive and negative.

Table 14
Percentage of ELI's for Which the *dutot* has Positive Bias for a Walsh Target, for Two Sampling Schemes

	pps$(\sqrt{q_y q_{y+1}})$			*maxminq*		
	psu 2	*psu* 3	*psu* 4	*psu* 2	*psu* 3	*psu* 4
'95–'96	75.0	86.2	75.9	64.3	61.1	61.1
'96–'97	60.7	72.4	65.5	53.6	65.5	51.8
'97–'98	65.5	75.9	78.6	41.4	27.6	42.9
'98–'99	72.4	75.9	70.4	48.3	75.9	40.8
'99–'00	89.7	72.4	75.9	48.3	20.7	44.9

Table 15 (a) gives the percent of ELI's in which the absolute bias from using *maxminq* is bigger than that from *pps* ($\sqrt{q_y\, q_{y+1}}$). In this regard, *pps* sampling is better. However, Table 15 (b) gives the percentage of ELI's in which *maxminq* yielded a larger mean square error, and here *maxminq* does better in all but two time periods/*psu*'s. We regard the mean square error criterion as the more decisive, especially given the bi-directionality of *maxminq*'s biases.

Table 15
Percentage of ELI's for Which the *dutot*'s Bias and Mean Square Error for a Walsh Target is Less for Probability Proportional to Size (Size Variable = $\sqrt{q_y q_{y+1}}$) than for *maxminq* Sampling

	(a) *Bias of pps less*			(b) *MSE of pps less*		
	psu 2	*psu* 3	*psu* 4	*psu* 2	*psu* 3	*psu* 4
'95 – '96	82.1	93.1	86.2	32.1	58.6	41.4
'96 – '97	89.2	96.6	100.0	35.7	37.9	27.6
'97 – '98	89.7	86.2	100.0	41.4	24.1	64.3
'98 – '99	89.7	82.8	92.6	41.4	37.9	40.7
'99 – '00	89.7	96.6	41.4	34.5	31.0	37.9

We conclude that the good effects of *maxminq* sampling combined with the *dutot* estimator are *not* explainable in terms of approximate mimicry of *pps* sampling. They behave differently; and overall m*axminq* seems to be somewhat *better* than *pps* ($\sqrt{q_y\, q_{y+1}}$).

We can see no alternative to explain why the *dutot* sample index should target the Walsh population index when the consistently largest sellers are sampled than that of this "brute force" mechanism: to the extent that the Walsh can be represented by a small sample of items, it is best represented by those with the consistently largest quantities, and these items are the ones the *maxminq* sampling scheme supplies.

References

Balk, B. (1999). On the use of unit values as consumer price subindices. *Proceedings of the Fourth Meeting of the International Working Group on Price Indices*, BLS, Washington, D.C.

Balk, B. (2003). Price indexes for elementary aggregates: The sampling approach. *Proceedings of the Seventh Meeting of the International Working Group on Price Indices* (*Ottawa Group*), Paris.

BLS Handbook of Methods (2005). http://stats.bls.gov/bls/descriptions.htm.

Consumer Price Indexes Technical Manual (2005). Office for National Statistics, London, http://www.statistics.gov.uk/downloads/theme_economy/CPI_Technical_Manual_2005.pdf.

De Haan, J., Opperdoes, E. and Schut, C. (1999). Item selection in the consumer price index: Cut-off versus probability sampling. *Survey Methodology*, 25, 1, 31-41.

Dalén, J. (1998). Studies on the comparability of consumer price indices. *International Statistical Review*, 66, 1, 83-113.

Diewert, E. (1997). "Commentary" [on 'Alternative Strategies for Aggregating Prices in the CPI' by M.D. Shapiro and D.W. Wilcox]. Federal Reserve Bank of St. Louis Review, 79, 3, 27-37.

Diewert, E. (2004). Index number theory: Past progress and future challenges. Presented at the SSHRC Conference on Price Index Concepts and Measurement, Vancouver, Canada, at http://www.econ.ubc.ca/diewert/concepts.pdf.

Dorfman, A.H., Leaver, S.G. and Lent, J. (1999). Some observations on price index estimators. *Proceedings of the Federal Committee on Statistical Methodology Research Conference*, *Monday B Sessions*, 56-65.

Reinsdorf, M., and Triplett, J.E. (2005). A review of reviews: Ninety years of professional thinking about the consumer price index. To appear, *Proceedings of the June 2004 NBER-CRIW Conference on Price Indexes*, Vancouver.

The Retail Prices Index Technical Manual (1998). (Ed. M. Baxter, The Stationary Office, London, at http://www.statistics.gov.uk/downloads/theme_economy/RPI_TECHNICAL_MANUAL.pdf.

Richardson, D.H. (2000). Scanner indexes for the CPI. *Proceedings of the Conference on Scanner Data and Price Indexes*, NBER, Cambridge, http://www.nber.org/books/.

Royall, R.M. (1976). Current advances in sampling theory: Implications for human observational studies. *American Journal of Epidemiology*, 104, 463-473.

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

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

Survey Methodology, December 2006
Vol. 32, No. 2, pp. 217-231
Statistics Canada, Catalogue No. 12-001

An Evaluation of Matrix Sampling Methods Using Data from the National Health and Nutrition Examination Survey

Neal Thomas, Trivellore E. Raghunathan, Nathaniel Schenker, Myron J. Katzoff and Clifford L. Johnson [1]

Abstract

Researchers and policy makers often use data from nationally representative probability sample surveys. The number of topics covered by such surveys, and hence the amount of interviewing time involved, have typically increased over the years, resulting in increased costs and respondent burden. A potential solution to this problem is to carefully form subsets of the items in a survey and administer one such subset to each respondent. Designs of this type are called "split-questionnaire" designs or "matrix sampling" designs. The administration of only a subset of the survey items to each respondent in a matrix sampling design creates what can be considered missing data. Multiple imputation (Rubin 1987), a general-purpose approach developed for handling data with missing values, is appealing for the analysis of data from a matrix sample, because once the multiple imputations are created, data analysts can apply standard methods for analyzing complete data from a sample survey. This paper develops and evaluates a method for creating matrix sampling forms, each form containing a subset of items to be administered to randomly selected respondents. The method can be applied in complex settings, including situations in which skip patterns are present. Forms are created in such a way that each form includes items that are predictive of the excluded items, so that subsequent analyses based on multiple imputation can recover some of the information about the excluded items that would have been collected had there been no matrix sampling. The matrix sampling and multiple-imputation methods are evaluated using data from the National Health and Nutrition Examination Survey, one of many nationally representative probability sample surveys conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention. The study demonstrates the feasibility of the approach applied to a major national health survey with complex structure, and it provides practical advice about appropriate items to include in matrix sampling designs in future surveys.

Key Words: Missing data; Multiple imputation; Respondent burden; Split questionnaire; Sample survey.

1. Introduction

Data from sample surveys are used by researchers and policy makers in many fields. These surveys often involve nationally representative probability samples and extensive data collection based on questionnaires, and they must balance the competing goals of reasonable length and completeness in providing relevant information. The number of topics covered by such surveys, and correspondingly the amount of interviewing time involved, have typically increased over the years. The resultant increased respondent burden may be among the factors contributing to the declining response rates that have occurred. Such declining rates can result in reduced precision of survey estimates. They can also result in increased bias, if systematic differences between the nonrespondents and respondents are not accounted for in analyses of the incomplete data. Moreover, the expansion of topics covered, along with efforts to maintain high response rates, have increased the costs of conducting surveys.

One potential solution to the problem of providing the information that is needed while limiting respondent burden is to carefully form subsets of the items in a survey and administer one such subset to each respondent. Different subsets of questions (items) are administered to different subsets of respondents, so that each item is administered to at least some of the respondents. Questionnaire designs of this type are called "split-questionnaire" designs or "matrix sampling" designs, the latter name reflecting the idea that respondents (rows) and items (columns) are both "sampled" from a conceptual complete population data matrix. In many matrix sampling designs, some items (herein called "core" items) are administered to all respondents, whereas other items (herein called "split" items) are only administered to a subset of respondents. Typically, the items chosen to be core items either are especially important or are predictive of many of the split items.

The administration of only a subset of the survey items to each respondent in a matrix sampling design creates what can be considered missing data, with the missingness being

1. Neal Thomas, Datametrics Research, Inc., 61 Dream Lake Drive, Madison, CT 06443, U.S.A. E-mail: snthomas99@yahoo.com; Trivellore E. Raghunathan, Department of Biostatistics and Institute for Social Research, University of Michigan, Ann Arbor, MI 48106, U.S.A. E-mail: teraghu@umich.edu; Nathaniel Schenker, Office of Research and Methodology, National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Hyattsville, MD 20782, U.S.A. E-mail: nschenker@cdc.gov; Myron J. Katzoff, Office of Research and Methodology, National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Hyattsville, MD 20782, U.S.A. E-mail: mkatzoff@cdc.gov; Clifford L. Johnson, Division of Health and Nutrition Examination Surveys, National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Hyattsville, MD 20782, U.S.A. E-mail: cljohnson@cdc.gov.

at random or even completely at random (Rubin 1976), since the missing data are the result of a known probability mechanism based possibly on design variables. Multiple imputation (Rubin 1987), a general-purpose approach developed for handling data with missing values, is appealing for the analysis of data from a matrix sample, because once the multiple imputations are created, data analysts can apply standard sample survey methods to the analysis of the completed data. Moreover, if the matrix sample has been designed in such a way that the items that are administered to each respondent are predictive of the items that are not administered, then the multiple-imputation approach can utilize the included items to recover information about the excluded items. We focus on multiple imputation because it is well-suited for this situation: 1) the burden of applying complex multivariate methods can be performed once by the survey organization most familiar with the design; 2) it can be implemented with existing software; and 3) it does not require novel methods for each of the numerous estimands targeted in most studies. However, alternative estimation methods to multiple imputation, both model-based and design-based, can be developed and applied to data from matrix designs.

The matrix sampling approach has been applied or explored in various settings, such as educational assessment (Sirotnik and Wellington 1977; Beaton and Zwick 1992; Zeger and Thomas 1997), health research (Wacholder, Carroll, Pee and Gail 1994; Raghunathan and Grizzle 1995; Houseman and Milton 2006), the US Census (Navarro and Griffin 1993), and business (Shoemaker 1973). Moreover, a type of matrix sampling was also used in the National Health Interview Survey of the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention prior to 1997. In that survey, chronic conditions were divided into six lists, and information about the conditions on each list was obtained from about one-sixth of the respondents (Schenker, Gentleman, Rose, Hing and Shimizu 2002). In the context of data collection for a general purpose national health survey, however, an approach to creating matrix sample designs that exploits the inherent associations among the items has not been studied.

This paper develops and evaluates a method for creating matrix sampling forms, each form containing a subset of items to be administered to randomly selected respondents. The method can be applied in complex settings, including situations in which skip patterns are present. Forms are created in such a way that each form includes items that are predictive of the excluded items, so that subsequent analyses based on multiple imputation can recover information about the excluded items that would have been collected had there been no matrix sampling. The method assumes that a training sample is available. The training sample may be from a previous administration of a complete survey or from a pilot sample collected to support survey design. The matrix sampling method is evaluated in a study using data from the National Health and Nutrition Examination Survey (NHANES), one of many nationally representative surveys conducted by NCHS (http://www.cdc.gov/nchs/nhanes.htm). The NHANES, a cross-sectional survey which has been repeated several times during different time periods, obtains a large amount of data from respondents via a household questionnaire, a linked mobile-site medical examination, and laboratory analysis of biological specimens. It is of interest to examine the feasibility of matrix sampling designs for surveys such as the NHANES, which has intricate structural dependencies among its items reflected in its numerous skip patterns, along with its multiple survey components. For purposes of realism, the form-design method is applied using pilot data from the Second NHANES (NHANES II), and then the resulting design, together with multiple-imputation methods, are evaluated in a simulation study based on NHANES III data. Section 2 describes the method for designing matrix sampling forms. In Section 3, the design and results of the study based on the NHANES are described. A concluding discussion is given in Section 4.

2. Designing Matrix Sampling Forms

This section develops a method for creating matrix sampling forms, each form containing a subset of items to be administered to selected respondents.

In designing a matrix sample, it is necessary to decide which items will be core items to be included on all forms, and which items will be split items to be included on only some forms. Typically, the core items are selected based on substantive judgment as well as other considerations about the relative importance of items. Key items, for which precision of certain estimators is to be maximized, should be designated as core items, whereas less important items can be designated as split items. In addition, it is useful to select core items that are predictive of many of the split items, so that information about split items that are excluded from a form can be recovered from the core items in conjunction with the split items that are included in the form. Finally, the cost and respondent burden associated with an item are a consideration, since it can be beneficial to designate expensive and/or burdensome items as split items. The emphasis in this development is on how to allocate the split items to forms once the core items have been chosen, so it is assumed here that the core items have already been selected. However, it will be seen that the method for allocating split items uses a measure that also accounts for the usefulness of

the core items for predicting the split items. The potential to predict split items is estimated from a training sample.

It is also necessary to select a format for organizing the split items. To ensure that every pair of split items appears together on some form, so that direct estimation of all two-way associations between variables is possible, the split items are divided into blocks, and matrix sampling forms are created by putting two or more blocks of split items together (Raghunathan and Grizzle 1995). The size and number of blocks determine the length and number of forms. For example, in the study involving the NHANES to be discussed in Section 3, the split items are divided into four blocks, and each form contains two blocks (along with the core items), so there are a total of six (4 choose 2) forms. In the method developed here, the blocks are of approximately equal size, and each split item is assigned to only one block. Using blocks of the same length yields similar reduced burden for all study participants. It also yields similar precision for items of the same type. These features are not requirements for all matrix sampling designs, however. If additional precision of estimation were desired for an item, it could be included on more than one form, or it could be designated as a core item to be included on all forms.

A good matrix sampling design allocates split items to blocks in such a way that for each split item excluded from a block, there are split items included in the block that, together with the core items, are predictive of the excluded item; this facilitates the recovery of information about the excluded item during analyses of the data. The discussion below develops a method aimed at achieving this goal. The development is in two parts. First, in Section 2.1, an index is formulated for ranking how well each split item is predicted by every other split item, with predictive utility assessed as relative gain in precision conditional on the core items being included. Methods are also given for estimating the values of the index from a training sample. Second, in Section 2.2, an algorithm for assigning split items to blocks based on the index of predictive value is described.

2.1 An Index of Predictive Value

2.1.1 Preliminary Notation for Matrix Sampling Designs

Let Y denote a split item to be predicted, $\boldsymbol{X} = (X_1, ..., X_c)$ denote the core items, and Z denote a split item used to predict Y.

As mentioned above, a matrix sampling design creates what can be considered missing data. Thus, the subjects in a potential matrix sampling design can be ordered so that the n_{obs} subjects with observed values of Y are listed first, the $Y-$values being denoted by $Y_1, ..., Y_{n_{\text{obs}}}$, and the n_{mis} subjects with missing $Y-$values follow, the $Y-$values being denoted by $Y_{n_{\text{obs}}+1}, ..., Y_{n_{\text{tot}}}$, where $n_{\text{tot}} = n_{\text{obs}} + n_{\text{mis}}$ is the total number of observations.

The expectation and variance of Y in the population targeted for matrix sampling are denoted by $\text{E}(Y)$ and $\text{V}(Y)$.

2.1.2 Simplifying Assumptions

Several assumptions are made to simplify the calculation of the index. The assumptions are used when computing the index, but not in subsequent data analyses. Each assumption could be weakened or eliminated if additional research indicates that it results in substantial degradation to the assessment of potential matrix sampling designs.

1. Each split predictor Z is considered separately when added to the core items $\boldsymbol{X}$. If there are several items with high mutual correlation, the assignment algorithm attempts to put the items in different blocks as required for an effective matrix design. A multivariate approach based on partial correlations accounting for other split items would be anticipated to yield similar properties, and would require much more computation.
2. Each split predictor Z is assumed to be fully observed, when in practice, it will not always be available to predict Y because Z is itself a split item. Also, the occurrence of unplanned missing data (*i.e.*, missing data not created by matrix sampling) is not considered. Although these assumptions may overstate the usefulness of Z for improving estimates of $\text{E}(Y)$, such overstatement may be ameliorated via multivariate methods that utilize several variables Z. Moreover, each Z will be administered the same number of times, so any systematic bias in predictive value should be approximately the same for each split item; the primary use of the index is to order items, which is not changed by a common bias.
3. For derivation of the index values, simple random sampling is assumed for both the respondents in the matrix sample and for the training sample. Again, consistent overestimation of precision is not anticipated to substantially diminish the performance of the index.
4. It is assumed that matrix sampling produces missing data that are missing completely at random. This assumption is satisfied for all of the matrix designs considered.
5. For purposes of deriving approximations below, it is assumed that n_{tot} is large and that the ratio $n_{\text{obs}} / n_{\text{tot}}$ is fixed as n_{tot} increases. This approximation should be adequate for most estimands in national surveys.

2.1.3 Estimation Based on Multiple Imputation

The index of predictive value to be developed is based on the goal of estimating $\mathrm{E}(Y)$ via multiple imputation applied to the matrix sample. A multiple-imputation-based estimator, $\bar{\bar{y}}$, of $\mathrm{E}(Y)$ is approximated, under the assumption of an infinite number of imputations, as

$$\bar{\bar{y}} = \lim_{M\to\infty} M^{-1} \sum_{j=1}^{M} \bar{y}_j .$$

In this expression: M is the number of imputations; and $\bar{y}_j$ is the mean from the j^{th} completed data set with imputed values $Y_{i,j}$, $i = n_{\text{obs}} + 1, \ldots, n_{\text{tot}}$, and observed values $Y_{i,j} = Y_i$, $i = 1, \ldots, n_{\text{obs}}$ (which do not change across completed data sets), that is,

$$\bar{y}_j = n_{\text{tot}}^{-1} \sum_{i=1}^{n_{\text{tot}}} Y_{i,j} = n_{\text{tot}}^{-1} \left(\sum_{i=1}^{n_{\text{obs}}} Y_i + \sum_{i=n_{\text{obs}}+1}^{n_{\text{tot}}} Y_{i,j} \right).$$

An estimator of the variance of $\bar{\bar{y}}$ when the imputations are created using $\boldsymbol{X}$ and Z, based on the common variance formula in Rubin (1987, Section 3.1), is

$$\mathrm{V}_{\text{MI}} = \mathrm{V}_{\text{comp}} + \mathrm{V}_{\text{imp}}, \qquad (1)$$

where the first term is an estimate of the variance that would be obtained with complete data, and the second term is an estimate of the variance between imputed data sets. With large samples, so that the variance of Y can be treated as known, $\mathrm{V}_{\text{comp}} = \mathrm{V}(Y)/n_{\text{tot}}$, and

$$\mathrm{V}_{\text{imp}} = \lim_{M\to\infty} M^{-1} \sum_{j=1}^{M} (\bar{y}_j - \bar{\bar{y}})^2 . \qquad (2)$$

It is assumed throughout that the imputation model is compatible with the complete-data model, so the variance estimator in (2) is consistent (Rubin 1987, Section 3.6; Meng 1994).

2.1.4 Defining the Index

When matrix samples are collected, simple but potentially inefficient estimators of univariate summaries of a split item, Y, can be obtained from the observed data without any imputation (that is, using just the observed values of Y), because the subjects the missing Y – values are missing completely at random; the variance of the no-imputation estimator of $\mathrm{E}(Y)$ is denoted by $\mathrm{V}_{\text{NI}} = \mathrm{V}(Y)/n_{\text{obs}}$.

The proposed index is the proportion of the difference between V_{NI} and V_{comp} that is recovered by the multiple-imputation estimator, which incorporates the information contained in $\boldsymbol{X}$ and Z:

$$I(Y \mid \boldsymbol{X}, Z) = \frac{\mathrm{V}_{\text{NI}} - \mathrm{V}_{\text{MI}}}{\mathrm{V}_{\text{NI}} - \mathrm{V}_{\text{comp}}} . \qquad (3)$$

The index $I(Y \mid \boldsymbol{X}, Z)$ takes the value 1 when $\boldsymbol{X}$ and Z perfectly predict the omitted values of Y (so that $\mathrm{V}_{\text{MI}} = \mathrm{V}_{\text{comp}}$), and it takes the value 0 when $\boldsymbol{X}$ and Z do not predict the omitted values of Y at all, so that the multiple-imputation estimator is not an improvement over the no-imputation estimator (*i.e.*, $\mathrm{V}_{\text{MI}} = \mathrm{V}_{\text{NI}}$).

The index can be used to assess the potential contribution of each split item Z to the estimation of the mean of every other split item Y. A desirable matrix sampling design ensures that for each split item Y that is excluded from a block, there are other split items Z included in the block with high index values for predicting Y, so that information about Y can be recovered during analyses of data from the matrix sample.

Note:

1. The variances V_{NI}, V_{comp}, and V_{imp} are proportional to n_{tot}^{-1}, so $I(Y \mid \boldsymbol{X}, Z)$ is independent of n_{tot}.
2. If the core items $\boldsymbol{X}$ are highly predictive of Y, the index will not differentiate much between the remaining split items Z; but in this situation, the selection of appropriate Z for predicting Y is less important, since Y is already predicted well by $\boldsymbol{X}$.

2.1.5 Approximating V_{imp}

To facilitate the computation of the index $I(Y \mid \boldsymbol{X}, Z)$, it is useful to approximate the variance V_{imp}. The approximation developed here refers to a specific matrix sampling design, presuming that one has been chosen.

Assume that the distribution of Y given $(\boldsymbol{X}, Z)$ follows a generalized linear model with a link function μ that depends on unknown parameters $\boldsymbol{\beta}$,

$$\mathrm{E}(Y \mid \boldsymbol{X}, Z) = \mu((\boldsymbol{X}^T, Z)\boldsymbol{\beta}),$$

where the link function is equal to the identity for continuous Y, $\mu(Y) = Y$, and the logistic function for binary Y, $\mu(Y) = \mathrm{logit}^{-1}(Y)$. For continuous Y, a constant residual variance, σ^2, is also assumed. Although they are not developed here, extensions of these models and methods can be developed for categorical and ordered categorical variables. The individual categories can be represented by binary variables, or summaries can be formed when there are numerous categories.

Schafer and Schenker (2000) derived an approximation to the variance between imputed data sets, that is, V_{imp}, when the estimate computed from each completed data set is a smooth function g of the means of the variables involved. (In the current development, g is the identity.) Their approximation, which is based on first-order Taylor series expansions of g and μ and large-sample results

from the theory of sample surveys (*e.g.*, Wolter 1985, Chapter 6), will be used here.

Maximum likelihood (or quasi-likelihood, McCullagh and Nelder 1989) estimation of $\boldsymbol{\beta}$ based on the n_{obs} subjects with observed values of Y yields an estimator, $\hat{\boldsymbol{\beta}}$, with variance-covariance matrix $V_{obs}(\hat{\boldsymbol{\beta}})$ (recall simplifying assumption 4 of Section 2.1.2). Set $\bar{\mu}_{mis}(\hat{\boldsymbol{\beta}}) \equiv n_{mis}^{-1} \sum_{i=n_{obs}+1}^{n_{tot}} \mu((\boldsymbol{X}_i^T, Z_i)\hat{\boldsymbol{\beta}})$, and denote its derivative with respect to the j^{th} component of $\boldsymbol{\beta}$ evaluated at $\hat{\boldsymbol{\beta}}$ by $\bar{\mu}'_{mis,\,j}(\hat{\boldsymbol{\beta}})$, $j = 1, \ldots, (c+1)$. The derivative has the form

$$\bar{\mu}'_{mis,\,j}(\hat{\boldsymbol{\beta}}) = n_{mis}^{-1} \sum_{i=n_{obs}+1}^{n_{tot}} X_{ij}\, f(\hat{\boldsymbol{\beta}}, \boldsymbol{X}_i, Z_i) \qquad j = 1, \ldots, c$$

and

$$\bar{\mu}'_{mis,\,c+1}(\hat{\boldsymbol{\beta}}) = n_{mis}^{-1} \sum_{i=n_{obs}+1}^{n_{tot}} Z_i\, f(\hat{\boldsymbol{\beta}}, \boldsymbol{X}_i, Z_i),$$

where $f(\hat{\boldsymbol{\beta}}, \boldsymbol{X}_i, Z_i) = \mu((\boldsymbol{X}_i^T, Z_i)\hat{\boldsymbol{\beta}})[1 - \mu((\boldsymbol{X}_i^T, Z_i)\hat{\boldsymbol{\beta}})]$ when Y is binary. When Y is continuous, $f(\hat{\boldsymbol{\beta}}, \boldsymbol{X}_i, Z_i) = 1$, which implies that the derivatives $\bar{\mu}'_{mis,\,j}(\hat{\boldsymbol{\beta}})$ are equal to the means of the core items $\boldsymbol{X}$ and the split item Z.

Now let $\bar{\mu}'_{mis}(\hat{\boldsymbol{\beta}})$ denote the vector of derivatives and P_{mis} denote the proportion of subjects with missing Y. Applying equation (10) of Schafer and Schenker (2000), with their function g equal to the identity, and their general parameter $\boldsymbol{\theta}$ equal to $\boldsymbol{\beta}$, yields

$$V_{imp} \approx P_{mis}^2 \left[n_{mis}^{-1} \left\{ n_{mis}^{-1} \sum_{i=n_{obs}+1}^{n_{tot}} \mu((\boldsymbol{X}_i^T, Z_i)\hat{\boldsymbol{\beta}}) \left[1 - \mu((\boldsymbol{X}_i^T, Z_i)\hat{\boldsymbol{\beta}})\right] \right\} + (\bar{\mu}'_{mis}(\hat{\boldsymbol{\beta}}))^T V_{obs}(\hat{\boldsymbol{\beta}})\, \bar{\mu}'_{mis}(\hat{\boldsymbol{\beta}}) \right] \quad (4)$$

when Y is binary, and

$$V_{imp} \approx P_{mis}^2 [\sigma^2 / n_{mis} + (\bar{\mu}'_{mis}(\hat{\boldsymbol{\beta}}))^T V_{obs}(\hat{\boldsymbol{\beta}})\, \bar{\mu}'_{mis}(\hat{\boldsymbol{\beta}})] \quad (5)$$

when Y is continuous.

2.1.6 Estimating the Index of Predictive Value from a Training Sample

Because the planned missing data in our matrix sampling designs are assumed to be missing completely at random, sample moments and other parameter estimates from a training sample can be used to estimate the corresponding moments and parameters in both the subsamples with observed and missing values of Y, under the assumption that the training sample is drawn from the same target population. The moments and parameters include: $V(Y)$; the residual variance σ^2, which can be estimated by $\hat{\sigma}_{tr}^2$, the estimated residual variance from the regression fitted to the training sample; the regression coefficients, with estimates $\hat{\boldsymbol{\beta}}_{tr}$ from the training sample; and the variance-covariance matrix of the regression coefficients, which can be approximated by rescaling the estimate $V_{tr}(\hat{\boldsymbol{\beta}}_{tr})$ from the training sample to obtain $V_{obs}(\hat{\boldsymbol{\beta}}) \approx (n_{tr} / n_{obs}) V_{tr}(\hat{\boldsymbol{\beta}}_{tr})$, where n_{tr} is the size of the training sample. The derivatives $\bar{\mu}'_{mis}(\hat{\boldsymbol{\beta}})$ and the function involving μ in (4) are also in the form of subsample means, and thus can be estimated by the corresponding means in the training sample. Denoting the derivatives in the training sample by $\bar{\mu}'_{tr}(\hat{\boldsymbol{\beta}}_{tr})$, and substituting the training sample estimators into (4) and (5), yields

$$V_{imp} \approx P_{mis}^2 \left[n_{mis}^{-1} \left\{ n_{tr}^{-1} \sum_{i=1}^{n_{tr}} \mu((\boldsymbol{X}_i^T, Z_i)\hat{\boldsymbol{\beta}}_{tr}) \left[1 - \mu((\boldsymbol{X}_i^T, Z_i)\hat{\boldsymbol{\beta}}_{tr})\right] \right\} + \frac{n_{tr}}{n_{obs}} (\bar{\mu}'_{tr}(\hat{\boldsymbol{\beta}}_{tr}))^T V_{tr}(\hat{\boldsymbol{\beta}}_{tr})\, \bar{\mu}'_{tr}(\hat{\boldsymbol{\beta}}_{tr}) \right] \quad (6)$$

for binary variables, and

$$V_{imp} = P_{mis}^2 (\hat{\sigma}_{tr}^2 / n_{mis} + \hat{\sigma}_{tr}^2 / n_{obs}), \quad (7)$$

for continuous variables, with the latter expression following from the fact that $(\bar{\mu}'_{tr}(\hat{\boldsymbol{\beta}}_{tr}))^T V_{tr}(\hat{\boldsymbol{\beta}}_{tr})\, \bar{\mu}'_{tr}(\hat{\boldsymbol{\beta}}_{tr})$ reduces to the simple form $\hat{\sigma}_{tr}^2 / n_{tr}$.

2.2 Assigning Split Items to Blocks

2.2.1 Design Criteria

Matrix sampling forms are created by allocating split items to different blocks, as described at the beginning of Section 2. Four design goals guide the assignment of items: 1) assign each split item to a single block; 2) assign an approximately equal number of items to each block; 3) assign logically linked items to the same block; and 4) assign one or more items to each block that predict the items omitted from the block. Denote the number of blocks by n_{block} ($n_{block} = 4$ in the NHANES simulation study).

A quantitative criterion for the fourth goal is specified separately for each split item Y by finding the $(n_{block} - 1)$ other split items Z with the highest predictive index values $I(Y \mid \boldsymbol{X}, Z)$, for the potential allocation of one of them to the $(n_{block} - 1)$ blocks not containing Y. The items Z exclude those items linked to Y, which must appear with Y in a block. The $(n_{block} - 1)$ values of $I(Y \mid \boldsymbol{X}, Z)$ for the items Z provide an upper limit on the predictive indices that could be achieved for Y. Because these optimal index values are determined separately for each split item Y, they may not be achievable for all items Y simultaneously.

To evaluate a given matrix sampling design, the highest index value $I(Y \mid \boldsymbol{X}, Z)$ actually achieved for each of the $(n_{block} - 1)$ blocks not containing a split item Y is

determined. The average of the $(n_{\text{block}} - 1)$ differences between these indices and the corresponding optimal predictive indices for Y is computed. These average differences are then averaged across all of the split items Y to yield an overall measure for the design.

2.2.2 An Assignment Algorithm

The criteria in Section 2.2.1 require maximization over a set of integer inputs (block assignments) to a function subject to a set of linear constraints imposed by the need to create approximately equal-length blocks with some items potentially linked together. Although integer programming methods could be applied to this maximization, the following algorithm is much simpler, and it achieved nearly optimal results for the NHANES application, as demonstrated in Section 3.1.

Step 1. Randomly order the split items. The assignment of items to blocks proceeds sequentially, via repetition of steps 2 and 3 below, until all of the items have been assigned.

Step 2. Assign the next (or first) unassigned item, say $Y^{(0)}$, to the block with the fewest items. If multiple blocks are tied, assign $Y^{(0)}$ to the block with the lowest maximum predictive index $I(Y^{(0)} \mid \boldsymbol{X}, Z)$ for $Y^{(0)}$. If a tie still remains, assign $Y^{(0)}$ to any of the eligible blocks. If there are items linked to $Y^{(0)}$, also assign them to the selected block.

Step 3. For each item assigned in Step 2 ($Y^{(0)}$ or its linked items), find the remaining unassigned item, say $Y^{(1)}$, most predictive of it. Assign $Y^{(1)}$ (and any items linked to $Y^{(1)}$) to a block other than the block selected in step 2, by following the same procedure that was used for $Y^{(0)}$ in step 2.

Experience with the NHANES data suggests moderate sensitivity of the algorithm to the initial ordering of the items (in step 1). To reduce the dependence, 1,000 designs were generated with randomly selected orderings, and the one yielding the best overall measure of predictive value (as defined at the end of Section 2.2.1) was selected.

3. A Study Using Data from the NHANES

To assess the feasibility of a matrix sampling design for a survey like the NHANES, an evaluation study was conducted. First, NHANES II (*i.e.*, the second NHANES) was used to create a matrix sampling design via the method described in Section 2. This simulates the realistic situation in which data from a previous survey are used in designing the questionnaire for a new survey. The design so developed was then applied to several simulated samples created from NHANES III data. The subjects in NHANES III with complete data on a selected set of variables were treated as a large finite population. One hundred samples were drawn from the NHANES III finite population using a stratified two-stage sample design with unequal probabilities of selection. The complete data are available for each simulated sample, providing a "gold standard." The matrix sampling design was then imposed on each sample, and the missing values due to matrix sampling were multiply imputed. Several analyses were conducted using the matrix samples without imputation, the multiply imputed matrix samples, and the samples of complete data (*i.e.*, the gold standard). Results summarized across the simulated samples yield estimates of the repeated sampling properties of the different methods.

The matrix sampling design created using NHANES II data is summarized in Section 3.1. The design of the simulation study using NHANES III data is described in Section 3.2. Results of the study are presented in Section 3.3. Some limitations of the study that are not discussed in Sections 3.1 – 3.3 are covered in Section 3.4.

3.1 A Matrix Sampling Design Based on Training Data from NHANES II

Given the time that would have been required to extract and analyze all of the NHANES III variables, only a subset were included in the study to keep it manageable, although the software utilized in the study could be applied with many more variables. The variables in the study include items representing many of the topics included in the survey and were selected in consultation with substantive experts. The data types include binary and continuous variables representing survey questions and laboratory measurements. One pair of items forming a skip pattern was included: "Have you smoked 100+ cigarettes?" followed by "Do you smoke now?" The algorithm for assigning split items to blocks, described in Section 2.2, forced these items to be in the same block.

Table 1 gives brief descriptions of the variables included. Variables that appeared in NHANES III but not in NHANES II (again, a realistic situation) have asterisks next to their names.

As mentioned earlier, the matrix sampling design was constructed with four blocks. Each block contained all of the core items. In addition, the split items that appeared in NHANES II were allocated to the blocks by applying the methods developed in Section 2 to data from NHANES II. (In estimating the necessary indices, missing values in the NHANES II data were handled by analyzing only the complete cases.) The split items that did not appear in NHANES II were randomly divided and assigned to the blocks to keep the block lengths approximately equal. The "Type" column of Table 1 identifies the core and split variables and indicates the block assignments for the split variables.

For each split item that appeared in NHANES II, Table 2 displays the following: the block to which the item was assigned ("Block"); the three highest predictive indices for other split items as predictors of the item in question ("Optimal"); and the highest index values actually achieved by the selected design in the three blocks not containing the item in question ("Achieved"). The index values are sorted from low to high for each item in question, so the columns in the table containing index values do not correspond to specific items or blocks. Table 2 shows that the selected design is nearly optimal for the criteria of Section 2.2.1. For example, the average difference between the optimal predictive indices and the corresponding indices actually achieved is only 0.002.

The column of Table 2 labeled "Low" under "Achieved" provides lower bounds on the anticipated improvement in estimators of univariate means for the split items. Nineteen of the twenty-one predictive indices in this column are less than 0.20, suggesting relatively low efficiency for multiple-imputation estimators in this matrix sampling design. For more discussion of this issue, see Sections 3.3 and 4.

Table 1
Variables from NHANES III that were Included in the Evaluation.
Items Marked with Asterisks did not Appear in NHANES II

Variable Name	Description of the Variable	Type
BMPBMI	Body Mass Index	Core
CHP*	Serum Cholesterol (MG/DL)	Core
DMARETHN	Race-Ethnicity	Core
DMPCREGN	Census Region, Weighting(Texas in South)	Core
DMPMETRO	Rural/Urban Code Based on Usda Code	Core
GHP*	Glycated Hemoglobin: (%)	Core
HAB1	Is Health in General Excellent, ..., Poor	Core
HAB2*	Go to Particular Place for Health Care	Core
HAB5*	Past 12 Months, # Times Saw Doctor	Core
HAC1C	Doctor Told: Congestive Heart Failure	Core
HAC1L*	Doctor Ever Told you Had: Lupus	Core
HAC1M	Doctor Ever Told you Had: Gout	Core
HAD1	Ever Been Told you Have Sugar/Diabetes	Core
HAD10	Are you Now Taking Diabetes Pills	Core
HAE3	Told 2+ Times you Had Hypertension/HBP	Core
HAF10	Doctor Ever Told you Had a Heart Attack	Core
HAF26	Severe Dizziness for More Than 5 Minutes	Core
HAL1	Cough Most Days, 3+ Consecutive mo in YR	Core
HAL6	Had Wheezing,Whistle in Chest Past 12 MO	Core
HAL14E	Symptoms Brought on by: Pollen	Core
HAZMNK1R	Average K1 BP from Household and MEC	Core
HAZMNK5R	Average K5 BP from Household and MEC	Core
HFA12	Marital Status	Core
HFA8R	Highest Grade or YR of School Completed	Core
HSAGEIR	Age at Interview (Screener) – Qty	Core
HSSEX	Sex	Core
I1P	Serum Insulin (UU/ML)	Core
G1P	Plasma Glucose (MG/DL)	Split – 1
HAC1J	Doctor Ever Told you Had: Goiter	Split – 1
HAC1N*	Doctor Ever Told you Had: Skin Cancer	Split – 1
HAC1O	Doctor Ever Told you Had: Other Cancer	Split – 1
HAF14*	Get Pain in Either Leg While Walking	Split – 1
HAL11A	Stuffy, Itchy, or Runny Nose, Past 12 MO	Split – 1
BMPWHR*	Waist to Hip Ratio	Split – 2
HAC1E	Doctor Ever Told you Had: Asthma	Split – 2
HAC1K	Doctor Ever Told you Had:Thyroid Disease	Split – 2
HAF24	Numbness etc,1 Side Face/Body for > 5 Min	Split – 2
HAL11B	Watery, Itchy Eyes in Past 12 Months	Split – 2
HAL19A*	In Past 12 Months Had: Cold or Flu	Split – 2
HAL19C*	In Past 12 Months Had: Pneumonia	Split – 2
HAT28	Active Compared with Men/Women your Age	Split – 2
PBP	Lead (UG/DL)	Split – 2
SPPFVC*	FVC, Largest Value (ML)	Split – 2

Table 1 (Continued)
Variables from NHANES III that were Included in the Evaluation.
Items Marked with Asterisks did not Appear in NHANES II

Variable Name	Description of the Variable	Type
FEP	Serum Iron (UG/DL)	Split – 3
HAF1	Ever Had Any Pain or Discomfort in Chest	Split – 3
HAF23	Weak/Paralysis on Face, Arm, Leg For > 5 Min	Split – 3
HAL19B	In Past 12 MO Had: Sinusitis/Sinus Prob	Split – 3
HAR1	Have you Smoked 100+ Cigarettes In Life	Split – 3
HAR3	Do you Smoke Cigarettes Now	Split – 3
SPPPEAK*	Peak Expiratory Flow	Split – 3
BDPTOBMD*	Bone Mineral Density Total Region-GM/CM SQ	Split – 4
HAB4	Past 12 MOS, # Times Stayed in Hospital	Split – 4
HAC1D	Doctor Ever Told you Had: Stroke	Split – 4
HAC1F	Doctor Ever Told Had: Chronic Bronchitis	Split – 4
HAC1H	Doctor Ever Told you Had: Hay Fever	Split – 4
HAC1I	Doctor Ever Told you Had: Cataracts	Split – 4
HAE6*	Ever Had Blood Cholesterol Checked	Split – 4
HAM11*	Consider Self Over/Under/Right Weight	Split – 4
HAE7*	Doctor Told Blood Cholesterol Level High	Split – 4

Table 2
Indices of Predictive Value Based on NHANES II Data for the Split Items in the Matrix Sampling Design

		Optimal			Achieved		
Item	Block	Low	Medium	High	Low	Medium	High
HAC1J(GOITER)	1	0.04	0.04	0.15	0.04	0.04	0.15
HAC1O(OTHER CANCER)	1	0.05	0.06	0.13	0.05	0.06	0.13
HAL11A(NASAL SYMPTOMS)	1	0.17	0.27	0.29	0.17	0.27	0.29
G1P(PLASMA GLUCOSE)	1	0.26	0.30	0.43	0.26	0.30	0.43
HAC1E(ASTHMA)	2	0.09	0.10	0.13	0.08	0.09	0.13
HAC1K(THYROID DISEASE)	2	0.07	0.07	0.15	0.07	0.07	0.15
HAF24(NUMBNESS)	2	0.12	0.12	0.12	0.11	0.12	0.12
HAL11B(WATERY EYES)	2	0.14	0.15	0.25	0.14	0.15	0.25
HAT28(ACTIVE FOR AGE)	2	0.12	0.13	0.16	0.11	0.13	0.16
PBP(LEAD (UG/DL))	2	0.19	0.20	0.21	0.18	0.20	0.21
HAF1(PAIN IN CHEST)	3	0.25	0.29	0.29	0.23	0.25	0.29
HAF23(WEAK/PARALYSIS)	3	0.08	0.12	0.12	0.08	0.12	0.12
HAL19B(SINUSITIS/SINUS)	3	0.07	0.12	0.21	0.07	0.12	0.21
HAR1(100+ CIGARETTES)	3	0.13	0.14	0.14	0.13	0.14	0.14
HAR3(SMOKE NOW)	3	0.10	0.11	0.12	0.10	0.11	0.12
FEP(SERUM IRON)	3	0.05	0.05	0.08	0.05	0.05	0.08
HAB4(# HOSP STAYS)	4	0.07	0.11	0.19	0.07	0.11	0.19
HAC1D(STROKE)	4	0.19	0.20	0.24	0.18	0.20	0.24
HAC1F(BRONCHITIS)	4	0.10	0.12	0.12	0.10	0.10	0.12
HAC1H(HAY FEVER)	4	0.07	0.07	0.09	0.04	0.07	0.09
HAC1I(CATARACTS)	4	0.08	0.09	0.12	0.08	0.09	0.12

Note: Optimal predictive indices are determined for each item separately and may not be achievable for all items simultaneously.

3.2 Design of the Simulation Study Based on NHANES III Data

3.2.1 Population and Sample Design

The matrix sampling design and multiple-imputation analysis could be applied to the entire NHANES III sample. Although this would be informative, a study based on a single data set would not allow the assessment of repeated-sampling statistical properties of the methods studied. Therefore, the 11,759 subjects from the NHANES III survey who had complete data on the variables listed in Table 1 were treated as a finite population, and repeated samples were drawn from this population. In selecting samples, a complex sample design was used instead of simple random sampling to create a more realistic simulation study. To achieve this objective, three design variables were added to the finite population: (1) simulation stratum; (2) simulation cluster; and (3) simulation sample weight (here, the modifier "simulation" is used to distinguish these quantities from the original NHANES III design variables).

1. **Simulation strata**: The NHANES III public-use sample has 49 strata with two clusters per stratum. The strategy for the simulation study was to create a smaller number of strata with a larger number of clusters within the strata, to ensure sufficient sample-to-sample variation between the simulated samples. The 49 original strata were collapsed into 20 simulation strata as follows. Each of the 49 original strata were classified into one of eight categories formed from the cross-classification of census region (4 levels) and rural/urban status based on the United States Department of Agriculture code (2 levels). Within each of these eight categories, a cluster analysis was performed using the stratum-level proportions of non-Whites to select the original strata to combine. Combining the original strata created two or three simulation strata within each of the eight categories, yielding a total of 20 simulation strata. This method of creating larger strata also increased the racial heterogeneity between the resulting simulation strata, which increases the importance of weighting in the analyses.

2. **Simulation clusters**: The NHANES III public-use sample has 98 clusters, with two clusters in each of the original 49 strata. After the 49 original strata were collapsed into 20 simulation strata, the original clusters were subdivided based on another cluster analysis using systolic and diastolic blood pressure readings and body mass index (BMI). Subjects with similar values were grouped together to create a setting with intraclass correlation for these three variables within each simulation cluster. The number of simulation clusters per simulation stratum ranged from 3 to 25, and the number of subjects per simulation cluster ranged from 30 to 98.

3. **Simulation sampling weights**: The simulation sampling weights were determined by the following two-stage sample design. First, from each simulation stratum, two simulation clusters were drawn via simple random sampling without replacement. Because there were unequal numbers of simulation clusters across the 20 simulation strata, the simulation sampling weight corresponding to this stage was $w_{1h} = A_h / 2, h = 1, 2, ..., 20$, where A_h is the number of simulation clusters in simulation stratum h.

   Second, from each selected simulation cluster, 30 subjects were drawn at random without replacement with varying probabilities of selection. If the cluster size was 30, then all subjects were included in the sample. For clusters with more than 30 subjects, the first-draw selection probabilities were computed by normalizing the reciprocals of the original weights from the NHANES III public-use sample to sum to 1 within each simulation cluster, with the normalized reciprocal for each subject used as the selection probability for that subject. The first-draw selection probabilities within simulation clusters ranged from 0.0003 to 0.2756.

   Let i index sampled subjects within a simulation cluster, c denote sampled clusters within a simulation stratum, and h denote simulation strata as above, $i = 1, 2, ..., 30$, $c = 1, 2$, $h = 1, 2, ..., 20$. If the size of cluster c in stratum h was 30, then the second-stage simulation weight for subject i in cluster c was $w_{2ich} = 1$. If the size of cluster c in stratum h was greater than 30, then the second-stage simulation weight for subject i in cluster c was $w_{2ich} \propto \pi_{ich}^{-1}$, where π_{ich} denotes the first-draw selection probability for subject i. The final simulation sampling weight for each sampled subject was $w_{ich} = w_{1h} \times w_{2ich}$, $i = 1, 2, ..., 30, c = 1, 2$, $h = 1, 2, ..., 20$.

The design effects for estimating population means averaged approximately 2.1 in this simulation study. The complex sample design features in the study are informative in the sense that ignoring the design features in analyses of data may result in biased estimates and underestimation of sampling variances. This is due in particular to the use of data on race, blood pressure, and BMI in the simulation sampling design, and the well-documented connection between race/ethnicity and blood pressure or BMI.

3.2.2 Simulating Matrix Samples

One hundred independent probability samples were drawn from the finite population. Each simulated sample included 1,200 subjects (20 simulation strata, 2 simulation clusters per simulation stratum, 30 subjects per simulation cluster).

Matrix sampling was overlayed on each simulated sample by assigning each of the 1,200 subjects randomly to one of the six forms containing the core items and one of the block pairs (1, 2), (1, 3), (1, 4), (2, 3), (2, 4), or (3, 4). The random assignment was carried out such that 200 subjects were assigned to each form. Thus, for each matrix sample, the core items were available for all 1,200 sampled subjects, whereas each split item was available for 600 sampled subjects.

3.2.3 Estimation Methods Compared

Point estimates from each sample in the simulation study were obtained using three methods: analyzing the complete

data for a gold standard; analyzing the matrix sampled data with no imputation; and applying multiple imputation to fill in the missing values caused by matrix sampling, followed by multiple-imputation analyses. For the complete-data and no-imputation analyses, the point estimates were weighted. For the multiple-imputation analyses, the same weights were used in calculating the point estimate from each of the multiple completed data sets, and then the usual averaging of the multiple point estimates was carried out (Rubin and Schenker 1986; Rubin 1987, Section 3.1).

Multiple imputation of the missing split items was carried out using the sequential regression approach (Kennickell 1991; Oudshoorn, Van Buuren and Van Rijckevorsel 1999; Raghunathan, Lepkowski, Van Hoewyk and Solenberger 2001), as implemented by the software package IVEware (http://www.isr.umich.edu/src/smp/ive). Five sets of imputations were created by independently applying the sequential regression approach five times, with ten iterations of the sequential regression algorithm for each set of imputations. The number of imputations is based on theory and experience showing that five imputations is usually adequate, especially if the fraction of missing information is not large (Rubin 1996). With missing-data rates for the split items of 50%, the fraction of missing information, which is roughly $1 - V_{comp} / V_{imp}$, is expected to be at most 50%, as is borne out in the simulation results. Rubin (1987, Table 4.1) gave the large-sample relative efficiency of five imputations relative to an infinite number of imputations as 90% when there is 50% missing information. A larger number of imputations would increase precision for estimating the between-imputation variance (V_{imp}) and the fraction of missing information.

To account for the complex simulation sample design, main effects were included in the imputation model for simulation stratum and simulation cluster nested within simulation stratum. The logarithm of the simulation sampling weight was also included as a predictor in the imputation model, along with the core and split items.

3.3 Results of the Simulation Study

To evaluate estimates based on the matrix sampling design, two types of analysis problems were considered: estimating the population means of the split items; and regression analyses involving the split and core items. Properties of the no-imputation, multiple-imputation, and complete-data estimators across the 100 simulated data sets were compared with each other to assess bias and loss of efficiency due to matrix sampling combined with multiple imputation.

3.3.1 Estimating Population Means of Split Items

For the population mean of a split item, the simulated standardized bias of the no-imputation estimator was defined as $(Ave_{NI} - Ave_{comp}) / SD_{NI}$, where Ave_{NI}, Ave_{comp}, and SD_{NI} denote, respectively, the averages of the no-imputation and complete-data estimates and the standard deviation of no-imputation estimates across the 100 simulated data sets. An analogous simulated standardized bias was defined for the multiple-imputation (MI) estimator. Table 3 summarizes the simulated standardized biases for the 32 split items.

Table 3
Simulated Standardized Biases of the No-Imputation and Multiple-Imputation Estimators of the Population Means for the 32 Split Items

	Frequency	
Standardized Biases	No Imputation	Multiple Imputation
-1.4		1
$(-1, -0.6]$		4
$(-0.6, -0.4]$		5
$(-0.4, -0.2]$		4
$(-0.2, 0)$	15	10
$(0, 0.2)$	17	4
$[0.2, 0.4)$		
$[0.4, 0.6)$		2
$[0.6, 1)$		
1.4		1
4.6		1
Total	32	32

Because our matrix sampling mechanism results in missing data that are missing completely at random, the no-imputation estimators are close to unbiased. This is reflected in the simulation results by the fact that none of the absolute standardized biases is larger than 0.2. The multiple-imputation estimators generally have somewhat higher simulated standardized biases than do the no-imputation estimators, although the absolute standardized biases are less than one for all but three split items and less than 0.6 for all but seven. As a guideline for judging standardized biases, Cochran (1977, page 14) shows that a standardized bias of 0.6 produces nominal 95% confidence intervals having roughly 91% actual coverage. Any substantial biases observed in this study when matrix sampling is used in conjunction with multiple imputation are likely due to deficiencies in the imputation models and not to the matrix sampling itself, given that the no-imputation analyses were seen to be approximately unbiased. With larger sample sizes in an application to an actual survey, the corresponding standardized biases would tend to be moved upward because of the smaller denominators; but the standardized biases might also be moved downward because of improved large-sample approximations.

Loss of efficiency due to matrix sampling rather than using the full questionnaire can be assessed by comparing the sampling error of the no-imputation, multiple-imputation, and complete-data estimators (computed as standard deviations across the 100 simulated data sets). Table 4 summarizes the ratios of the simulated standard deviations

of the multiple-imputation estimators to those of the no-imputation estimators, and the ratios of the simulated standard deviations of the complete-data estimators to those of the multiple-imputation estimators (the term "simulated standard deviation" of an estimator is used rather than "simulated standard error" to avoid confusion with the estimated standard error that could be obtained from the analysis of each simulated data set).

Table 4
Ratios of the Simulated Standard Deviations of the No-Imputation (NI), Multiple-Imputation (MI), and Complete-Data (comp) Estimators of the Population Means for the 32 Split Items

	Frequency	
Ratios	SD_{MI}/SD_{NI}	SD_{comp}/SD_{MI}
(0.5, 0.6]		2
(0.6, 0.7]		9
(0.7, 0.8]		14
(0.8, 0.9]		6
(0.9, 0.95]	7	
(0.95, 1]	18	1
(1, 1.03]	7	
Total	32	32

Typically, the multiple-imputation estimators are more efficient than the no-imputation estimators, but the gain in efficiency is only modest, as indicated by the fact that most of the ratios SD_{MI}/SD_{NI} in Table 4 are between 0.9 and 1. Such modest gains in efficiency can be predicted roughly from the indices of predictive value based on data from NHANES II (displayed in Table 2), as follows. Because each split item is included in only half of the matrix sampling forms, it follows that the variance of a complete-data estimator of the mean of a split item should be about one-half the size of the variance of the corresponding no-imputation estimator. Dividing the numerator and denominator of expression (3) by V_{NI}, and setting $V_{comp}/V_{NI} = 0.5$, yields $2(1 - V_{MI}/V_{NI})$ as an approximate expression for the index of predictive value in this simulation study. For an index of 0.12, which is the median of the "Achieved Medium" indices in Table 2, it follows that V_{MI}/V_{NI} should be about 0.94. This ratio of variances is equivalent to a ratio of standard deviations of about $\sqrt{0.94} = 0.97$, which is near the middle of the range of ratios summarized in Table 4. In this study, because the multiple-imputation estimators are only modestly more efficient than the no-imputation estimators, and because the multiple-imputation estimators have some biases associated with them, the mean square errors for the multiple-imputation estimators are higher than those for the no-imputation estimators in 22 out of 32 cases.

The simulation results on the efficiency of the multiple-imputation estimators relative to the complete-data estimators also conform with theory. Since V_{comp}/V_{NI} should be about 0.5, and since V_{MI} should be slightly smaller than V_{NI}, it follows that V_{comp}/V_{MI} should be slightly larger than 0.5, or equivalently, that the typical ratio of standard deviations SD_{comp}/SD_{MI} should be slightly larger than $\sqrt{0.5} = 0.71$. Indeed, the median of the ratios summarized in Table 4 is 0.75. An alternative to the multiple imputation estimation is two-phase weighting based on core item estimators and their differences between blocks. Any advantage in efficiency from multiple-imputation estimation would be due to the additional information from the split items.

3.3.2 Estimating Regression Coefficients

The matrix sampling and multiple-imputation methods were also evaluated for estimation of the coefficients of eight regression models, which were specified to be similar to models that have appeared in the literature. The regression models, which are listed in Table 5, had a total of 115 coefficients. No-imputation estimators for the regression coefficients were not included in the simulation study, although some theoretical results on their efficiency are discussed in this section.

For each regression coefficient, the simulated standardized bias was defined analogously to the definition used for each mean in Section 3.3.1. Table 6 summarizes the standardized biases for the 115 regression coefficients. Most of the standardized biases are small, with absolute values greater than one for only five coefficients and absolute values of 0.6 or greater for only seven.

Table 7 summarizes the ratios of the standard deviations of the complete-data estimates across the 100 simulated data sets to those of the multiple-imputation estimates, for the 115 regression coefficients. Separate summaries are displayed by whether the regression models involve split variables from only one block (Models 1, 2, 6, and 7) versus two blocks (Models 3, 4, 5, and 8). A larger proportion of the ratios are close to one than was the case for estimating means (Table 4). In addition, for several regression coefficients (particularly from Models 3, 6, 7, and 8), the simulated standard deviations of the complete-data estimators are moderately larger than those of the multiple-imputation estimators, and for one coefficient, the ratio is about two. Finally, there are four regression coefficients for which there appears to be a substantial loss of efficiency due to matrix sampling, with ratios less than 0.3 (one each from Models 1, 2, 5, and 8). The ratios close to or larger than one could be due in part to a lack of fit of some regression models to the complete data and a better fit of the models to the data completed by imputation, with the latter resulting from an imputation process that is based on regression models. Moreover, the two smallest ratios occur for regression models involving split variables from two blocks, for which the fraction of subjects in the matrix sample with no missing data is only one-sixth, as discussed further below.

Table 5
Regression Models Used in the Evaluation

Type of regression model	Dependent variable	Variables recoded to create predictors including interaction terms. Each model also includes an intercept term. For each variable, the number in the parentheses indicates the number of regression coefficients associated with the variable	Split variables in the regression models. For each variable, the number in the parentheses indicates the block containing the variable
1. Linear	G1P	HSSEX(1) ,HSAGEIR(1), DMARETHN(3), and GHP(1)	G1P(1)
2. Logistic	HAF10	HSSEX(1), HSAGEIR(1), DMARETHN(3), FEP(1), and BMPBMI(1)	FEP(3)
3. Logistic	HAF10	HSSEX(1), HSAGEIR(1), DMARETHN(3), HAD1(1), HAE3(1), PBP(1), FEP(1), CHP(1), and G1P(1)	FEP(3) and G1P(1)
4 and 5. Linear	SPPFVC	HSAGEIR(1), DMARETHN (3), HFA8R(2), and BMPBMI(1) [By gender (HSSEX), and restricted to never smokers (HAR1, HAR3)]	SPPFVC(1), HAR1(3), and HAR3(3)
6 and 7. Logistic	HCHP (1 IF CHP>=240 AND 0 OTHERWISE)	HSAGEIR(2), DMARETHN(3), HFA8R(1), BMPBMI(3), (HAR3,HAR1)(2), BMPBMI*HSAGEIR(6), and DMARETHN*BMPBMI(9) [By Gender (HSSEX)]	(HAR3, HAR1)(3)
8. Logistic	HAC1E	HSAGEIR(5), HSSEX(1), DMARETHN(3), BMPBMI(4), (HAR3, HAR1)(2), SPPPEAK(1), and SPPFVC(1)	HACIE(2), (HAR1,HAR3)(3), SPPPEAK(3), and SPPFVC(2)

Table 6
Simulated Standardized Biases of the Multiple-Imputation Estimators for the 115 Regression Coefficients

Range of Standardized Biases	Frequency
– 5.2	1
– 1.5	1
– 1.3	1
– 1.1	1
(– 1, -0.6]	2
(– 0.6, – 0.4]	2
(– 0.4, – 0.2]	3
(– 0.2, 0)	52
(0, 0.2)	44
[0.2, 0.4)	6
[0.4, 0.6)	1
[0.6, 1)	
3.7	1
Total	115

Table 7
Ratios of the Simulated Standard Deviations of the Complete-Data Estimators to those of the Corresponding Multiple-Imputation Estimators, for the 115 Regression Coefficients, by Whether the Regression Models Involve Split Variables from Only One Block Versus Two Blocks

	Frequency	
Range of Ratios	One Block	Two Blocks
(0, 0.1]		1
(0.1, 0.2]		1
(0.2, 0.3]	2	
(0.3, 0.4]		
(0.4, 0.5]	1	
(0.5, 0.6]		3
(0.6, 0.7]	2	3
(0.7, 0.8]	2	7
(0.8, 0.9]	4	4
(0.9, 0.95]	2	3
(0.95, 1]	29	8
(1, 1.05]	20	5
(1.05, 1.1]	4	2
(1.1, 1.2]	3	2
(1.2, 1.4]		4
(1.4, 1.6]		2
2.0		1
Total	69	46

For regression models involving split variables from only one block, the theoretical efficiency of the complete-data estimator relative to the no-imputation estimator, that is, the ratio of the variance of the latter to the former, is approximately two because only half of the subjects in the matrix sample will have complete data on those variables; and for regression models involving split variables from two blocks, the theoretical relative efficiency is approximately six. In contrast, the respective simulated relative efficiencies of the complete-data estimator relative to the multiple-imputation estimator, that is, the inverses of the squared ratios summarized in Table 7, are less than two for 64 out of 69 coefficients when only one block is involved; and they are less than six for 44 out of 46 coefficients when two blocks are involved. Thus, the multiple-imputation estimators are generally more efficient than the no-imputation estimators for regression problems. Nevertheless, the large losses of efficiency of the multiple-imputation estimators relative to the complete-data estimators for some coefficients as well as the apparent gains in efficiency for other coefficients are worth further investigation.

3.4 Additional Limitations of the Simulation Study

This section briefly discusses some additional limitations of the simulation study and adjustments required during the implementation of the study.

Originally, two questions about two conditions, gout and lupus (HAC1M and HAC1L) were designated as split items. Due to low prevalence of these two conditions in the constructed finite population, many of the simulated samples had no subjects with these conditions. After a few preliminary runs, the designations for these two items were changed from split to core. In general, in situations with limited sample sizes, conditions with very low prevalence rates may need to be designated as core items. In addition, due to issues such as some split items appearing in NHANES III but not in NHANES II, as well as logical linkages between some split items, the number of split items per block in the simulation study varied slightly more than intended (from 6 to 10).

In the regression models listed in Table 3, the number of predictors ranged from 8 to 27, because some of the regression models included interaction terms as predictors. Even with the sample size of 1,200 for the simulated samples, some of the complete-data estimators were unstable. This was due in part to small sample sizes for some combinations of variables that affected the estimation of interactions. Note that in many applications of matrix sampling to large surveys, complete-data sample sizes would be substantially larger than the size of 1,200 used in our simulation study.

The Monte Carlo standard errors of the simulated averages in this study are approximately one-tenth of the standard deviations of the individual quantities across the 100 samples. However, the standard deviations across the

samples varied widely from one estimand to another, due to differences in scaling. For example, the simulated standard deviations of the complete-data estimators of the 115 regression coefficients ranged from 9.8×10^{-5} to 1169.6. More precise estimates of bias and efficiency could be computed based on a larger number of simulated samples than was used in this study.

4. Discussion

In this paper, a method was developed for creating matrix sampling designs that have the property that the items included on forms are predictive of the items that have been excluded. The feasibility of implementing such designs in a complex, large-scale health survey was demonstrated via an example involving the National Health and Nutrition Examination Survey. Matrix sampling designs, in conjunction with multiple imputation, can be used to expand the scope of a survey without increasing respondent burden or unduly increasing the burden to subsequent data analysts.

In the study involving NHANES data, the multiple-imputation analyses of data from the matrix samples were modestly effective, with minor evidence of bias and with greater efficiency than simply analyzing the matrix sampled data without imputation. The increased efficiency was especially evident in the context of regression analyses.

Matrix sampling in the NHANES example typically resulted in large losses of precision compared to what could have been achieved with a longer, complete survey (*i.e.*, no matrix sampling), however. This finding, which is in contrast with the more promising results obtained in other applications of matrix sampling, highlights the importance of including good predictors of the split items in a survey. For example, an application of matrix sampling to an educational survey (*e.g.*, Beaton and Zwick 1992) has been much more successful, because the split items are highly correlated responses to questions designed to measure the same trait. Raghunathan and Grizzle (1995) also demonstrated much greater recovery of information on omitted items in the context of a health survey.

The items in the NHANES example were chosen mainly to represent a variety of important health characteristics, without much consideration given to their ability to predict or be predicted by other variables. Many of the split items represented rare illnesses that are not well predicted by common medical conditions and standard laboratory measurements; in hindsight, these variables were not good candidates for split items. Variables representing rare events may also cause difficulties with many common statistical methods that rely on large-sample approximations, making them less amenable to model-based imputation.

Better candidates for matrix sampling designs are "panels" of inter-related items. For example, matrix sampling techniques can be useful when there are multiple measurements of the same (or closely related) quantities, and it is desired to collect some of the measurements for subsets of the survey respondents due to cost and time considerations. Some rudimentary forms of matrix sampling are already being applied in such settings, and there may be substantial improvements possible by applying methods, such as those developed in this paper, that aim to exploit the associations among the variables.

Acknowledgments

The work of Neal Thomas and Trivellore E. Raghunathan was supported in part by a professional services contract between NCHS and Datametrics Research, Inc. The authors thank Randy Curtin of NCHS for helpful suggestions. The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.

References

Beaton, A., and Zwick, R. (1992). Overview of the National Assessment of Educational Progress. *Journal of Educational Statistics*, 17, 95-109.

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

Houseman, E., and Milton, D. (2006). Partial questionnaire designs, questionnaire nonresponse, and attributable fraction: Applications to adult onset asthma. *Statistics in Medicine*, 25, 1499-1519.

Kennickell, A.B. (1991). Imputation of the 1989 Survey of Consumer Finances: stochastic relaxation and multiple imputation. *Proceedings the Survey Research Methods Section*, American Statistical Association, 112-121.

McCullagh, P., and Nelder, J. (1989). *Generalized Linear Models*. Second Edition, London: Chapman Hall.

Meng, X.-L. (1994). Multiple imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, 9, 538-573.

Navarro, A., and Griffin, R. (1993). Matrix sampling designs for the year 2000 Census. *Proceedings the Survey Research Methods Section*, American Statistical Association, 480-485.

Oudshoorn, K., Van Buuren, S. and Van Rijckevorsel, J. (1999). Flexible multiple imputation by chained equations of the AVO-95 Survey. Leiden: TNO Prevention and Health, Report PG/VGZ/99.045.

Raghunathan, T.E., and Grizzle, J.E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, 90, 54-63.

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, 85-95.

Rubin, D.B. (1976). Inference and missing data (with discussion). *Biometrika*, 63, 581-592.

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.

Rubin, D.B., and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.

Schafer, J.L., and Schenker, N. (2000). Inference with imputed conditional means. *Journal of the American Statistical Association*, 95, 144-154.

Schenker, N., Gentleman, J.F., Rose, D., Hing, E. and Shimizu, I.M. (2002). Combining estimates from complementary surveys: A case study using prevalence estimates from national health surveys of households and nursing homes. *Public Health Reports*, 117, 393-407.

Shoemaker, D.M. (1973). *Principles and Procedures of Matrix Sampling*. Cambridge, MA: Ballinger.

Sirotnik, K., and Wellington, R. (1977). Incidence sampling: An integrated theory for matrix sampling. *Journal of Educational Measurement*, 14, 343-399.

Wacholder, S., Carroll, R.J., Pee, D. and Gail, M.H. (1994). The partial questionnaire design for case-control studies (with discussion). *Statistics in Medicine*, 13, 623-649.

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

Zeger, L.M., and Thomas, N. (1997). Efficient matrix sampling for correlated latent traits: Examples from the National Assessment of Educational Progress. *Journal of the American Statistical Association*, 92, 416-425.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2006.

J.-F. Beaumont, *Statistics Canada*
Y. Berger, *The University of Reading, UK*
C. Boudreau, *Medical College of Wisconsin*
L. Burck, *Central Bureau of Statistics, Israël*
F. Butar, *Sam Houston University*
J. Chipperfield, *Australian Bureau of Statistics*
S.R. Chowdhury, *Westat Inc.*
G. Datta, *University of Georgia*
P. Duchesne, *Université de Montréal*
K. Duncan, *Dominican University, Chicago*
F. Dupont, *INSEE*
G.B. Durrant, *Southampton Statistical Sciences Research Institute, University of Southampton, UK*
M. Elliott, *University of Michigan*
J. Eltinge, *United States Bureau of Labor Statistics*
M. Feder, *Research Triangle Institute*
R. Folsom, *Research Triangle Institute*
O. Frank, *Stockholm University*
J. Gambino, *Statistics Canada*
C. Girard, *Statistics Canada*
M. Gosh, *University of Florida*
B. Graubard, *National Cancer Institute*
G. Griffiths, *Australian Bureau of Statistics*
D. Haziza, *Statistics Canada*
J. Horgan, *Dublin City University*
V.G. Iannacchione, *RTI International*
J. Jiang, *University of California at Davis*
J.-K. Kim, *Department of Applied Statistics, Korea*
P. Kokic, *Australian Bureau of Agriculture and Resource Economics*
P. Kott, *United States Department of Agriculture*
M. Kovačevič, *Statistics Canada*
F. Kreuter, *Joint Program in Survey Methodology*
M.D. Larsen, *Iowa State University*
P. Lavallée, *Statistics Canada*
H. Lee, *Westat, Inc.*
R. Lehtonen, *University of Helsinki*
C. Leon, *Statistics Canada*
W.W. Lu, *Department of Mathematics and Statistics*
A. Matei, *Université de Neuchâtel, Suisse*
D. Melec, *United States Bureau of the Census*
J.M. Montaquila, *Westat, Inc.*
R. Munnich, *University of Tubingen*
J. Opsomer, *Iowa State University*
Z. Patak, *Statistics Canada*
D. Pfeffermann, *Israël and University of Southampton*
N. Prasad, *University of Alberta*
L. Qualité, *Université de Neuchâtel, Suisse*
M.G. Ranalli, *Universita' degli Studi di Perugia*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
O. Sautory, *Insee-Cepe*
J. Schafer, *Pennsylvania State University*
A. Scott, *University of Auckland*
R. Singh, *U.S. Census Bureau*
C. Skinner, *University of Southampton*
E. Stuart, *Mathematica Policy Research Inc*
C.J. Swartz, *Simon Fraser University*
R. Valliant, *University of Michigan*
Z. Wang, *Wilfrid Laurier University, Waterloo*
M. Winglee, *Westat*
C. Wu, *University of Waterloo*
W. Yung, *Statistics Canada*
E. Zanutto, *Department of Statistics, The Wharton School, University of Pennsylvania*

Acknowledgements are also due to those who assisted during the production of the 2006 issues: Cécile Bourque, Louise Demers, Anne-Marie Fleury, Roberto Guido, Liliane Lanoie, Micheal Pelchat and Isabelle Poliquin (Dissemination Division), Nadine Lacroix (Client Services Division), Sheri Buck (Systems Development Division), François Beaudin (Official Languages and Translation Division) and Sophie Chartier (Business Survey Methods Division). Finally we wish to acknowledge Christine Cousineau, Céline Ethier and Denis Lemire of Household Survey Methods Division, for their support with coordination, typing and copy editing.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 22, No. 1, 2006

Volume 22, No. 2, 2006

All inquires about submissions and subscriptions should be directed to jos@scb.se

CONTENTS TABLE DES MATIÈRES

Volume 34, No. 2, June/juin 2006

CONTENTS TABLE DES MATIÈRES

Volume 34, No. 3, September/septembre 2006

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. A paper copy may be required for formulas and figures.

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.

1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.

1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.

1.4 Acknowledgements should appear at the end of the text.

1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.

3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", *etc*.

3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.

3.4 Write fractions in the text using a solidus.

3.5 Distinguish between ambiguous characters, (*e.g.*, w, ω ; o, O, 0; l, 1).

3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.

4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, *e.g.*, Cochran (1977, page 164).

5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

6. Short Notes

6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.