03

SURVEY METHODOLOGY

STATISTICS CANADA
JAN 27 2004

Catalogue No. 12-001-XPB

A JOURNAL PUBLISHED BY STATISTICS CANADA

DECEMBER 2003

•

VOLUME 29

•

NUMBER 2

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A JOURNAL PUBLISHED BY STATISTICS CANADA

DECEMBER 2003 • VOLUME 29 • NUMBER 2

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry, 2004

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without prior written permission from Licence Services, Marketing Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

January 2004

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa

Statistics Canada Statistique Canada

Canada

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.

MANAGEMENT BOARD

Chairman G.J. Brackstone

Members
D.A. Binder
G.J.C. Hole
C. Patrick
R. Platek (Past Chairman)
E. Rancourt (Production Manager)
D. Roy
M.P. Singh

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

D.R. Bellhouse, *University of Western Ontario*
D.A. Binder, *Statistics Canada*
J.M. Brick, *Westat, Inc.*
C. Clark, *U.S. Bureau of the Census*
J. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
M.A. Hidiroglou, *Statistics Canada*
G. Kalton, *Westat, Inc.*
P. Kott, *National Agricultural Statistics Service*
P. Lahiri, *JPSM, University of Maryland*
S. Linacre, *Australian Bureau of Statistics*
G. Nathan, *Hebrew University, Israel*
D. Norris, *Statistics Canada*
D. Pfeffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*
T.J. Rao, *Indian Statistical Institute*
L.-P. Rivest, *Université Laval*
N. Schenker, *National Center for Health Statistics*
F.J. Scheuren, *National Opinion Research Center*
R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
E. Stasny, *Ohio State University*
R. Valliant, *JPSM, University of Michigan*
J. Waksberg, *Westat, Inc.*
K.M. Wolter, *Iowa State University*
A. Zaslavsky, *Harvard University*

Assistant Editors J.-F. Beaumont, P. Dick, H. Mantel and W. Yung, *Statistics Canada*

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 manuscripts prepared following the guidelines given in the Journal in either English or French to the Editor, Dr. M.P. Singh, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. E-mail: singhmp@statcan.ca. Four nonreturnable printed copies of each manuscript can also be sent.

Subscription Rates

The price 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 x 2 issues); Other Countries, CDN $30 ($15 x 2 issues). Subscription order should be sent to Statistics Canada, Dissemination Division, Circulation Management, 120 Parkdale Avenue, 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.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 29, Number 2, December 2003

CONTENTS

Survey Methodology, December 2003
Vol. 29, No. 2, pp. 105-106
Statistics Canada

In This Issue

This issue of *Survey Methodology* includes a special section on Census Coverage Error which presents four papers, including two papers on the coverage survey used in the United States, one from Turkey, and one from Italy. The special section is preceded by a discussed paper, and followed by four papers on various topics.

In the first paper of this issue, Rao, Scott and Benhin study the repeated inverse sampling method proposed by Hinkins, Oh and Scheuren. In this approach, random subsamples are drawn from a complex sample in such a way that each subsample is unconditionally a simple random sample from the population. Rao, Scott and Benhin present some theoretical results for the expectation and variance of the repeated inverse sampling estimator. They then explore some conditions under which the repeated inverse sampling estimator converges to the original full sample estimator. They finally propose an approach based on estimating equations that avoids some of the potential bias of the repeated inverse sampling estimator for nonlinear parameters. The paper is followed by two fascinating discussions by Eltinge and Hinkins, and a rejoinder by the authors.

Hogan, in the first paper of the special section on Census Coverage Errors, presents a concise overview of the survey used to provide estimates for net undercoverage in the 2000 Census. He presents the Accuracy and Coverage Evaluation (ACE) study in the context of general post enumeration surveys and dual system estimators. He also presents the assumptions needed for these types of surveys to produce unbiased estimates and a detailed discussion where these assumptions failed in the 2000 ACE. The results are very interesting.

The next paper is also concerned with the 2000 ACE. Cantwell and Ikeda examine the crucial assumptions made when some data is missing. One of the points the authors note is that when a rare characteristic – persons missed by the Census in this case – is being estimated the methods used to adjust for missing data are very important. The authors point out the changes made from the methods used in previous post enumeration surveys for the 2000 ACE.

Ayhan and Ekni present the coverage procedures used in a different census context. While the basic post enumeration survey design is used in Turkey, there are some interesting differences between their experiences and those of the United States. Since Turkey uses a de facto approach to Census residence as opposed to the de jure approach used in the United States, there are some operational differences in the post enumeration surveys. These are clearly pointed out by the authors.

The final paper in the special section on Census Coverage Errors, by Cocchi, Fabrizi and Trivisano, describes the 1991 Italian Population Census and the the Post Enumeration Survey (PES) used to measure undercount. Since the census is administered by municipalities, data on the statistical quality of municipalities are used as auxiliary information for PES modelling and estimation. Poisson regression trees and hierarchical Poisson models are used to analyze the data. Results are summarized and discussed, and some recommendations are given.

Skinner and Carter extend estimation for Skinner and Elliot's measure of disclosure risk for survey microdata from the equal probability sampling case to the unequal probability sampling case under an assumption of Poisson sampling. Effects of possible departures from Poisson sampling are also considered.

The problem of inference for partially synthetic microdata sets is considered by Reiter. Statistical agencies may release microdata sets with completely synthetic data in order to protect respondent confidentiality. Methods for inference when the complete dataset is synthetic have been developed but most agencies release only partially synthetic datasets, that is, datasets for which only sensitive variables are imputed. There has been little reported in the literature under this situation. Reiter's proposed method is shown to be valid under a Bayesian framework and under a design-based framework and is illustrated by simulation studies.

In Brewer and Donadio, a variance estimator for the Horvitz-Thompson estimator that does not require the calculation of the second-order inclusions probabilities is obtained under high entropy situations. High entropy situations occur when there is the absence of any detectable pattern or ordering in the selected sample units. Under high entropy situations, an approximate variance formula is derived and verified through a model-assisted approach. A sample estimator for this approximate design-variance of the Horvitz-Thompson estimator is then developed. Finally, the proposed estimator is empirically compared with several other estimators using several populations.

Finally, Chow and Thompson present a Bayesian approach to designs where social links are exploited to obtain a sample of hidden or hard-to-access human populations. The authors provide an accessible introduction to the Bayesian approach in which the social links from one person to another are used to create the prior distribution. It is easy to adjust these priors when information is vague. The result is that from the resulting posterior distribution a large number of questions can be answered.

M.P. Singh

Survey Methodology, December 2003
Vol. 29, No. 2, pp. 107-128
Statistics Canada

Undoing Complex Survey Data Structures: Some Theory and Applications of Inverse Sampling

J.N.K. RAO, A.J. SCOTT and E. BENHIN[1]

ABSTRACT

Application of classical statistical methods to data from complex sample surveys without making allowance for the survey design features can lead to erroneous inferences. Methods have been developed that account for the survey design, but these methods require additional information such as survey weights, design effects or cluster identification for microdata. Inverse sampling (Hinkins, Oh and Scheuren 1997) provides an alternative approach by undoing the complex survey data structures so that standard methods can be applied. Repeated subsamples with unconditional simple random sampling structure are drawn and each subsample analysed by standard methods and then combined to increase the efficiency. This method has the potential to preserve confidentiality of microdata, although computer-intensive. We present some theory of inverse sampling and explore its limitations. A combined estimating equations approach is proposed for handling complex parameters such as ratios and "census" linear regression and logistic regression parameters. The method is applied to a cluster correlated data set reported in Battese, Harter and Fuller (1988).

KEY WORDS: Combined estimating equations; Confidentiality; Repeated subsampling.

1. INTRODUCTION

There is a fairly clear distinction between the focus of traditional sample survey methodology and that of the rest of applied statistics. Survey samplers have concentrated on developing efficient (but complicated) ways of drawing samples to estimate rather simple quantities (population means, proportions, totals, *etc.*). Most other applied statisticians, by contrast, have concentrated on developing sophisticated methods for fitting very complicated models, but assuming a rather simple sampling structure (often that the observations are independent).

In reality, data from complicated surveys are often used to fit complicated models. For example, people may want to use data from a Labour Force Survey to characterize the association between education and unemployment levels. They might want to use data from health surveys to study the association between housing conditions or poverty and morbidity, and so on. Extending the range of application of standard methods so that they can be applied to data from complicated sample surveys, involving multi-stage sampling and variable selection probabilities, is difficult and cumbersome; see *e.g.*, Skinner, Holt and Smith (1989).

How do practitioners deal with the complexity of survey data structures? Adapting a quote from Hinkins, Oh and Scheuren (1997) (abbreviated HOS hereafter): "If your only tool is a hammer, every problem looks like a nail!"; the hammer available to most people is one of the big statistical packages (SAS, Splus, SPSS, *etc.*). Most people still just push their data through a standard program and ignore the survey design features. This is in spite of the fact that a great deal of effort over the last two decades has been spent on developing methods to analyze survey data that take account of design features, and specialized programs such as SUDAAN or WesVar are now available to implement some of these methods.

An alternative to developing complex new tools (which may rarely be used in practice anyway!) is to work backwards: instead of tailoring the methods to fit the data, tailor the data to fit the methods. One approach along these lines was developed in Rao and Scott (1992; 1999). Another approach has been suggested in HOS. Their basic idea is to avoid the pain caused by a complicated sample by choosing a subsample (inverse-sample) that has a simple random sample structure unconditionally (or at least has a structure that is considerably simpler to handle than the original sample). Obviously this involves some loss in efficiency, especially if the subsample is very much smaller than the original sample, as often turns out to be necessary. However, we can increase the efficiency by repeating the process independently many times and averaging the results.

Is it possible to produce subsamples with the desired properties? The answer is often "yes", although the resulting subsample size, m, might have to be small (in fact, no more than $m = 2$ for some standard stratified multistage designs). HOS give algorithms for producing simple random inverse-samples for a number of standard designs. We summarize the inverse sampling schemes in section 2 for ready reference. These schemes include both exact and approximate methods in terms of matching simple random sampling. In this paper we look at some of the properties of

[1] J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Canada, K1S 5B6. E-mail: jrao@math.carleton.ca; A.J. Scott, Department of Statistics, University of Auckland, Auckland, New Zealand. E-mail: scott@stat.auckland.ac.nz; E. Benhin, Household Survey Methods Division, Statistics Canada, Ottawa, Canada, K1A 0T6. E-mail: emmanuel.benhin@statcan.ca.

the repeated inverse sampling procedures given in section 2. In particular, we develop some basic theory of inverse sampling in section 3, and illustrate some of the strengths and weaknesses of the procedure. In section 4, we study the special case of a population total. We propose a combined estimating equations (CEE) approach in section 5 for handling complex parameters such as ratios and "census" regression parameters. Finally, some concluding remarks are given in section 6. Proofs of theorems are given in the appendix.

2. INVERSE SAMPLING ALGORITHMS

In this section we summarize the inverse sampling schemes, proposed by Hinkins *et al.* (1997), for ready reference. These schemes include both exact and approximate methods in terms of matching simple random sampling (SRS) unconditionally.

Suppose we have a sample s_0 of observations drawn from the finite population of size N according to a specified complex design. We wish to draw a subsample s^* of size m from s_0 such that the unconditional probability of s^*, $p(s^*)$, matches simple random sampling with $p(s^*) = 1/\binom{N}{m}$, either exactly or approximately. We have

$$p(s^*) = \sum_{s_0 \supset s^*} p_0(s_0)p(s^*|s_0), \tag{2.1}$$

where $p_0(s_0)$ is the probability of selecting s_0 and $p(s^*|s_0)$ is the conditional probability of choosing s^*. If $p(s^*|s_0)$ does not depend on s_0, then it follows from (2.1) that

$$p(s^*|s_0) = p_2(s^*) = \frac{p(s^*)}{\sum_{s_0 \supset s^*} p_0(s_0)}. \tag{2.2}$$

Denote the first-order and second-order inclusion probabilities corresponding to s^* and s_0 as (π_i^*, π_{il}^*) and (π_i, π_{il}) respectively, where $\pi_i^* = m/N$ and $\pi_{il}^* = m(m-1)/N(N-1), i \neq l$. Similarly, denote the conditional inclusion probabilities as $(\tilde{\pi}_i(s_0), \tilde{\pi}_{il}(s_0))$. If the conditional inclusion probabilities do not depend on s_0, then we write them as $(\tilde{\pi}_i, \tilde{\pi}_{il})$. It is readily seen that

$$\pi_i^* = \sum_{s_0 \ni i} p_0(s_0)\tilde{\pi}_i(s_0); \quad \pi_{il}^* = \sum_{s_0 \ni i,l} p_0(s_0)\tilde{\pi}_{il}(s_0). \tag{2.3}$$

If $\tilde{\pi}_i(s_0) = \tilde{\pi}_i$ and $\tilde{\pi}_{il}(s_0) = \tilde{\pi}_{il}$, then it follows from (2.3) that

$$\pi_i^* = \pi_i \tilde{\pi}_i, \quad \pi_{il}^* = \pi_{il} \tilde{\pi}_{il}. \tag{2.4}$$

In section 4 we use (2.4) to study the properties of inverse sampling for estimating a population total. Note that (π_i^*, π_{il}^*) may correspond to some other simpler sampling design if it is not feasible to match simple random sampling (SRS), *e.g.*, stratified simple random sampling.

2.1 Stratified Simple Random Sampling

Suppose that the original sample s_0 is a stratified simple random sample, *i.e.*,

$$p_0(s_0) = \prod_{h=1}^{L} \binom{N_h}{n_h}^{-1}, \tag{2.5}$$

where $N_h(n_h)$ denotes the number of population (sample) units in stratum $h(=1, ..., L)$. We wish to draw a subsample s^* of size m such that $p(s^*) = 1/\binom{N}{m}$, where $N = \sum_{h=1}^{L} N_h$. Clearly, m cannot be larger than $\min(n_h)$. Let $\boldsymbol{m} = (m_1, ..., m_L)^T$ denote the (random) number of units in each stratum that belong to s^*, $0 \le m_h \le m$, $\sum_{h=1}^{L} m_h = m$. Noting that the number of terms in $\sum_{s_0 \supset s^*}$ equals $\prod_{h=1}^{L}\binom{N_h - m_h}{n_h - m_h}$, it follows from (2.2) that

$$p(s^*|s_0) = \frac{\prod_{h=1}^{L}\binom{N_h}{m_h}}{\binom{N}{m}} \frac{1}{\prod_{h=1}^{L}\binom{n_h}{m_h}}. \tag{2.6}$$

The subsampling scheme readily follows from (2.6): (i) Generate $\boldsymbol{m}$ from the hypergeometric distribution $f(\boldsymbol{m}) = \prod_{h=1}^{L}\binom{N_h}{m_h}/\binom{N}{m}$; (ii) Draw a simple random sample of size m_h, without replacement, from the n_h sample units in stratum h, independently across strata $h(=1, ..., L)$. HOS specify $p(s^*|s_0)$ first and then verify that it gives $p(s^*) = \binom{N}{n}^{-1}$. Our approach provides the subsampling scheme from the specification of $p_0(s_0)$ and $p(s^*)$.

2.2 One-stage Cluster Sampling

HOS studied the case of one-stage cluster sampling in detail. Three sampling designs for s_0 were investigated: (1) Equal cluster sizes, M, and simple random sampling of clusters; (2) Unequal cluster sizes, M_i, and simple random sampling of clusters; (3) Unequal cluster sizes, M_i, and clusters sampled with probability proportional to size M_i and with replacement.

Case 1. Exact matching with SRS is difficult to implement in the case of equal cluster sizes, M, and simple random sampling of clusters. Suppose s_0 contains k clusters drawn from K clusters in the population ($N = KM$). A simple approximate method of subsampling selects one element at random from each sample cluster so that the size of s^* is k. Hoffman, Sen and Weinberg (2001) used a similar method for biostatistical applications. HOS used systematic sampling to select one case from each sample cluster.

Case 2. Hoffman *et al.* (2001) selected one unit at random from each cluster in the case of unequal cluster sizes, under a model-based framework for clustered data. For sampling applications, this method does not work in the sense that it is not possible to obtain SRS of fixed sizes by subsampling, even approximately. HOS proposed an alternative method that artificially enlarges the population to equal cluster size case and then applies subsampling used in Case 1. We first force all clusters to have the same size by adding an

appropriate number of pseudo-unit to bring them up to the size of the largest sample cluster. Then we take one unit at random from each sample cluster, and discard any pseudo-units to obtain the final sample. This approximate method makes $p(s^*|s_0)$ depend on s_0 because the conditional probability depends on $M(s_0)$, the size of the largest sample cluster.

Case 3. For the case of unequal cluster sizes M_i and probability proportional to size (PPS) sampling with replacement, HOS proposed a simple method of subsampling which gives $p(s^*) = (1/N)^k$, where s^* now denotes an ordered simple random sample drawn with replacement from the $N = \sum_{i=1}^{K} M_i$, units in the population, *i.e.*, $s^* = (i_1, ..., i_j, ..., i_k)$, where i_j denotes the unit drawn in the j-th draw ($j = 1, ..., k$). Viewing the sample clusters as ordered, we select one unit at random from each sample cluster. Note that the same cluster might appear more than once in the ordered sample. Denote the size of the cluster drawn in the i-th PPS draw by M_i', then

$$p(s^*) = \left[\prod_{i=1}^{k} \frac{M_i'}{N}\right]\left[\prod_{i=1}^{k} \frac{1}{M_i'}\right] = \left(\frac{1}{N}\right)^k, \qquad (2.7)$$

where $\prod_{i=1}^{k}(M_i'/N)$ is the probability of drawing the ordered cluster sample. Note that s_0 is the ordered PPS sample and we have only one term in the summation in (2.1).

If the clusters are drawn with inclusion probabilities $\pi_i = kM_i/N$ and without replacement, then it is not possible to match SRS. However, we can treat the clusters as if they were drawn with replacement, as done in practice, and then apply the scheme for Case 3. This will lead to overestimation of variance if the variance of the estimator is smaller than the variance of the estimator under PPS sampling with replacement (see *e.g.*, Wolter 1985, page 45). However, the overestimation is not serious if the sampling fraction k/K is small (see Section 4.3).

2.3 Two-stage Cluster Sampling

HOS also studied two-stage sampling for the following cases: (1) Equal cluster sizes, M, and k clusters sampled with equal probability in the first stage; simple random subsample of equal size, m, drawn independently within each sampled cluster (PSU). (2) Unequal cluster sizes, M_i, and k clusters sampled with PPS and with replacement; simple random subsamples of unequal sizes, m_i, drawn independently within each cluster in the with replacement sample.

Case 1. As in the case of one-stage cluster sampling, exact method of inverse sampling is difficult to implement. A simple approximate method of inverse sampling selects one unit at random from each of the k subsamples.

Case 2. As in Case 3 of uni-stage cluster sampling, we simply select one unit at random from each of the ordered subsamples. HOS suggested a different method: Take a simple random sample with replacement of k clusters first and then with each selected cluster take one unit at random from the corresponding subsample. It appears that the first stage inverse sampling of clusters is not necessary. To see this, we note that

$$p_0(s_0) = \prod_{i=1}^{k}\left[\left(\frac{M_i'}{N}\right)\frac{1}{\binom{M_i'}{m_i'}}\right],$$

where m_i' is the subsample size associated with the cluster selected in the i-th draw ($i = 1, ..., k$). We wish to draw a subsample s^* of size k such that $p(s^*) = (1/N)^k$, where $N = \sum_{i=1}^{K} M_i$. Also the number of terms in $\sum_{s_0 \supset s^*}$ equals $\prod_{i=1}^{k}\binom{M_i' - 1}{m_i' - 1}$ and

$$\sum_{s_0 \supset s^*} p_0(s_0) = \prod_{i=1}^{k}\left[\left(\frac{M_i'}{N}\right)\frac{\binom{M_i' - 1}{m_i' - 1}}{\binom{M_i'}{m_i'}}\right] = \prod_{i=1}^{k}\frac{m_i'}{N}.$$

It follows from (2.2) that $p(s^*|s_0) = \prod_{i=1}^{k}(1/m_i')$ and hence the subsampling scheme readily follows.

2.4 Stratified Two-stage Sampling

Suppose we have a two-stage sample from each stratum, where the clusters are sampled with PPS with replacement and subsampling is done independently within each sampled cluster. Using the inverse sampling procedure of Case 2, section 2.3, we get simple random samples from each stratum. We can then apply the method of section 2.1, treating the inverse-samples as if drawn without replacement to get an inverse-sample of size $k_0 = \min_h(k_h)$, where k_h is the number of sampled clusters in stratum h. In the important case of $k_h = 2$ psu's sampled from each stratum, the inverse-sample size, k_0, is only two.

3. BASIC PROPERTIES

The results in this section are quite general and apply equally to sample surveys and the type of clustered situation considered by Hoffman *et al.* (2001). Suppose that we are interested in estimating some population parameter, θ, and we have a sample, s_0, of observations drawn from the population according to some complex design. We assume that we have a subsampling algorithm that can produce samples from some simpler design. This design will often be simple random sampling, but we can extend the range of

applications considerably by allowing for the possibility of more general (sub-)designs; for example, stratified SRS when the original sample is a stratified two-stage sample. Our only requirement for the simpler design is that we can produce an estimator of the quantity of interest, θ, together with an estimator of its variance. Let $\hat{\theta}_j^*$ and $\hat{V}_j^*$ denote the estimator and variance estimator produced from the j-th subsample when we generate a sequence of g conditionally independent subsamples s_j^* $(j = 1, ..., g)$. Note that the $\hat{\theta}_j^*$'s are not unconditionally independent when averaged over the distribution of the initial sample, s_0. A "separate" inverse-sampling estimator of θ based on the g subsamples is given by

$$\hat{\theta}_g = \frac{1}{g}\sum_{j=1}^{g} \hat{\theta}_j^*. \tag{3.1}$$

We denote the estimator based on s_0 as $\hat{\theta}$. Theorem 1 below gives basic results on $\hat{\theta}_g$ and its variance.

Theorem 1

1. Conditional on the original sample, s_0, $\hat{\theta}_g$ converges almost surely to $E(\hat{\theta}_1^* \mid s_0) = \hat{\theta}_\infty$, say, as $g \to \infty$.
2. $E(\hat{\theta}_g) = E(\hat{\theta}_1^*)$.
3. $\mathrm{Var}(\hat{\theta}_g) = \mathrm{Var}(\hat{\theta}_\infty) + \frac{1}{g}E[\mathrm{Var}(\hat{\theta}_1^* \mid s_0)]$.
4. If $r_g = \dfrac{\mathrm{Var}(\hat{\theta}_g)}{\mathrm{Var}(\hat{\theta}_\infty)}$, then $r_g = 1 + \dfrac{r_1 - 1}{g}$.

Result 4 of Theorem 1 demonstrates that increasing the number of subsamples, g, does indeed increase the efficiency of $\hat{\theta}_g$. More precisely, the variance ratio r_g has the form $a + b / g$. If the subsample estimator, $\hat{\theta}_1^*$, is unbiased for θ, then so is the inverse-sampling estimator, $\hat{\theta}_g$. However, if $\hat{\theta}_1^*$ has bias of order m^{-1}, where m denotes the subsample size, then $\hat{\theta}_g$ has exactly the same bias. Since m will usually be very much smaller than the original sample size, this bias can be appreciable. This is a serious limitation of $\hat{\theta}_g$ in the nonlinear cases, such as ratios and regression coefficients. In section 5, we propose an alternative estimator of θ based on the estimating equations (EE) approach. This estimator is asymptotically unbiased for any m as the size of s_0 increases, unlike $\hat{\theta}_g$.

Result 4 of Theorem 1 can be used to determine the number of subsamples, g needed to obtain reasonable efficiency. For example, HOS give an example in which $r_1 = 29.3$. The original sample was a very efficient stratified random sample with $n = 15{,}618$ observations taken from the Statistics of Income corporate survey, while the subsample was a simple random sample of $m = 2{,}224$ observations. A single subsample is relatively inefficient. However, in this case, repeated inverse sampling recovers all the information in the original sample in the limit. Applying Result 4 of Theorem 1 leads immediately to the following table:

g	1	10	100	1000
r_g	29.3	3.83	1.28	1.03

(HOS produced these same results by simulation but this is unnecessary in view of Result 4 of Theorem 1.) We see that $g = 100$ subsamples would be adequate for many purposes and that we obtain almost full efficiency with $g = 1{,}000$.

The fact that $\hat{\theta}_1^*, ..., \hat{\theta}_g^*$ are not unconditionally independent means that estimating $\mathrm{Var}(\hat{\theta}_g)$ is not completely straightforward. However, a relatively simple variance estimator may be obtained using Theorem 2 below.

Theorem 2

The variance of $\hat{\theta}_g$ may be expressed as

$$\mathrm{Var}(\hat{\theta}_g) = \mathrm{Var}(\hat{\theta}_1^*) - \frac{g-1}{g}E\left[\mathrm{Var}\left(\hat{\theta}_1^* \mid s_0\right)\right]. \tag{3.2}$$

We can estimate the first term of (3.2) by $\hat{V}_j^*$ for $j = 1, ..., g$, and hence by their average $g^{-1}\sum_{j=1}^{g} \hat{V}_j^*$. In addition, the quantity

$$s_{\theta g}^2 = \frac{1}{g-1}\sum_{j=1}^{g} (\hat{\theta}_j^* - \hat{\theta}_g)^2$$

gives an unbiased estimator of $E[\mathrm{Var}(\hat{\theta}_1^* \mid s_0)]$ because $\hat{\theta}_1^*, ..., \hat{\theta}_g^*$ are conditionally independent given the initial sample, s_0. This leads to an estimator of $\mathrm{Var}(\hat{\theta}_g)$ of the form

$$\hat{V}_g = \frac{1}{g}\sum_{j=1}^{g} \hat{V}_j^* - \frac{1}{g}\sum_{j=1}^{g} (\hat{\theta}_j^* - \hat{\theta}_g)^2. \tag{3.3}$$

The properties of the variance estimator $\hat{V}_g$ depend on the properties of the subsample estimator $\hat{V}_j^*$. For example, if $\hat{V}_j^*$ is unbiased, then $\hat{V}_g$ is also unbiased.

For the special case of a population total $\theta = Y$ and simple random subsampling, *i.e.*, $p(s^*) = 1/\binom{N}{m}$, we have $\hat{\theta}_j^* = \hat{Y}_j^* = N\bar{y}_j^*$ and $\hat{Y}_j^*$ is unbiased for Y with unbiased variance estimator $\hat{V}_j^* = N^2(m^{-1} - N^{-1})s_{jy}^{*2}$, where $\bar{y}_j^*$ is the mean and s_{jy}^{*2} is the variance of the j-th subsample. The variance estimator $\hat{V}_g$ of $\hat{\theta}_g = \hat{Y}_g = g^{-1}\sum_{j=1}^{g}(N\bar{y}_j^*)$, given by (3.3), is unbiased, and it reduces to

$$\hat{V}_g = \frac{1}{g}\sum_{j=1}^{g} \hat{V}_j^* - \frac{N^2}{g}\sum_{j=1}^{g} \left(\bar{y}_j^* - \bar{y}_g\right)^2, \tag{3.4}$$

where $\bar{y}_g = g^{-1}\sum_{j=1}^{g} \bar{y}_j^*$. HOS derived a variance estimator by first expressing $\mathrm{Var}(\hat{Y}_g)$ as

$$\mathrm{Var}\left(\hat{Y}_g\right) = N^2\frac{m-1}{m}S_y^2 + \frac{1}{g}\sum_{j=1}^{g}\mathrm{Var}\left(\hat{Y}_j^*\right) - N^2\frac{mg-1}{mg}E\left(s_{cy}^{*2}\right), \tag{3.5}$$

where S_y^2 is the population variance and s_{cy}^{*2} is the sample variance using all gm subsample units. In the second step, they remarked that we can generate an approximately unbiased estimator of $\mathrm{Var}(\hat{Y}_g)$ from (3.5) by replacing S_y^2 and $\mathrm{Var}(\hat{Y}_j^*)$ with unbiased estimators and replacing $E(s_{cy}^{*2})$ by s_{cy}^{*2}. We now follow this recipe and obtain an explicit form for the HOS variance estimator, denoted $\hat{V}_{g(\mathrm{HOS})}$. Noting that each s_{jy}^{*2} is unbiased for S_y^2, a pooled unbiased estimator of S_y^2 is obtained as $g^{-1}\sum_{j=1}^{g} s_{jy}^{*2}$. Further, s_{cy}^{*2} may be decomposed as $(mg-1)s_{cy}^{*2} = (m-1)\sum_{j=1}^{g} s_{jy}^{*2} + m\sum_{j=1}^{g}(\bar{y}_j^* - \bar{y}_g)^2$. Hence,

$$\hat{V}_{g(\mathrm{HOS})} = N^2\left(\frac{m-1}{m}\right)\frac{1}{g}\sum_{j=1}^{g} s_{jy}^{*2} + \frac{1}{g}\sum_{j=1}^{g}\hat{V}_j^* - N^2\left\{\left(\frac{m-1}{m}\right)\frac{1}{g}\sum_{j=1}^{g} s_{jy}^{*2} + \frac{1}{g}\sum_{j=1}^{g}\left(\bar{y}_j^* - \bar{y}_g\right)^2\right\} = \hat{V}_g. \tag{3.6}$$

It follows from (3.6) that the variance estimator of HOS is in fact identical to our variance estimator (3.4) and also exactly unbiased.

4. ESTIMATION OF A TOTAL

4.1 Exact Matching

As shown in section 3, repeated subsampling increases the efficiency of an estimator, but this does not necessarily mean that the inverse-sampling estimator, $\hat{\theta}_g$, converges to the original full sample estimator, $\hat{\theta}$, as $g \to \infty$, even when we start with an unbiased estimator for the subsample. In this section, we study the special case of a total $\theta = Y$ and consider the Horvitz-Thompson (H-T) unbiased estimator, $\hat{Y} = \sum_{i\in s_0} y_i/\pi_i$, based on the original full-sample. Theorem 3 below establishes conditions under which the corresponding inverse-sampling estimator

$$\hat{Y}_g = \frac{1}{g}\sum_{j=1}^{g}\hat{Y}_j^* \tag{4.1}$$

converges to the H-T estimator, $\hat{Y}$, for the original design as $g \to \infty$, where

$$\hat{Y}_j^* = \sum_{i\in s_j^*}\frac{y_i}{\pi_i^*}$$

and π_i^* is the unconditional inclusion probability for the i-th unit. If the subsample s_j^* is a simple random sample unconditionally, then $\pi_i^* = m/N$, where m is the subsample size.

Theorem 3

Let $\tilde{\pi}_i(s_0)$ be the conditional probability that the i-th unit is selected in the subsample for a given initial sample, s_0. Suppose that $\hat{\theta}_j^* = \hat{Y}_j^*$ is the H-T estimator of a total $\theta = Y$ for the j-th subsample. Then the limiting inverse-sampling estimator, $\hat{\theta}_\infty^* = \hat{Y}_\infty$, will be the H-T estimator, $\hat{Y}$, for the original design if and only if the conditional inclusion probabilities $\tilde{\pi}_i(s_0)$ are constant for all s_0 containing the i-th unit, *i.e.*, $\tilde{\pi}_i(s_0) = \tilde{\pi}_i$ for all $s_0 \supset i$.

The condition $\tilde{\pi}_i(s_0) = \tilde{\pi}_i$ is a fairly natural one for most sampling designs for which the H-T estimator is used. If the subsamples are all simple random samples of fixed size m, then the estimator for a subsample is simply $N\bar{y}_j^*$, which is the natural estimator under simple random sampling.

Theorem 4 below establishes conditions under which the inverse-sampling variance estimator, $\hat{V}_{g,\mathrm{HT}}$, of $\hat{Y}_g$ converges to $\hat{V}_{\mathrm{HT}}$, the H-T variance estimator of the full-sample estimator $\hat{Y}$, as $g \to \infty$. We have

$$\hat{V}_{\mathrm{HT}} = \sum_{i,l\in s_0}\sum\frac{\pi_{il} - \pi_i\pi_l}{\pi_i\pi_l\pi_{il}}y_i y_l \tag{4.2}$$

(see Cochran 1977, page 261) and

$$\hat{V}_{g,\mathrm{HT}} = \frac{1}{g}\sum_{j=1}^{g}\hat{V}_{j,\mathrm{HT}}^* - \frac{1}{g}\sum_{j=1}^{g}(\hat{Y}_j^* - \hat{Y}_g)^2$$

with

$$\hat{V}_{j,\mathrm{HT}}^* = \sum_{i,l\in s_j^*}\sum\frac{\pi_{il}^* - \pi_i^*\pi_l^*}{\pi_i^*\pi_l^*\pi_{il}^*}y_i y_l, \tag{4.3}$$

where π_{il}^* is the unconditional joint inclusion probability for the i-th and l-th units. If the subsample s_j^* is a simple random subsample unconditionally, then $\pi_{il}^* = m(m-1)/[N(N-1)]$, $i \neq l$. Note that $\hat{V}_{j,\mathrm{HT}}^*$ is the H-T variance estimator of $\hat{Y}_j^*$, and $\pi_{ii}^* = \pi_i^*$, $\pi_{ii} = \pi_i$.

Theorem 4

If $\hat{V}_{j,\mathrm{HT}}^*$ is the Horvitz-Thompson (H-T) variance estimator of $\hat{Y}_j^*$ for the j-th subsample, then conditional on s_0, $\hat{V}_{g,\mathrm{HT}}$, converges to the Horvitz-Thompson (H-T) variance estimator of $\hat{Y}$ for the original design, as $g \to \infty$, if the conditional joint inclusion probabilities are constant for all s_0 containing a given pair (i, l) of units, *i.e.*, $\tilde{\pi}_{il}(s_0) = \tilde{\pi}_{il}$ for all $s_0 \supset \{i, l\}$.

In Theorem 4 we considered the H-T variance estimator. But the Sen-Yates-Grundy(S-Y-G) variance estimator, $\hat{V}_{\mathrm{SYG}}$, is often preferred over the H-T variance estimator, $\hat{V}_{\mathrm{HT}}$, because it is more stable and several designs for which it is always nonnegative are known, while $\hat{V}_{\mathrm{HT}}$ frequently takes negative values (Cochran 1977, page 261). The S-Y-G variance estimator of $\hat{Y}$ exists for fixed sample size designs and it is given by

$$\hat{V}_{\mathrm{SYG}} = \sum_{i<l\in s_0}\sum\frac{\pi_i\pi_l - \pi_{il}}{\pi_{il}}\left(\frac{y_i}{\pi_i} - \frac{y_l}{\pi_l}\right)^2, \tag{4.4}$$

for the full-sample design. Similarly, the S-Y-G variance estimator of $\hat{Y}_j^*$ is

$$\hat{V}^*_{j,\mathrm{SYG}} = \sum\sum_{i<l\in s_j^*} \frac{\pi_i^*\pi_l^* - \pi_{il}^*}{\pi_{il}^*}\left(\frac{y_i}{\pi_i^*} - \frac{y_l}{\pi_l^*}\right)^2. \qquad (4.5)$$

The inverse-sampling variance estimator is given by

$$\hat{V}_{g,\mathrm{SYG}} = \frac{1}{g}\sum_{j=1}^{g} \hat{V}^*_{j,\mathrm{SYG}} - \frac{1}{g}\sum_{j=1}^{g} (\hat{Y}_j^* - \hat{Y}_g)^2. \qquad (4.6)$$

Theorem 5 below shows that $\hat{V}_{g,\mathrm{SYG}}$ does not converge to $\hat{V}_{\mathrm{SYG}}$ as $g \to \infty$, *i.e.*, $\hat{V}_{\infty,\mathrm{SYG}} \neq \hat{V}_{\mathrm{SYG}}$. If the subsample is a simple random sample unconditionally, *i.e.*, $\pi_i^* = m/N$ and $\pi_{il}^* = m(m-1)/[N(N-1)]$; $i \neq l$, then $\hat{V}^*_{j,\mathrm{HT}} = \hat{V}^*_{j,\mathrm{SYG}}$ and $\hat{V}_{\infty,\mathrm{SYG}} = \hat{V}_{\infty,\mathrm{HT}} = \hat{V}_{\mathrm{HT}}$, the H-T variance estimator of $\hat{Y}$.

Theorem 5

The inverse-sampling variance estimator (4.6) does not converge to the S-Y-G variance estimator (4.4) as $g \to \infty$.

4.2 Exact Matching: PPS Estimates

(i) Unistage cluster sampling

For the case of PPS sampling with replacement of clusters with unequal sizes M_i, we have exact matching with SRS with replacement. The estimates of Y is given by $\hat{Y}_{\mathrm{pps}} = (N/k)\sum_{i=1}^{k} \bar{Y}_i'$, where N is the total number of population elements and $\bar{Y}_i'$ is the mean of the cluster selected on the i-th draw. The estimator $\hat{Y}_{\mathrm{pps}}$ is not equal to the H-T estimator of Y. The variance estimator of $\hat{Y}_{\mathrm{pps}}$ is given by

$$\hat{V}_{\mathrm{pps}} = \frac{N^2}{k}\frac{1}{k-1}\sum_{i=1}^{k}\left(\bar{Y}_i' - \frac{1}{k}\sum_{i=1}^{k}\bar{Y}_i'\right)^2.$$

The inverse-sampling estimator corresponding to $\hat{Y}_{\mathrm{pps}}$ is given by $\hat{Y}_g = g^{-1}\sum_{j=1}^{g}\hat{Y}_j^*$, where $\hat{Y}_j^*$ denotes the estimator of Y from the j-th inverse sample. It is easy to verify that $\hat{Y}_\infty = \hat{Y}_{\mathrm{pps}}$, and $\hat{V}_\infty = \hat{V}_{\mathrm{pps}}$, noting that $\hat{Y}_j^* = (N/k)\sum_{i\in s_j^*} y_i'$ where y_i' denotes the value of the element of an inverse-sample selected from the cluster in the i-th draw. Thus, inverse sampling preserves both the estimator and the variance estimator.

(ii) Two-stage cluster sampling

Turning to the case of unequal cluster sizes, M_i, we select the clusters with PPS and with replacement, and then draw simple random subsampling of equal size, m, independently within each cluster in the with-replacement sample. The estimator of Y is $\hat{Y}_{\mathrm{pps}} = (N/k)\sum_{i=1}^{k}\bar{y}_i'$ where $\bar{y}_i'$ is the sample mean of the cluster selected in the i-th draw. The variance estimator of $\hat{Y}_{\mathrm{pps}}$ is given by

$$\hat{V}_{\mathrm{pps}} = \frac{N^2}{k}\frac{1}{k-1}\sum_{i=1}^{k}\left(\bar{y}_i' - \frac{1}{k}\sum_{i=1}^{k}\bar{y}_i'\right)^2.$$

The inverse-sampling estimator is given by $\hat{Y}_g = g^{-1}\sum_{j=1}^{g}\hat{Y}_j^*$, where $\hat{Y}_1^* = (N/k)\sum_{i\in s_j^*} y_i'$, and y_i' is defined as above. It is easy to verify that $\hat{Y}_\infty = \hat{Y}_{\mathrm{pps}}$ and $\hat{V}_\infty = \hat{V}_{\mathrm{pps}}$. Thus, inverse sampling preserves both the estimator and the variance estimator.

4.3 Approximate Matching

In section 2 we noted that exact matching with SRS is difficult to implement when the original sampling design involves clusters. We proposed several approximate matching methods to overcome this difficulty. In this subsection we study the properties of the approximate matching methods.

4.3.1 Unistage Cluster Sampling

In section 2.2, Case 1, we considered the case of equal cluster sizes, M, and simple random sampling of clusters. The estimator of a total Y is given by $\hat{Y} = (K/k)\sum_{i=1}^{k} Y_i$, where Y_i is the i-th sample cluster total and K is the number of population clusters. The variance estimator of $\hat{Y}$ is

$$\hat{V} = \frac{K^2}{k}\left(1 - \frac{k}{K}\right)\frac{1}{k-1}\sum_{i=1}^{k}\left[Y_i - \frac{1}{k}\sum_{i=1}^{k} Y_i\right]^2.$$

For inverse sampling, we proposed approximate matching by selecting one unit at random from each sample cluster, $i(=1,\ldots,k)$. The inverse-sampling estimator is given by $\hat{Y}_g = g^{-1}\sum_{j=1}^{g}\hat{Y}_j^*$ with $\hat{Y}_j^* = N\bar{y}_j^*$ denoting the estimator of total Y from the j-th inverse-sample. The inverse-sampling variance estimator, $\hat{V}_g$, is given by (3.4).

It is easy to verify that $\hat{Y}_\infty = \hat{Y}$ so that approximate matching preserves the original estimator $\hat{Y}$ in the limit. On the other hand, it can be shown that

$$\hat{V}/\hat{V}_\infty = 1 - k/K. \qquad (4.7)$$

It now follows from (4.7) that $\hat{V}_\infty$ leads to overestimation of the variance if the sampling fraction k/K is not small.

4.3.2 Two-stage Cluster Sampling

In section 2.3, Case 1, we considered the case of two-stage cluster sampling with equal cluster sizes, M, and SRS without replacement in both stages. The H-T estimator of the total Y is given by $\hat{Y} = (K/k)\sum_{i=1}^{k}\hat{Y}_i$, where $\hat{Y}_i = M\bar{y}_i$ and $\bar{y}_i$ is the sample mean of the i-th sample cluster. The variance estimator of $\hat{Y}$ is given by

$$\hat{V} = N^2\left\{\frac{1}{k}\left(1 - \frac{k}{K}\right)s_{1y}^2 + \frac{k}{K}\left(1 - \frac{m}{M}\right)\frac{1}{km}s_{2y}^2\right\}, \qquad (4.8)$$

where $s_{1y}^2 = \sum_{i=1}^{k}(\bar{y}_i - \bar{y})^2/(k-1)$, $s_{2y}^2 = \sum_{i=1}^{k} s_{2i}^2/k$ with s_{2i}^2 denoting the sample variance in the i-th cluster, $\bar{y}_i$ is the i-th cluster sample mean and $\bar{y} = \sum_{i=1}^{k}\bar{y}_i/k$ is the overall sample mean (see Cochran 1977, pages 276-278).

For inverse sampling, we proposed approximate matching by selecting one element at random from the m sample elements in each sample cluster $i(=1,\ldots,k)$. Denote the values of the elements by $y_1',\ldots,y_k'$. The inverse-sampling estimator of the total is given by

$\hat{Y}_g = g^{-1}\sum_{j=1}^{g}\hat{Y}_j^*$, where $\hat{Y}_j^* = (N/k)\sum_{i\in s_j^*} y_i'$. The inverse-sampling variance estimator, $\hat{V}_g$, is given by (3.4).

It is easy to verify that $\hat{Y}_\infty = \hat{Y}$ so that approximate matching preserves the original estimator $\hat{Y}$ in the limit. On the other hand, it can be shown that $\hat{V}_g$ tends to

$$\hat{V}_\infty = N^2\frac{1}{k}s_{1y}^2 \tag{4.9}$$

as $g \to \infty$. It follows from (4.8) and (4.9) that

$$\frac{\hat{V}}{\hat{V}_\infty} = 1 - \frac{k}{K}\left[1 - \left(1 - \frac{m}{M}\right)\frac{1}{km}\frac{s_{2y}^2}{s_{1y}^2}\right]$$

$$\approx 1 - \frac{k}{K}, \tag{4.10}$$

because the neglected term in (4.10) is of order $(mK)^{-1}$. It follows that $\hat{V}_\infty$ again leads to overestimation of the variance if the sampling fraction k/K is not small.

5. COMBINED ESTIMATING EQUATIONS APPROACH

In this section, we study an estimating equations approach to inverse sampling. This approach permits valid inferences on nonlinear parameters such as ratios and "census" linear regression and logistic regression parameters. As noted in section 3, the inverse-sampling estimator $\hat{\theta}_g$, given by (3.1), has exactly the same bias as the subsample estimator $\hat{\theta}_1^*$, and the bias of $\hat{\theta}_1^*$ is of order m^{-1}, where m is the subsample size. As a result, the bias of $\hat{\theta}_g$ can be appreciable because m is usually very much smaller than the original sample size n. In fact, m could be as small as 2 for stratified two-stage cluster sampling designs with two sample clusters in each stratum. Moreover, for logistic regression and other cases, the calculation of $\hat{\theta}_j^*$ and $\hat{\theta}$ involves iterative solutions. As a result, the implementation of $\hat{\theta}_g$, and the inverse-sampling variance estimator $\hat{V}_g$, given by (3.3), could become computationally very cumbersome when the number of inverse-samples, g, is large. We avoid these difficulties using a combined estimating equations (CEE) approach.

In section 5.1, we consider the special case of a ratio of totals, $R = Y / X$, and spell out the "combined approach" suggested by HOS towards the end of section 3.1 of their paper. Section 5.2 gives the general theory and discusses special cases. The results of section 5.2 are applied in section 5.3 to a cluster correlated data set reported in Battese, Harter and Fuller (1988).

5.1 Ratio of Totals

HOS suggested a "combined approach" to estimate the ratio, R, of totals Y and X. We now explain this approach and relate it to the CEE approach in section 5.2.

Denote the estimator of R based on the j-th inverse-sample as $\hat{R}_j^* = \hat{Y}_j^*/\hat{X}_j^*$. The separate inverse-sampling estimator of R is then given by $\hat{R}_g = g^{-1}\sum_{j=1}^{g}\hat{R}_j^*$. HOS noted that the bias of $\hat{R}_g$ can be large when the subsample size is small. They proposed to estimate the numerator and denominator of R separately, using the g subsamples. This leads to the "combined" inverse-sample estimator

$$\hat{R}_{gc} = \frac{\hat{Y}_g}{\hat{X}_g}, \tag{5.1}$$

where $\hat{Y}_g = g^{-1}\sum_{j=1}^{g}\hat{Y}_j^*$ and $\hat{X}_g = g^{-1}\sum_{j=1}^{g}\hat{X}_j^*$. Now, assuming that the final size of the "combined" sample is sufficiently large, it follows from (5.1) that

$$E(\hat{R}_{gc}) \approx \frac{E(\hat{Y}_g)}{E(\hat{X}_g)} = \frac{Y}{X} = R$$

under the conditions of Theorem 3. That is, $\hat{R}_{gc}$ is approximately unbiased for R, regardless of the subsample size, provided g is sufficiently large.

Similarly, using the Taylor linearization approximation, we obtain the variance of $\hat{R}_{gc}$ as

$$V(\hat{R}_{gc}) \approx \frac{1}{X^2}V(\tilde{U}_g), \tag{5.2}$$

where $\tilde{U}_g = g^{-1}\sum_{j=1}^{g}\tilde{U}_j^*$ is the inverse-sampling estimator of the total U of the residuals $u_i = y_i - Rx_i$, $i = 1, \ldots, N$. Noting that $\tilde{U}_g$ is the inverse-sampling estimator of a total, it follows from (3.3) that an inverse-sampling estimator of $V(\tilde{U}_g)$ is given by

$$\tilde{V}_{gU} = \frac{1}{g}\sum_{j=1}^{g}\tilde{V}_{jU}^* - \frac{1}{g}\sum_{j=1}^{g}\left(\tilde{U}_j^* - \tilde{U}_g\right)^2, \tag{5.3}$$

where $\tilde{V}_{jU}^*$ is the variance estimator produced from the j-th subsample. Since R is unknown, we replace R by $\hat{R}_{gc}$ in (5.3) to get the variance estimator $\hat{V}_{gU}$. Now, replacing X by its estimator $\hat{X}_g$ and $V(\tilde{U}_g)$ by $\hat{V}_{gU}$ in (5.2), we get the inverse-sampling linearization variance estimator of $\hat{R}_{gc}$ as

$$\hat{V}_L(\hat{R}_{gc}) = \frac{1}{\hat{X}_g^2}\hat{V}_{gU}. \tag{5.4}$$

Under the conditions of Theorem 4, $\hat{V}_L(\hat{R}_{gc})$ converges to the customary linearization variance estimator of the full-sample estimator $\hat{R} = \hat{Y}/\hat{X}$.

5.2 Nonlinear Parameters

(i) Full-sample estimating equations

A finite population parameter vector $\boldsymbol{\theta}_N$ may be regarded as the solution to "census" estimating equations (EE's):

$$U(\boldsymbol{\theta}) = \sum_{k \in U} \mathbf{u}_k(\boldsymbol{\theta}) = \mathbf{0}, \tag{5.5}$$

where $\sum_{k \in U}$ denotes the summation over the finite population U of size N, and the estimating functions $\mathbf{u}_k(\boldsymbol{\theta})$ are suitably chosen (Binder 1983; Godambe and Thompson 1986). For example, consider the scalar case of (5.5) and let $u_k(\theta) = y_k - \theta$ in (5.5). This gives the population mean $\theta_N = \bar{Y}$. Similarly, letting $u_k(\theta) = y_k - \theta x_k$ we get the ratio of totals: $\theta_N = R = Y/X$. The choice $\mathbf{u}_k(\boldsymbol{\theta}) = \mathbf{x}_k(y_k - \mu_k(\boldsymbol{\theta}))$ with $\mu_k(\boldsymbol{\theta}) = \mathbf{x}_k^T\boldsymbol{\theta}$ gives the census linear regression parameters

$$\boldsymbol{\theta}_N = \left(\sum_{k \in U} \mathbf{x}_k \mathbf{x}_k^T\right)^{-1} \sum_{k \in U} \mathbf{x}_k y_k.$$

The choice $\mathbf{u}_k(\boldsymbol{\theta}) = \mathbf{x}_k(y_k - \mu_k(\boldsymbol{\theta}))$ with $\log[\mu_k(\boldsymbol{\theta})/(1-\mu_k(\boldsymbol{\theta}))] = \mathbf{x}_k^T\boldsymbol{\theta}$ gives the census logistic regression parameters $\boldsymbol{\theta}_N$. Kovacevic and Binder (1997) give estimating functions, $\mathbf{u}_k(\boldsymbol{\theta})$, that lead to various measures of income inequality, such as the Gini index and the polarization index.

The full-sample estimating equations are given by

$$\hat{\mathbf{U}}(\boldsymbol{\theta}) = \sum_{k \in s_0} w_k \mathbf{u}_k(\boldsymbol{\theta}) = \mathbf{0}, \tag{5.6}$$

where w_k is the survey weight attached to $k \in s_0$; in particular, $w_k = 1/\pi_k$ if the H-T estimator of $\mathbf{U}(\boldsymbol{\theta})$ is used. The solution to (5.6) gives the full-sample estimator $\hat{\boldsymbol{\theta}}$ which, in general, is nonlinear and hence biased. We assume that the size of the original sample, s_0, is large enough to neglect the bias of $\hat{\boldsymbol{\theta}}$. For logistic regression and other complex cases, it is necessary to solve (5.6) iteratively to obtain the full-sample estimator $\hat{\boldsymbol{\theta}}$. The Newton-Raphson (N-R) algorithm is commonly used to solve (5.6). The r-th step of the N-R algorithm is given by

$$\hat{\boldsymbol{\theta}}^{(r)} = \hat{\boldsymbol{\theta}}^{(r-1)} + \hat{\mathbf{J}}^{-1}(\hat{\boldsymbol{\theta}}^{(r-1)})\,\hat{\mathbf{U}}(\hat{\boldsymbol{\theta}}^{(r-1)}), \tag{5.7}$$

where $\hat{\boldsymbol{\theta}}^{(r-1)}$ is the value of $\hat{\boldsymbol{\theta}}$ obtained at the $(r-1)$-th iteration, and $\hat{\mathbf{U}}(\hat{\boldsymbol{\theta}}^{(r-1)})$ and $\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}}^{(r-1)})$ are the values of $\hat{\mathbf{U}}(\boldsymbol{\theta})$ and $\hat{\mathbf{J}}(\boldsymbol{\theta}) = -\partial\hat{\mathbf{U}}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T = -\sum_{k \in s_0} w_k \partial\mathbf{u}_k(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(r-1)}$ Iterating the N-R algorithm to convergence produces the estimator $\hat{\boldsymbol{\theta}}$ as well as the observed information matrix $\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})$.

Under regularity conditions, Binder (1983) obtained a Taylor linearization estimator of the covariance matrix, $\mathbf{V}(\hat{\boldsymbol{\theta}})$, of $\hat{\boldsymbol{\theta}}$ as

$$\hat{\mathbf{V}}_L(\hat{\boldsymbol{\theta}}) = \left[\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})\right]^{-1} \hat{\mathbf{V}}\left[\hat{\mathbf{U}}(\hat{\boldsymbol{\theta}})\right]\left[\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})\right]^{-1}, \tag{5.8}$$

where $\hat{\mathbf{V}}[\hat{\mathbf{U}}(\hat{\boldsymbol{\theta}})]$ is a variance estimator of the estimated total, $\hat{\mathbf{U}}(\boldsymbol{\theta})$, of the $\mathbf{u}_k(\boldsymbol{\theta})$'s evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. For example, if $u_k(\theta) = y_k - \theta x_k$ then $\hat{\theta} = \sum_{k \in s_0} w_k y_k / \sum_{k \in s_0} w_k x_k = \hat{Y}/\hat{X} = \hat{R}$ is the ratio estimator, and (5.8) reduces to the customary linearization variance estimator

$$\hat{V}_L(\hat{\theta}) = \frac{1}{\hat{X}^2}\hat{V}\left[\sum_{k \in s_0} w_k u_k(\hat{\theta})\right], \tag{5.9}$$

noting that $\hat{J}(\hat{\theta}) = \sum_{k \in s_0} w_k x_k = \hat{X}$.

(ii) **Separate estimating equations**

The separate inverse-sampling estimators $\hat{\boldsymbol{\theta}}_j^*$, $j = 1, ..., g$ are obtained by solving the separate estimating equations (SEE)

$$\hat{\mathbf{U}}_j^*(\boldsymbol{\theta}) = \frac{N}{m}\sum_{k \in s_j^*} \mathbf{u}_k(\boldsymbol{\theta}) = 0;\ j = 1, ..., g. \tag{5.10}$$

In general, we require g iterative solutions to get $\hat{\boldsymbol{\theta}}_1^*, ..., \hat{\boldsymbol{\theta}}_g^*$. The inverse-sampling estimator of θ is then given by

$$\hat{\boldsymbol{\theta}}_g = \frac{1}{g}\sum_{j=1}^{g} \hat{\boldsymbol{\theta}}_j^*. \tag{5.11}$$

It follows from (5.11) that $\hat{\boldsymbol{\theta}}_\infty = E(\hat{\boldsymbol{\theta}}_1^* \mid s_0)$ and $E(\hat{\boldsymbol{\theta}}_\infty) = E(\hat{\boldsymbol{\theta}}_1^*)$. Assuming first moment matching with SRS, it follows from (5.10) that the bias $E(\hat{\boldsymbol{\theta}}_1^*) - \boldsymbol{\theta}$ is of order m^{-1}, where m is the subsample size. The inverse-sampling estimator of $V(\hat{\boldsymbol{\theta}}_g)$ is given by

$$\hat{V}_g = \frac{1}{g}\sum_{j=1}^{g} \hat{V}_j^* - \frac{1}{g}\sum_{j=1}^{g} \left(\hat{\boldsymbol{\theta}}_j^* - \hat{\boldsymbol{\theta}}_g\right)\left(\hat{\boldsymbol{\theta}}_j^* - \hat{\boldsymbol{\theta}}_g\right)^T, \tag{5.12}$$

where $\hat{\mathbf{V}}_j^*$ is given by

$$\hat{\mathbf{V}}_j^* = \left[\hat{\mathbf{J}}_j^*(\hat{\boldsymbol{\theta}}_j^*)\right]^{-1} \hat{\mathbf{V}}\left[\hat{\mathbf{U}}_j^*(\hat{\boldsymbol{\theta}}_j^*)\right]\left[\hat{\mathbf{J}}_j^*(\hat{\boldsymbol{\theta}}_j^*)\right]^{-1}, \tag{5.13}$$

$\hat{\mathbf{V}}[\hat{\mathbf{U}}_j^*(\hat{\boldsymbol{\theta}}_j^*)]$ is the variance estimator of the j-th subsample total $\hat{\mathbf{U}}_j^*(\boldsymbol{\theta})$, denoted $\hat{\mathbf{V}}_{jU}^*$ (see equation (5.19) below), evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_j^*$, and $\hat{\mathbf{J}}_j^*(\hat{\boldsymbol{\theta}}_j^*)$ is $\hat{\mathbf{J}}_j^*(\boldsymbol{\theta}) = -\partial\hat{\mathbf{U}}_j^*(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_j^*$.

(iii) **Combined estimating equations**

We now obtain a combined estimating equations (CEE) estimator $\hat{\boldsymbol{\theta}}_{gc}$ that leads to valid inference regardless of the subsample size m. We simply combine the g equations in (5.10) before solving for $\boldsymbol{\theta}$. This leads to combined estimating equations

$$\hat{\mathbf{U}}_{gc}(\boldsymbol{\theta}) = \frac{1}{g}\sum_{j=1}^{g} \hat{\mathbf{U}}_j^*(\boldsymbol{\theta}) = \frac{1}{g}\sum_{j=1}^{g} \frac{N}{m}\sum_{k \in s_j^*} \mathbf{u}_k(\boldsymbol{\theta}) = \mathbf{0}. \tag{5.14}$$

In general, we solve (5.14) using the N-R iterations (5.7) with $\hat{\mathbf{U}}(\hat{\boldsymbol{\theta}}^{(r-1)})$ changed to $\hat{\mathbf{U}}_{gc}(\hat{\boldsymbol{\theta}}^{(r-1)})$ and $\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}}^{(r-1)})$ changed to $\hat{\mathbf{J}}_{gc}(\hat{\boldsymbol{\theta}}^{(r-1)})$, where

$$\hat{\mathbf{J}}_{gc}(\boldsymbol{\theta}) = -\frac{\partial\hat{\mathbf{U}}_{gc}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T} = -\frac{1}{g}\sum_{j=1}^{g} \frac{N}{m}\sum_{k \in s_j^*} \frac{\partial\mathbf{u}_k(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T}. \tag{5.15}$$

At convergence, we obtain the CEE estimator $\hat{\boldsymbol{\theta}}_{gc}$ as well as the observed information matrix $\hat{\mathbf{J}}_{gc}(\hat{\boldsymbol{\theta}}_{gc})$. Note that we solve the combined estimating equations (5.14) only once to get $\hat{\boldsymbol{\theta}}_{gc}$, unlike the separate estimating equations method

that solves the g equations (5.10) to get $\hat{\boldsymbol{\theta}}_1^*, ..., \hat{\boldsymbol{\theta}}_g^*$ and $\hat{\boldsymbol{\theta}}_g = \sum_{j=1}^g \hat{\boldsymbol{\theta}}_j^* / g$.

To illustrate the proposed CEE method, consider the special case of ratio $\theta_N = R$, in which case $u_k(\theta) = y_k - \theta x_k$. The combined estimating equations (5.14) reduce to $\hat{Y}_g - \theta \hat{X}_g = 0$ and the solution $\hat{\theta}_{gc}$ is identical to the combined inverse-sampling estimator $\hat{R}_{gc}$ given by (5.1).

Assuming first moment matching with SRS, it follows from (5.14) that $\hat{\boldsymbol{\theta}}_{\infty c}$ is a solution of

$$\hat{\mathbf{U}}_{\infty c}(\boldsymbol{\theta}) = E\left[\hat{\mathbf{U}}_1^*(\boldsymbol{\theta}) \mid s_0\right] = \hat{\mathbf{U}}(\boldsymbol{\theta}) = \mathbf{0}. \tag{5.16}$$

As a result, $\hat{\boldsymbol{\theta}}_{\infty c} = \hat{\boldsymbol{\theta}}$ regardless of the subsample size m. Thus, the bias of $\hat{\boldsymbol{\theta}}_{gc}$ is of the same order as the bias of $\hat{\boldsymbol{\theta}}$ for sufficiently large g, regardless of the subsample size, m.

We now apply Binder'(1983) method to $\hat{\mathbf{U}}_{gc}(\boldsymbol{\theta})$ to get a linearization inverse-sampling estimator of $\mathbf{V}(\hat{\boldsymbol{\theta}}_{gc})$. It follows from (5.8) that

$$\hat{\mathbf{V}}_L(\hat{\boldsymbol{\theta}}_{gc}) = \left[\hat{\mathbf{J}}_{gc}(\hat{\boldsymbol{\theta}}_{gc})\right]^{-1} \hat{\mathbf{V}}\left[\hat{\mathbf{U}}_{gc}(\hat{\boldsymbol{\theta}}_{gc})\right]\left[\hat{\mathbf{J}}_{gc}(\hat{\boldsymbol{\theta}}_{gc})\right]^{-1}, \tag{5.17}$$

where $\hat{\mathbf{V}}\left[\hat{\mathbf{U}}_{gc}(\hat{\boldsymbol{\theta}}_{gc})\right]$ is the variance estimator of the estimated total, $\hat{\mathbf{U}}_{gc}(\boldsymbol{\theta})$, of the $\mathbf{u}_k(\boldsymbol{\theta})$'s evaluated at $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{gc}$. Note that $\hat{\mathbf{J}}_{gc}(\hat{\boldsymbol{\theta}}_{gc})$ is obtained at the convergence of the N-R algorithm applied to (5.14).

Since $\hat{\mathbf{U}}_{gc}(\boldsymbol{\theta})$ is the inverse-sampling estimator of the total $\mathbf{U}(\boldsymbol{\theta})$, it follows that the inverse-sampling estimator of $\mathbf{V}[\hat{\mathbf{U}}_{gc}(\boldsymbol{\theta})]$ is given by

$$\tilde{\mathbf{V}}_{gU} = \frac{1}{g}\sum_{j=1}^{g} \tilde{\mathbf{V}}_{jU}^* - \frac{1}{g}\sum_{j=1}^{g} \left[\hat{\mathbf{U}}_j^*(\boldsymbol{\theta}) - \hat{\mathbf{U}}_{gc}(\boldsymbol{\theta})\right]\left[\hat{\mathbf{U}}_j^*(\boldsymbol{\theta}) - \hat{\mathbf{U}}_{gc}(\boldsymbol{\theta})\right]^T, \tag{5.18}$$

where $\tilde{\mathbf{V}}_{jU}^*$ is the SRS variance estimator from the j-th subsample, assuming second moment matching. If the matching is with respect to SRS without replacement, then

$$\tilde{\mathbf{V}}_{jU}^* = \frac{N^2}{m}\left(1 - \frac{m}{N}\right)\frac{1}{m-1} \times \sum_{k \in s_j^*} \left[\mathbf{u}_k(\boldsymbol{\theta}) - \frac{1}{m}\sum_{k \in s_j^*} \mathbf{u}_k(\boldsymbol{\theta})\right]\left[\mathbf{u}_k(\boldsymbol{\theta}) - \frac{1}{m}\sum_{k \in s_j^*} \mathbf{u}_k(\boldsymbol{\theta})\right]^T \tag{5.19}$$

In the case of matching to SRS with replacement, we replace $1 - m/N$ by 1 in (5.19). Now substituting $\hat{\boldsymbol{\theta}}_{gc}$ for $\boldsymbol{\theta}$ in (5.18) we get

$$\hat{\mathbf{V}}\left[\hat{\mathbf{U}}_{gc}(\hat{\boldsymbol{\theta}}_{gc})\right] = \frac{1}{g}\sum_{j=1}^{g} \hat{\mathbf{V}}_{jU}^* - \frac{1}{g}\sum_{j=1}^{g} \hat{\mathbf{U}}_j^*(\hat{\boldsymbol{\theta}}_{gc})\hat{\mathbf{U}}_j^*(\hat{\boldsymbol{\theta}}_{gc})^T = \hat{\mathbf{V}}_{gU}, \tag{5.20}$$

where $\hat{\mathbf{V}}_{jU}^*$ is obtained from (5.19) by substituting $\hat{\boldsymbol{\theta}}_{gc}$ for $\boldsymbol{\theta}$. Note that $\hat{\mathbf{U}}_{gc}(\hat{\boldsymbol{\theta}}_{gc}) = \mathbf{0}$.

Under second moment matching with SRS, as $g \to \infty$, it is easy to verify that $\hat{\mathbf{V}}_L(\hat{\boldsymbol{\theta}}_{gc})$ converges to Binder's estimator $\hat{\mathbf{V}}_L(\hat{\boldsymbol{\theta}})$ given by (5.8). This follows by noting that $\hat{\boldsymbol{\theta}}_{\infty c} = \hat{\boldsymbol{\theta}}$, $\hat{\mathbf{J}}_{\infty c}(\boldsymbol{\theta}) = \hat{\mathbf{J}}(\boldsymbol{\theta})$ and $\mathbf{V}_{\infty U} = \hat{\mathbf{V}}[\hat{\mathbf{U}}(\boldsymbol{\theta})]$ under second moment matching with SRS. Thus, the covariance estimator $\hat{\mathbf{V}}_L(\hat{\boldsymbol{\theta}}_{gc})$ provides valid inferences on $\boldsymbol{\theta}$ for large number of subsamples, g, regardless of the subsample size, m.

To illustrate the calculation of the linearization inverse-sampling estimator $\hat{\mathbf{V}}_L(\hat{\boldsymbol{\theta}}_{gc})$, given by (5.17), consider the special case of a ratio $\theta_N = R$ with $u_k(\theta) = y_k - \theta x_k$. We have

$$\tilde{\mathbf{V}}_{jU}^* = \frac{N^2}{m}\left(1 - \frac{m}{N}\right)\frac{1}{m-1}\sum_{k \in s_j^*}\left[u_k(\theta) - \bar{u}_j^*(\theta)\right]^2, \tag{5.21}$$

where $\bar{u}_j^*(\theta) = \bar{y}_j^* - \theta \bar{x}_j^*$ and $(\bar{y}_j^*, \bar{x}_j^*)$ are the j-th subsample means. Further,

$$\hat{J}_{gc}(\theta) = \frac{N}{g}\sum_{j=1}^{g} \bar{x}_j^* = \hat{X}_g \tag{5.22}$$

and

$$\hat{U}_j^*(\theta) = N(\bar{y}_j^* - \theta \bar{x}_j^*). \tag{5.23}$$

It now follows from (5.21) – (5.23) that the CEE-based linearization estimator (5.17) is identical to the inverse-sampling linearization variance estimator (5.4).

Turning to linear regression with $\mathbf{u}_k(\boldsymbol{\theta}) = \mathbf{x}_k(y_k - \mathbf{x}_k^T\boldsymbol{\theta})$, we have

$$\tilde{\mathbf{V}}_{jU}^* = \frac{N^2}{m}\left(1 - \frac{m}{N}\right)\frac{1}{m-1} \times \sum_{k \in s_j^*}\left[\mathbf{u}_k(\boldsymbol{\theta}) - \bar{\mathbf{u}}_j^*(\boldsymbol{\theta})\right]\left[\mathbf{u}_k(\boldsymbol{\theta}) - \bar{\mathbf{u}}_j^*(\boldsymbol{\theta})\right]^T, \tag{5.24}$$

where $\bar{\mathbf{u}}_j^*(\boldsymbol{\theta}) = m^{-1}\sum_{k \in s_j^*} \mathbf{u}_k(\boldsymbol{\theta})$. Also,

$$\hat{\mathbf{J}}_{gc}(\boldsymbol{\theta}) = \frac{N}{g}\sum_{j=1}^{g}\frac{1}{m}\sum_{k \in s_j^*} \mathbf{x}_k\mathbf{x}_k^T$$

and

$$\hat{\mathbf{U}}_j^*(\boldsymbol{\theta}) = \frac{N}{m}\sum_{k \in s_j^*} \mathbf{x}_k(y_k - \mathbf{x}_k^T\boldsymbol{\theta}).$$

Finally, consider the case of logistic regression with $\mathbf{u}_k(\boldsymbol{\theta}) = \mathbf{x}_k(y_k - \mu_k(\boldsymbol{\theta}))$. In this case, $\tilde{\mathbf{V}}_{jU}^*$ is given by (5.24) with $\mathbf{u}_k(\boldsymbol{\theta}) = \mathbf{x}_k(y_k - \mu_k(\boldsymbol{\theta}))$. Also,

$$\hat{\mathbf{J}}_{gc}(\boldsymbol{\theta}) = \frac{N}{g}\sum_{j=1}^{g}\frac{1}{m}\sum_{k \in s_j^*} \mu_k(\boldsymbol{\theta})(1 - \mu_k(\boldsymbol{\theta}))\mathbf{x}_k\mathbf{x}_k^T,$$

and

$$\hat{\mathbf{U}}_j^*(\boldsymbol{\theta}) = \frac{N}{m} \sum_{k \in s_j^*} \mathbf{x}_k (y_k - \mu_k(\boldsymbol{\theta})).$$

It is important to note again that the estimator $\hat{\boldsymbol{\theta}}_{gc}$ and the associated covariance estimator $\hat{\mathbf{V}}_L(\hat{\boldsymbol{\theta}}_{gc})$ can be implemented from a microdata with data from g subsamples, each of size m. Neither the survey weights w_k nor the cluster identifiers are needed so that confidentiality of microdata may be preserved.

5.3 An Example

We now use a data set reported in Battese, Harter and Fuller (1988) to illustrate how the separate and combined estimating equations methods perform. The data were collected from $k = 12$ counties in north-central Iowa. The counties were divided into area segments and a sample of area segments was selected from each county. Here counties represent clusters and sample area segments within a county represent elements. The number of sample area segments (m_i) ranged from 1 to 5 giving a total of $n = \sum_{i=1}^{k} m_i = 37$ sample elements. For each sample element (i, j), Battese *et al.* (1988) gave the number of reported hectares of corn (y_{ij}) obtained by interviewing farm operators and the number of pixels classified as corn (x_{1ij}) and soybeans (x_{2ij}) obtained from remote sensing satellite readings ($j = 1, ..., m_i$; $i = 1, ..., k$). Data from one of the sample area segments were suspected to be erroneous and hence excluded from the analysis. Thus we have $n = 36$ observations (y_{ij}, x_{ij}).

For illustration, we treat the sample as if it was selected by the following two stage cluster sampling: (i) In the first stage, counties were selected with replacement and with probabilities proportional to the number of area segments M_i in the counties. (ii) In the second stage, sample area segments were selected by simple random sampling without replacement from each selected county. We consider two parameters: (i) population ratio $\theta = R = Y/X$, where Y and X are the population totals of y and x; (ii) census regression coefficient of y on x, $\boldsymbol{\theta} = \mathbf{B} = (\sum_{l \in U} \mathbf{x}_l \mathbf{x}_l^T)^{-1} (\sum_{l \in U} \mathbf{x}_l y_l)$, where $\mathbf{x}_l = (1, x_{1l}, x_{2l})^T$ and l denotes a population element.

For selected values of g, we generated g inverse-samples independently using the procedure for Case 2 in section 2.3. We then used the g subsamples to estimate R using the separate estimating equations (SEE) method and the combined estimating equations (CEE) method given in section 5. The corresponding variance estimates and the linearization variance estimates of the full-sample estimates $\hat{\boldsymbol{\theta}}$ were computed.

Table 1 reports the full-sample estimate $\hat{R}$, the SEE estimate $\hat{R}_g$, the CEE estimate $\hat{R}_{gc}$ and the corresponding variance estimates. It is clear from Table 1 that both CEE and SEE perform well in tracking the full-sample estimate $\hat{R}$ and the corresponding linearization full-sample variance estimate even for $g = 500$.

Table 2 reports the results for the regression coefficients $\mathbf{B} = (B_0, B_1, B_2)^T$. As g increases, both SEE and CEE seem to track the full-sample estimates $\hat{B}_1$ and $\hat{B}_2$, while SEE leads to slightly larger value for $\hat{B}_0$. However, the SEE variance estimates perform poorly, even for very large $g = 10{,}000$ in tracking the linearization full-sample variance estimates, with SEE value about one-half of the corresponding full-sample value for B_0 and B_1. On the other hand, the CEE variance estimates perform very well in tracking the full-sample variance estimates, confirming the theory.

Table 1
Estimation of Population Ratio R

		$g = 500$		$g = 1{,}000$		$g = 5{,}000$	
	Full-sample	CEE	SEE	CEE	SEE	CEE	SEE
Estimate	0.4096	0.4101	0.41	0.4096	0.4095	0.4095	0.4094
Variance Estimate $\times 10^{-4}$	1.9513	1.8769	1.8508	1.8482	1.8302	1.932	1.9178

Table 2
Estimation of Census Regression Parameters, B_0, B_1 and B_2

		$g = 500$		$g = 1{,}000$		$g = 10{,}000$	
	Full-sample	CEE	SEE	CEE	SEE	CEE	SEE
Est. of B_0	53.3588	49.9532	52.6649	53.5876	56.7143	53.2401	56.3196
Est. of B_1	0.3176	0.3251	0.318	0.3171	0.3086	0.3179	0.31
Est. of B_2	-0.1326	-0.1258	-0.1302	-0.133	-0.1378	-0.1324	-0.1377
B_0: Var. Est.	416.1609	457.5178	293.8789	407.3107	224.0846	437.961	251.395
B_1: Var. Est. $\times 10^{-3}$	2.1153	2.2925	1.164	1.9127	0.5354	2.2366	0.8882
B_2: Var. Est. $\times 10^{-3}$	2.7369	3.0352	2.4811	2.7226	2.3174	2.8028	2.3229

6. CONCLUDING REMARKS

In this paper we have presented some theory of inverse sampling. Efficiency of inverse sampling is increased by drawing repeated subsamples and then combining the results from the subsamples.

For estimating a total, we obtained conditions for the limiting inverse-sampling estimator to approach the full-sample estimator (Theorem 3) and for the limiting inverse-sampling variance estimator to approach the full-sample variance estimator (Theorem 4). For estimating complex parameters, we proposed a combined estimating equations (CEE) approach and demonstrated its advantages over separate estimating equations (SEE) approach (section 5).

We have studied inverse sampling algorithms for some sampling designs in section 2. But further work is needed to cover other sampling designs and also to avoid the limitations noted in section 2.

We are studying various extensions to include post-stratified full-sample estimators, analysis of categorical survey data, clustered survival data (Binder 1992) and longitudinal survey data.

ACKNOWLEDGEMENTS

The authors wish to thank the Associate Editor and the referees for constructive suggestions. This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

APPENDIX

Proofs of Theorems

Proof of theorem 1

Result 1 follows directly from (3.1) on noting that conditional on $s_0, \hat{\theta}_1^*, ..., \hat{\theta}_g^*$ are independent identically distributed (i.i.d.) bounded random variables.

Result 2 follows from the standard relationship between conditional and unconditional expectations:

$$E(\hat{\theta}_g) = E\left[E\left(\hat{\theta}_g \mid s_0\right)\right] = E\left[E\left(\hat{\theta}_1^* \mid s_0\right)\right] = E\left(\hat{\theta}_1^*\right).$$

Result 3 follows from the corresponding result for variances, and the conditional independence of the $\hat{\theta}_j^*$'s given s_0:

$$\begin{aligned}\mathrm{Var}\left(\hat{\theta}_g\right) &= \mathrm{Var}\left[E\left(\hat{\theta}_g \mid s_0\right)\right] + E\left[\mathrm{Var}\left(\hat{\theta}_g \mid s_0\right)\right] \\ &= \mathrm{Var}\left(\hat{\theta}_\infty\right) + \frac{1}{g} E\left[\mathrm{Var}\left(\hat{\theta}_1^* \mid s_0\right)\right].\end{aligned}$$

Result 4 follows directly from Result 3.

Proof of theorem 2

Theorem 2 follows from applying Results 3 of Theorem 1 with $g = 1$ to obtain

$$\mathrm{Var}(\hat{\theta}_\infty) = \mathrm{Var}(\hat{\theta}_1^*) - E\left[\mathrm{Var}(\hat{\theta}_1^* \mid s_0)\right],$$

and then substituting this expression for $\mathrm{Var}(\hat{\theta}_\infty)$ in Result 3 of Theorem 1 for general g.

Proof of theorem 3

We have

$$\hat{Y}_j^* = \sum_{i \in s_j^*} \frac{y_i}{\pi_i^*} = \sum_{i \in s_0} \frac{y_i I_{ij}^*(s_0)}{\pi_i^*},$$

where $I_{ij}^*(s_0)$ takes the value 1 if the i-th unit is included in the j-th subsample s_j^* and 0 otherwise, and π_i^* is the corresponding (unconditional) inclusion probability. Thus

$$\hat{Y}_\infty = E\left[\hat{Y}_1^* \mid s_0\right] = \sum_{i \in s_0} \frac{y_i \tilde{\pi}_i(s_0)}{\pi_i^*}.$$

This is equal to $\hat{Y} = \sum_{i \in s_0}(y_i/\pi_i)$, the H-T estimator for the original design, if and only if $\tilde{\pi}_i(s_0) = \tilde{\pi}_i = \pi_i^*/\pi_i$.

Proof of theorem 4

Conditional on s_0, it follows from (3.3) that $\hat{V}_{g,\mathrm{HT}}$ converges almost surely to

$$\hat{V}_{\infty,\mathrm{HT}} = E\left(\hat{V}_{1,\mathrm{HT}}^* \mid s_0\right) - \mathrm{Var}\left(\hat{Y}_1^* \mid s_0\right) \qquad (A.1)$$

as $g \to \infty$. Now, noting that $\tilde{\pi}_{il}(s_0) = \tilde{\pi}_{il} = \pi_{il}^*/\pi_{il}$, we get

$$\begin{aligned}E\left(\hat{V}_{1,\mathrm{HT}}^* \mid s_0\right) &= \sum_{i,l \in s_0} \sum \frac{\pi_{il}^* - \pi_i^* \pi_l^*}{\pi_i^* \pi_l^* \pi_{il}} \tilde{\pi}_{il} y_i y_l \\ &= \sum_{i,l \in s_0} \sum \left(\frac{\tilde{\pi}_{il}}{\pi_i^* \pi_l^*} - \frac{1}{\pi_{il}}\right) y_i y_l. \qquad (A.2)\end{aligned}$$

Further,

$$\begin{aligned}\mathrm{Var}\left(\hat{Y}_1^* \mid s_0\right) &= \sum_{i,l \in s_0} \sum \left(\tilde{\pi}_{il} - \tilde{\pi}_i \tilde{\pi}_l\right) \frac{y_i}{\pi_i^*} \frac{y_l}{\pi_l^*} \\ &= \sum_{i,l \in s_0} \sum \left(\frac{\tilde{\pi}_{il}}{\pi_i^* \pi_l^*} - \frac{1}{\pi_i \pi_l}\right) y_i y_l. \qquad (A.3)\end{aligned}$$

It now follows from (A.1) - (A.3) that $\hat{V}_{\infty,\mathrm{HT}} = \hat{V}_{\mathrm{HT}}$.

Proof of theorem 5

Conditional on s_0, it follows from (3.3) that

$$\hat{V}_{\infty,\mathrm{SYG}} = E\left(\hat{V}_{1,\mathrm{SYG}}^* \mid s_0\right) - \mathrm{Var}\left(\hat{Y}_1^* \mid s_0\right) \qquad (A.4)$$

where

$$\mathrm{Var}\left(\hat{Y}_1^* \mid s_0\right) = \sum_{i<l \in s_0} \sum \left(\tilde{\pi}_i \tilde{\pi}_l - \tilde{\pi}_{il}\right)\left(\frac{y_i}{\pi_i^*} - \frac{y_l}{\pi_l^*}\right)^2, \qquad (4.5)$$

provided the subsample size is also fixed (Cochran 1977, page 260). Further,

$$E\left(\hat{V}^{*}_{1,\mathrm{SYG}} \mid s_0\right) = \sum_{i<l\in s_0}\sum \frac{\left(\pi^{*}_i \pi^{*}_l - \pi^{*}_{il}\right)}{\pi_{il}}\left(\frac{y_i}{\pi^{*}_i} - \frac{y_l}{\pi^{*}_l}\right)^2. \quad \text{(A.6)}$$

It now follows that

$$\hat{V}_{\infty,\mathrm{SYG}} = \sum_{i<l\in s_0}\sum \frac{\pi_i \pi_l - \pi_{il}}{\pi_{il}} \tilde{\pi}_i \tilde{\pi}_l \left(\frac{y_i}{\pi_i \tilde{\pi}_i} - \frac{y_l}{\pi_l \tilde{\pi}_l}\right)^2, \quad \text{(A.7)}$$

Comparing (A.7) and (A.4) we see that $\hat{V}_{\infty,\mathrm{SYG}} \neq \hat{V}_{\mathrm{SYG}}$.

REFERENCES

BATTESE, G.E., HARTER, R.M. and FULLER W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*. 83, 28-36.

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

BINDER, D.A. (1992). Fitting Cox's proportional hazard models from survey data. *Biometrika*. 79, 139-147.

COCHRAN, W.G. (1977). *Sampling Techniques*. Third Edition; New York: John Wiley & Sons, Inc.

HINKINS, S., OH, H.L. and SCHEUREN, F. (1997). Inverse sampling design algorithms. *Survey Methodology*. 23, 11-21.

HOFFMAN, E.B., SEN, P.K. and WEINBERG, C.R. (2001). Within-cluster resampling. *Biometrika*. 88, 1121-34.

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

RAO., J.N.K., and SCOTT, A.J. (1992). A simple method for the analysis of clustered binary data. *Biometrics*. 48, 577-585.

RAO., J.N.K., and SCOTT, A.J. (1999). A simple method for analysing overdispersion in clustered Poisson data. *Statistics in Medicine*. 18, 1373-1385.

SKINNER, C.J., HOLT, D. and SMITH, T.M.F. (Eds.)(1989). *Analysis of Complex Surveys*. Chichester: Wiley.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Comment

JOHN L. ELTINGE[1]

1. OVERVIEW

Rao, Scott and Benhin (henceforth RSB), in conjunction with Hinkins, Oh and Scheuren (1997) (henceforth HOS), have produced a fascinating set of ideas and methods for inverse sampling. This discussion will highlight several related ideas and practical issues that the survey community is likely to encounter as it considers practical applications of inverse sampling. Section 2 notes some relationships between standard probability weights and the random weights implicitly constructed through repeated inverse sampling. Section 3 discusses two types of approximations that may arise in variance estimation from inverse sample data. Section 4 considers the practical operational simplifications that may result from inverse sampling in some cases. Section 5 discusses the use of inverse sample data with standard (simple random sample-based) graphical methods. Section 6 explores the potential benefits and limitations of inverse sampling in attempts to reduce identification risk in public-use datasets.

2. POINT ESTIMATION: INVERSE SAMPLING AS A FORM OF FILTERING

Borrowing some ideas from the sampling, signal processing and confidentiality literature (*e.g.*, Duncan and Pearson 1991), we can think of a point estimator as the result of multiple steps of "filtering" of observations from a population. For example, in construction of a standard Horvitz-Thompson estimator of a population total, a set of population values can be viewed as passing through two stages of filters corresponding, respectively, to the selection of sample units and to the inverse-probability weighting of those units. Similarly, the point estimator (4.1) in RSB may be viewed as the result of two stages of filtering, where the second stage now corresponds to weighting by a random factor determined by the number of times a given sample unit appears in the g repeated inverse samples. Under conditions, the filter weights in (4.1) converge to the inverse-probability weights in the Horvitz-Thompson point estimator as g increases. In this sense, we can view the point estimator (4.1) as an approximation to the Horvitz-Thompson estimator. Similar comments apply to the general nonlinear point estimators and general inverse samples considered in RSB.

In addition, single inverse sampling can be viewed as a special type of two-phase sampling in which the second-phase selection rates are proportional to the inverses of the first-phase sampling rates. This leads naturally to the question of whether standard ideas from two-phase sampling can lead to efficiency gains in either single or multiple inverse sampling. For example, recall that in standard two-phase sampling, one can often improve efficiency by using ratio or regression-based adjustments in conjunction with auxiliary variables X observed for all first-phase sampling units. See *e.g.*, Särndal, Swensson and Wretman (1992, Chapter 9). Similarly here, one could construct a public-use dataset consisting of a single or multiple inverse sample dataset accompanied by estimated totals (based on the full complex sample) for a vector of auxiliary variables X. Also, some additional auxiliary information would be required for consistent variance estimation. Given sufficiently strong auxiliary variables X, the resulting ratio or regression-based adjusted point estimators could help to improve the precision of inverse-sample-based analyses. This in turn could reduce the number of inverse subsamples required to ensure that the regression-adjusted multiple-inverse-sample point estimator has a variance that is sufficiently small.

More generally, in many complex-survey cases (outside of two-phase designs), standard weighted point estimators also go beyond direct use of inverse-probability weights to incorporate auxiliary information through, *e.g.*, ratio or regression adjustments. Also, in some cases, one reduces the numerical values of certain extreme probability weights, in an attempt to avoid problems with variance inflation induced by influential observations. See, *e.g.*, Zaslavsky, Schenker and Belin (2001). A natural question is whether one could modify the inverse sampling algorithm so that the inverse design is "tuned" to the adjusted weights rather than the direct inverse-probability weights. This would be of serious interest for cases in which adjusted-weight point estimators are expected to have a substantially smaller mean squared error than inverse-probability-weight point estimators. For cases in which this modified approach is advisable, it would be of interest to study corresponding ways in which to extend the RSB approach to variance estimation.

[1] John L. Eltinge, Office of Survey Methods Research, U.S. Bureau of Labor Statistics. E-mail: Eltinge_J@bls.gov.

3. APPROXIMATIONS EMPLOYED IN VARIANCE ESTIMATION AND INFERENCE

For some complex designs, HOS and RSB noted that exact extraction of a simple random sample may be impossible, or may lead to a very small inverse sample, which in turn requires compensation through the use of a very large g. Consequently, sections 2 and 4.3 of RSB consider approximate matching methods, and section 4.1 considers inverse sampling that may produce a design that is simpler than the original complex design, but is more complex than a simple random design.

In parallel with this, recall that some of the sampling literature considers variance estimators that are based on approximations to the true sample design. One example is variance estimation based on stratum collapse. See, *e.g.*, Rust and Kalton (1987) and references cited therein. In addition, Korn and Graubard (1995, sections 4.2 and 4.3) consider variance estimators that ignore the original primary-sample-unit-level clustering and treat secondary sample units as if they were primary sample units.

In some cases, these approaches may be problematic, while in other cases they may produce satisfactory variance estimators. For the latter cases, one could consider development of an inverse sample procedure based on the approximate "variance estimation design" rather than on the true sample design. Under that approach, it would be of special interest to consider the relative magnitudes of errors associated with, respectively, sampling under the original design, the approximation error in the "variance estimation design," and the additional error induced through use of a finite number of inverse samples.

4. OPERATIONAL SIMPLICITY

In principle, most point estimation, variance estimation and inference methods that have been developed for simple random sample data can be extended to handle complex sample data. However, the work required for such extensions is often nontrivial, and may discourage many potential analysts from making efficient use of the available data. In an informal sense, data analysts often appear to choose their analytic approaches based on a rough cost-benefit evaluation, in which they will focus on analyses that they believe will offer them most or all of the scientific insights available from the data, while not requiring an investment in analytic effort that they consider disproportionate to the potential scientific benefit. Statisticians and subject-area data analysts may often have different views regarding the relative costs and scientific benefits of a given analytic effort. In some cases, inverse sampling may help to ameliorate the effects of these differing views.

In particular, as indicated by RSB and HOS, an investment by a statistical agency in construction of inverse samples may lead to some reduction in the burden encountered by a given analyst. This investment may be especially worthwhile if both of the following conditions are satisfied.

(a) An analyst intends to carry out a large number of different analyses on a single survey dataset; lacks appropriate complex-survey software for many (or all) of the intended analyses; and perceives the programming of complex-survey procedures to require a major investment of effort.

(b) The additional computational steps required for point estimation (*e.g.*, the averaging carried out in the point estimators (3.1) or (4.1), or the combined estimating equation (5.14)) or variance estimation (*e.g.*, the variance estimators (3.3), (3.4), (5.18) or (5.20)) impose a relatively low incremental burden on the analyst, or can be absorbed into the analytic software in a form that is transparent to the analyst.

5. GRAPHICAL DISPLAYS

Hinkins *et al.* (1997, page 19) and Scheuren (1997) have noted the potential for application of inverse sampling to statistical graphics for complex survey data. For example, Scheuren (1997) noted that many methods of statistical graphics (*e.g.*, scatterplots) have been developed primarily for sets of independent and identically distributed observations. Direct application of these methods to complex survey data may produce misleading graphs, due to the effects of, *e.g.*, differential sampling rates or intracluster correlation. Since a given inverse sample is a simple random sample from the original population, the above mentioned problems would not arise when standard graphical methods were applied to data from a single inverse sample.

However, for inverse samples with small or moderate m, a scatterplot from a single inverse sample may not suffice for many purposes. An alternative approach would be to use several inverse samples in conjunction with local smoothing methods, *e.g.*, bivariate density estimation. For purposes of optimization, it may be useful to consider adjustment of some features of standard (simple random sample based) bivariate density estimators (*e.g.*, bandwidth) to account for unconditional correlation across the multiple inverse samples. Within this context, note that at a given point on the plane, a customary (simple random sample based) density estimator can be viewed as a solution to an estimating equation. Consequently, it would be of interest to study specific ways in which the RSB results on estimating equation methods may shed light on efficient approaches to bivariate density estimation based on inverse samples.

6. IDENTIFICATION RISK

As noted by RSB and HOS, a major potential attraction of inverse sampling is that it allows the computation of approximately design unbiased point estimators and variance estimators without explicit use of weights, stratum labels or cluster labels. This is of considerable practical interest in the preparation of public-use datasets because release of these types of design information can increase the risk that a sample unit can be identified by a data user. This in turn may constitute a violation of statistical agency pledges of respondent confidentiality. See, for example, de Waal and Willenborg (1997) and Chen and Keller-McNulty (1998) for detailed discussion of confidentiality issues associated with the release of weights.

In addition, in many household surveys in North America, strata and primary sample units are defined largely through geographical factors. For example, a primary sample unit in the U.S. is often a county or a group of contiguous counties. Release of nominally uninformative primary sample unit labels, accompanied by demographic and household-level observations Y, can lead to identification of the primary sample unit if the PSU-level aggregates of the observations Y vary in distinctive patterns that are publicly known. For example, a given county may have an unusual demographic profile, or may have a distinctive pattern of expenditures, *e.g.*, for natural gas or electricity.

For this reason, it would be of interest to evaluate the extent to which public release of multiple inverse samples may provide information that would allow a data user to reconstruct weights or PSU-level groupings that are informative. For instance, in keeping with comments by Mantel (2002), suppose that a given measured variable Y is reported on a continuous scale, and that for many responding units, the numerical value of Y is unique. Then (in keeping with the comments in section 2) matching of the reported Y values across a very large number g of multiple inverse samples would allow a data user to estimate the probability weights associated with a given respondent i. This in turn could lead back to the abovementioned identification problems considered by de Waal and Willenborg (1997) and Chen and Keller-McNulty (1998). For certain extreme cases, similar problems may arise with the identifiability of primary sample units. The extent to which these issues are of practical concern depend on the relative empirical magnitudes of various error sources (including error induced by the use of finite g), and would be of interest to study for specific agency cases.

ACKNOWLEDGEMENTS

The author thanks Van Parsons, Fritz Scheuren and Al Zarate for many useful discussions of inverse sampling and its possible use in reduction of identification risk. The views expressed here are those of the author and do not necessarily reflect the policies of the U.S. Bureau of Labor Statistics.

ADDITIONAL REFERENCES

CHEN, G., and KELLER-MCNULTY, S. (1998). Estimation of identification disclosure risk in microdata. *Journal of Official Statistics*. 14, 79-95.

DUNCAN, G.T., and PEARSON, R.W. (1991). Enhancing access to microdata while protecting confidentiality (with discussion). *Statistical Science*. 6, 219-239.

KORN, E.J., and GRAUBARD, B.I. (1995). Analysis of large health surveys: Accounting for the sampling design. *Journal of the Royal Statistical Society, Series A* 158, 263-295.

MANTEL, H. (2002). Floor discussion at Statistics Canada Symposium, November 8, 2002.

RUST, K., and KALTON, G. (1987). Strategies for collapsing strata for variance estimation. *Journal of Official Statistics*. 3, 69-81.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model-Assisted Survey Sampling*. New York: Springer-Verlag.

SCHEUREN, F.J. (1997). Personal communication.

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

ZASLAVSKY, A.M., SCHENKER, N. and BELIN, T.R. (2001). Downweighting influential cluters in surveys: Application to the 1990 Post Enumeration Survey. *Journal of the American Statistical Association*. 96, 858-869.

Comment

SUSAN HINKINS[1]

Rao, Scott, and Benhin (RSB) have done an excellent job of summarizing our results on inverting complex samples and they have moved the subject substantially further with an impressive body of theoretical results. Their paper develops valuable new insight for statisticians who wish to consider, at the design stage, the option of using resampling techniques in the analysis. In this way, invertible designs can be used. As the authors point out, there are still many interesting problems to be considered. We discuss here some specific points from the paper and also some of the open problems raised in our applied research into the employment of inverse sampling.

How to Use the Resulting Samples – The estimation of totals or means has the advantage that the combined and separate estimates are identical. Once one moves beyond "simple" estimation problems, there are many open questions as to the best use of the resulting samples, but combining the samples is a most reasonable approach. For a parameter such as a ratio, that is a function of totals, it would seem intuitive to calculate the best estimate of each total, and apply the function to the estimates and this is what we would recommend. In fact, because the ratio estimator is used in many situations, we did comment briefly on this in the 1997 HOS paper. However, RSB have made this point explicitly and in addition they have provided a coherent methodology for the estimation of variance from combined samples. This provides researchers with valuable tools for applying the inverse sampling techniques to a wider range of problems.

In Hinkins, Oh and Scheuren (1995), we considered the use of inverse sampling for the problem of calculating tests of independence from a 2×2 contingency table when the data come from a stratified sample. Contingency table analysis and regression analysis were both developed largely in the IID world and, therefore, adjustments are needed to use them in complex survey settings. We drew multiple simple random samples and calculated the simple Pearson chi-square test from the combined data. As the number of samples increases, the probability of rejecting the null hypothesis also increases, so one cannot take an arbitrarily large number of simple random samples. The problem was how to calibrate the tests, so that the desired level (*e.g.*, a 0.05 significance level for example) is achieved. Preliminary results indicated that one could determine the number of simple random samples to combine to achieve the desired level for the test, and using the Pearson chi-square on the combined samples compared well to the Fellegi (1980) methodology applied to the original stratified sample, while perhaps being more user friendly.

In work that Hinkins, Liu, and Scheuren presented at the 1998 Statistical Society of Canada Conference, simulation results were shown for regression fits to inverse samples from a complex design (stratified median balanced design). In this case, the original design selected 100 replicates; in each replicate, one observation was selected from each of six strata, so that the observations were median balanced (Liu 1999). The selection was with replacement across replicates. The inverse sample consisted of selecting one unit from each replicate. We looked at regression fits to individual inverse samples, and at the regression fit to the combination of several inverse samples. The population regression line had a slope of 0.842 and $R^2 = 0.64$. Using single inverse samples, the estimated slopes ranged from 0.70 to 1.13. Combining six inverse samples, the estimated slope was 0.845 with $R^2 = 0.64$.

Variance Estimation – The estimation of variance in the HOS 1997 formulation is an interesting problem because the samples are not unconditionally independent. In our 1997 paper we suggested for ratio estimates that if the combined sample is sufficiently large so that a Taylor Series approximation is acceptable, then the "usual" approximation to the variance for a ratio could be used. That is, the variance could be estimated using the approximation

$$\mathrm{Var}(\hat{R}) \doteq \frac{1}{\bar{X}^2}\mathrm{Var}(\bar{e}) \text{ where } R = \frac{Y}{X} \text{ and } e_i = y_i - Rx_i.$$

The estimated variance for the ratio estimate based on the combined samples can then be calculated in the "usual" manner as

$$\mathrm{var}(\hat{R}_c) = \mathrm{var}\left(\frac{\bar{y}_c}{\bar{x}_c}\right) = \frac{1}{\bar{x}_c^2}\mathrm{Var}(\bar{e}_c)$$

where $\bar{e}_c = (1/g)\sum_{j=1}^{g}\bar{e}_j$ and the mean in the j^{th} resample is $\bar{e}_j = \bar{x}_j - \hat{R}_c\,\bar{y}_j = \bar{x}_j - (\bar{y}_c/\bar{x}_c)\,\bar{y}_j$.

Using the estimate of variance generalized by the RSB equation (3.4) to estimate the variance of $\bar{e}_c$ results in the following variance estimate for the combined ratio estimate:

$$\mathrm{var}(\hat{R}_c) = \frac{1}{\bar{x}_c^2}\left(\frac{1}{g}\sum_j\left(\frac{1}{m}-\frac{1}{N}\right)s_{je}^2 - \frac{1}{g}\sum_j \bar{e}_j^2\right)$$

$$\text{where } s_{je}^2 = \frac{1}{m-1}\sum_{i=1}^{m}(e_{ji} - \bar{e}_j)^2.$$

[1] Susan Hinkins, Senior Statistician, National Opinion Research Center, 1122 South Fifth Ave, Bozeman, MT 5975. E-mail: hinkins-susan@norc.net.

As one would expect, this is the same variance estimator for the combined ratio estimate as Rao, Scott and Benhin construct using their estimating equations technique.

The use of the combined resampled samples for estimating regression coefficients is also addressed by RSB. They have developed a variance estimate, using the estimating equations technique, which appears to work well and further expands the possibilities for the use of resampling techniques. Their result also allows further research on the properties of the estimated variance in combined samples.

In the RSB regression example, it is not clear whether the variance estimate for B_0 has converged. The question of convergence of the estimates of error for nonlinear parameters is an interesting one, especially since these estimates are likely to be used to calibrate the process. (By calibration we mean the determination of when "enough" samples have been drawn, based on the desired use.) In the case of estimating a parameter, the only information available for calibration may be the comparison of the combined estimate, for example, to the original estimate from the complex design, and the comparison of their estimated standard errors. That is, while we may know that the variance will converge, only the estimates of variance are available for calibration.

Consider the following example where the inverse sample algorithm is used to invert a design with three strata and the minimum stratum sample size is two. Therefore, each re-sample is of size $m = 2$ and one would not expect fast convergence. Two ratios are estimated. Using 1,000 re-samples, the point estimates from the combined samples are within ±1.0% of the original estimates; using 10,000 re-samples the point estimates are within ±0.3% of the original estimates.

The estimates of the standard errors behave quite differently, however. For each parameter, Table 1 shows the ratio of the estimated standard error for the combined simple random samples to the estimated standard error of the original stratified estimate. The estimate of variance for the combined estimates was calculated using the method described above.

Table 1
Ratio of Estimated Standard Errors: Combined Estimate to Original Estimate

	Parameter	1,000 samples	10,000 samples
Totals	X_1	1.22	1.03
	Y_1	1.21	0.99
Ratio Estimate	$R_1 = Y_1/X_1$	1.07	1.07
Totals	X_2	1.02	0.95
	Y_2	0.94	0.93
Ratio Estimate	$R_2 = Y_2/X_2$	0.46	0.98

Using 1,000 re-sampled simple random samples, the estimated standard error of the combined estimate of X_1 is 22% larger than the estimated standard error for the original stratified estimate of X_1. Incidentally, this was not surprising to us. With 10,000 re-samples, the standard error for the combined estimate is reasonably close to that of the original stratified estimate. Similar results are seen for the estimate of Y_1. The standard error for the combined estimate of the ratio R_1 however converges more quickly, and appears to be relatively stable.

Consider the second set of variables. This time the standard errors for the combined estimates of the totals X_2 and Y_2 appear to have converged with only 1,000 samples. On the other hand, the standard error for the estimate of R_2 is severely under-estimated, as compared to the standard error of the original stratified estimate. An additional 9,000 draws, however, increases the estimated standard error for the ratio so that it is approximately equal to that of the original estimator.

Clearly, more analysis on the use of inverse sampling and the variance estimation for ratio and regression estimates is needed. Also, this example points out that the calibration of the inverse sample must consider all parameters of interest.

The remainder of the discussion considers two areas of interest where inverse sampling may be useful: providing public use data, and modeling or regression analysis. These two problems also illustrate two general types of data usage that may require different approaches to calibration.

Public-Use Data – The goal of using inverse sampling may be to provide public use data that will give substantially similar estimates as the estimates from the complex design, while permitting implementation of commonly available data analysis procedures using traditional computer software. If inference based on inverse-sample techniques can be demonstrated to be consistent with full complex-sample techniques, then data users with limited computer resources can perform select design-based analyses using mainstream statistical software. The results in the RSB paper expand the theory, providing conditions where the use of such resampling techniques is applicable.

A necessary feature in public-use data is the protection of confidentiality. For federal statistical agencies in the United States, public use files have been one of the responses to achieving the goal of "openness" (*e.g.*, Duncan, Jabine and deWolf 1993). However, the growing electronic availability of data of all sorts through the Internet and the advances in record linkage software can be seen to endanger this openness (*e.g.*, Doyle, Lane, Theeuwes and Zayatz 2001).

The goals of public use data can come into conflict when, for example, the information on the nature of the sample selection must be provided, implicitly or explicitly, for the calculation of design-based variances, but this information significantly increases the likelihood of

identifying an individual. In many surveys, geographical location plays an important part in the sampling, but the finer details of the geographical sampling structures cannot be released with the data without endangering the confidentiality of the individuals. If the geographical sampling structures are deleted to maintain confidentiality, then the data become difficult to analyze using the standard design-based methods. In this case, the use of inverse sampling would allow the release of data without the finer details of the geographical structures, for example, while still allowing analysis using the standard methods.

For example, the US National Health Interview Survey (NHIS) uses state-level stratification and selects counties and metropolitan areas for the sample. A public use file is released for the NHIS data in a form where the complex sample structure is simplified to that of a stratified design with two PSUs imbedded within each stratum. The original design strata and PSUs were masked in part using some of the techniques discussed in Eltinge (1999) and Parsons and Eltinge (1999). This masked "2 PSUs per stratum" design can be used to calculate variances. We investigated the NHIS design to see if inverse sampling was applicable for providing public use data (Hinkins and Scheuren 2001) and we found that it was not possible to invert the design down to the level of detail that was useful to data analysts. We still believe that inverse sampling can be an attractive option for providing public use data sets, when the design is invertible. It is not necessarily a viable option, however, unless its use is anticipated in the original design, so that invertible designs are used.

Another possible use of inverse sampling should be mentioned with respect to this example. For analytical domains covering most of the strata, the variance estimators from the NHIS public use data will be stable, *i.e.*, the estimators have large associated degrees of freedom. But for subpopulations that are less geographically dispersed, that cover few strata, the resulting degrees of freedom may be very small, and the variance estimate may be quite unstable. In such instances, it may be possible to produce a more stable variance estimator by drawing many, many samples from the public use design. In this case, rather than providing public use data, the inverse sampling might be used as a "black box" variance calculator that would provide more stable variance estimators for rare items in the population.

Modeling and Graphical Applications – Inverse sampling can be used to provide data in a form that allows greater analysis potential. This may be particularly valuable when there are multiple uses for the data. A natural example grows out of our initial proposal for using a resampling approach for the Statistics of Income (SOI) stratified samples of corporate tax returns. The underlying population is highly skewed (a relatively few large units accounting for a large percentage of the total value) and in order to provide efficient estimates of annual totals, a highly stratified sample design is used. However for economists, another important use of the data is modeling economic activity and developing tax models, which is not the same problem as calculating a finite population regression estimate.

Another such example, from EPA, is a large stratified sample of US lakes from which water chemistry measurements were made in order to provide background measurements relating to acid rain. These data were also of great interest to biologists who were interested in modeling certain aspects of the chemical and physical relationships.

Interpreting regression models in finite population sampling can be confusing. There are many well thought-out approaches to regression in a complex sample setting, but the simplified rule of thumb is that you generally can't ignore the design structure (for example the sample weights.) To analysts interested in modeling the underlying parametric structure, this can seem counter-intuitive. And if the design is ignored, one can get the wrong answer unless either there are no missing regressors or the design is not confounded with regressors (both unlikely in our experience for complex designs). A simple random sample satisfies the second requirement. In the case of the SOI sample, if economists were interested in modeling the structure of the small to medium corporations, for example, then fairly large simple random samples could be generated from the stratified design. And a combination of multiple draws might provide a reasonable data base.

Finally, the use of graphical techniques in modeling and regression analysis is very important for understanding how a variable depends on other predictor variables. Even in the simple problem with one or two predictors of a dependent variable, the graphical display of relationships using weighted sample data is difficult. The analysis of residuals and the detection of outliers are more difficult with weighted data. Graphing is a powerful tool for extracting information from data. This would seem to be an area where the use of inverse sampling specifically for producing simple random samples should be considered.

As RSB rightly note in their conclusions, there are still many opportunities for further research and analysis. Their paper makes significant steps in advancing the theory and the application potential for the use of resampling procedures, opening doors to more opportunities.

ADDITIONAL REFERENCES

DOYLE, P., LANE, J., THEEUWES, J. and ZAYATZ, L. (2001). *Confidentiality, Disclosure and Data Access*. North-Holland: New York.

DUNCAN, G., JABINE, T. and DEWOLF, V. (1993). *Private Lives and Public Policies*. National Academy Press: Washington.

ELTINGE, J.L. (1999). Use of stratum mixing to reduce primary-unit-level identification risk in public-use survey datasets. *Proceedings of the 1999 Federal Committee on Statistical Methodology Research Conference.*

FELLEGI, I. (1980). Approximate tests of independence and goodness of fit based on multistage samples. *Journal of the American Statistical Association*. 75, 261-268. See also Scheuren, F. (1972), Topics in Multivariate Finite Population Sampling and Data Analysis: George Washington University Doctoral Dissertation.

HINKINS, S., and SCHEUREN, F. (2001). *Increasing Public Accessibility to National Health Interview Survey Data (NHIS) Using Inverse Sampling*. Report prepared for NCHS under a Professional Services Contract.

HINKINS, S., OH, H.L. and SCHEUREN, F. (1995). Using an inverse sampling algorithm for tests of independence based on stratified samples. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

HINKINS, S., PARSONS, V. and SCHEUREN, F. (2000). Increasing Public Accessibility to Complex Survey Data by Using Inverse Sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

HINKINS, S., LIU, Y. and SCHEUREN, F. (1998). Presentation at the Annual Statistical Society of Canada Meeting in June 1998.

LIU, Y. (1999). Balanced Sampling Design: An Improvement over the Classical Sampling Design. Ph.D Dissertation. The George Washington University.

MULROW, J., and SCHEUREN, F. (1998). The Confidentiality Beasties. *Turning Administrative Systems into Information Systems*. Internal Revenue Service.

PARSONS, V.L., and ELTINGE, J.L. (1999). Stratum partition, collapse and mixing in construction of balanced repeated replication variance estimators. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Response From the Authors

1. INTRODUCTION

We thank the discussants, John Eltinge and Susan Hinkins, for their insightful comments and for suggesting some topics for further research on inverse sampling. In our rejoinder, we will attempt to address some of the issues raised by the discussants.

Our research on inverse sampling was motivated by the pioneering work of Hinkins, Oh and Scheuren (1997) (henceforth HOS). The latter authors developed several inverse sampling algorithms and provided some applications. They also noted the potential of inverse sampling in providing public-use microdata files, consisting of multiple simple random subsamples, that can be used to make valid inferences, such as regression and categorical data analysis, and to develop graphical displays of the data. The main contributions of our article is to provide some theoretical support (Theorems 1 – 5) and to develop the combined estimating equations (CEE) approach (section 5) to handle a variety of analyses of the data, such as linear and logistic regression, even when the subsample sizes are small. We have developed a linearization inverse-sampling variance estimator (equations (5.17) and (5.20)) that can be computed from the microdata file, and provided conditions for its convergence to the full-sample linearization variance estimator as the number of subsamples, g tends to ∞.

(i) Point estimation of a total

In the context of estimating a total $\theta = Y$, we proposed the inverse-sampling estimator $\hat{Y}_g$ given by (4.1) and showed that as $g \to \infty$, it converges to the full-sample Horvitz-Thompson estimator under the condition $\tilde{\pi}_i(s_0) = \tilde{\pi}_i$ for all $s_0 \supset i$ (see Theorem 3). Eltinge raised the important issue of improving the efficiency of $\hat{Y}_g$, for a given g. To this end, he suggested that single inverse sampling may be viewed as special type of two-phase sampling, and that using this analogy one could implement ratio or regression-based inverse-sampling estimators by constructing a public-use data set consisting of g subsamples, $\{(y_i, \boldsymbol{x}_i); i \in s_j^*\}, j = 1, ..., g$, supplemented by the full-sample estimated totals $\hat{\boldsymbol{X}}$ for a vector of auxiliary variables, $\boldsymbol{x}$. For example, a ratio inverse-sampling estimator is given by $\hat{Y}_{rg} = (\hat{Y}_g / \hat{X}_g)\hat{X}$, where $\hat{X}_g$ is the inverse-sampling estimator of the total X. Eltinge remarked that some additional auxiliary information may be required for variance estimation. It would be useful to pursue Eltinge's suggestions; one of us (E. Benhin) is looking into variance estimation. Benhin is also studying the analogues of full-sample calibration (or generalized regression) estimators constructed from multiple inverse samples (subsamples).

Eltinge also noted that in some cases the full-sample weights are adjusted to avoid problems with variance inflation induced by influential observations. He raised the question whether it is possible to modify the inverse sampling algorithms such that the resulting inverse-sampling estimator, say $\tilde{Y}_g$, converges to the adjusted-weight full-sample estimator, say, $\tilde{Y}$, as $g \to \infty$. This appears to be a challenging problem, but it may be possible to achieve approximate solutions.

(ii) Nonlinear parameters

In section 3 we considered a "separate" inverse sampling estimator, $\hat{\theta}_g$, of a nonlinear parameter θ, such as a ratio of totals $\theta = Y/X = R$, and noted that $\hat{\theta}_g$ can lead to large bias if the subsample size, m, is small. This is due to the fact that the bias of $\hat{\theta}_g$ is of the order m^{-1}. In her discussion, Hinkins noted that HOS were in fact aware of this problem and that HOS commented briefly on estimating the ratio R (page 18 of HOS). In particular, HOS suggested the estimation of the numerator Y and the denominator X separately, leading to the "combined" inverse-sampling estimator, $\hat{R}_{gc} = \hat{Y}_g / \hat{X}_g$, which follows as a special case of our CEE approach (see section 5.2). In section 5.1, we have spelled out the combined approach of HOS for the ratio R, at the suggestion of the Associate Editor, Fritz Scheuren.

(iii) Approximate variance estimator

Eltinge noted that approximate full-sample variance estimators, such as those based on stratum collapse, have been proposed in the literature and that it may be possible to develop inverse sampling procedures based on the approximate "variance estimation design" rather than the original sampling design. Such procedures may lead to larger subsample sizes, m. For example, in the case of stratified two-stage sampling with two clusters per stratum, we have $m = 2$ and m can be increased by stratum collapsing. This in turn may require a smaller number of subsamples, g, compared to the number of subsamples for the original design. Alternatively, for a given g, we may be able to obtain a more stable variance estimator, provided the full-sample approximate variance estimator is deemed to be satisfactory.

For PPS sampling without replacement, practitioners often assume that the sampling was with replacement to estimate the variance. In this case, HOS noted that "an inverse algorithm would exist to the same order of approximation as was being assumed to estimate variances" (page 16 of HOS).

(iv) **Number of subsamples**

The stability of the inverse-sampling variance estimator depends on the number of subsample, g, drawn from the full-sample and the function (or parameter) being estimated. For smaller g, the variance estimator can even take negative values. Also, when m is very small (as in the case of stratified two-stage sampling with two clusters per stratum), we will need a very large g to obtain a stable inverse-sampling variance estimator. We can increase m either by the approximate methods noted in (iii) or by drawing stratified random subsamples, provided confidentiality requirements or other considerations do not preclude the use of stratified subsamples.

Hinkins noted that the number of subsamples, g, may be determined by "calibrating" the inverse-sampling estimates and variance estimates to the corresponding full-sample values, and that the resulting g might vary significantly across parameters of interest. To illustrate the latter point, Hinkins studied the case of three strata and minimum stratum sample size of two, and computed the ratio, r, of the inverse-sampling variance estimator to the full-sample variance estimator for two ratios $R_1 = Y_1/X_1$ and $R_2 = Y_2/X_2$. Hinkins showed that the use of $g = 1{,}000$ subsamples leads to poor calibration for $R_2 (r = 0.46$ compare to $r = 0.98$ with $g = 10{,}000)$. This result is somewhat surprising, but it could be attributed to the instability of the inverse-sampling variance estimator with subsample $m = 2$. Hinkins noted that the inverse-sampling variance estimator for the intercept term B_0 in our Table 2 (denote CEE) may be behaving somewhat erratically as g increases. We agree with her, but it is difficult to address the question of convergence for nonlinear parameters such as B_0. Clearly, we need more work on the choice of g for variance estimation under inverse sampling. Fritz Scheuren noted in private correspondence that "the data user does know, however, what the main users are going to do, so g can be chosen with the important parameters in mind. But, of course, not all".

(v) **Analysis of survey data**

Computations of valid standard errors of parameter estimators from a full-sample microdata set may not be feasible in the context of stratified multistage sampling without the identification of clusters and strata on the data file. Even when the necessary information for standard error calculations is available on the data set, an analyst may lack appropriate complete-survey software for many (or all) of the intended analyses, as noted by Eltinge. On the other hand, valid standard errors may be obtained via the CEE approach using microdata files containing multiple simple random subsamples without the need for survey weights, clusters identifiers, *etc.* Moreover, as noted by Eltinge, the additional computational steps for implementing the CEE approach "impose a relatively low incremental burden on the analyst, or can be absorbed into the analytic software that is transparent to the analyst". However, we need further work on providing the necessary enhancements to standard software in order to implement the CEE method in practice.

Hinkins, Oh and Scheuren (1995) combined the subsamples to test independence in a 2×2 contingency table. Their Pearson chi-squared statistic is of the form

$$X^2 = (gm) \sum_{i=1}^{2} \sum_{j=1}^{2} (\hat{P}_{ijg} - \hat{P}_{i+g} \hat{P}_{+jg})^2 / (\hat{P}_{i+g} \hat{P}_{+jg}),$$

where $\hat{P}_{ijg}$ is the inverse-sampling combined estimator of the (i,j)-th cell proportion P_{ij} calculated from g subsamples each of size m, and $\hat{P}_{i+g} = \sum_j \hat{P}_{ijg}$, $\hat{P}_{+jg} = \sum_i \hat{P}_{ijg}$. It is clear from the form of X^2 that it increases with g so that the probability of rejecting the null hypothesis also increases with g. Hinkins, Oh and Scheuren (1995) noted that it may be possible to determine the number of subsampling, g, to combine to achieve the desired test level using X^2. This idea looks interesting, but actual implementation of the method needs further study, especially for testing hypotheses in multi-way tables. Instead of using this approach, it is possible to develop first- and second-order Rao-Scott corrections to X^2 by using the multiple subsamples to implement Rao and Scott (1984) corrections, based on the concept of design effects. These adjusted X^2 will be valid for any g. Benhin is currently studying the Rao-Scott corrections in the context of inverse sampling. As $g \to \infty$, the corrected X^2 will converge to the Rao-Scott adjusted X^2 based on the full-sample.

(vi) **Graphical displays and modeling**

Direct application of standard methods for statistical graphics and modeling to complex survey data may produce misleading graphs and models, as noted by Eltinge, due to the effects of clustering, unequal weights, stratification and other features of the survey data. On the other hand, it is appropriate to apply standard methods to data from a single inverse sample (or subsample), provided the subsample is simple random sample unconditionally. However, the subsample size, m, is typically small and hence the subsample data set is not informative for graphical displays or modeling. The size of the data set may be increased to gm by combining the g subsamples, but the application of standard methods (*e.g.*, scatter plots) to the combined data set can produce misleading displays and inferences because the subsamples are unconditionally correlated. Eltinge made some useful suggestions on accounting for the unconditional correlation in the context of bivariate density estimation, but much work remains to be done in the area of statistical graphics and modeling using multiple inverse samples.

(vii) **Confidentiality of microdata**

As noted by Eltinge, a major potential attraction of inverse sampling is that it allows the calculation of point estimators, standard errors, *etc.* from the microdata file, consisting of multiple subsamples, without the knowledge of weights, cluster labels or stratum labels. This feature

allows the reduction of identification risk induced by the knowledge of cluster labels etc. It could be a challenging task to evaluate the extent to which the data file of multiple subsamples allows data users to reconstruct weights or cluster labels. Note that the characteristic values reported on the data file of inverse samples are real in the sense of corresponding to the values in the full sample.

If the full sample is a PPS cluster sample and the subsamples are obtained by selecting one element from each cluster, then cluster identification may be avoided by first randomly permuting the data vectors within each subsample and then reporting the permuted subsamples. The CEE approach is invariant to permutations of data vectors within each subsample.

It should be noted that the confidentiality protection provided by the data set with multiple subsamples is never more than the protection provided by a simple random full sample. Various methods have been proposed in the literature for limiting disclosure in microdata obtained from simple random sampling, such as microdata masking (see *e.g.*, Cox 1994). We can use similar methods on the data set with multiple subsamples, if necessary. Raghunathan, Reiter and Rubin (2002) proposed multiple imputation for statistical disclosure limitation in the context of simple random sampling. The basic idea behind their proposal is to simulate multiple copies of the population by imputing for the nonsampled values using an imputation model based on auxiliary variables available for all the units in the population and then releasing a random sample from each of the synthetic populations. They used a parametric model-based approach and an approximate Bayesian bootstrap method for imputing the nonsampled values. The parametric approach protects confidentiality more effectively since the imputed values do not contain observed records, unlike the approximate Bayesian bootstrap, but it is far more susceptible to misspecifications of the imputation models. Note that the Raghunathan *et al.* (2002) method is fundamentally different from our method for complex full-samples. However, it is interesting to note that the variance estimator of Raghunathan et al. is given by the variance between the imputed data estimators **minus** the average of the imputed data variance estimators, whereas our variance estimator (3.3) is given by the average of the subsample variance estimators **minus** the variance between the subsample estimators. In the case of multiple imputation for missing data, the variance estimator is given by the average of the imputed data variance estimators **plus** the variance between the imputed data estimators, treating the imputed values as the true values.

(viii) Concluding remarks

As noted by Hinkins, inverse sampling is not necessarily a viable operation unless its use is anticipated at the full-sample design stage to permit the use of invertible designs. Currently, we do not have inverse sampling procedures for several commonly used full-sample designs. For example, consider single stage cluster sampling with probability proportional to a measure of cluster size $\tilde{M}_i$, not necessarily equal to the actual cluster size M_i. In this case, we cannot apply the algorithm in Case 3 of section 2 to get a simple random subsample.

Further work is clearly needed on developing suitable algorithms to achieve exact matching or at least approximate matching with simple random sampling or stratified random sampling. As Fritz Scheuren noted in private communication, "this stuff is fun, but lots of fence to paint yet".

We thank the Associate Editor, Fritz Scheuren, for his interesting observations on our rejoinder.

ADDITIONAL REFERENCES

COX, L.H. (1994). Matrix matching methods for disclosure limitation in microdata. *Survey Methodology*. 20, 165-169.

RAGHUNATHAN, T.E., REITER, J.P. and RUBIN, D.B. (2002). Multiple imputation for statistical disclosure limitation. Technical report, Department of Biostatistics, University of Michigan, Ann Arbor.

Survey Methodology, December 2003
Vol. 29, No. 2, pp. 129-138
Statistics Canada

The Accuracy and Coverage Evaluation: Theory and Design

HOWARD HOGAN[1]

ABSTRACT

This paper discusses both the general question of designing a post-enumeration survey, and how these general questions were addressed in the U.S. Census Bureau's coverage measurement planned as part of Census 2000. It relates the basic concepts of the Dual System Estimator to questions of the definition and measurement of correct enumerations, the measurement of census omissions, operational independence, reporting of residence, and the role of after-matching reinterview. It discusses estimation issues such as the treatment of movers, missing data, and synthetic estimation of local corrected population size. It also discusses where the design failed in Census 2000.

KEY WORDS: Dual system estimation; Census adjustment; Undercount.

1. INTRODUCTION

The U.S. Census Bureau attempted to correct the initial Census 2000 population figures for measured net undercount (U.S. Census Bureau 2000.) This correction was to be based on the Accuracy and Coverage Evaluation (A.C.E.). The A.C.E. is a post-enumeration survey based on the dual system estimator (DSE). Although seemingly well designed and well executed, the initial A.C.E. production estimates were badly flawed. The A.C.E. produced an estimate of 3.3 million net undercount (378,000 s.e.). This contrasts sharply with the current demographic analysis estimate of only 340 thousand (Robinson 2001) as well as a later revised survey estimate of a 1.3 million overcount (542,000 s.e.) (U.S. Census Bureau 2003).

This paper discusses both the general question of designing a post-enumeration survey (PES), and how these general questions were addressed in the U.S. Census Bureau's plans for the A.C.E. Where applicable, it discusses where the assumptions underlying the design failed in 2000. Throughout, I will use the terms DSE and PES when a general question is discussed and A.C.E. for specific details of the U.S. 2000 design. The next section defines the dual system model as applied to census coverage measurement. Section 3 discusses the definition and measurement of census correct and erroneous enumerations. Section 4 presents the issues in defining and measuring omissions. Section 5 deals with small area estimation. The paper ends with a discussion of some of the problems encountered in implementing the A.C.E together with some concluding remarks.

2. THE DUAL SYSTEM ESTIMATION MODEL

The use of the dual system model is well known either for measuring the completeness of vital events registration (Sekar and Deming 1949; Marks, Seltzer and Krotki 1974) or for use in measuring coverage errors in census data (Marks 1979; Wolter 1986; U.S. Bureau of the Census 1985.) Application of the dual system model in the context of the 1990 Census, including the issue of census adjustment, is documented in Hogan (1992, 1993.)

The standard Petersen (1896), Sekar-Deming or dual system estimator (DSE) can be expressed as:

$$\hat{N}_{++} = N_{+1}\left(N_{1+}/N_{11}\right) \qquad (1)$$

where

- N_{11} is the number of people counted in both the census and the survey,
- N_{+1} is the number of people correctly counted in the census,
- N_{1+} is the number of people counted in the survey, and
- N_{++} is the total number of people.

That is, the total population is estimated by the number captured in the census multiplied by the ratio of those in the survey to those in both systems (*i.e.*, the inverse of the coverage rate of the census, as measured by the survey).

The DSE will yield a direct estimate of the population of class j, as well as any sum of classes. The class j might be the household population of a state, of a district, of an ethnic group, or perhaps of an ethnic group within a state.

Requirements for estimating small or local populations, for example, age by sex, by race, by town, often far exceed the capacity of even a very large sample. To meet this need, the DSE is combined with a synthetic assumption to produce estimates for areas of geography smaller than that defined by the domain j. The synthetic estimator assumes that a proportion or ratio measured at an aggregate level applies equally to all sub-groupings (Gonzalez 1973; Gonzalez and Hoza 1978.) Using a synthetic assumption, we write

$$\hat{N}^{s}_{jkh} = CCF_j C_{jkh} \qquad (2)$$

[1] Howard Hogan, Chief, Economic Statistical Methods and Programming Division, Census Bureau, Washington, D.C. 20233.

$$CCF_j = \frac{\hat{N}_j}{C_j} \quad (3)$$

where,

$\hat{N}^s_{jkh}$ is the estimated population in domain j, available at the level of geography k and demographic subclass h.

CCF_j is the net coverage correction factor

$\hat{N}_j$ is the DSE for domain j

C_{jkh} is the measure (usually census count) of the population in domain j available at the level of geography k and demographic subclass h, and

$$C_j = \sum_k \sum_h C_{jkh}. \quad (4)$$

C_j need not equal the number of people correctly included in the census (N_{+1}). N_{+1} is estimated from sample data and is not available for all small areas. C is normally the census count, including imputations and erroneous inclusions (duplicates, *etc.*).

Summing over group j and subclass h yields a measured population for the given geographic area k, we have

$$\hat{N}^s_k = \sum_j \sum_h CCF_j C_{jkh}. \quad (5)$$

For example, j may define all 0-17 year-old Asians in owner-occupied housing units while k may define Orange County, California, and h may define 11-year-old girls.

While this produces a small-area and small-group estimate, this calculation can generate fractions. The typical user of census data prefers whole person records. The U.S. Census uses controlled rounding and person record imputation to create integer number of person records for ease of tabulation and data acceptance.

3. MEASURING CORRECT ENUMERATIONS

3.1 Defining and Measuring Correct and Erroneous Enumerations

The first step in operationalizing Equation 1 is to define and estimate the set of individuals "correctly" in the census. In this context "correctly" has four dimensions:

1. Appropriateness
2. Uniqueness
3. Completeness
4. Geographic correctness

"Appropriateness" means that the person should be included in the census. People who die before or who were born after the census reference date (April 1 in the U.S.) are not part of the population (universe) to be measured. Similarly, records that refer to fictitious "people," tourists, or animals are out-of-scope.

"Uniqueness" refers to the fact that we wish to measure the number of people included in the census, not the number of census records. If more than one record refers to a single person, the count of records must be reduced for purposes of the DSE.

"Completeness" means that the census record must be sufficient to identify a single person. If it lacks sufficient identifying information, we cannot determine whether the person was appropriately and uniquely included in the census, nor can we determine whether he or she was also included in the survey.

Although completeness is necessary for the DSE, the census count includes imputations and other incomplete enumerations. Census operations normally have a requirement for a "data-defined person record." In Census 2000, the requirement was two characteristics where name counts as a characteristic. The name field must have at least three characters in the first and last name fields combined. The characteristics that are included in the counting are relationship to the householder, sex, race, Hispanic origin, and either age or year of birth. (Childers 2001)

When a record does not meet these requirements census processing substitutes (imputes) a data-defined record. Since the census processing identifies all these whole-person imputations, the quantities are known and need not be estimated. Traditionally, the number of whole person imputations is denoted by II, for "insufficient information."

Additionally, there are person records that are acceptable for census processing but insufficient for use in the DSE. This group includes records with reasonably complete data but without a person's name. Accurate matching or additional interviewing is not possible for these cases. For A.C.E. 2000, the definition for "sufficient information for matching" was complete name and two characteristics. (Childers 2001)

"Geographic correctness" means that people are included in the census where they should be included. Enumerations outside that defined search area (or areas) are counted in the census but not correctly included in the census. This area must be searched during the matching process as well as searched for census duplicates. As the number of addresses in the search area increases, the complexity of matching increases and the chance of matching error grows. This increased complexity and possible levels of error will affect both the matching between the survey and the census and the search for census duplicates. The more addresses that must be searched, the more likely a true match will be missed. Equally importantly, the chance of a false match increases. For example, the chance of finding two people with similar names and ages living in the same block is small. The chances of finding two such people in a large city is considerable.

Two dimensions must be defined to operationalize a search area: (1) correct location and (2) the search area around the correct location.

The "correct location" defines where, under the DSE residence rules, the person should be included in the census. These rules may differ from the rules used in the census. The only requirement is that the location be precisely defined and consistently applied during PES processing. More than one location may be defined as correct so long as the rule is consistently applied. However, usually only one location is defined as correct. This was the rule in the A.C.E.

In the 1990 PES and 2000 A.C.E. the Census Bureau adopted the following rule:

> The person is correctly included in the census if he or she is included at the location where the person considers, at the time of the survey interview, to have been his or her usual residence as of April 1.

This definition generally follows the census rules. However, it makes an explicit allowance for the fact that the concept of "usual residence" is somewhat subjective. Because of this subjectivity, where the person considers his/her usual (April 1) residence may have changed by the time of the survey interview. This, by itself, does not bias the DSE. However, it does require consistent reporting of the "correct location."

The second dimension of geographic correctness is the area of search around the correct location, *i.e.*, the search area. The concept of a search area is to accommodate errors in either the census or survey assignment of residents to a particular geography. It has the effect of lowering the variance and can, in some circumstances, lower the bias as well.

The A.C.E. used the following definition:

> A person was correctly enumerated if the person was counted in the block cluster containing his/her usual residence; or if he/she was included by the census in the housing unit where he/she usually resides, and the housing unit was included in a block adjacent to the correct block cluster.

An important part of this design is that enumerations of people in the "wrong" location are to be classified as erroneous, whether or not the people are also enumerated in the correct location. Thus a person counted only once, but in the wrong location, should be measured, on average, as contributing one erroneous enumeration (in the wrong location) while being missed (one omission) in the correct location. This approach obviates the need to search widely for possible duplicates, but does require that the field interview determine a unique correct location for each person.

The definition of "correctly included" does not depend on the correctness of classification j. For example, if a person was really 19 years-old, but was counted in the census as 17, he/she is still considered as correctly included. This is discussed in section 5.2.

To estimate the number of people correctly included in the census, one must take a sample of all data-defined census enumerations. This sample is called the enumeration (or E) sample. Census whole-person imputations (Π's) are not part of the E-sample frame.

To maximize correlation with the population sample (see below), the A.C.E. first defines a set of sample areas. These are either a single block or a group of contiguous blocks and are known as block clusters. If a block is sampled, all census records coded to that block, even incorrectly, fall into sample. If the block contains many census housing unit records it may be subsampled.

The records in the E-sample will be checked for completeness. Only records that meet the minimum completeness requirement can be considered as correctly enumerated in the census. Records are then searched throughout the search area to see if the person was counted more than once within the sample block (uniqueness). Duplicate search is done using computer-assisted clerical matching. If more than one record is found, the extra records are coded as duplicates.

Appropriateness and geographic location cannot be determined from the census enumeration alone, but require additional interviewing. If interviewing locates a member of the household, or an acceptable respondent who can confirm the person's existence and that the person had his/her usual residence there on April 1, the enumeration is accepted as correct.

If the respondent reports that the person did not live in the block or search area on April 1, the enumeration is excluded from the correct enumerations. This can occur when the person responded to the census but moved before April 1; the person moved in after April 1 but was enumerated by the census nonresponse follow up operation; or when a parent incorrectly reports a college student as living at home.

The interviewers may determine that the person never existed or was never associated with the block. These records are considered erroneous. It is difficult in some cases to prove that a "person" was not real, especially in a large block. The A.C.E. required the interviewers to find at least three knowledgeable respondents before coding a record as fictitious. However, since the person might have lived somewhere else in the block, it can be difficult in some situations to code the record fictitious.

An important source of error arises from the need to accept proxy responses to verify many enumerations. If the proxy reports a different "correct" residence than the person himself would, an enumeration could be miscoded, since the requirement of a unique "correct" residence would be violated. The A.C.E. used proxy interviews for households that moved between the time of the census and the time of the A.C.E. interviews. Even within a household, different members may hold different views of a person's "correct" residence on Census Day. Proxy respondents, both household and non-household, were responsible for many

of the errors in reporting residence in the A.C.E. and thus, the underestimation of census error.

After missing-data estimation and sample weighting, we can estimate the number of people correctly counted in the census as

$$N_{+1} = (C - II)\frac{CE}{N_e}. \quad (6)$$

Where

C = Census total records, including imputed, duplicate, fictitious, *etc.* (the Census count),

II = number of whole-person census imputations,

CE = weighted estimate of appropriate, unique, complete and correct enumerations,

N_e = weighted E-sample estimate of total, including duplicate, fictitious, *etc.*

Occasionally, due to processing errors or timing constraints there may be a group of census enumerations that are excluded from both the E-sample processing and from the searching and matching process. Thus, while these records may be processed in time to be included in the official census results, they arrived too late be included in coverage measurement processing. These cases are sometimes known as "Late Census Adds" (LCA). These cases can be handled analogously to the treatment of census whole person imputations, that is replace (C - II) in Equation 6 with (C - II - LCA). Excluding the LCAs will not affect the DSE of the true population if the number of matches is reduced proportionally to the number of census correct enumerations. Said another way, the assumption is that the probability of a LCA being excluded from the A.C.E. processing must be statistically independent of its inclusion probability in the A.C.E. This is, of course, the traditional dual system independence assumption. (See Hogan 2001 for the supporting theory.) Although there were 2.3 million LCAs in Census 2000, analysis of the A.C.E. results by Raglin (2002) showed a trivial impact on the final DSE results.

In situations where the number of whole person imputations (II) was small, $(CE/N_e - 1)$ would be a measure of census gross overcoverage. That measure, however, is a function of the operational definitions of "correctly enumerated" adopted by the coverage measurement design. Definitions adopted to produce a good measure of net coverage, especially with respect to completeness and geographic correctness, may differ from those most appropriate for studying the quality of Census field operations. In any case, Census 2000 included 5.8 million whole-person imputations, of which 1.2 million were for housing units where the interviewer was unable to obtain even the number of residents (see Table 1 in Nash 2001, and page ii of Wetrogan and Cresce 2001.)

4. MEASURING THE PROPORTION OF PEOPLE CORRECTLY ENUMERATED

Having defined the set of correctly enumerated people, the next step in the DSE is to estimate the census coverage rate, N_{11}/N_{1+}.

Conceptually, estimating the rate entails (1) taking a sample of people, (2) determining whether they should be enumerated in the census, and (3) determining whether they were, indeed, correctly enumerated, using the same definitions as were used to measure N_{+1}. If an unbiased sample can be drawn of people who should have been enumerated and, if we can determine whether they actually were correctly enumerated (included in the census), then the DSE will produce asymptotically unbiased estimates. If each step can be approximately correct, the results will approach an unbiased estimate.

The first step in the process is, normally, to draw a random area sample. The A.C.E. uses the same set of block clusters for this purpose that it uses to define the E-sample.

Interviewers then canvass the block and prepare an independent list of people who should have been enumerated. This list constitutes the population or P-sample. The (weighted) sum of the people on this list, denoted $\hat{N}_p$, estimates N_{1+}. However, it is not the number which is of interest, but the ratio of N_{11} to N_{1+}, which we approximate by the ratio of correct matches, $\hat{M}$, to $\hat{N}_p$.

Operationally, the "correctly enumerated" census records are searched to see if the P-sample people were enumerated. The (weighted) number who were matched $(\hat{M})$ estimates N_{11}.

The DSE model will work if we can approximate:

1. Operational independence
2. Consistent reporting of residence
3. Accurate matching
4. Homogeneity within post-stratum

4.1 Operational Independence

Operational independence is the easiest assumption to approximate, but still requires vigilance. In Census 2000, the A.C.E. sample was drawn and the housing units listed before the delivery of the census questionnaires. Although personal contact was minimal, some people may react differently to the census because of their inclusion in survey listing. Early telephone interviews were allowed for independently listed housing units linked to a census address with a completed census questionnaire. This operation occurred while census nonresponse follow up was still being conducted in the area. Personal visit interviewing took place concurrently with some census "coverage improvement" interviewing. Clearly, some contamination could occur. Great care was taken to prevent the same field staff from working the same area in both Census and A.C.E. and to prevent the sharing of information. Still, some people may react differently to the survey because

they were enumerated, for example, by a very polite or very surly enumerator. Others may believe that they have a duty to provide the information once, but not twice.

Operational independence must also be preserved in office procedures. Definitions of "nonresponse" or "sufficient information" are sometimes applied differently to matched and non-matched *P*-sample records. The A.C.E. guarded against unnecessarily introducing operational dependence by forcing the processing system to first decide whether a case is acceptable for matching and only then attempt matching. The philosophy is "Do not attempt to find a match unless you would be satisfied that, if no match is found, the person was not enumerated!"

Before beginning the matching, *P*-sample records first are reviewed for:

(1) Appropriateness
(2) Uniqueness
(3) Completeness
(4) Geographic correctness

The A.C.E. contained no obviously fictitious records. One important safeguard is the use of Computer Assisted Personal Interviewing (CAPI). The CAPI instrument makes falsification difficult by "time stamping" the interview and recording every key stroke. We have instituted a quality assurance process to minimize other sloppy or dishonest A.C.E. interviewing. In addition, one important exception to the "no follow up" rule are cases where A.C.E. fabrication is possible, *e.g.*, cases where no one in the household matches, implying possible fabrication.

Out of scope records, *e.g.*, group quarters, are screened out. Occasionally, survey duplicates occur and these are eliminated (uniqueness). Finally, if the survey interview does not meet minimal standards, the case is converted to nonresponse and is later imputed.

4.2 Consistent Reporting of Residence

To measure the number of people correctly in both systems, we must determine whether or not a *P*-sample person was correctly enumerated in the census. This is done by searching the correct census records in the area where the person should have been enumerated.

The same definition of geographic correctness must apply both to whether an enumeration (in the *E*-sample) was correct and to whether the person (in the *P*-sample) was correctly enumerated. Failure to make these concepts agree is termed "balancing error."

Specifically, we must have the same definition of "correct" location and the same search area around the correct location. Errors can result in both erroneous non-matches and erroneous matches. Difficulty comes primarily from two sources. First, both the *P* and *E*-sample accept proxy responses. Thus, even though the person might have a clear and consistent understanding of his usual residence, the proxy respondent may not. Secondly, the way in which the question is posed in each interview could lead to different responses even from the same person. This might result in false non-match/not correctly enumerated status. On the other hand, if the person was incorrectly included by the census, we could incorrectly count the person as "correctly enumerated." Both errors clearly occurred on a relatively large scale in the A.C.E. (See section 6.)

The other dimension of geographic correctness is, again, the extent of search. The same area must be used to define the correct residence for determining both whether an enumeration was correct and whether a person was correctly enumerated. This is achieved by consistently applying the same search area definitions as in section 3.

4.3 Accurate Matching

The purpose of matching is to determine whether a person interviewed in the *P*-sample was also enumerated in the census within the defined search area. Much of the matching is now done by a computerized matching system. The system produces matches, possible matches, and non-matched cases. Repeated tests have shown that cases matched by the computer are nearly certainly correctly linked (Belin 1993). Nearly all clerical matching is now computer-assisted and largely paperless. This new system makes searching easier, including duplicate search. It restricts the codes clerks can apply to only those appropriate for the situation. The almost paperless system eliminated lost and misfiled A.C.E. questionnaires.

The first-level clerks were backed up by a team of 46 technicians. Training for these technicians began in September 1999. They were supported by a team of seven permanent analysts, most of whom have been matching for many years. Each level of matching acts as quality assurance for the level before. In addition, each level could refer problem cases to the next higher level. All matching was done in one location by one staff. The 1980 and 1990 matching operations were done in three and seven sites, respectively.

The use of the A.C.E. procedures for movers also greatly simplified the matching. Information about those who had moved was gathered from current residents. Under the procedures used in 1980 and 1990, movers were interviewed at their residence at the time of the PES interview. It was necessary then to code the reported correct Census Day residence to the correct census geography before beginning matching. This procedure was difficult, especially in rural areas. Mover matching was never before automated. In A.C.E., all matching, including for movers, was done in the *E*-sample block cluster or an adjacent block, using the same computer and computer-assisted clerical matching system. The change in the treatment of movers is discussed below.

4.4 The Role of After-Matching Reinterview

Some cases are sent to the field to gather further information after the initial matching is complete. This

after-matching reinterview is often termed "follow up interview."

The follow up interview process, like all PES activities, must fit into the overall framework of the DSE. Specifically, it must account for:

1. Appropriate, unique and correct response
2. Independence between census and survey inclusion probabilities
3. Balancing *P* and *E*-sample concepts
4. Search area and unique location matching rules
5. Treatment of missing data.

Follow up is only useful if it provides more accurate or consistent responses. Simply obtaining a different response is not justification. Since follow up takes place further from the census reference date than the initial interview, it is more difficult to obtain accurate responses. This is equally true for *E*-sample follow up and *P*-sample follow up. To provide better responses, follow up must use better resources, for example: (1) better respondents (household *vs.* proxy), (2) a better trained, supervised or quality-controlled interviewer, or (3) better questions or interview procedures.

The census data collection period extends from mid-March through mid-summer. Because of the huge scale of the operation, little emphasis is placed on verifying that the people were residents of the household on April 1. Quality assurance reinterview to prevent fabrication is minimal. Because of better training and supervision, and more complete questioning, the A.C.E. follow up interviewing can, in general, obtain more accurate information on residence and location than that gathered during the census process itself. Thus all non-matched *E*-sample cases were sent to follow up.

Follow up can, however, compromise independence. If all cases were sent to follow up, independence would not necessarily be compromised. However, cases that are matched during initial matching are seldom sent to follow up. To do so would stress the resources available for follow up. Instead, only non-matches or "possibly matched" cases are usually selected for follow up. This can introduce operational dependence.

The biases that can be introduced by follow up can occur even if the follow up interview was successfully conducted, since follow up may selectively change the defined "correct location" for non-matches but not for matches. If the follow up operation results in a non-interview, further biases can be introduced depending upon the missing data models applied to these cases.

Choosing cases for follow up requires balancing the need for accurate and consistent information with the need for independence. The *P*-sample only followed up cases when better information was likely. Cases sent to follow up included:

1. Possible matches, since with the information at hand the interviewers can resolve the situation,
2. Initial non-household proxy interviews that result in non-matches. Since we have not spoken to a household member, we have reason to doubt the accuracy,
3. Non-matched cases where, for the same housing unit, the census reports one family and the A.C.E. reports another. In order to ensure consistent reporting of Census Day address between the *P*-sample and the *E*-sample, these cases are sent out together,
4. Partial-household non-matches.

Cases that match and some other non-matched cases were generally not sent to follow up. For example, the A.C.E. did not follow up whole-household nonmatched cases where the census missed the unit, reported it as vacant, or could not obtain an interview (last resort information only).

4.5 Homogeneity Within Post-stratum

The DSE requires that the capture probabilities be independent for all individuals within estimation domains called post-strata. This is approximated by making the post-strata as homogeneous as possible with respect to the census capture probabilities, and then striving for as uniform as possible inclusion probabilities for the survey.

Dividing the population into many relatively small post-strata can increase within strata homogeneity. However, small strata can have high sampling variance and ratio bias. Ratio bias follows from the fact that the DSE is inherently a ratio estimator. This bias tends to decrease as the size of the post-stratum increases. In addition, our treatment of movers adds an additional ratio (see below). For this reason, we designed post-strata with a minimum expected sample size of 100.

For the A.C.E. we post-stratified based on the following variables:

1. Race / Hispanic Origin (7)
2. Age / sex (7)
3. Tenure (2)
4. Metropolitan area size and type of enumeration area (4)
5. Return rates (2)
6. Region (4)

where the number in parenthesis refers to the number of categories. More details on the post-strata are found in Haines (2001).

Coverage differences between racial and ethnic groups is well documented. (See for example Robinson, Ahmed, Das Gupta and Woodrow 1993; Hogan 1993.) Social, cultural, linguistic and economic differences may lead different racial and ethnic groups to react differently to the census procedures.

Demographic analysis and previous coverage surveys have demonstrated that people are differentially missed in different age groups and that the pattern is different for males and females. Most important in this pattern is young adults (Robinson *et al.* 1993.)

The importance of tenure was first measured following the 1980 Census and then implemented in the 1990 post-stratification. Those who live in owner-occupied houses are less mobile. They may feel that they have more of a stake in their community and thus, are more influenced by the census outreach program.

Metropolitan area size obviously affects housing patterns and is correlated with the way the Census Bureau builds its address lists. The combined variable "metropolitan area size and type of enumeration area" isolates differences in housing unit coverage. It may, in addition, measure some aspects of social and economic isolation.

The census return rate measures public cooperation with the census, an important predictor of coverage. It also measures directly the proportion of the enumeration that must be done in the census nonresponse follow up. One difficulty in this variable is that not all areas of the country are in the mailback universe. A small proportion is done by direct interview, and obviously have no "return rate." We have chosen to group these areas with "high" mail response areas.

Census Region picks up, among other things, broad differences in settlement patterns and housing stock. "Brown stone walk ups" are more common in the Northeast. Mobile homes are more common in the South.

Obviously, the complete cross-classifications can lead to very small cells. The maximum set of post-strata the sample was designed to support was 448. In fact, after collapsing small cells, there were 416 post-strata.

4.6 Treatment of Movers

People who move between the census reference date and the time of the survey interview present a challenge for designing a DSE for census application. First, people who move are more likely to be missed by the census and by the survey. Secondly, if a person has a different "usual residence" at the time of the survey than he did at the time of the census, one must decide where to sample him.

In the 1990 PES, movers were sampled where they lived at the time of the survey interview. We then searched the census records at, and only at, their April 1 usual residence. This is known as procedure B (Marks 1979). This approach requires both coding the address to the correct Census Day geography and then matching. These activities are complex and time consuming.

The A.C.E. used a different procedure known as procedure C. The A.C.E. estimated the number of movers by the number of people who moved into the sample blocks between April 1 and the time of the A.C.E. interview (in-movers). If the population was closed to international migration, deaths, movement to group quarters, *etc.*, then the number of people who moved in must equal the number who moved out (out-movers). They are the same people in the population, if not in the sample. It is normally easier to find people where they are, so the measured number of in-movers is normally a better estimate of the total number of movers than the measured number of out-movers.

The proportion of movers who are correctly enumerated is estimated by matching the out-movers to the census records for the sample block and extended search area, if appropriate. The estimated number of correctly enumerated movers is then $\hat{M}_t = (\hat{M}_o/\hat{N}_o)\,\hat{N}_i$ where $\hat{M}$ denotes the weighted number of correct matches; $\hat{N}$ denotes the weighted population number; and the subscripts denote total movers (t), out-movers (o) and in-movers (i).

If we denote those who do not move by the subscript n, the overall coverage rate becomes

$$\frac{N_{11}}{N_{+1}} = \frac{\hat{M}_n + \hat{M}_t}{\hat{N}_n + \hat{N}_i}.$$

The effect of procedure C is to increase the effective capture probabilities in the survey for movers and thus increase homogeneity of inclusion in the survey with respect to mover status (*i.e.*, mover/nonmover) (Griffin 2000).

There will be nonresponse and incomplete response at various steps. The goal of the missing data process is to improve the estimate of the number of people correctly counted (from the E-sample) or the estimate of the coverage ratio (from the P-sample). In designing missing data procedures, we choose methods that support the underlying DSE assumptions. Starting with the 1990 PES, the U.S. has estimated the probability a nonresponse record was correct rather than assigned a "zero/one" classification. (Schenker 1988, Belin 1993) The methods used for the A.C.E. are described in Cantwell and Ikeda (in this volume).

5. SYNTHETIC ESTIMATION

5.1 The Synthetic and Dual System Model

To this point, we have been dealing with the actual DSE. However, as noted in section 2, we use a synthetic estimator to distribute the measured net undercount to local areas and small groups.

In the A.C.E. the carrying-down was based on the same post-stratification variables as the DSE itself. The synthetic estimation is based on the assumptions that (1) the DSE estimates the true population, and (2) within post-strata, the true population is distributed proportionally to the pre-adjustment census count.

Clearly, at some level the second assumption can be only true with respect to the expected census counts. That is, even if within post-strata all people had identical probabilities of being enumerated in the census, we would observe different outcomes across blocks. The underlying

DSE explicitly models the undercount as a stochastic process.

As areas get larger, two things happen. First, the stochastic effect, or the random "block effect" begins to average out. Secondly, the effect of the actual undercount from a collection of blocks becomes positively correlated with the post-stratum's coverage correction factor. That is, the larger the area, the more the area's undercount determines the net correction factor.

The stochastic effect would be trivial for all but the smallest areas if Wolter's (1986) autonomous independence assumption held in practice, that is, if each person was included or missed independently of all other people. In fact, it is well known whole families are often missed or duplicated. Indeed, the whole building (or sometimes even block) might be missed or duplicated by the census address listing procedure. The failure of the autonomous independence assumption does not cause a bias in the dual system model as long as the underlying probabilities are equal within post-strata. This failure can mean that observed coverage for a block is inconsistent with the estimated undercount adjustment. However, as attention is turned to larger areas, the stochastic effect diminishes and is replaced with the problem of true heterogeneity of the underlying capture probabilities (see Haines 2001 for synthetic estimation details.)

5.2 Misclassification Error

In the discussion so far, we have accepted the post-stratum classification, j, as fixed. In practice, some people will be classified in different post-strata in the census and in the survey. For example, a woman may be reported as age 28 in the census and 31 in the survey, placing her in different post-strata.

Such misreporting is normally not important for matching. Name, address, month and day of birth, relation and household composition are far more important than age, race or sometimes even sex. So, assuming a match, in the above example we would have one correctly enumerated 28 year-old in the E-sample and one correctly enumerated 31 year-old in the P-sample. Misclassification can be seen to have two effects. To the extent the true undercount probabilities are homogeneous with respect to the true characteristics, misclassification introduces heterogeneity (and heterogeneity bias) into the observed estimation cells. This is true even if reporting is consistent between the census and the survey, because it can introduce unobserved subgroups within post-strata where the probabilities of inclusion in each system are correlated.

Inconsistent reporting between the census and the survey poses a problem for the synthetic estimator as well as for the DSE. This is easily seen by ignoring census imputations and erroneous enumerations. In this case, the coverage correction factor is the inverse of the matching rate (N_{11j}/N_{1+j}) where j represents the post-stratum. If the classification into the post-strata is inconsistent between the census and survey, we would be applying the rate, estimated from one group, to a somewhat different group. While misclassification may be ignorable at the poststratum level, it may be important locally. The A.C.E. protected itself against the general problem by avoiding, when possible, post-stratum definitions based on variables with high reporting variability.

6. FAILURE OF THE A.C.E. DESIGN AND CONCLUDING REMARKS

In spite of being seemingly well designed and well executed, the A.C.E. failed to even approximately measure the coverage error in the 2000 U.S. Census. The chief reason seems to have been a failure of the assumption of consistent reporting of Census Day residence. In other words, depending upon when and where and with whom the interview was conducted two or more residences were reported as the correct one for a large number of people in sample.

We know that this happened because, after the both the census and the A.C.E. were completed, we were able to search and match nationally. This allowed us to search for census duplicates, even when the pair was miles apart. This was possible because, for the first time, practically all names in the census were data captured. (See Fay 2002; Mule 2001, 2002.) We could see, for example, how many of the people who were classified by the A.C.E. E-sample as "correctly enumerated" were also enumerated somewhere else, including at an other household or in a group quarters.

In one study, of the 1.3 million (weighted) E-sample people linked to a duplicate census enumeration outside the search area, only 14 percent were coded as erroneous enumerations by the A.C.E. (Feldpausch 2001, Table 1.) Since the A.C.E. E-sample was a random sample, one would expect that for any pair of duplicates it would pick up the erroneous enumeration roughly half the time.

Another 521 thousand E-sample cases (weighted) were linked to census enumerations in group quarters. Of these, only 31 percent were classified as erroneous by the A.C.E. (Feldpausch 2001, Table 3.) Roughly half, 271 thousand, of these linked E-sample cases were linked to an enumeration in a college dormitory. Under census residence rules, those living in a dormitory should be counted there, and not at home. However, the A.C.E. classified only 45 percent of these E-sample cases as erroneous enumerations. Since the proportions coded as correctly enumerated by the A.C.E. are significantly different from what would be reasonable, one must conclude that the A.C.E. had a strong tendency to misclassify enumeration status. Interestingly, many of these misclassified cases, the exact number is hard to determine, must have been A.C.E. matches. This is certainly due to the tendency of respondents to confirm people as living at an address who should be counted as living somewhere else.

We now have clear evidence that large number of parents of college students living in dormitories will consistently report their child as living at home even though census instructions clearly say not to. Further, both parents in a "joint custody" situation may consistently report the child as living in each of two households. Neighbors, no doubt, will report someone as "living there" who is in fact away at college, in the military, in jail, or at a second home. This misreporting occurred in spite of the numerous, detailed and specific probing questions about ususal residence asked by the A.C.E.

The extended search for census duplicates discussed above formed the principal evidence for A.C.E. error. However, other evidence was also gathered, including a re-interview study. These evaluations are discussed in detail in the Census Bureau's "Executive Steering Committee on A.C.E. Policy" (ESCAP) documentation. (See ESCAP I 2001, ESCAP II 2001).

The results of these evaluations is that the A.C.E. failed to correctly identify 4.7 million erroneous enumerations (U.S. Census Bureau 2003, page iv). In addition, it probably mis-identified the residences of large numbers of people in the *P*-sample, leading to both false matches and false non-matches. An extensive program by the Census Bureau of analysis and estimation produced the 1.3 million overcount estimate cited above. However, this program was uniquely tailored to the special circumstances of the 2000 post-census rematching, reinterviewing and duplicate search. Those interested are directed to U.S. Census Bureau (2003).

This paper has described the theory of the DSE, and has discussed how PES in general, and A.C.E. in particular, have implemented that theory. It has described the approximations necessary in real applications and the types of errors that can occur.

It discussed how carefully each of these approximations must be controlled. Obviously, the A.C.E did not successfully measure the large number of duplicates in the 2000 Census. Failure of even extensive probing questions to elicit accurate reports of ususal residence was the principal cause. However, the theory and design developed here should be of value in any future coverage measurement program.

ACKNOWLEDGEMENTS

This paper reports the results of research and analysis undertaken by Census Bureau staff. The opinions expressed are those of the author and do not necessarily reflect those of the Census Bureau.

REFERENCES

BELIN, T.R. (1993). Evaluating sources of variation in record linkage through a factorial experiment. *Survey Methodology*. 19, 13-29.

CANTWELL, P., and IKEDA, M. (2003). Handling missing data in the 2000 accuracy and coverage evaluation survey. *Survey Methodology*. 29, 2, in press.

CHILDERS, D. (2001). Accuracy and Coverage Evaluation: The Design Document. DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-DT-1 (Revised).

ESCAP I (2001). Report of the Executive Steering Committee for Accuracy and overage Evaluation Policy. March 1, 2001. (See www.census.gov/dmd/www/pdf/Escap2.pdf)

ESCAP II (2001). Report of the Executive Steering Committee for Accuracy and Coverage Evaluation Policy on Adjustment for Non-Redistricting Uses. October 17, 2001. (See www.census.gov/dmd/www/pdf/Recommend2.pdf)

ESCAP II (2001). Census Person Duplication and the Corresponding A.C.E. Enumeration Status. Executive Steering Committee for A.C.E. Policy II, Report 6.

FAY, R. (2002). Probabilistic models for detecting census person duplication. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

FELDPAUSCH, R. (2001). ESCAP II: Census Person Duplication and the Corresponding A.C.E. Enumeration Status. Executive Steering Committee for A.C.E. Policy II, report 6.

GONZALEZ, M. (1973). Use and evaluation of synthetic estimators. *Proceedings of the Social Statistics Section*, American Statistical Association. 73, 7-15.

GONZALEZ, M., and HOZA, C. (1978). Small-area estimation with application to unemployment and housing estimates. *Journal of the American Statistical Association*. 73, 361, 7-15.

GRIFFIN, R. (2000). Accuracy and Coverage Evaluation Survey: Dual System Estimation. DSSD Census 2000 Procedures and Operations Memorandum Series Q-20.

HAINES, D. (2001). Accuracy and Coverage Evaluation Survey: Computer Specifications for Person Synthetic Estimation (U.S.) Re-issue of Q-30. DSSD Census 2000 Procedures and Operations Memorandum Series Q-46.

HOGAN, H. (1992). The 1990 Post-Enumeration Survey: An overview. *The American Statistician*. 46, 261-269.

HOGAN, H. (1993). The Post-Enumeration Survey: Operations and Results. *Journal of American Statistical Association*. 88, 423.

HOGAN, H. (2001). Accuracy and Coverage Evaluation Survey: Effect of Excluding 'Late Census Adds'. DSSD Census 2000 Procedures and Operations Memorandum Series Q-43. http://www.census.gov/dmd/www/pdf/Q-43.pdf

MARKS, E.S., SELTZER, W. and KROTKI, K.J. (1974). *Population Growth Estimation*. New York: Population Council.

MARKS, E.S. (1979). The Role of Dual System Estimation in Census Evaluation. In Recent Developments in PGE, (K. Krotki). University of Alberta Press. 156-188.

MULE, T. (2001). ESCAP II: Person Duplication in Census 2000. Executive Steering Committee for A.C.E. Policy II, Report 20.

MULE, T. (2002). Further Study of Person Duplication Statistical Matching and Modeling Methodology. DSSD A.C.E. Revision II Memorandum Series PP-51.

NASH, F.F. (2001). ESCAP II: Analysis of Census Imputations. Executive Steering Committee for A.C.E. Policy II, Report 21.

PETERSEN, C.G.J. (1896). The Yearly Immigration of Young Plaice into the Limfjord from the German Sea. Report of the Danish Biological Station. 6, 1-48.

RAGLIN, D. (2002). ESCAP II: Effect of Excluding Reinstated Census People from the A.C.E. Person Process. Report 13, http://www.census.gov/dmd/www/pdf/Report13.PDF

ROBINSON, J.G., AHMED, B., DAS GUPTA, P. and WOODROW, K. (1993). Estimates of population coverage in the 1990 united states census based on demographic analysis. *Journal of the American Statistical Association*. 88, 1061-77.

ROBINSON, J.G. (2001). ESCAP II: Demographic Analysis Results. Executive Steering Committee for A.C.E. Policy II, Report 1.

SCHENKER, N. (1988). Handling missing data in coverage estimation with application to the 1986 test of adjustment related operations. *Survey Methodology*. 14, 87-97.

SEKAR, C.C., and DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*. 44, 101-115.

U.S. CENSUS BUREAU (1985). Evaluating Census of Population and Housing, Statistical Training Document, ISP-TR-5, Washington, D.C.

U. S. CENSUS BUREAU (2000). Statement on the Feasibility of Using Statistical Methods to Improve the Accuracy of Census 2000.

U. S. CENSUS BUREAU (2003). Technical Assessment of A.C.E. Revision II, March 12, 2003, http://www.census.gov/dmd/www/pdf/ACETechAssess.pdf

WETROGAN S.I., and CRESCE A.R. (2001). ESCAP II: Characteristics of Census Imputations. Executive Steering Committee for A.C.E. Policy II, Report 22.

WOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*. 81, 338-346.

Survey Methodology, December 2003
Vol. 29, No. 2, pp. 139-153
Statistics Canada

Handling Missing Data in the 2000 Accuracy and Coverage Evaluation Survey

PATRICK J. CANTWELL and MICHAEL IKEDA[1]

ABSTRACT

The Accuracy and Coverage Evaluation survey was conducted to estimate the coverage in the 2000 U.S. Census. After field procedures were completed, several types of missing data had to be addressed to apply dual-system estimation. Some housing units were not interviewed. Two noninterview adjustments were devised from the same set of interviews, one for each of two points in time. In addition, the resident, match, or enumeration status of some respondents was not determined. Methods applied in the past were replaced to accommodate a tighter schedule to compute and verify the estimates. This paper presents the extent of missing data in the survey, describes the procedures applied, comparing them to past and current alternatives, and provides analytical summaries of the procedures, including comparisons of dual-system estimates of population under alternatives. Because the resulting levels of missing data were low, it appears that alternative procedures would not have affected the results substantially. However some changes in the estimates are noted.

KEY WORDS: Cell Imputation; Noninterview Adjustment; Logistic Regression; Dual-System Estimation.

1. INTRODUCTION

Following the 2000 Census in the United States, the Census Bureau conducted the Accuracy and Coverage Evaluation (A.C.E.) survey. The survey had two goals: (1) to measure the level of net undercoverage across the nation and in various demographic and geographic domains through a statistical technique called dual-system estimation, and (2) to produce revised population counts that could be used to adjust for this net undercoverage – if the adjusted numbers were deemed to be more accurate than the initial census counts (Hogan 2003).

In the process of interviewing and following up respondents in the A.C.E., some households were missed, and certain information needed to calculate the dual-system estimates was not collected from other sample respondents. This paper describes the levels of missing data, discusses the procedures used to address the problem in the A.C.E., and provides some results and evaluations. It should be noted that the term "missing data" applies after all follow-up attempts were made in the field. These activities included multiple attempts at interviews, the use of highly trained clerks and technicians to resolve cases, and the follow-up of cases where a second interview could provide additional required information.

The A.C.E. realized three main types of missing data. First, some households were not interviewed because they could not be contacted or the interview was refused. What makes the situation different in the A.C.E. is that to each sample housing unit, *two* noninterview adjustments were applied; one corrected for noninterviews on Census Day, while the other corrected for noninterviews on the day of the A.C.E. interview. As will be shown, the need for two adjustments reflects the different ways that out-movers and in-movers were treated in the dual-system estimation.

The second type of missing data occurred when information for a household or person was available but specific demographic characteristics needed for dual-system estimation were not collected. For missing tenure (owner vs. non-owner), race, and Hispanic origin, a form of nearest-neighbor hot-deck imputation was used to take advantage of the correlations often found among people living in geographic proximity. In general, the values of age and sex are geographically less clustered, but often well predicted by specific conditions, such as the person's relationship (*e.g.*, spouse, child) to the household's reference person, or whether information is available on the person's spouse. Therefore, national donor distributions conditioned on relevant covariates were used to impute for age and sex. Because characteristic imputation for the A.C.E. was similar to that done in the Post-Enumeration Survey following the 1990 Census, the methods and results are not discussed further in this paper.

The third type also involved item missing data. For a small number of people in the A.C.E., not enough information was collected to determine the resident status (whether or not the person was living in the sampled block cluster on Census Day) or the match status (whether or not the person actually matched to someone in the census). Similarly, some people counted in the census lacked sufficient information to determine whether they were correctly enumerated. The status in such cases is said to be "unresolved." Yet this information is required to compute dual-system estimates. To resolve such cases, a probability of resident (or match or correct enumeration) was assigned as the average weighted value from a set of resolved cases with similar characteristics.

Some of these procedures – described in greater detail below – were applied in similar forms in the 1990 Post-Enumeration Survey, as well as in tests conducted during

[1] Patrick J. Cantwell and Michael Ikeda, Mathematical Statisticians, U.S. Census Bureau, Statistical Research Division, Washington, D.C. 20233-9100.

the 1990s. The main exception is the assignment of a probability in the case of unresolved resident, match, or enumeration status. In the Post-Enumeration Survey and at times for specific tests in the 1990s, these probabilities were computed using a logistic regression model. The method applied in the 2000 A.C.E. used less information than some alternatives such as logistic regression, but was simpler to implement and verify in the tight A.C.E. schedule.

The levels of missing data in the A.C.E. were relatively low, which helped to reduce the potential for additional error in the estimates.

- The household noninterview rates were 3.0% and 1.1% (unweighted), respectively, on Census Day and Interview Day.
- The imputation rates for the five A.C.E. characteristics required for dual-system estimation ranged from 1.4% to 2.5% (unweighted and weighted).
- Among people in the A.C.E., the rates of unresolved resident and match status were 2.3% and 1.2% (unweighted), respectively; among census enumerations, only 3.0% (2.6% weighted) of the sample had unresolved enumeration status.

When assigning probabilities for unresolved status, the success of the variables used to define imputation cells was mixed. Variables that used information related to an individual's processing in the survey operations discriminated well among cells. However, variables describing the person's demographic characteristics appear to have been generally less successful.

Section 2 contains background information about the A.C.E. and dual-system estimation. The A.C.E. noninterview adjustment is discussed in section 3. For persons with unresolved resident, match, or enumeration status, a probability was assigned according to procedures described in section 4. Section 5 examines the effect of some alternatives to the A.C.E. missing data procedures on the dual-system estimates of the population. Finally, a few observations are recounted in section 6. For a detailed description of the missing data procedures for the 2000 A.C.E., see Cantwell (2001). Summaries of missing data can be found in Cantwell *et al.* (2001).

In what follows, unweighted frequencies and proportions are generally given. Unless noted otherwise, the weighted numbers are very close. However, the probabilities assigned to unresolved cases in Tables 4, 5, and 6 are the actual weighted ones used in the estimation.

2. A BRIEF ACCOUNT OF THE SURVEY AND DUAL-SYSTEM ESTIMATION

Through the Accuracy and Coverage Evaluation (A.C.E.), the Census Bureau attempted to measure and adjust for the historical differential net undercount observed in the U.S. Census (Anderson and Fienberg 1999, page 29). Like the 2000 Census, the A.C.E. covered the entire nation. (A separate sample and analysis were conducted for Puerto Rico.) A sample of about 300,000 housing units in 11,303 block clusters was selected (Fenstermaker 2000, Childers 2000).

To estimate coverage of the population, the A.C.E. relied on dual-system estimation, a method based on capture-recapture methodology (Peterson 1896, Sekar and Deming 1949). Suppose one considers only those housing units contained in the sample of block clusters selected for the A.C.E. After the census enumeration – but without using *any* information collected in the census – the Census Bureau independently interviewed people in the A.C.E sample and obtained a roster of people living in the units on Census Day, April 1, 2000. These results were then matched to (compared with) the census enumeration in those block clusters to estimate how many people were missed. Within the sample block clusters, the units enumerated independently in the A.C.E. were defined as the *P*-Sample, and those enumerated in the census as the *E*-Sample.

In the same sample of block clusters, comparisons and analyses were made to estimate the proportion of census enumerations that were correct, that is, complete, unique, and recorded in the proper location. Erroneous enumerations include people who are duplicated or fictitious, or should not be counted at that address, for example, because their usual residence was elsewhere, such as in a college dormitory. The resulting dual-system estimator is

$$\hat{N} = (C - I)\,\hat{p}_{ce}\left(\frac{1}{\hat{p}_{\text{match}}}\right), \qquad (1)$$

where C is the official census count, including imputed persons and erroneous enumerations; I is the number of whole-person imputations; $\hat{p}_{ce}$ is the weighted estimate of the proportion of correct enumerations in the census; and $\hat{p}_{\text{match}}$ is the weighted estimate of the proportion of *P*-Sample people who match to someone enumerated in the census. People are imputed, for example, when a census enumerator confirms that a certain number of people live at an eligible address, but sufficient additional information cannot be gathered. The actual number of whole-person imputations is known and can be removed from C in the estimate.

Dual-system estimates were calculated separately within population subgroups called post-strata. Post-stratum estimates were then used to determine adjustment factors to be applied to all people counted in the census according to their specific post-stratum. Finally, adjusted counts for any geographic area were calculated by summing the adjusted counts across post-strata in the area. For more detailed information on A.C.E. field operations and dual-system estimation in general, see Childers (2000) and Hogan (1993, 2003), respectively.

3. NONINTERVIEW ADJUSTMENT

Noninterview adjustment was performed only on the *P*-Sample; in the census (and, thus, in the *E*-Sample), people in all known housing units were accounted for through a variety of procedures. The small number of housing units whose information was collected by a proxy respondent, often a neighbor or building manager, were treated as valid interviews and are not the subject of the noninterview adjustment. Because people moved in and out of housing units between Census Day and the time of the A.C.E. interview, the Census Bureau had to consider the mover status – out-mover, in-mover, or non-mover – of all people in the *P*-Sample, as well as the interview situation at the two different moments. Out-movers were living in the housing unit in question on Census Day, but had moved out before Interview Day. The situation was reversed for in-movers. Non-movers lived in the unit on both days. At the time of the A.C.E. interview, *in one interview* questions were asked to determine who lived in the household on Interview Day and who lived there on Census Day. Mover status was assigned to each person in the sample, and two rosters were created for each household – the Census Day roster and the Interview Day roster.

The A.C.E. used in-movers to estimate the *number* of *P*-Sample movers, while using out-movers to estimate the *match rate* of the movers. The weighted *P*-Sample total, that is, the denominator of $\hat{p}_{\text{match}}$ in equation (2), is estimated as the weighted total of all non-movers and in-movers. Yet the weighted number of *P*-Sample matches is estimated by adding the number of matches among non-movers to the product of the number of in-movers and the match rate for out-movers:

$$\hat{p}_{\text{match}} = \frac{M_{nm} + N_{im} \times \dfrac{M_{om}}{N_{om}}}{N_{nm} + N_{im}}, \qquad (2)$$

where N (people) and M (matches) are indexed by *nm*, *im*, and *om*, representing non-movers, in-movers, and out-movers, respectively. All in-movers and non-movers were generally assumed to be A.C.E. Interview Day residents. (People living in group quarters, such as college students in dormitories, were not eligible for the *P*-Sample.)

The mover procedure used in the A.C.E. differed from that used in the 1990 Post-Enumeration Survey. In 1990 in-movers were used to estimate the number of movers *and* their match rate. For the latter, the in-movers had to be matched back to their address on Census Day. That procedure was changed for the census tests conducted during the 1990's to accommodate the planned use of sampling for census nonrespondents. When the U.S. Supreme Court ruled against the sampling plan in 1999 (*Department of Commerce v. United States House of Representatives*, 525 U.S. 316, 1999), it was thought that changing the mover procedure again so late before the census would introduce unacceptable risks.

Due to the mover procedure described above, each housing unit had two interview statuses – one based on the housing unit's situation as of Census Day, and the other based on the day of the A.C.E. interview. A unit that was vacant, removed from the list of eligible housing units (because, for example, it was demolished or used only as a business), or in certain special places was not considered an interview or a noninterview. Table 1 provides a fictional but illustrative block cluster. It demonstrates how the status of a housing unit on Census Day and Interview Day would have been determined.

Results of the A.C.E. interviewing operation are shown in Table 2. Of the 261,969 housing units occupied on Census Day, 7,794 (3.0 percent) were noninterviews. The corresponding numbers for Interview Day were 267,155 and 3,052 (1.1 percent).

As different interview statuses were possible for a housing unit on Census Day and Interview Day, different noninterview adjustments were required for each day. Each of the two adjustments generally spread the weights of noninterviewed units over interviewed units in the same noninterview cell: the sample block cluster crossed with the type of basic address, defined as single-family, multi-unit (such as apartments and condominiums), or all others. Other characteristics, known for all housing units, could have been used to define the cells. However, the cells were defined to take advantage of the typical local homogeneity, and of the fact that people living in, for example, apartments share many of the characteristics – household size, propensity to move, *etc.* – that are related to capture probabilities in the census.

The noninterview adjustment based on the Census Day status of housing units was used to adjust the person weights of non-movers and out-movers. Similarly, the Interview Day noninterview adjustment was used to adjust the person weights of in-movers. Within a noninterview cell, the adjustment factor for *Census Day* was computed as the weighted sum of interviews and noninterviews for Census Day divided by the weighted sum of interviews for Census Day. A housing unit's weight was the inverse of the final selection probability of its block cluster into the A.C.E. sample. (These weights were trimmed in a very small number of clusters.)

The noninterview adjustment factor for Interview Day was computed as above, but with its status – interview, noninterview, vacant, or delete – being considered for Interview Day rather than for Census Day. The example in Table 1 demonstrates the calculation of the noninterview adjustment for the fictional block cluster. Because the non-interview rates were so small, the noninterview adjustment factors were close to 1 for most housing units in the sample. For Census Day, the factors were less than 1.10 for more than 92% of the units; for Interview Day, the factors were less than 1.10 for over 98% of the units.

Table 1
An Example of Adjustment for Noninterviews

Consider a block cluster with nine housing units, all having the same type of basic address, for example, all single family homes, as depicted below

Housing Unit	Weight	Actual Situation	Status of (and Information from) A.C.E. Interview	Census Day Status	A.C.E. Interview Day Status
1	100	Resident on 4/1/00 and at time of A.C.E. interview	Interviewed in A.C.E.	Interview	Interview
2	100	Resident on 4/1 and at time of A.C.E. interview	Neighbor (proxy) interviewed in A.C.E.	Interview	Interview
3	100	Resident on 4/1 and at time of A.C.E. interview	No one interviewed in A.C.E.	Noninterview	Noninterview
4	100	Vacant on 4/1, resident at time of A.C.E. interview	Interviewed in A.C.E., knows of 4/1 status	Vacant	Interview
5	100	Vacant on 4/1, resident at time of A.C.E. interview	Interviewed in A.C.E., no knowledge of 4/1 status	Noninterview	Interview
6	100	Vacant on 4/1, resident at time of A.C.E. interview	No one interviewed in A.C.E.	Noninterview	Noninterview
7	100	Resident on 4/1, vacant at time of A.C.E. interview	Information obtained from proxy	Interview	Vacant
8	100	Resident on 4/1, vacant at time of A.C.E. interview	No info on 4/1 status; Census staff determines vacant at time of A.C.E.	Noninterview	Vacant
9	100	Resident on 4/1, different resident at time of A.C.E. interview	Interviewed in A.C.E., knows of 4/1status	Interview	Interview

In this noninterview cell (sample block cluster × type of basic address), people in interviewed housing units would have received the following noninterview adjustments:

(1) to the person weights of non-movers and out-movers, the Census Day noninterview adjustment = 800 / 400 = 2.0

(2) to the person weights of in-movers, the A.C.E. Interview Day noninterview adjustment = 700 / 500 = 1.4

Table 2
Status of Household Interviews in the A.C.E. (Unweighted)

	Census Day		A.C.E. Interview Day	
	Number	Percent	Number	Percent
Total Housing Units	300,913	100.0	300,913	100.0
Interviews	254,175	84.5	264,103	87.8
Noninterviews	7,794	2.6	3,052	1.0
Vacant Units	28,472	9.5	29,662	9.9
Deleted Units	10,472	3.5	4,096	1.4
Noninterview rate[1]	3.0%		1.1%	

[1] Noninterview rate = Noninterviews / (Interviews + Noninterviews)

When the unweighted number of noninterviewed units in a given noninterview cell was more than twice the unweighted number of interviewed units, the weights of the noninterviewed units in this cell were spread over the interviewed units in a broader set of noninterview cells. This remedy was needed for only 65 cells for the Census Day adjustment, and 13 cells for the Interview Day adjustment. The prescribed procedure differs from the usual collapsing of sparse cells, but allowed us to address such cells in a simple automated fashion. This capability was important under a very tight schedule when it was impossible to predict which cells would have too few interviews. For evaluation purposes, the housing-unit weights were later re-computed under a collapsing scheme, and compared to the weights as determined in the A.C.E. Again, due to the low rates of noninterview, the weights were the same for most units, and close for the rest. The effect on the resulting dual-system estimates is shown in section 5.2.

4. ASSIGNING PROBABILITIES FOR UNRESOLVED CASES

After all A.C.E. follow-up activities were completed, there remained a small fraction of the A.C.E. sample without enough information to compute the components of the dual-system estimator given in equation (1). Their status was said to be "unresolved."

4.1 Unresolved Cases and Their Frequencies

One component of the dual-system estimator in equation (1) is $\hat{p}_{\text{match}}$, the estimated proportion of the P-Sample who match to someone enumerated in the census. In (2) for $\hat{p}_{\text{match}}$, when estimating the number of people (N_{nm}, N_{om}) or matches (M_{nm}, M_{om}) among non-movers and out-movers, only Census Day residents of the sample block clusters were considered; someone who usually lives in a nursing home, for example, was omitted from the computation.

Thus, for each person in the *P*-Sample, determining resident status and match status was required.

After follow-up operations were completed, all people in the *P*-Sample who were eligible to be matched to the census were classified into three types according to their status as a resident in their sampled block *on Census Day:* residents, nonresidents, and unresolved persons – those for whom there was not enough information to determine the resident status. Further, each confirmed or possible (unresolved) Census Day resident in the *P*-Sample was determined to be a match, a nonmatch, or unresolved match. The match status for confirmed Census Day nonresidents, such as in-movers, was not used in the estimation. The estimator in (1) also requires an estimate of the proportion of correct enumerations in the census, $\hat{p}_{ce}$. After whole-person imputations were removed from the *E*-Sample, each remaining person had one of three types of enumeration status: correct, erroneous, or unresolved.

Table 3 summarizes the frequencies of resident and match status in the *P*-Sample, and enumeration status in the *E*-Sample. The table also shows results for non-movers and out-movers in the *P*-Sample. One can see that the extent of unresolved cases is quite small: 2.3% for resident status, 1.2% for match status, and 3.0% for enumeration status. (The weighted rates are 2.2%, 1.2%, and 2.6%, respectively.) In the 1990 Post-Enumeration Survey, the rate of unresolved matches was 1.9%, and unresolved enumerations was 2.4%. (Resident status was not defined in a manner comparable to 2000.) Care must be taken, however, as the definitions of the several statuses were slightly different in 1990.

4.2 Assigning Probabilities to Unresolved Cases

In the A.C.E., a form of cell imputation was used to assign probabilities for sample cases with unresolved resident, match, or enumeration status. All people in the sample – resolved and unresolved – were placed into groups called imputation cells based on operational and demographic characteristics. Different variables were used to define cells for each type of status. Within each imputation cell the weighted average of 1's and 0's (representing, *e.g.*, match and non-match, respectively) among the resolved cases was calculated, and that average was imputed for all unresolved persons in the cell. More details are provided below.

In the 1990 Post-Enumeration Survey, hierarchical logistic regression was used to calculate probabilities of match and correct enumeration for cases with missing information. (Due to the procedure used to treat movers in 1990, resident status played a different role then.) The model and some results are discussed in Belin *et al.* (1993).

During the 1990s, the Census Bureau originally planned to produce in 2000 adjusted census estimates for each of the 50 states (and the District of Columbia) using data collected only from that state. This approach affected the strategy for treating unresolved status in two ways. First, within each state, there would be far fewer data – resolved cases – on which to build a logistic regression model. Second, there would be 153 different models to examine and verify, separate models for resident, match, and enumeration status in each state. Because the production schedule for the A.C.E. provided only about three weeks for addressing all

Table 3
Final Status Frequencies for the *P* and *E*-Samples (Unweighted)

		Final resident status			
P-Sample	Total people[1]	Confirmed resident	Confirmed nonresident	Unresolved resident	Resident rate for resolved cases
U.S. Total	653,337	95.8%	1.9%	2.3%	98.1%
Mover status					
Non-mover	627,992	96.6%	1.7%	1.7%	98.3%
Out-mover	25,345	75.2%	7.5%	17.4%	91.0%
		Final match status			
P-Sample	Total people[2]	Match	Nonmatch	Unresolved match	Match rate for resolved cases
U.S. Total	640,945	90.3%	8.5%	1.2%	91.4%
Mover status					
Non-mover	617,490	91.1%	8.0%	0.9%	91.9%
Out-mover	23,455	67.8%	21.7%	10.5%	75.8%
		Final enumeration status			
E-Sample	Total people	Correct enumeration	Erroneous enumeration	Unresolved enumeration	Correct enumeration rate for resolved cases
U.S. Total	704,602	92.6%	4.4%	3.0%	95.5%

[1] Those in the *P*-Sample eligible to be matched to the census.
[2] Confirmed or possible residents in the *P*-Sample.

aspects of missing data, it was believed that a procedure to handle unresolved status that was simpler to implement and verify would reduce the risk of not completing the dual-system estimates under the imposed deadline. Cell imputation provided the desired simplicity, but its accuracy relative to that under logistic regression modeling had to be evaluated in subsequent testing.

During census tests in 1995 and 1996, certain types of unresolved status were addressed using logistic regression, while cell imputation was used for other types. The latter procedure was used exclusively in the Census Dress Rehearsal in 1998 (Ikeda, Kearney and Petroni 1998), when the Census Bureau was still preparing to produce estimates independently within each state. Data from these tests indicated that the exact method of calculating probabilities for unresolved status (match, resident, or correct enumeration) had only a minor effect on the dual-system estimates. Details of this research can be found in Petroni (1997, 1998a, 1998b, and 1998c).

With the decision by the U.S. Supreme Court in 1999 (*Dept. of Commerce v. U.S. House of Representatives*), the Census Bureau changed the design of the survey and removed the restriction that adjusted estimates be based solely on data from within each state. However, there remained concerns about implementing a logistic regression approach that had not been tested in the Dress Rehearsal. Further, there was no guarantee that available software would adequately run logistic models on data sets the size of the entire A.C.E. sample (between 640,000 and 750,000 people). Based on these concerns and research findings on relative accuracy, a decision was made to use the simpler procedure, cell imputation, to resolve missing status in the A.C.E.

To demonstrate how cell imputation was applied in the A.C.E., one can look at resident status; the method was applied analogously to match and enumeration status. First, all non-movers and out-movers in the *P* Sample were placed into a number of imputation cells according to operational and demographic characteristics, as defined in Table 4; in-movers were left out, as their Census Day resident probability was 0 by definition. *Among the resolved cases* in cell i, denoted by the set $R(i)$, an indicator variable for resident status was defined as $1_{res,j} = 1$, if person j was a resident in the household on Census Day, or 0, otherwise. Then within cell i, the weighted proportion of Census Day residents, was computed:

$$P(res)_i = \frac{\sum_{j \in R(i)} w_j 1_{res,j}}{\sum_{j \in R(i)} w_j} \tag{3}$$

where w_j is the weight of person j incorporating all stages of sampling. $P(res)_i$ was then assigned to each unresolved person in cell i, that is, each of the 15,082 people (2.3% of 653,337) with unresolved resident status. (The exception is for match code group 7, as explained below.) Table 4 provides the resident probabilities assigned within the cells. This assignment defines for all cases – resolved and unresolved – an "extended" indicator, allowing values between 0 and 1:

$$1'_{res,j} = \begin{cases} 1_{res,j}, & \text{if } j \in R(i) \\ P(res)_i, & \text{otherwise} \end{cases} \tag{4}$$

The estimated numbers of non-movers and out-movers in the *P*-Sample in (2), N_{nm} and N_{om}, respectively, are then computed by attaching the person weight and summing the indicator $1'_{res,j}$ over the non-movers and out-movers, respectively, in all cells. The number of matches, M_{nm} or

Table 4
Imputation Cells for Resolving Resident Status in the *P*-Sample

P Sample Match Code Group	Owner		Non-Owner	
	Nonhispanic White	Others	NonhispanicWhite	Others
1. Matches needing follow-up	0.982	0.986	0.993	0.991
2. Possible matches	0.973	0.968	0.966	0.972
3a. Partial household nonmatches needing follow-up; aged 18-29, child of reference person	0.755	0.901	0.883	0.928
3b. Partial household nonmatches needing follow-up; others not in 3a	0.956	0.971	0.959	0.969
4. Whole household nonmatches needing follow-up, not conflicting households	0.920	0.943	0.911	0.914
5. Nonmatches from conflicting households	0.910	0.927	0.945	0.954
6. Resolved before follow-up	0.993	0.990	0.990	0.988
7. Insufficient information for matching (Weighted column average of groups 1-5 and 8)	0.813	0.867	0.844	0.872
8. Potentially fictitious or said to be living elsewhere on Census Day	0.119	0.123	0.177	0.157

M_{om}, and thus, $\hat{p}_{match}$, are determined analogously, as is $\hat{p}_{ce}$, in the case of enumeration status.

In the Census Dress Rehearsal of 1998, cell imputation for unresolved resident probability was used with only three cells: persons sent to follow-up, persons not needing follow-up, and persons with insufficient information for matching. For the third cell, which contained no resolved cases, a proportion based on all resolved cases in the first two cells was assigned. Results from the Dress Rehearsal (Kearney and Ikeda 1999) suggested that dividing the *P*-Sample into the various match code groups would be helpful. Further research and discussion suggested adding other demographic variables within match code group. The larger A.C.E. sample size in 2000 made it possible to support a larger set of imputation cells.

For the A.C.E. in 2000, match code groups 1 through 7 were determined from the match codes and other variables derived *before* the follow-up operation, as explained in Childers (2000). Group 8 was formed differently. Some information from the follow-up operation was coded in time for the A.C.E. missing data procedures. (Under the original schedule, this information would have become available too late to be of use.) *After* the follow-up operation, a small number of people in the *P*-Sample were coded as being potentially fictitious or said to be living elsewhere on Census Day. Among the resolved cases in this group, the probability of being a resident was much lower than for resolved people in other groups. Thus, people satisfying the conditions for group 8 were placed there first, and each of the remaining people was placed appropriately in one of the first seven groups.

Two tenure categories were used: owners and non-owners. Persons were also placed into one of two race-ethnicity categories: Nonhispanic white, and all others. People of multiple races (for example, a person responding as White and Asian) were placed in the latter group. Match code group 3, partial household nonmatches, was split into two subgroups. The first, 3a, included persons in group 3 who were 18 to 29 years of age and were listed on the A.C.E. household roster as a child of the reference person. These were young people many of whom were attending college, sharing residence with colleagues, or moving in and out of their parents' residence. Classification and regression tree analysis, applied to data from the Census Dress Rehearsal of 1998, suggested that this combination of characteristics would discriminate well with respect to resident status. The group 3b included all other persons in group 3.

The resident probability for unresolved *P*-Sample persons was computed as described above, except for those in match code group 7 – people with insufficient information for matching. Within this row in Table 4, there were essentially no resolved cases from which to extract a probability of being a Census Day resident. Because of their lack of information – most of these cases did not even have a valid name – these people did not go through the matching operation and were not sent to follow-up. To determine a resident probability for these cases, a weighted proportion of Census Day residents (1's and 0's) was computed among the resolved cases in match code groups 1 through 5 and 8, separately for each of the four tenure × race/ethnicity classes. This probability was then assigned to those in group 7. Left out of this computation were those people who were resolved before follow-up (group 6). Observations from the Dress Rehearsal indicated that, in terms of their demographic and operational characteristics, people in group 7 tended to be more like those in groups 1-5 and 8, than like those in group 6.

The issue of unresolved matches was treated like that for unresolved resident status in (3) and (4), with resident status replaced by match status, but with a different set of cells, as is seen in Table 5. Confirmed nonresidents were excluded from the computations of match probabilities.

For unresolved match probability in the Dress Rehearsal, only one imputation cell was used within each of the geographic sites. Subsequent analysis (Kearney and Ikeda 1999) showed that mover status (non-mover *vs.* out-mover) discriminated well between matches and nonmatches among the resolved cases. Thus, for the 2000 A.C.E. mover status was used to define imputation cells for match status. The housing-unit address match code refers to the initial match between housing units on the independent (A.C.E.) listing and the census address list; conflicting housing units, determined during A.C.E. person match activities, were those where the census and A.C.E. rosters had two completely different lists of residents for Census Day (Childers 2000).

It should be noted that 98.3% of the unresolved matches (7,693 of 7,826) were people with insufficient information for matching. As mentioned above, most of them did not have a valid name, and almost all (7,506) were not sent to follow-up. Further, their rate of missing characteristics was much higher than average. Therefore, little useful predictive information was available when forming imputation cells for match status. Variables such as age and ethnicity – that had a higher chance of being imputed and might be of questionable quality – were avoided.

People with at least one imputed demographic variable (age, sex, tenure, race, or Hispanic origin) were grouped when resolving match status. Unpublished studies indicated that – at least among resolved cases in the Dress Rehearsal – the presence of these imputed characteristics is negatively associated with the propensity to be a match. Out-movers from a unit that was a nonmatch or a conflicting household were not separated according to this variable to ensure a reasonable number of resolved cases in each cell from which to estimate the proportion of matches.

In the *E*-Sample, unresolved enumeration status was addressed as discussed above. See Table 6.

Table 5
Imputation Cells for Resolving Match Status in the *P*-Sample

Mover Status	Housing-Unit Address Match Code			
	Housing unit was a match		Housing unit was a nonmatch or the household was conflicting	
Non-mover	No imputed characteristics[1] 0.945	1 or more imputed characteristics 0.901	No imputed characteristics 0.690	1 or more imputed characteristics 0.567
Out-mover	No imputed characteristics 0.798	1 or more imputed characteristics 0.791	0.516	

[1] Among the characteristics age, sex, tenure, race, or Hispanic origin.

Table 6
Imputation Cells for Resolving Enumeration Status in the *E*-Sample

	E-Sample Match Code Group	No Imputed Characteristics[1]		1 or More Imputed Characteristics
1.	Matches needing follow-up	0.977		0.977
2.	Possible matches	0.968		0.968
3a.	Partial household nonmatches; aged 18-29, child of reference person	0.871		0.908
3b.	Partial household nonmatches; others not in 3a	0.974		0.960
4.	Whole household nonmatches where the housing unit matched; not conflicting households	Nonhispanic White 0.965	Others 0.974	0.958
5.	Nonmatches from conflicting households; for housing units not in regular nonresponse follow-up	0.975		0.965
6.	Nonmatches from conflicting households; housing units in regular nonresponse follow-up	0.914		0.926
7.	Whole household nonmatches, where the housing did not match in housing-unit matching	Nonhispanic White 0.959	Others 0.947	0.950
8.	Resolved before follow-up	Nonhispanic White 0.995	Others 0.990	0.979
9.	Insufficient information for matching	0 (assigned by definition)		
10.	Targeted extended search cases[2]	0.928		0.858
11.	Potentially fictitious people	0.058		0.088
12.	People said to be living elsewhere on Census Day	0.229		0.210

[1] Among the characteristics age, sex, tenure, race, or Hispanic origin.

[2] Targeted extended search refers to a field operation conducted to reduce the variance in the dual-system estimates caused by clustered geocoding errors. For more information, see Navarro and Olson (2001).

As with resident status for *P*-Sample people, a key factor in determining enumeration status was the *E*-Sample person's match code group, although the match code groups were defined differently for the two samples. Similar to the *P*-Sample, people coded as potentially fictitious or said to be living elsewhere on Census Day during the follow-up operation were first placed in groups 11 or 12, respectively. The remainder of the *E*-Sample was then placed in the appropriate match code group, as defined in the table. Group 3 was split into two subgroups, as when determining resident status in the *P*-Sample. That is, people aged 18 to 29 who were children of the reference person were separated. Other characteristics used to define cells were the presence or absence of imputed characteristics, as defined in the imputation cells for match status; and whether the person was Nonhispanic white or any other race-ethnicity combination. It should be noted that, according to A.C.E. procedures, anyone in the *E*-Sample with insufficient information for matching (group 9) was automatically assigned an enumeration probability of 0.

4.3 Comparing Probabilities Under Cell Imputation and Logistic Regression

It can be insightful to compare the probabilities assigned to cases with unresolved status under alternative procedures. Belin (2001) presents such a comparison under a logistic regression model that considered 186 predictors for resident and match status, and 202 predictors for enumeration status. The variables included most of those used in the cell estimation described in section 4.2, as well as individual demographic characteristics, such as age, gender, and relationship to the household's reference person; information about the A.C.E. interview, such as whether the respondent was a proxy; information derived from the sampling process; local-area features, such as whether the area was urban or non-urban; and the interactions among the variables. As the models were fit to the resolved cases sent to follow-up, and then applied to unresolved cases to predict a probability, the models are ignorable in the sense that unresolved status is not considered as a covariate in the underlying model. (See Rubin 1976.)

Tables 7 and 8 summarize the probabilities assigned to unresolved cases under A.C.E. cell imputation and the logistic modeling averaged over the different match code groups. Recall that cell imputation probabilities were computed from weighted data as in (3), while the logistic regression models were run on unweighted data. The predicted probabilities for the two procedures were averaged across all unresolved people unweighted. With an exception to be mentioned later, probabilities and estimates in the A.C.E. were typically similar when using unweighted and weighted data, as the sample was designed to avoid a wide range of weights.

Table 7
Average Resident and Match Probabilities Under Cell Imputation and Logistic Regression

P-Sample / Match Code Group	Resident Status			Match Status		
		Avg. Probability Assigned			Avg. Probability Assigned	
	Number Unresolved	Cell Imputation	Logistic Regression	Number Unresolved	Cell Imputation	Logistic Regression
1. Matches needing follow-up	767	0.989	0.983	4	0.848	0.941
2. Possible matches	352	0.970	0.962	131	0.889	0.837
3. Partial household nonmatches	1,306	0.956	0.951	71	0.893	0.050
4. Whole household nonmatches	1,610	0.917	0.926	36	0.770	0.010
5. Nonmatches, conflicting household	1,455	0.940	0.927	49	0.616	0.070
6. Resolved before follow-up	129	0.990	0.990	23	0.842	0.940
7. Insufficient information	7,506	0.844	0.851	7,506	0.835	0.880
8. Fictitious, living elsewhere	2,402	0.148	0.167	6	0.655	0.041

Table 8
Average Enumeration Probabilities Under Cell Imputation and Logistic Regression

E-Sample / Match Code Group	Enumeration Status		
		Avg. Probability Assigned	
	Number Unresolved	Cell Imputation	Logistic Regression
1. Matches needing follow-up	711	0.977	0.986
2. Possible matches	305	0.968	0.967
3. Partial household nonmatches	2,191	0.962	0.963
4. Whole household nonmatches where the housing unit matched; not conflicting	4,813	0.967	0.974
5. Nonmatches from conflicting households; housing units not in nonresponse follow-up	532	0.973	0.973
6. Nonmatches from conflicting households; housing units in nonresponse follow-up	779	0.917	0.926
7. Whole household nonmatches, where the housing unit did not match	3,881	0.954	0.961
8. Resolved before follow-up	179	0.990	0.982
9. Insufficient information for matching	0	----	----
10 Targeted extended search cases	2,902	0.918	0.679
11. Potentially fictitious people	1,690	0.064	0.077
12. People said to be living elsewhere on Census Day	3,152	0.225	0.280

Comparing procedures, one sees almost no difference in the average probabilities assigned for resident status. This is not surprising, as cell imputation used the match code group (among other variables) to define cells. Match status presents a different story. To recall, match code group was not used in the cell imputation, as almost all unresolved matches (98.3% of 7,826; 7,506 before the follow-up operation, and 187 more after follow-up) had insufficient information for matching. The first two groups have slightly different probabilities assigned under the two procedures. But in groups 3, 4, and 5, all nonmatches before follow-up, the average probabilities are high under cell imputation (0.893, 0.770, and 0.616), and very low under logistic regression (0.050, 0.010, and 0.070). Of the 156 cases in the three cells, 134 were people each of whom was given an initial code indicating a "nonmatch"; later it was determined correctly that the person had insufficient information for matching. In almost every case, the A.C.E. interviewer recorded a name like "Child Jones", "José Don't Know", or "Unknown Smith". Such cases should have been caught before matching by a clerk, and assigned an initial code of insufficient information. Instead, a match to the census was attempted and failed. If not for this error, such people would have been placed in group 7, where their match probability under logistic regression would have been much higher. Thus, for this small set of 134 cases, the logistic variable, match code group, takes on an incorrect value, and the model predicts a probability – much too low – based on the many resolved cases in group 3, 4, or *5 who really were nonmatches,* but were sent to follow-up primarily to resolve their resident status, not their match status.

The predicted match probabilities in group 8 were also very different. However, with only six unresolved cases, the effect on estimation would be minimal.

Comparing average enumeration probabilities by match code group in Table 8, one sees almost no difference except in group 10, targeted extended search cases. There, the average probability assigned by cell imputation, 0.918, is much higher than that predicted by logistic regression, 0.679. The difference can be explained by the weighting. In the *E*-Sample, of 32,334 people eligible for the targeted extended search operation, 8,298 (all in match code group 10) were sampled out to contain costs and given an A.C.E. weight of 0. The matching operation did not try to determine whether the 8,298 cases were enumerated correctly or not, but simply left them on the data file as erroneous enumerations. Probabilities based on cell imputation were assigned as in equations (3) and (4), incorporating the A.C.E. weight. This removed from the computation those who were sampled out of the A.C.E. The logistic regression model was run on unweighted data and included the 8,298 cases in group 10, bringing down the probability of a correct enumeration predicted for the 2,902 people with unresolved enumeration status.

5. THE EFFECT OF SOME ALTERNATIVE MISSING DATA PROCEDURES ON DUAL-SYSTEM ESTIMATES

In the last section, predicted probabilities were compared across two options for treating cases with unresolved status. But the ultimate effect of competing procedures is seen in the resulting dual-system estimates. In this section, several alternatives to those used in the A.C.E. for addressing missing data are compared via the resulting estimates. When they differ significantly, it is not clear which procedure is to be preferred. It should be noted that the A.C.E. estimates released by the U.S. Census Bureau in March of 2001 have been revised following further analyses (Haines 2003). Even though the A.C.E. data are flawed and A.C.E. estimates should generally not be used, it is believed that they are adequate to evaluate the differences in the estimates caused by alternative missing data approaches.

5.1 Results from an Early Evaluation

In the months after initial dual-system estimates from the A.C.E. were released, alternatives to the applied missing data procedures were studied. There were several reasons: estimating the variation that might result from the alternatives, incorporating this variation into total error and loss function analysis for the A.C.E. dual-system estimates, and investigating the viability of non-ignorable missingness procedures for addressing unresolved status. As the results are available in Keathley, Kearney, and Bell (2001), only a summary will be provided here.

Three alternatives involving the noninterview adjustment were examined. The first defined cells differently for the adjustment, adding variables such as race, Hispanic origin, tenure, and household size, as determined from a match to the census file. This procedure tended to produce larger dual-system estimates. Two other noninterview alternatives had no apparent affect on the estimates. In one, a nearest-neighbor noninterview adjustment, the weight of a non-interviewed household was added to that of the nearest interviewed household in the sorted file. In the second, the last 30% of A.C.E. interviews completed were labeled as "late" interviews. The weights of noninterviewed units were added only to the weights of late interviews. These alternatives tried to take advantage of the anticipated homogeneity of units induced by geographic proximity or time of response to the A.C.E.

The other alternatives described in Keathley *et al.* (2001) address unresolved resident, match, or enumeration status. A "late" data approach used information collected only from the last 30% of interviews in the *P*-Sample, or housing units that required nonresponse follow-up in the *E*-Sample. By itself, this approach did not appear to affect the dual-system estimates. The remaining alternatives involved logistic regression models to predict probabilities for

unresolved cases. First, an ignorable logistic model, the one described above (Section 4.3) in Belin (2001), was applied to unresolved resident, match, and enumeration status and tended to produce smaller dual-system estimates (47,481 smaller for the U.S. total). However, it appears that the lowered (on average) enumeration probabilities assigned to the 2902 unresolved cases in the *E*-Sample match code group 10 (see section 4.3) would have more than accounted for this decrease.

Perhaps more interesting are three alternatives that attempted to construct non-ignorable logistic models by lowering the probabilities assigned to unresolved cases, on the premise that ignorable models may overstate the underlying probabilities (Belin 2001). Data from the 1990 Post-Enumeration Survey and its evaluation follow-up were used to estimate non-ignorable effects and incorporate them into the 2000 logistic models. This strategy tended to produce larger dual-system estimates when applied to unresolved match probabilities, and smaller estimates when applied to resident or enumeration probabilities. This result is not surprising, based on equation (1) and the fact that the average match probability assigned to cases with unresolved resident status is less than that for cases with resolved resident status. Although the study's authors conclude that "[t]here is no evidence to suggest that the non-ignorable missingness procedures that we considered are or are not viable alternative missing data procedures" (Keathley *et al.* 2001, page 2), Belin's approach takes a promising step toward addressing the non-ignorability of the missing status.

5.2 Analyses on Other Alternative Procedures

In this section, differences in the dual-system estimates are presented under six numbered alternatives described and motivated below. The results are provided in Table 9 for the U.S. total and for breakdowns by race-ethnicity, tenure, and age. For a precise definition of the race-ethnicity domains, see Kostanich (2001). (Note that a small part of the U.S. population was not part of the A.C.E. universe.) For each alternative, the three numbers given are (a) the difference: the alternative estimate minus the A.C.E. estimate; (b) the standard error of that difference; and (c) the percent relative difference.

Alternative (1) reconsidered the noninterview procedure as applied in the A.C.E. to adjustment cells with a relatively small number of completed interviews. (See section 3.) In this alternative, instead of spreading weights from non-interviewed units over a wider range of cells, cells with too few interviews were collapsed with nearby cells, and noninterview adjustment factors were computed afresh in the newly created cells. Except for Nonhispanic Blacks, none of the estimated differences in Table 9 under this alternative are statistically significant (greater than two standard errors). Similarly, except for several race-ethnicity domains less than two million in size, none of the relative differences are greater than 0.01%.

Alternatives (2), (3), and (4) were derived after examining the effects of the variables used in the imputation cells on the resulting assigned probabilities. From the probabilities assigned in Tables 4 and 6, it is clear that the match code groups discriminated well with regard to resident and enumeration status. Yet it appears that dividing the cells based on demographic variables, such as "Nonhispanic white" *vs.* "Other," made less of a difference. To investigate the effect of demographic variables on the imputation, new probabilities were assigned for unresolved status without using them. Specifically, all resolved and unresolved cases were combined across cells for Nonhispanic white and Other (resident and enumeration status), for match code groups 3a and 3b (resident and enumeration), and for "No imputed characteristics" and "1 or more imputed characteristics" (match and enumeration); the variables derived from A.C.E. operations – match code group, housing-unit address match code, and mover status – were retained. Alternative (2) applies the smaller set of cells only in the *P*-Sample, that is, only for unresolved resident and match status; alternative (3) applies it only in the *E*-Sample (enumeration status); and alternative (4) applies it to both samples.

Under alternative (2), the greatest change in the resident probabilities assigned to unresolved cases occured in the four (original) imputation cells in group 3a, affecting only 96 people with unresolved status. In most other cells for resident status (over 99% of the cases), the probabilities changed very little. A large difference in match probabilities occurred only in the cell "non-mover, nonmatched unit or conflicting household, one or more imputed characteristics," containing 421 unresolved cases. The variable differentiating the number of imputes appears to have had an effect here; if its two "impute" subcells are collapsed, the probability assigned to the "one or more" cell is dominated by the much larger number of resolved people with no imputes, raising the value from 0.567 to 0.684. As is seen in Table 9, the effect on the dual-system estimates is statistically significant for the U.S. total and almost all the breakdowns shown, except for two race-ethnicity groups with sizes under one million people. The relative differences do not appear to be very large, however, ranging from 0.01% to 0.04%. It is not obvious which missing data option produces estimates closer to the unknown true values.

Table 9
Dual-System Estimates Under Alternative Missing Data Procedures

Each cell to the right of the vertical bar contains, in order, estimates of (a) the difference: the alternative estimate minus the A.C.E. estimate, (b) the standard error of that difference, and (c) the relative difference as a percent.

		Estimated Differences Based on Six Alternatives to A.C.E. Missing Data Procedures					
	A.C.E. Estimate (Standard Error)	(1) Noninterview Adjustment With Collapsed Cells	(2) Collapsed Imputation Cells: *P*-Sample Only	(3) Collapsed Imputation Cells: *E*-Sample Only	(4) Collapsed Imputation Cells: *P* and *E*-Samples	(5) Imputing Probabilities Based on the MES	(6) Imputing Probabilities Based on the MER
U.S. Total	276,848,873	-4,299	-55,284	-568	-55,852	-63,632	385,969
	(366,543)	(7,423)	(1,623)	(2,581)	(3,045)	(5,368)	(24,358)
		0.00%	-0.02%	0.00%	-0.02%	-0.02%	0.14%
Race-Ethnicity Domains							
Nonhispanic White	194,226,285	-2,467	-32,324	-1,677	-34,000	-61,817	108,604
	(265,893)	(6,247)	(1,055)	(1,870)	(2,163)	(4,534)	(13,026)
		0.00%	-0.02%	0.00%	-0.02%	-0.03%	0.06%
Nonhispanic Black	34,210,774	-3,495	-11,136	-119	-11,255	-1,303	124,710
	(118,415)	(1,290)	(753)	(1,328)	(1,528)	(1,417)	(11,343)
		-0.01%	-0.03%	0.00%	-0.03%	0.00%	0.36%
Hispanic	35,552,109	725	-8,132	1,432	-6,700	196	124,937
	(138,870)	(3,016)	(857)	(973)	(1,297)	(1,577)	(10,657)
		0.00%	-0.02%	0.00%	-0.02%	0.00%	0.35%
Native Hawaiian or Pacific Islander	618,698	-98	-73	88	15	-107	1,330
	(17,873)	(81)	(72)	(43)	(85)	(74)	(616)
		-0.02%	-0.01%	0.01%	0.00%	-0.02%	0.22%
Nonhispanic Asian	10,056,009	709	-3,175	-257	-3,431	-414	19,556
	(64,372)	(571)	(356)	(439)	(567)	(576)	(3,704)
		0.01%	-0.03%	0.00%	-0.03%	0.00%	0.19%
American Indian on Reservation	567,053	-245	-59	61	2	-38	1,402
	(7,235)	(300)	(49)	(17)	(52)	(73)	(250)
		-0.04%	-0.01%	0.01%	0.00%	-0.01%	0.25%
American Indian *not* on Reservation	1,617,944	572	-386	-96	-482	-148	5,430
	(22,032)	(661)	(68)	(174)	(186)	(144)	(1,446)
		0.04%	-0.02%	-0.01%	-0.03%	-0.01%	0.34%
Tenure							
Owner	188,764,543	-2,237	-34,503	933	-33,570	-7,816	125,058
	(260,408)	(3,805)	(1,205)	(1,971)	(2,317)	(1,942)	(10,063)
		0.00%	-0.02%	0.00%	-0.02%	0.00%	0.07%
Non-Owner	88,084,330	-2,063	-20,782	-1,501	-22,282	-55,816	260,911
	(226,108)	(6,057)	(1,121)	(1,607)	(1,935)	(5,071)	(21,684)
		0.00%	-0.02%	0.00%	-0.03%	-0.06%	0.30%
Age Group							
0 - 17	73,076,071	2,924	-21,872	-3,315	-25,186	-8,559	107,308
	(137,126)	(2,624)	(625)	(1,324)	(1,474)	(2,047)	(9,785)
		0.00%	-0.03%	0.00%	-0.03%	-0.01%	0.15%
18 - 49	129,785,393	-2,721	-23,304	3,247	-20,057	-44,534	244,070
	(208,070)	(4,714)	(1,143)	(1,565)	(1,930)	(3,777)	(16,245)
		0.00%	-0.02%	0.00%	-0.02%	-0.03%	0.19%
50 and Over	73,987,409	-4,502	-10,108	-500	-10,608	-10,538	34,591
	(111,125)	(2,766)	(563)	(670)	(877)	(1,421)	(4,561)
		-0.01%	-0.01%	0.00%	-0.01%	-0.01%	0.05%

Under alternative (3), the enumeration probabilities were re-computed using only the match code groups as imputation cells. Noticeable changes were detected in the probabilities in the (original) cells for match code group 3a. In the dual-system estimates, the only significant differences were found in two of the three age categories and some of the small race-ethnicity domains. Except for the latter domains, all the percent differences were under 0.01%. As alternative (4) uses the re-computed probabilities from the *P* and *E*-Samples, the resulting estimates here were dominated by the *P*-Sample results and thus were similar to those under alternative (2).

The final two alternative procedures employed the same set of imputation cells as those used in the A.C.E., but assigned to unresolved cases in both the *P* and *E*-Samples potentially improved probabilities, as determined from one of two evaluations conducted by the Census Bureau following the A.C.E. Alternative (5) secured its probabilities from the Matching Error Study (MES), while alternative (6) based them on the Measurement Error Reinterview (MER). Each study took place in a set of evaluation clusters, a roughly one-in-five subsample of the A.C.E. sample block clusters. Information on the MES and MER sample designs can be found in Petroni (2001) and Killion (2000).

The primary purpose of the MES was to evaluate the A.C.E. person matching operation. The evaluation clusters were rematched by expert matchers, and appropriate changes were made to final match codes and person status. No additional data were collected for the MES. Imputation cell probabilities based on MES data were generally similar to those assigned in the A.C.E. One exception, for resident status, was in the cell for match code group 4, Nonhispanic white, non-owner. Here, the MES probability, 0.712, was much lower than the A.C.E. value of 0.911. This resulted from one cluster in the cell that had 24 persons with large weights geocoded incorrectly, as detected in the MES. The MES enumeration probability for match code group 11, "1 or more imputed characteristics," 0.176, was a bit higher than that for the A.C.E., 0.088. Most other probabilities for resident, match, and enumeration status were close (within 0.03) between the A.C.E. and MES; all others were within 0.07.

In contrast, the MER was designed to evaluate the *data collection* error arising from the A.C.E. matching process. People in the MER were reinterviewed about nine months after Census Day to collect information analogous to that collected in the A.C.E. follow-up operation, but in greater detail. Based on the MER, resident probabilities tended to be substantially higher for the cells in match code group 8, but to be lower for the cells in groups 3, 4, and 5 (denoting nonmatches). The reductions tended to be larger in cells where group 8 took more cases away from groups 3, 4, and 5. One might note that the MER cells in subgroup 3a were fairly small. The "Nonhispanic white, non-owner" cell had only 34 unweighted resolved persons, while the other three cells in group 3a ranged from 125 to 140 unweighted resolved persons. The MER probabilities for enumeration status exhibited similar behavior, with probabilities in groups 11 and 12 raised, and those in the nonmatch groups (3 through 7) lowered. Match probabilities were similar between A.C.E. and MER, mostly differing by 0.01 to 0.05.

Before looking at the dual-system estimates under alternatives (5) (MES probabilities) and (6) (MER probabilities), one should note that, *for the comparison in Table 9*, only the probabilities assigned to unresolved cases were changed based on data collected through the MES or MER. Although the evaluated status of some people may have changed (for example, from nonmatch to match, or confirmed resident to unresolved resident) based on the evaluations, their status was not changed when computing these estimates, as the goal of this exercise was only to explore different methods or information *as they affect the missing data procedures component* of the dual-system estimates.

Under alternative (5), based on MES data and probabilities, the estimates decreased in almost all population domains in Table 9, although never more than 0.1%. Yet this decrease can be attributed almost exclusively to the domain Nonhispanic White. With alternative (6) based on MER data and probabilities, there were significant increases in the estimates of every domain. The relative differences under alternative (6) are larger in magnitude than for earlier alternatives, but all have an absolute magnitude of less than 0.4%. There are several relative differences greater than 0.3% in absolute value: for Nonhispanic Black, Hispanic, and American Indian not on Reservation.

6. OBSERVATIONS

The observations given here pertain to the third type of missing data, assigning probabilities to unresolved people in the A.C.E. It is important to note that the A.C.E. procedures were specified well before the conduct of the census and the A.C.E. The early deadlines were due to (1) the very tight schedule coordinating many separate but interrelated activities, and (2) the need for a process open to the scrutiny of policy makers as well as statistical experts. Although one can learn much about the missing data and the relevant correlation structures by examining the responses as they are collected, making decisions after seeing the data might have been construed as manipulating the results of an operation that had serious political implications.

In this light, one can look back and realize various ways to improve the process – too late to change the procedures.

This does not imply that we did not react to information made available unexpectedly during the processing of the data. We knew that the post-match follow-up operation would help resolve some cases, especially those whose true residence on Census Day was uncertain. Much other information was collected in these interviews, but we did not anticipate seeing the details. However, due to an intensive keying of the follow-up interview forms at the Bureau's processing center, some additional information was made available during the missing data operation. At that time, we added several match code groups not originally in the plan: group 8 for resident status; 11 and 12 for enumeration status. Separating the people in these groups allowed us to assign probabilities that were quite different – and, we believe, more accurate – from what they would have received.

Different models, imputation cells, or data could have been used to assign probabilities for unresolved cases. The values determined through logistic regression were quite similar on average, and may or may not have had an effect on the resulting population estimates. In section 5 it was shown that ignoring some of the demographic variables would have made a difference in the match rate, but probably not in the rate of correct enumeration. Basing the probabilities on data collected in the Matching Error Study or the Measurement Error Reinterview (not yet available during the A.C.E.) could have made a larger difference still. But it is unclear which one might have made an improvement; using MES data would have lowered the population estimates, while using MER data would have increased them.

Weighing the various results, one is constantly reminded that, when assigning probabilities to people with unresolved status, match code group was the most important variable. It worked well for resident and enumeration status, but could not be effectively used for match status. The problem there – perhaps the biggest hole in our procedures – is once again that almost all of the unresolved matches, and over half of the unresolved residents, were people with insufficient information for matching. Little information was collected on these cases, and almost all of them were not sent through the matching process or follow-up. Further, almost none of these people were included in any post-A.C.E. evaluations. In future tests a concerted effort should be made to obtain real information about the status of such people.

ACKNOWLEDGEMENTS

The authors thank Eric Schindler and Doug Olson for computing dual-system estimates and their standard errors under alternative procedures; Tom Belin, UCLA, for making available imputation probabilities under logistic regression models; and Mary Frances Zelenak and Ha Nguyen for compiling summaries of the extent of missing data in the A.C.E. This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U. S. Census Bureau.

REFERENCES

ANDERSON, M.J., and FIENBERG, S.E. (1999). *Who Counts? The Politics of Census-Taking in Contemporary America*. New York: The Russell Sage Foundation.

BELIN, T. (2001). Evaluation of unresolved enumeration status in 2000 Census Accuracy and Coverage Evaluation program. Unpublished report, prepared by Datametrics, Inc., for the U.S. Census Bureau.

BELIN, T., DIFFENDAL, G., MACK, S., RUBIN, D., SCHAFER, J. and ZASLAVSKY A. (1993). Hierarchical logistic regression models for imputation of unresolved enumeration status in undercount estimation. *Journal of the American Statistical Association*. 88, 1149-1166.

CANTWELL, P.J. (2001). Accuracy and Coverage Evaluation Survey: Specifications for the missing data procedures. *DSSD Census 2000 Procedures and Operations Memorandum Series*. Q-62.

CANTWELL, P.J., MCGRATH, D., NGUYEN, N. and ZELENAK, M.F. (2001). Accuracy and Coverage Evaluation: missing data results. *DSSD Census 2000 Procedures and Operations Memorandum Series*. B-7*.

CHILDERS, D. (2000). The Design of the Census 2000 Accuracy and Coverage Evaluation. DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-DT-1.

FENSTERMAKER, D. (2000). The Accuracy And Coverage Evaluation: sample design summary. *DSSD Census 2000 Procedures and Operations Memorandum Series*. R-33.

HAINES, D. (2003). A.C.E. Revision II results: changes in estimated net undercount. *DSSD A.C.E. Revision II Memorandum Series*. PP-58

HOGAN, H. (1993). The Post-Enumeration Survey: Operations and results. *Journal of American Statistical Association*. 88, 1047-1060.

HOGAN, H. (2003). The Accuracy and Coverage Evaluation: Theory and design. *Survey Methodology*. 29, 129-138.

IKEDA, M., KEARNEY, A. and PETRONI, R. (1998). Missing data procedures in the Census 2000 Dress Rehearsal Integrated Coverage Measurement sample. *Proceedings of the Survey Research Methods Section*, American Statistical Association. 617-622.

KEARNEY, A., and IKEDA, M. (1999). Handling of missing data in the Census 2000 Dress Rehearsal Integrated coverage measurement sample. *Proceedings of the Survey Research Section*, American Statistical Association. 468-473.

KEATHLEY, D., KEARNEY, A. and BELL, W. (2001). ESCAP II, Analysis of missing data alternatives for the Accuracy and Coverage Evaluation. Executive Steering Committee for A.C.E. Policy II (ESCAP II) Report 12.

KILLION, R.A. (2000). Measurement Error Reinterview Sample Selection. *Planning, Research, and Evaluation Division TXE/2010 Memorandum Series.* CM-MER-S-01.

KOSTANICH, D. (2001). Accuracy and Coverage Evaluation Survey: computer specifications for Person Dual System Estimation (U.S.) - Re-issue of Q-29. *DSSD Census 2000 Procedures and Operations Memorandum Series.* Q-37.

NAVARRO, A., and OLSON, D. (2001). Accuracy and Coverage Evaluation: effect of targeted extended search. *DSSD Census 2000 Procedures and Operations Memorandum Series.* B-18*.

PETERSON, C.G.J. (1896). The yearly immigration of young plaice into the Limfjord from the German Sea. *Report of the Danish Biological Station.* 6. 1-48.

PETRONI, R. (1997). Effect of using the 1996 ICM characteristic imputation and probability modeling methodology on the 1995 ICM *P* and *E*-sample data. *DSSD Census 2000 Dress Rehearsal Memorandum Series.* A-20.

PETRONI, R. (1998a). Effect of different methods for calculating match and residence probabilities for the 1995 P-sample data. *DSSD Census 2000 Dress Rehearsal Memorandum Series.* A-23.

PETRONI, R. (1998b). Effect of different methods for calculating correct enumeration probabilities for the 1995 *E*-sample data. *DSSD Census 2000 Dress Rehearsal Memorandum Series.* A-28.

PETRONI, R. (1998c). Effect of using simple ratio methods to calculate P-sample residence probabilities and E-sample correct enumeration probabilities for the 1995 data. *DSSD Census 2000 Dress Rehearsal Memorandum Series.* A-30.

PETRONI, R. (2001). EFU Sample Design, Stratification, Selection, and Weighting. Planning, Research, and Evaluation Division *TXE/2010 Memorandum Series.* CM-GES-S-02-R2.

RUBIN, D.B. (1976). Inference and Missing Data. *Biometrika. 63,* 581-592.

SEKAR, C.C. and DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association.* 44, 101-115.

Survey Methodology, December 2003
Vol. 29, No. 2, pp. 155-165
Statistics Canada

Coverage Error in Population Censuses: The Case of Turkey

H. ÖZTAŞ AYHAN and SÜHENDAN EKNI[1]

ABSTRACT

Coverage errors and other coverage issues related to the population censuses are examined in the light of the recent literature. Especially, when the actual population census count of persons are matched with their corresponding post enumeration survey counts, the aggregated results in a dual record system setting can provide some coverage error statistics. In this paper, the coverage error issues are evaluated and alternative solutions are discussed in the light of the results from the latest Population Census of Turkey. By using the Census and post enumeration survey data, regional comparison of census coverage was also made and has shown greater variability among regions. Some methodological remarks are also made on the possible improvements on the current enumeration procedures.

KEY WORDS: Census coverage error; Coverage error measures; Coverage error estimation; Dual record system estimate; Population census; Post enumeration survey.

1. INTRODUCTION

Coverage has been an important issue in censuses as well as in sample surveys. The difference between the census count and the target population count is the coverage error. When the census count is less than the target population count, it creates an undercount as is common in many countries.

Several techniques are available to understand the problem of coverage errors in censuses. Dual record system (DRS) estimator (Chandra Sekar and Deming 1949) was also extended by many researchers (Ayhan 2000; Casady, Nathan and Sirken 1985; Hogan 1990, 1993a and 1993b; Isaki 1992; Marks, Seltzer and Krotki 1974).

Dual record system estimates based on the census enumeration and a post enumeration survey (PES) are used by the U.S. Census Bureau to measure census coverage error (Hogan 1993a and 1993b; Mulry and Spencer 1988, 1990 and 1993). Post enumeration surveys can be used to improve the population estimate (Ayhan and Ekni 1991; Diffendal 1988; Hogan 1990; Hogan and Wolter 1988).

For the United States, the 1980 Census Post Enumeration Program attempted to measure census coverage through direct measurement using sample survey models (Fay, Passel, Robinson and Cowan 1988). Several methods are also proposed for the adjustment of census count for under enumeration (Choi, Steel and Skinner 1988; Cressie 1988 and 1990).

Recently, models for population coverage error have been studied extensively (Isaki 1992; Wolter 1986). A method of overlapping data systems or multiple frame methodology was used to improve the population estimates (Goodman 1949; Hartley 1962 and 1974; Bankier 1986).

This study highlights the methodological problems related to the population census coverage and proposes remedies to some of the issues covered. In addition, it proposes and discusses alternative estimates for the population census coverage errors. To achieve the above goals, coverage evaluation issues are included in the design of the PES.

In this paper, section 2 discusses methods of census enumeration and section 3 covers post enumeration survey procedures. Methods of coverage error estimation is presented in section 4 of the paper. Estimators of the population total is given in section 5 and comparison of the coverage error statistics are presented in section 6. Important findings are summarized in the conclusion.

2. METHODS OF CENSUS ENUMERATION

Population censuses have many common features in most countries. The method of enumeration can either be based on *de facto* or *de jure* system. In *de jure* system people are enumerated at their normal residence, while *de facto* system enumerates people actually there. *De facto* system is widely used in developing countries, and the *de jure* system is generally used in developed countries. Countries that are using *de facto* system of enumeration seem to have more problems related to coverage, than the countries which are using *de jure* system of enumeration. These problems stem mainly from their existing imperfect frames for their target population.

De facto based population censuses are generally conducted on a single day, as a complete count, to determine the total population within the country on the day of enumeration. The citizens of the country who are living outside the country were excluded from the census, whereas alien population who are present within the country were included in the census.

[1] H. Öztaş Ayhan, Department of Statistics, Middle East Technical University, 06531 Ankara, Turkey; Sühendan Ekni, State Institute of Statistics, 06100 Ankara, Turkey.

Design of the Census Operations

The Population Census of Turkey was conducted, on the basis of *de facto* system of enumeration, by the State Institute of Statistic (SIS) to determine the quantitative, social and economic characteristics of the population.

For the Census enumeration, the *list of buildings* are created by the local authorities and send to the local Census Committees. Based on the list of buildings (Forms 1 or 2), *enumeration districts* (EDs) *list of buildings* (Form C) are formed (see Appendix 1 for details). Due to the lack of timely availability of the complete list of buildings in the SIS Central Office, the number of EDs are estimated by projection techniques for advance fieldwork planning of the Census, as well as PES. EDs are obtained by assigning 100 persons per enumerator in province and district centers, and 200 persons per enumerator in sub-districts and villages, based on average daily workloads. They are then numbered sequentially. In the Census, the listed addresses were taken as the base for identifying the "dwelling units (DUs)", while the "individual persons" within the household(s) of the dwelling unit is considered as the unit of enumeration.

The workload of each enumerator is taken as an ED, which contains a list of addresses to be covered within a specified close interval. Instructions are given to the enumerator to treat this interval as a compact segment. If an enumerator encountered an address not on the list, it is included in the enumeration by work definition. For vacant and nonexistent units the related information is also recorded. There was no special procedure for dealing with reluctant respondents or in general any non-interviewed units, due to the compulsory nature of response by the related Act. The enumerator's workload is set in such a way that, they will complete all the interviews in a given day. For very special cases, the instruction is given to complete the enumeration of the segment during the extended hours in the same census day.

Additional enumerators were assigned to enumerate the "*special enumeration districts*" such as the places of the mobile populations (travelers, persons on duty, nomadic tribes, *etc.*) and institutional populations (hospitals, prisons, factories, military establishments, *etc*).

Institutional population are covered by additional enumerators, who are assigned for these special EDs. The mobile populations travelling by vehicles are stopped and were enumerated as a group when they first appeared within the borders of the provinces. The passengers continues their journey after enumeration and duplicate enumerations are avoided by placing an "*Enumerated*" sign on the vehicle after the census operation, and later their individual identification is also checked by manual and computerized algorithms, against the other records of the relevant settlement.

The Census was conducted on a Sunday, and the enumeration was completed on the same day. On the Census day, a national curfew was declared. The enumerators visited each household (HH) within the dwelling units listed in their enumeration district building lists and completed the census questionnaire (see Appendix 2 for details). For the *Household Module*, the information is collected from an adult household member for the general household characteristics, while for the *Individual Person's Module* the information is obtained from self respondents on their personal characteristics.

The following type of errors occurred during the different stages of the census operations;

(1) Omission errors and erroneous inclusions has occurred during the construction of the List of Buildings. However, due to the use of compact segment approach in the enumeration process of the census operation, these errors are mostly eliminated.

(2) Response errors based on memory recall error, cheating, and inadequate answer for coding has occurred during the census enumeration. These are measured as the response inconsistency in the *Response Reliability Study* (Ayhan and Ekni 1991; SIS 1994) of the Population Census, which was based on the PES.

(3) Some enumerator errors (failure to probe, inadequate perception of response, and recording errors) are also observed during the census operation. These are also covered by the *Response Reliability Study*.

(4) Processing errors such as, coder and verifier errors also occur during the data processing and these are eliminated later during the data handling in the office.

3. POST ENUMERATION SURVEY PROCEDURES

The objectives of the PES are to determine coverage error in the population census as well as obtain measures of response reliability of the questions in the census. In this paper, the first objective is discussed for the Population Census of Turkey, and the preliminary findings for both objectives are summarized by Ayhan and Ekni (1991).

3.1 Sample Selection Procedures

The sample design for the PES is initiated 3 months before the Census operation. At this stage, creation of the Population Census EDs was not complete yet.

Stratification and estimation of population EDs. The previous Population Census enumeration district lists of the State Institute of Statistics is used as the base for sampling frame for PES operation. The population of people is first stratified into 5 *geographical-socio economical regions* of Turkey. A second explicit stratification variable is also used, which is based on the 8 *size groups for the place of the settlements*, in a nested structure within regions. Here,

urban-rural boundary corresponds to a population size of 10,000. The number of census enumeration districts were estimated for 40 *design strata* for the Census day by using forward population projection method, which were based on the person counts of the two previous population censuses. For the census enumeration, EDs were created by using Form C within the Central Office. A total of 479 251 EDs are established for the Census. Sampling frame information is given in Table 1.

Table 1
Estimated Number of Population and Sample EDs by Regions and Urban-Rural Strata

	URBAN		RURAL		TOTAL	
Region	Popu. ED	Samp. ED	Popu. ED	Samp. ED	Popu. ED	Samp. ED
h	$M_h^{(U)}$	$m_h^{(U)}$	$M_h^{(R)}$	$m_h^{(R)}$	M_h	$m_h^{(1)}$
1	125,726	125	40,333	32	166,059	157
2	42,442	42	24,992	20	67,434	62
3	65,466	76	45,925	36	111,391	112
4	15,790	16	30,459	22	46,249	38
5	39,358	40	48,760	34	88,118	74
Total	288,782	299	190,469	144	479,251	443

The expansion factors: $F_h^{(1)} = N_h^{(1)}/n_h^{(1)} \neq M_h/m_h^{(1)} = F_h^{(2)}$

The coverage of the number of dwelling units in the Census and PES were achieved by the following procedures. The number of population EDs for each province was determined and numbered sequentially. Then, the number of population EDs in each population strata was estimated by, dividing the projected strata population ($N_{h.}$) to the fixed daily workload of enumerators ($B_{h.}$). Population EDs were estimated for urban areas as $M_{hi} = N_{hi}/B_{hi}$ and for rural areas as $M_{hj} = N_{hj}/B_{hj}$, where the ED sizes are taken as fixed daily workload, $B_{hi} = 100$ persons in urban strata and $B_{hj} = 200$ persons in rural strata. The results of the population projections for each strata by urban-rural aggregation are also obtained. The estimated number of population EDs and expansion factors for regions and urban-rural strata are also computed.

Selection of sample EDs. A stratified multistage sample of localities and blocks are selected systematically for PES from the available *master sampling frame* of the State Institute of Statistics at the Central Office. The blocks of the master sampling frame is periodically updated for the multi purpose selection of other samples on routine basis. The interviewers of the PES is recruited and trained in the Central Office, and then interviewers are send as a team to the local sample settlements for the independent enumeration of the selected PES sample. For the identification purpose, the selected sample blocks are linked to their corresponding Population Census EDs of the settlement in the field by previously given instructions to the PES interviewers.

For the use of Dual Record System estimation, the sample enumeration districts for the PES should be determined independently from the census frame. This is an absolutely crucial assumption of the DRS model, which was emphasised by many researchers during the past 50 years (Ayhan 2000; Chandra *et al.* 1949).

Due to the use of unwanted old ED lists in some areas, the range of the planned workload per ED per enumerator may have changed and consequently the selected sizes of the EDs may be different from the actually enumerated sizes. This will effect the achieved sampling fractions, which will naturally be different from the selected.

A total of $m = 443$ sample enumeration districts are selected in 16 province centers, 23 districts, 16 sub-districts and 43 villages within the 40 strata. For the PES, a sample of 443 EDs are selected from the created ED list by systematic sampling.

The sampling fractions and sample allocation was achieved in the following way. Equal probability selection method was used to select the sample enumeration districts in all strata. Sampling fractions were planned to be $f_h \cong 0.001$ for all strata. However, the sampling fractions are also varied among strata. Technical details of the sampling fractions and the sample allocation are given below. The sampling fractions $[f_h^{(.)}]$ and sample allocation $[n_h^{(.)}]$ can be achieved as,

$$f_h^{(1)} = n_h^{(1)}/N_h^{(1)} = 1/F_h^{(1)} \text{ and } f_h^{(2)} = m_h^{(1)}/M_h = 1/F_h^{(2)}. \quad (1)$$

The total population sizes of urban (U) and rural (R) EDs are,

$$N_h^{(U)} = M_h^{(U)} B_{hi} = \left[\sum_i^I M_{hi}\right] B_{hi} \quad \forall\, h \,\&\, i \quad \text{and} \quad (2)$$

$$N_h^{(R)} = M_h^{(R)} B_{hj} = \left[\sum_j^J M_{hj}\right] B_{hj} \quad \forall\, h \,\&\, j \quad (3)$$

where the components are defined earlier.

Then the population size of each stratum was determined as

$$N_h^{(1)} = \left[N_h^{(U)} + N_h^{(R)}\right]. \quad (4)$$

Similarly, the corresponding sample sizes of each stratum are

$$n_h^{(1)} = \left[n_h^{(U)} + n_h^{(R)}\right] \quad (5)$$

$$\text{where } n_h^{(U)} = m_h^{(U)} B_{hi} \text{ and } n_h^{(R)} = m_h^{(R)} B_{hj}. \quad (6)$$

3.2 Design of the PES Operations

The fieldwork operation for PES was identical to the census, where the details are given in section 2.2. For operational purposes, each ED was defined as a close

interval of dwelling unit numbers within the streets. In terms of special enumeration districts (*i.e.*, institution) the total number of enumeration districts are checked with prior information which was obtained at province level.

Due to previously given instructions to the enumerators, PES starts in the sample enumeration districts an hour after the starting time of the census enumeration on the same day. PES enumerators visit the same households in the same (ascending) order as the census enumerators, so that PES enumerators did not visit the same household before the census enumerators. Results of the PES was used as a basis for evaluation, after matching the individual cases with the census records for the corresponding EDs.

4. METHODS OF COVERAGE ERROR ESTIMATION

This section addressed coverage error estimation, by stating data matching procedures and dual system estimation procedures and related findings. The evaluation and estimation of population coverage is obtained using the list of EDs from two independent sources. In this section, the data matching procedures, dual record system estimators, alternative population total estimators are proposed and the estimates are evaluated. A comparison of the computed coverage error statistics are also presented here.

4.1 Data Matching Procedures

Several models (Deming and Glasser 1959; Nathan 1967 and Tepping 1968) have been proposed for determining the optimum matching procedures. These are based on establishing procedures that minimise the *"estimated net matching error"* subject to given costs and other constraints (Marks *et al.* 1974). These models provided valuable concepts to the theory and practice of matching, but none of are completely satisfactory for all purposes.

The work of Tepping (1968), extended by Srinivasan and Muthiah (1968), required a minimum set of characteristics for the *"exact agreement"* in matching. Also, Ayhan and Ekni (1991) and SIS (1994) have used similar methods based on the following specifications;

(1) *Matching of the population of the EDs.* The total population of the ED was taken as the sum of the household population within the total DUs of the ED.

(2) *Matching of the households within the EDs.* Several sets of information (address of the dwelling unit, names of household head and number of persons in the household) was considered for matching of households.

(3) *Matching of the individual persons within the matched households.* A total of 4 Census / PES variables (names, age, sex and education level) were all used for exact agreement in matching of individuals.

(4) *Matching of non-matched individuals of the households.* This was achieved by matching with the other individuals in the neighboring households (from the other data source) by searching. The same criteria was used for exact agreement in matching of individuals.

The preliminary work of matching operation is done clerically, while matched households and persons are evaluated by automation. For the matching procedure of persons the frequencies $n(r, c)$ are shown in Table 2.

Table 2
The Layout of the Matching Procedure

		DATA SOURCE 2: (P E S)		
Matching Procedure		In	Not in	Total
DATA	In	$n(1, 1)$	$n(1, 2)$	$n(1, *)$
SOURCE 1:				
(CENSUS)	Not in	$n(2, 1)$	$\hat{n}(2, 2)$	$n(2, *)$
Total		$n(*, 1)$	$n(*, 2)$	$\hat{n}$

On the basis of the above specifications, the households are matched at the first stage, and within the matched households the persons are matched at the second stage. The results are presented in the following tables by regions. Enumeration districts are located in sample settlements within 19 provinces in 5 regions of the sample design. Out of 443 selected EDs, 437 were matched with their corresponding population census counterparts and other 6 EDs could not be matched due to differences in independently given instructions for their creation by the local offices. The information on the regional breakdown of the 6 non-matched EDs are provided in Table 3, while the information on the urban-rural breakdown was not obtained.

The matching procedure of households can be illustrated by $k(r, c)$ in the same way as presented for persons in Table 3. In the stratified case, the number of households in each strata can also be denoted by $k_h(r, c)$. The total number of households in the Census which are not matched with PES households can be estimated for each strata as

$$k_h(1, 2) = [k_h(1, *) - k_h(1, 1)] \tag{7}$$

and the total number of households in the PES which are not matched the Census households can also be estimated for each strata as

$$k_h(2, 1) = [k_h(*, 1) - k_h(1, 1)]. \tag{8}$$

Information on matched and non-matched households are given in Table 3.

Table 3
Matched and Non-matched Households in the Post Enumeration Survey and Census Enumeration Districts by Regions

Regions h	Selected no. of EDs $m_h^{(1)}$	Enumerated no. of EDs $m_h^{(2)}$	Matched HHs $k_h(1,1)$	Non-matched households Census $k_h(1,2)$	Non-matched households PES $k_h(2,1)$
1	157	154	3,320	168	144
2	62	62	1,262	27	30
3	112	112	2,636	262	259
4	38	38	645	204	80
5	74	71	995	170	175
Total	443	437	8,858	831	688

In these 437 EDs, a total of 8858 households were matched. In the Census 831 (9.38 %) and in PES 688 (7.77%) households could not be matched. The Census based household match rate was 90.62 %, while PES based match rate was 92.23 %, which is presented in Table 3.

Coverage rates for the Census and PES households are given by regions in Table 4. Comparison of Census and post enumeration results for the coverage rate of households (C_h) were higher for the Census in most regions (except Regions 2 and 5) and the total. Here all the coverage rates were greater than it was expected. In terms of persons within the covered households, the coverage rates (C_h^*) were higher for the Census for all regions and for the total. Total of matched persons were n (1, 1) = 41,020 in the Census and PES.

Differences in the coverage of EDs in the Turkish Census and PES comes due to the following reasons;

(1) Additional Forms of C and D are established by the Census Committee of the provinces through list of buildings (Forms 1 and 2). List of buildings are created by the local authorities and they are not reliable enough for some settlements.

(2) Numbering of EDs are also done at the local level, they are also effected by the insufficient numbering operation.

(3) Forms C and D may or may not contain 100 persons in urban and 200 persons in rural areas due to outdated listings.

(4) Due to different starting points by the Census and PES enumerators, the number of dwelling units covered were different.

(5) Application of the PES questionnaire was started at least 2 hours after the actual Census operation within the selected EDs. Coverage differences may be due to the mobility of the members of census completed households within the same ED.

(6) During the one day enumeration period, some of the planned Census and PES questionnaires could not be completed, resulting inconsistency during matching. Of course, this is a source of undercount, which happened rarely during the field enumeration.

(7) Because of the de facto enumeration base, the local visitors (from other dwellings of the apartment) for either data source were subject to change.

(8) Again, due to de facto enumeration, there will be counting errors for the mobile population for the Census. The PES only planned to cover the household population.

(9) The PES was not planned to cover the special EDs and mobile populations (*i.e.*, travelers, persons on duty, *etc*). By definition, international and domestic travelers are permitted to continue their travel after being counted, if their journey had started before the official census starting time. During this research, the mobile population was excluded from the analysis.

(10) Nomadic tribes (Special enumeration techniques are required for the census of nomadic tribes. *De jure* rather than *de facto* enumeration base, as well as mobile interviewers may be recommended for the enumeration of nomadic tribes in place of interviewers who are stationary.) will not be covered in the PES due to non-listings.

Table 4
Number of Households and Persons in the Census and Post Enumeration Survey by Regions

Regions h	Number of Household Census $k_h(1,*)$	Number of Household PES $k_h(*,1)$	Number of Household Coverage C_h	Number of Persons in Households Census $n_h(1,*)$	Number of Persons in Households PES $n_h(*,1)$	Number of Persons in Households Matched $n_h(1,1)$	Number of Persons in Households Coverage C_h^*
1	3,488	3,464	1.0069	14,035	13,926	13,393	1.0078
2	1,289	1,292	0.9977	6,587	6,582	6,400	1.0008
3	2,898	2,895	1.0010	13,058	12,984	11,644	1.0057
4	849	725	1.1710	4,233	3,580	3,134	1.1824
5	1,165	1,170	0.9957	7,898	7,888	6,449	1.0013
Total	9,689	9,546	1.0150	45,811	44,960	41,020	1.0189

Coverage rates: $C_h = k_h(1,*)/k_h(*,1)$ and $C_h^* = n(1,*)/n(*,1)$

(11) Both Census and PES EDs are enumerated with the same instruction for the previously defined close interval. However, due to the use of different quality of the frames (updated 1990 or outdated 1985 or even outdated 1980) the amount of workload of each interviewer was varying. Consequently, the amount of coverage in each ED may be different from both sources.

4.2 Dual Record System Estimation

Dual record system is used as a method for determining the estimated number of households and persons through a matching procedure. The results are used to estimate the total number of persons in each region and the total population. The model assumes independence of data collection from two sources, where the Census and the PES are used. In theory, all cells [n (r, c)] are observable except for $n(2, 2)$ and any of the totals that include $n(2, 2)$. Chandra *et al.* (1949) assumes that, there is no correlation bias with the estimate for cell $n(2, 2)$. For practical purposes, this paper also considers this assumption as valid. On the other hand, further discussion on the validity of such assumption is recently reported by Ayhan (2000).

The methodology and the estimation procedures are presented below. Estimation of the number of persons not in the Census or in PES

$$n(2,2) = [n(1,2)\,n(2,1)]/n(1,1). \tag{9}$$

Total number of persons is estimated as

$$n = n(1,1) + n(1,2) + n(2,1) + n(2,2) \tag{10}$$

or alternatively,

$$n = [n(*,1)\,n(1,*)]/n(1,1). \tag{11}$$

Table 2 earlier illustrated the matching procedure used for the dual record system method. The computational procedure presented here was repeated for each region separately. The estimates are given in Table 5. For each strata, n_h is computed as n previously.

Table 5
Matched and Non-matched Number of Persons in the Census and Post Enumeration Survey by Regions

Regions	Matched	Census non-match	PES non-match	Estimated omissions in both sources	Dual record system estimate
h	$n_h(1,1)$	$n_h(1,2)$	$n_h(2,1)$	$n_h(2,2)$	$n_h^{(D)}$
1	13,393	642	533	26	14,594
2	6,400	187	182	5	6,774
3	11,644	1,414	1,340	163	14,561
4	3,134	1,099	446	156	4,835
5	6,449	1,449	1,439	323	9,660
Total	41,020	4,791	3,940	673	50,424

4.3 Total Population versus Household Population

The total population was considered as the target population for the population projections, which was used to estimate the total number of EDs in the population. On the other hand, PES sample design only considered the household population as the target population. Because the PES design was based on the selected sample dwelling units only, which excluded the special enumeration districts (the institutional population).

As stated earlier, the PES sample design was taken as the base for the comparison of two different enumeration systems during the matching procedures. This naturally led us to consider the household population as the target population for the appropriate estimation of the population total by the proposed estimators. In order to achieve this, the institutional population was computed later, from the 1990 Census data, for regions and population size groups. The institutional population of regions are presented (by aggregating over the size groups) in Table 6.

Table 6
Determination of Household Population and Sample Sizes by Regions

Regions	Projected population size	Institutional population estimate	Household population size	Household survey sample size	Expansion factors	
h	$N_h^{(1)}$	$N_h^{(2)}$	$N_h^{(3)}$	$n_h^{(1)}$	$F_h^{(1)}$	$F_h^{(3)}$
1	20,639,200	367,184	20,272,016	18,900	1092.02	1072.59
2	9,242,600	89,934	9,152,666	8,200	1127.15	1116.18
3	15,731,600	176,031	15,555,569	14,800	1062.95	1051.05
4	7,670,800	55,104	7,615,696	6,000	1278.47	1269.28
5	13,687,800	249,309	13,438,491	10,800	1267.39	1244.30
Total	66,972,000	937,562	66,034,438	58,700	1140.92	1124.95

$N_h^{(3)} = N_h^{(1)} - N_h^{(2)}$ here $F_h^{(1)} = N_h^{(1)}/n_h^{(1)}$ and $F_h^{(3)} = N_h^{(3)}/n_h^{(1)}$

For the further use of the information on the institutional population, it was also assumed that, there were no coverage errors associated in measuring the institutional population during the 1990 Census enumeration. The household population of each region, are then computed by subtraction.

There were several reasons for removing the institutional population from the total population;

(1) The PES sample design only reflected the household population.

(2) The correct selection probabilities for the ideal coverage (representation) of each sample strata can only be based on the household population, not on the total population.

(3) The proposed coverage error estimates should only be based on the household population.

(4) The proposed estimators for the population total should also be based on the household population, where the PES results are household based.

(5) It will be wrong and misleading to make comparison of coverage error statistics, when the base populations are different.

(6) The census undercount is artificially inflated if the wrong population (namely, the total population) is taken as the target population.

4.4 Coverage Error Measures

Many coverage error statistics are proposed in the literature. Some of these error statistics are based on simple ratios or proportions, and others are based on more complex adjustment procedures. To simplify the solution to this problem, the following coverage error measures are proposed for the regional and total population. These are census coverage rate, census discrepancy rate and the amount of census discrepancy. The following coverage error measures are proposed which are based on the *household population*.

Census Coverage Rate:

Regional estimators:

$$\lambda_h^{(s)} = N_h^* / \hat{\mathbf{N}}_h^{(s)} \quad \forall h \quad h = 1, 2, ..., H \tag{12}$$

where N_h^* = Census count of the household population $[N_h^* = N_h - N_h^{(2)}]$ and $\hat{\mathbf{N}}_h^{(s)}$ = Estimate from source (or method) s.

Standard error of regional estimators: Making the following scale transformation $\lambda_h^{(s)}(0.5) = \tilde{\lambda}_h^{(s)}$ which is taken as a proportion, realizing that within each strata $\tilde{\lambda}_h^{(s)} + (1 - \tilde{\lambda}_h^{(s)}) = 1$, the standard error estimators of the census coverage rates of each region is computed as

$$\text{se}\left[\tilde{\lambda}_h^{(s)}\right] = \left[\frac{\tilde{\lambda}_h^{(s)} - \left(1 - \tilde{\lambda}_h^{(s)}\right)}{n_h^{(D)} - 1}\right]^{1/2}. \tag{13}$$

Total population estimator: $\lambda = N^* / \hat{\mathbf{N}}^{(s)}$ (14)

Census discrepancy rate:

Regional estimators: $\varphi_h^{(s)} = 1 - \lambda_h^{(s)} = \left[\delta_h^{(s)} / \hat{\mathbf{N}}_h^{(s)}\right]$. (15)

Total population estimator : $\varphi = 1 - \left[N^* / \hat{\mathbf{N}}^{(s)}\right] = 1 - \lambda$. (16)

Census discrepancy:

Regional estimators: $\delta_h^{(s)} = \mathbf{N}_h^{(s)} - N_h^* \quad \forall h$ (17)

Due to the limitations of the one day enumeration by the *de facto* system, other additional local coverage measures could not be considered for this study. Such additional coverage measures for the local areas could provide useful additional information for more complex coverage error models in countries who are employing *de jure* system of enumeration in their census taking.

Even with the limitations of the *de facto* census, one could compute coverage estimates for large domains (such as provinces), where the population would not likely to shift very much between Census and PES interview. This was not possible, due to the limited sample size of PES which did not provide independent provincial estimates to be made within the regions. In addition, the sample sizes might not be large enough to give sufficient precision.

5. ESTIMATORS OF POPULATION TOTAL

The *estimated population total* is taken as the weighted sum of the all regional estimates.

$$\hat{\mathbf{N}}^{(s)} = \sum_h^H \hat{\mathbf{N}}_h^{(s)}. \tag{18}$$

The *standard error estimators* for the total household population of each region is computed as

$$se\left[\hat{\mathbf{N}}_h^{(s)}\right] = \hat{\mathbf{N}}_h^{(s)} \left[\frac{p_h(1 - p_h)}{n_h^{(D)} - 1}\right]^{1/2} \tag{19}$$

while the proportion of each strata is computed as $p_h = n_h^{(D)} / \sum_h^H n_h^{(D)}$.

The determination of the coverage error of a given Census is not an easy task, especially when a perfect list of a target population is not available to compare the results. This is always the case in most countries of the world, except the ones with population registers.

Table 7
Estimates of the Regional and Total Household Population for 1990 by the Expanded Dual Record System Estimate and Their Standard Errors

h	$\hat{\mathbf{N}}_h^{(1)}$	se$[\hat{\mathbf{N}}_h^{(1)}]$	$\hat{\mathbf{N}}_h^{(2)}$	se$[\hat{\mathbf{N}}_h^{(2)}]$	$\hat{\mathbf{N}}_h^{(3)}$	se$[\hat{\mathbf{N}}_h^{(3)}]$
1	15,936,939	58,967*	15,436,073	57,113	15,653,378	57,917
2	7,635,314	31,305	7,367,741	30,208	7,561,003	31,000
3	15,573,280	57,621	14,571,298	53,914	15,398,933	56,976
4	6,181,402	38,943	5,884,582	37,073	6,136,969	38,663
5	12,242,987	48,972	11,502,934	46,012	12,019,938	48,080
Total	57,569,922	241,794	54,762,628	230,003	56,770,221	238,435

*: *Standard error estimates are rounded to the nearest integer.*

Comparison of the results of a population census with projection figures also creates some kind of comparison problems, due to the validity of the several assumptions relating to projection models. In order to avoid a single base of comparison, the following *expanded dual record system regional estimators* are proposed for the determination of the census coverage errors.

Estimator 1. $$\hat{\mathbf{N}}_h^{(1)} = F_h^{(1)} n_h^{(D)} \qquad (20)$$

where $F_h^{(1)} = N_h^{(1)} / n_h^{(1)}$ and $n_h^{(D)} = \sum_r \sum_c n_h(r,c)$.

Here $n_h^{(D)}$ refers to the unweighted DRS estimate and $n_h^{(1)}$ corresponds to the selected sample size.

Estimator 2. $\hat{\mathbf{N}}_h^{(2)} = F_h^{(2)} n_h^{(D)}$ where $F_h^{(2)} = M_h / m_h^{(1)}$. (21)

Estimator 3. $\hat{\mathbf{N}}_h^{(3)} = F_h^{(3)} n_h^{(D)}$ where $F_h^{(3)} = N_h^{(3)} / n_h^{(1)}$. (22)

The dual record system estimators are expected to yield higher estimated counts than a single round survey (*i.e.*, PES), by definition. Therefore, all the proposed estimators for the household population totals are DRS based. DRS estimates of the total household populations are given in Table 7.

Difference between the three proposed estimates, are only based on the type of expansion factors used. When we examine the expansion factors, $F_h^{(1)}$ is based on the projected population sizes over household survey sample sizes. On the other hand, $F_h^{(2)}$ is based on total population EDs over total sample EDs of the original PES design.

Finally, $F_h^{(3)}$ is based on the household population size over household survey sample size. The first two estimators include institutional population components $[N_h^{(2)}]$ in the numerator of their expansion factors $[N_h^{(1)}$ or $M_h]$, while only the third estimator uses household population information $[N_h^{(3)}]$ in its expansion factor. It is clear that, the expansion factor for the third estimator is derived from the ideal selection probabilities $[f_h^{(3)} = n_h^{(1)} / N_h^{(3)} = 1 / F_h^{(3)}]$ for the PES sample, which is based on household information. Therefore, *Estimator 3* can be considered as more representative of the household population.

6. COMPARISON OF COVERAGE ERROR STATISTICS

For the comparison of error statistics, the population counts should be of the same standard base. It will be recommended to use a household population count which matches the corresponding population estimate. The regional and total population counts are given in Table 8. As mentioned earlier, the institutional population counts are determined from the 1990 Census counts.

Table 8
Regional and Total Population Counts for Turkey, 1990

h	Census counts N_h	Institutional population counts $N_h^{(2)}$	Household population counts N_h^*
1	18,544,967	367,184	18,177,783
2	7,836,940	89,934	7,747,006
3	12,824,347	176,031	12,648,316
4	5,964,565	55,104	5,909,461
5	11,302,216	249,309	11,052,907
Total	56,473,035	937,562	55,535,473

where $N_h^* = N_h - N_h^{(2)}$

For the purpose of population coverage error evaluation, the *census coverage rate* and the amount of *census discrepancy* was used. The computed population coverage error rates are given by regions and the total in Table 9.

Table 9
Estimates of the Census Coverage Rates for Regional and Total Household Population in Turkey 1990 and Their Standard Errors

h	$\lambda_h^{(1)}$	$se[\tilde{\lambda}_h^{(1)}]$	$\lambda_h^{(2)}$	$se[\tilde{\lambda}_h^{(2)}]$	$\lambda_h^{(3)}$	$se[\tilde{\lambda}_h^{(3)}]$
1	1.14061	0.00410	1.17762	0.00407	1.16127	0.00408
2	1.01463	0.00607	1.05148	0.00607	1.02460	0.00607
3	0.81218	0.00407	0.86803	0.00411	0.82138	0.00408
4	0.95601	0.00718	1.00423	0.00719	0.96293	0.00719
5	0.90272	0.00506	0.96088	0.00508	0.91955	0.00507
Total	0.96466	0.00223	1.01411	0.00223	0.97825	0.00223

There is a clear pattern for certain regions, for all estimates. The census coverage rates can also be expressed as the amount of census discrepancy. A similar pattern is expected for the three estimators, since the estimators are highly correlated.

For the total population, estimates based on methods (1) and (3) has resulted in census undercount when compared with the corresponding actual population counts. Due to the computational procedures, *Estimate 3* can be recommended among others because *Estimate 3* is based on the projected household population, where the comparison base is the same as the selection.

There is also a pattern for regional estimates, regardless of the method of estimation. For regions 1 and 2, all estimates indicated census overcount, while census undercount was observed for all other regions by all estimates, except for *Estimate 2* in region 4.

7. CONCLUSIONS

The coverage error study of the population census had provided some useful information in evaluating the methodological issues which is summarised below.

Comparison of the three proposed population total estimates indicates that, the first estimate provided the highest value of the total count, while *Estimate 3* provided more representative result for the total household population.

The evaluation of the census coverage error rates and the amount of census discrepancy had shown that, for the total population, *Estimates 1 and 3* has resulted in census undercount. There is also a distinct pattern for regional estimates, regardless of the method of estimation. There seems to be a census overcount in the first two regions, while census undercount was observed for the other three regions by all estimates (except for *Estimate 2* in Region 4).

For the developing countries, the main problem of census taking is based on the undercount. In Turkey, the overcount issues in census taking only occur in very limited local areas and they are re-evaluated later and removed from the census data before release of the census results.

On the basis of these findings it is clear that, the comparison of several sample based estimates with the population census count indicated the existence of some methodological problems which are present in the enumeration procedures of the Turkish Population Census. The most important of these issues are the following;

(1) Improving and updating the list of possible EDs in rapidly growing peripheries of the large cities by the use of area methods.

(2) Obtaining a perfect list of all dwelling units within the EDs. This can be better achieved through a continuous screening operation by the local authorities, where they are responsible for this by law. Alternatively, a Census of Housing can be taken just before the population census by the SIS which will also provide a useful frame for the population census enumeration.

(3) There are many laws in the country which refers to the latest "population counts". This suggests that, major changes might be necessary on legal issues as well as in enumeration techniques.

(4) Enumeration of the mobile populations also requires special attention, methods and qualified personnel.

One would like to hope that, measuring the characteristics of the population through the Censuses may be considered important, by the responsible officials in time and the necessary developments will take place along these directions.

ACKNOWLEDGEMENTS

We would like to thank Professors Orhan Güvenen, Yalçin Tuncer, Vijay K. Verma, Moti Lal Tiku, M. Qamar Islam and Mr. Ömer Gücelioglu for their valuable comments. The contributions of Ms. Hasibe Dedes and Ms. Canan Bakici are also gratefully acknowledged. Finally, the comments and suggestions of the Editor, Associate Editor and the referee's of the Journal is very much appreciated. The views expressed are attributed to the authors and do not necessarily reflect those of the State Institute of Statistics, Turkey.

APPENDICES: TOOLS OF ENUMERATION

The following listing forms and questionnaires are used before and during the Census and PES operations.

APPENDIX 1. LISTING FORMS USED

Form 1: ***List of Buildings*** *(for localities with municipal organization).*

This list is created by the local municipality personnel and later produced in triplicate. Used for sequential numbering of DUs in urban areas.

Form 2: ***List of Buildings*** (for localities without municipal organization).

This list is created by the village head person and later produced in triplicate. Used for sequential numbering of dwelling units in rural areas.

Form C: ***Enumeration District List of Buildings.***

This list is based on Forms 1 or 2. The EDs are formed on the basis of this list in urban and rural areas, separately.

Form D: ***Census Control List.***

This is an update of Form C which was completed by the enumerator after the census field operation and returned to the Local Census Committees with the completed census questionnaires. This form and the completed census questionnaires are forwarded to the SIS after the census field operation.

APPENDIX 2. QUESTIONNAIRES USED

Form A: ***Population Census Questionnaire.***

The population census questionnaire consisted of four main parts. The information is collected through a personal interview by a paper and pencil approach.

Part 1. *Address details.*

Part 2. *Type of place of the residence.*

Part 3. *Household module* [contains 7 precoded household questions].

Information is collected to identify the household head, presence of head, total number of persons in HH, number of guests, number of HH members absent, ownership of present DU, and ownership of any other DU.

Part 4. *Individual person's module* [contains 26 precoded individual questions].

For each person present, information is obtained on sex, age, relation to HH head, place of birth, citizenship, permanent residence, educational background, marital status, fertility information, employment status, and main occupation.

Form B: ***Post Enumeration Survey Questionnaire.***

PES questionnaires are generally based on a subset of questions of the main study. However, for this study it was decided by the Census Advisory Committee to use the complete census questionnaire for the PES. The questionnaire for PES is completed in the same way as the Census.

REFERENCES

AYHAN, H.Ö., and EKNI, S. (1991). Coverage and response errors in 1990 Turkish Census of Population. *Bulletin of the International Statistical Institute.* 54, 45-46.

AYHAN, H.Ö. (2000). Estimators of vital events in dual-record systems. *Journal of Applied Statistics.* 27, 157-169.

BANKIER, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association.* 81, 1074-79.

CASADY, R.J., NATHAN, G. and SIRKEN, M.G. (1985). Alternative dual system network estimators. *International Statistical Review.* 53, 183-197.

CHANDRA SEKAR, C., and DEMING, W.E. (1949). On a method of estimating birth and death rates and the extend of registration. *Journal of the American Statistical Association.* 44, 101-115.

CHOI, C.Y. STEEL, D.G. and SKINNER, T.J. (1988). Adjusting the 1986 Australian census count for under enumeration. *Survey Methodology.* 14, 173-189.

CRESSIE, N. (1988). When are census counts improved by adjustment? *Survey Methodology.* 14, 191-208.

CRESSIE, N. (1990). Weighted smoothing of estimated undercount. U.S. Bureau of the Census, *1990 Annual Research Conference Proceedings.* 301-325.

DEMING, W.E., and GLASSER, G.J. (1959). On the problem of matching lists by samples. *Journal of the American Statistical Association.* 54, 403-415.

DIFFENDAL, G. (1988). The 1986 test of adjustment related operations in Los Angeles County. *Survey Methodology.* 14, 71-86.

FAY, R.E. PASSEL, J.S. ROBINSON, J.G. and COWAN, C.D. (1988). The Coverage of Population in the 1980 Census. *Evaluation and Research Reports,* PHC 80-E4, U.S. Bureau of the Census. 123.

GOODMAN, L.A. (1949). On the estimation of the number of classes in a population. *Annals of Mathematical Statistics.* 20, 572-579.

HARTLEY, H.O. (1962). Multiple frame surveys. In *Proceedings of the Social Statistics Section,* American Statistical Association. 203-206

HARTLEY, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā,* Series C. 36, 99-118.

HOGAN, H. (1990). The 1990 Post Enumeration Survey: An Overview. *U. S. Bureau of the Census Paper,* Washington DC. 6.

HOGAN, H. (1993a). The 1990 post enumeration survey: operations and results. *Journal of the American Statistical Association.* 88, 1047-1060.

HOGAN, H. (1993b). Planning for census correction: the 1990 United States experience. Invited Paper, 49th Session of the International Statistical Institute, Florence, Italy. *International Association of Survey Statisticians Booklet.* 133-150.

HOGAN, H., and WOLTER, K. (1988). Measuring accuracy in a post enumeration survey. *Survey Methodology.* 14, 99-116.

ISAKI, C.T. (1992). Model bias effects in small area coverage error estimation. *Communication in Statistics Serie A.* 21, 1213-1231.

MARKS, E.S., SELTZER, W. and KROTKI, K.J. (1974). Population Growth Estimation: *A Handbook of Vital Statistics Measurement.* New York: The Population Council.

MULRY, M.H., and SPENCER, B.D. (1988). Total error in the dual system estimator: the 1986 census of central Los Angeles county. *Survey Methodology*. 14, 241-263.

MULRY, M.H., and SPENCER, B.D. (1990). Total error in post enumeration survey estimates of population: the dress rehearsal census of 1988. U.S. Bureau of the Census, *1990 Annual Research Conference Proceedings*. 326-361 .

MULRY, M.H., and SPENCER, B.D. (1993). Accuracy of the 1990 census and undercount adjustments. *Journal of the American Statistical Association*. 88, 1080-1091.

NATHAN, G. (1967). Outcome probabilities for a record matching process with complete invariant information. *Journal of the American Statistical Association*. 62, 454-469.

S.I.S. (1994). 1990 Census of Population Response Reliability Survey. *State Institute of Statistics Publication* No. 1688, Ankara, 65.

SRINIVASAN, S.K., and MUTHIAH, S. A. (1968). Problems of matching births identified from two independent sources. *The Journal of Family Welfare*. 14, 13-22.

TEPPING, B.J. (1968). A model for optimal linkage of records. *Journal of the American Statistical Association*. 63, 1321-1332.

WOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*. 81, 338-346.

Survey Methodology, December 2003
Vol. 29, No. 2, pp. 167-175
Statistics Canada

A Hierarchical Model for the Analysis of Local Census Undercount in Italy

D. COCCHI, E. FABRIZI and C. TRIVISANO[1]

ABSTRACT

Census counts are known to be inexact based on comparisons of Census and Post Enumeration Survey (PES) figures. In Italy, the role of municipal administrations is crucial for both Census and PES field operations. In this paper we analyze the impact of municipality on Italian Census undercount rates by modeling data from the PES as well as from other sources using Poisson regression trees and hierarchical Poisson models. The Poisson regression trees cluster municipalities into homogeneous groups. The hierarchical Poisson models can be considered as tools for Small Area estimation.

KEY WORDS: Census undercount; Post enumeration survey; Bayesian hierarchical modelling; Gamma-Poisson regression models; Poisson regression trees.

1. INTRODUCTION

The Italian Population Census takes place every ten years and represents the most important institutional duty of the Italian National Institute of Statistics (ISTAT) (The work leading to this paper has been developed just before 2001 Italian Census and the subsequent PES. The results have been considered in performing the 2001 PES). In order to carry out the Census, ISTAT relies on municipal administrations who are responsible for all field operations (training of interviewers, planning of interviews, data gathering and basic data processing). During Census operations, each municipality works independently from the others under ISTAT supervision. The accuracy of the Census results therefore differ considerably from one municipality to another, even if contiguous. In Italy, the geographical area of a municipal borough is sub-divided into Census Enumeration Areas (EAs), which are assigned to a single interviewer during Census operations. The EAs differ in terms of shape, structure and difficulty of enumeration, as well as interviewer. It is likely that the undercount rate varies substantially among EAs within the same municipality.

After the 1991 Population Census, ISTAT conducted a Post Enumeration Survey (PES) to measure the phenomenon of undercount. Population Census counts are known to be generally incorrect because of missed, multiple and misplaced enumeration. Missed enumeration is the most important inaccuracy and typically yields a net population undercount that may vary geographically and between different social groups, and impacts the determination of the relative sizes of sub-populations (Abbate, Masselli, Signore 1993). Field operations of the PES were carried out by the sampled municipalities themselves. The 1991 Italian PES data have been analyzed by Abbate, Masselli and Signore (1993), who estimate the overall national undercount rate by means of a Lincoln-Petersen model (see Wolter 1986) using post-strata of municipalities based on large geographical areas (North, Center, South). Working on the same data, Fortini (1994) estimates the overall national undercount by means of latent class models.

Instead of estimating the undercount rate for the whole country or smaller domains, we propose models designed to explain the variation in undercount rate at the municipal level. The availability of factors accounting for the size of the net undercount may be a basis for creating homogeneous groups of municipalities, for planning a more efficient stratification in future Post Enumeration Surveys. Moreover, knowledge of those flaws in municipal organization which significantly influence the undercount may provide guidelines for actions designed to reduce its size.

Contributions which use disaggregated PES data are present in the literature. Alho, Mulry, Wurdeman and Kim (1993) consider a logistic regression model for the individual (household) probability of being censused. In keeping with Moura and Holt (1999), their model could be extended to include municipality or other group effects. We are in fact aware that our choice of modelling municipal data is not the same as the analysis of household level records, since many features determining individual propensity to be caught by the Census average out when dealing with aggregated data. A comprehensive analysis based on individual records is not feasible in the Italian case, since there were very few questions for individuals included in the 1991 PES schedule. Similarly, the 1991 PES provides very little auxiliary information on the EAs, with the consequence that models based on EA undercounts cannot be proposed.

Our analysis is based on combining different data sources. The auxiliary information comes from the abovementioned 1991 PES, two studies on the statistical quality of municipalities conducted by ISTAT during the early 90s (Di Pietro 1998, 1999) and demographic and social indicators obtained from the 1991 official Census results.

[1] D. Cocchi, E. Fabrizi and C. Trivisano, Dipartimento di Scienze Statistiche "P. Fortunati", Università di Bologna, Italy.

We face the problem of how to make efficient use of the information obtained from the various data sources. We have in fact a large number of variables, most of which are categorical or polychotomous. Instead of using a variable selection algorithm, we have chosen to build homogeneous groups of municipalities which are then introduced into the model by means of a design matrix for the random effects. These groups are constructed using Poisson regression trees (Therneau and Atkinson 1997). This hierarchical usage of information provides a natural basis for the design of strata of geographically non-contiguous municipalities.

Few EAs are re-censused within each sampled municipality in the PES; the average EAs sampling rate is 0.001. This is a typical Small Area setting where direct estimates of the municipal undercount rate are unreliable and ought to be replaced by synthetic or composite estimates based on a suitable model. The phenomenon of undercount is rare. Our data consist of counts and may show a large overdispersion with respect to a Poisson distributional assumption. We suggest the use of hierarchical Poisson regression models to manage overdispersion.

The hierarchical models here adopted manage explicitly overdispersion due to municipal heterogeneity. A further extra Poisson variability source is due to heterogeneity within municipalities, because of clustering of missed enumeration within EAs, or of clustering due to missed enumerations of individuals in the same family. This kind of overdispersion is not explicitly treated in the models.

We adopt a full Bayesian approach for specification and estimation purposes and base the solution of the models on Markov chain Monte Carlo simulation methods. Within this hierarchical framework, we deal with overdispersion by imposing a Gamma distribution on the rate of the first level Poisson distribution, thus marginally obtaining a Negative Binomial. Moreover, conditionally on the hyperparameters, the proposed model features posterior linearity and the corresponding posterior means for the municipal undercount rates are linear composite estimators. Thus, the amount of smoothing depends on how much information is provided by each municipal sample in the PES.

Our results show that the municipality stratification employed in designing the 1991 PES (based on geographical area and population size) can be improved, since the undercount rate is shown to be largely independent of geographical area. On the contrary, variables describing the statistical efficiency of local administrations are useful in discriminating between the different degrees of undercount among municipalities of similar size and demographic structure. Whilst leaving the design of the PES unchanged, our results may provide useful guidance when performing data analysis.

The present paper is organized as follows. Section 2 describes the basic features of the PES and of the other data sources we have taken into consideration. Section 3 looks at the Poisson regression trees used to build homogeneous groups of municipalities. In section 4 we introduce the hierarchical Poisson regression models, while empirical results and model comparisons are discussed in section 5.

2. THE PES DATA AND AUXILIARY INFORMATION

2.1 The Italian Post Enumeration Survey

The 1991 Italian Population Census took place on October 20th. The subsequent Post Enumeration Survey, based on a two stage stratified sampling design, was carried out a few weeks later. Municipalities constitute the primary units, whereas the secondary ones are represented by the Census EAs. An EA is the smallest area into which the municipal territory is partitioned for Census operations; each EA is assigned to a single interviewer.

The primary sampling units were stratified according to geographical area (North-West, North-East, Center, South, Islands) and demographic size (7 classes for the municipalities below 350,000 inhabitants), producing 35 strata. Within each stratum the sampled municipalities were selected without replacement and with probability proportional to their demographic size. The 10 municipalities with more than 350,000 inhabitants have been included in the sample as self-representative units. The secondary sampling units were selected with equal probabilities by systematic sampling. The final PES sample contains 85 municipalities and 638 EAs (out of a national total of 8,095 municipalities and 64,000 EAs) with a national design based estimate of 1.24% (Abbate, Masselli and Signore 1993).

The PES forms were filled out during face to face interviews and contained just a few simple questions. The characteristics of the sampled households are limited to the number and gender of household members. Other PES questions were designed to facilitate record linkage with the Census result, and therefore to reduce both misplaced enumeration and other non sampling errors in the evaluation of undercount (see Fortini 1994 for details).

2.2 The Surveys of the Statistical Quality of Municipalities

A data set on the statistical quality of Italian municipalities was constructed by ISTAT (see Di Pietro 1998, 1999). It integrates different sources: information from 1991 Census performance records, municipal population registers and Interior Ministry data. This data set contains also the results of three administrative surveys, conducted during the 90s, carried out in order to evaluate the performance of municipalities with regard to their commitments to ISTAT. The first survey is about the computerization of municipal Statistics Bureaus. The second survey, known with the acronym POSAS, is a post-Census survey of the demographic registers of the resident population, classified by year of birth, age and civil status. The third survey, known with the acronym ISCAN, regards the

appropriateness of registrations on the municipal population registers list. These surveys provide data for all Italian municipalities.

From this data set we selected a subset of variables related to the municipal activity at the time of the 1991 Census:

a) the percentage of noncoded fields of the Census household forms which had to be filled out, after the interview of the households, by the municipal Statistics Bureaus (PERCOD);

b) the ratio of the population temporarily abroad to the population present at the 1991 Census (PERCEST);

c) the ratio of the difference between the 1991 Census and population registers counts to the 1991 Census counts (PERDIFF);

d) the time needed to update municipal demographic registers on the basis of 1991 Census results (IND01);

e) delay in street name updating (IND11).

2.3 Demographic Variables

We also consider a set of demographic ratios from the 1991 Census results. In particular, we use the percentages of "single member" and "more than one family" households, and sex ratios (males/females) in the municipality. The municipal resident population – resulting from the uncorrected 1991 Census counts – is also a very important variable. The number of EAs sampled in each municipality for the PES is a further signal of the municipality importance.

3. POISSON REGRESSION TREES

The available data sources provide us with a large number of auxiliary variables, many of which are categorical or polychotomous. Before we fit the hierarchical models, we group municipalities with homogeneous household undercount rates using Poisson binary regression trees. Groups based on trees are included as factors in the models described in the next section. Our principal aim is to check the effectiveness of traditional stratifications, improving them ex post by hierarchical models with suitable covariates and to verify how they differ from comparable results based on optimal groupings.

The conditional regression models are based on the canonical logarithmic link. The splitting criterion is based on the usual deviance statistic (Therneau and Atkinson 1997):

$$\text{Deviance}_{\text{parent}} - (\text{Deviance}_{\text{child, left}} + \text{Deviance}_{\text{child, right}})$$

The basic idea for building a tree is to begin with a large tree T_0 constructed using a naive and mild stopping rule (as the minimum number of observations in the final nodes of the tree) and then to select the right-sized tree among the sub-trees of T_0 by pruning. The established methodology for pruning trees is cost-complexity pruning, first introduced by Breiman, Friedman, Olshen and Stone (1984). Let D_T be the deviance of a subtree T of T_0, *size* (T) the number of terminal nodes of T and $\alpha > 0$ a cost-complexity parameter for defining the cost-complexity measure:

$$D_T(\alpha) = D_T + \alpha \text{size}(T) \qquad (1)$$

For a specified α the tree $T(\alpha)$ that minimizes (1) can be found. It can be shown (Breiman *et al.* 1984) that a nested family of subtrees $\{T_0, T_1, ..., T_k, ..., T_{\text{root}}\}$ of T_0 exists such that each tree is optimal for a range of values of α.

The problem is now reduced to selecting one of these subtrees. The selection is carried out in order to minimize the prediction error defined as the deviance contribution for a new observation. To estimate the prediction error, the availability of an independent sample would be in principle the best option, but since it is advisable to use all data to "instruct" the tree in the best possible way, a cross-validation method is used. Usually, the tree T_{k_0} with the minimum estimated prediction error is selected. Here we use a more severe pruning rule which consists in selecting the smallest tree with an estimated prediction error not larger than the estimated prediction error of T_{k_0} plus its standard error. This pruning rule, known as the "1 SE rule" (Breiman *et al.* 1984), is adopted in order to avoid model overfitting.

Since the cross-validation of Poisson regression trees may give, in some nodes, infinite values for the deviance statistic, we use Bayesian shrinkage estimators of the true rates, based on a simple Poisson-Gamma model, as suggested in Thernau and Atkinson (1997).

We built three different trees based on different starting subsets of auxiliary variables.

Tree 1 (shown in Figure 1) is based on demographic variables only. The first split separates municipalities with population less than 100,100 from those with more than 100,100. This splitting value is almost coincident with the 100,000 demarcation value used in the stratification of municipalities for the 1991 PES. The second split isolates a sub-sample of small municipalities for which less than 4 EAs were sampled in the PES. A further split is made on the basis of the sex ratio.

Tree 2 (Figure 2) is based exclusively on variables concerning the quality of the statistical performance of municipalities. The first split is based on the timing in correcting demographic registers (IND01): those municipalities that were quickest in performing this activity have the lowest undercount rates. Lower level splits highlight the problem of people temporarily abroad (PERCEST) which in areas characterized by massive emigration may lead to serious undercounting of the municipal population and errors in the book-keeping of demographic registers (PERDIFF). In this tree, one half of the sample is classified in a single node which is likely to contain residual heterogeneity.

Tree 3 (Figure 3) is based on both demographic and quality variables. The first split is based on the municipal population exactly as was the case in Tree 1. Subsequently, the subset of municipalities with less than 100,100 inhabitants is split into small and middle sized municipalities at a threshold of 13,200. The quality variable included in this tree consists of timing in correcting demographic registers (IND01).

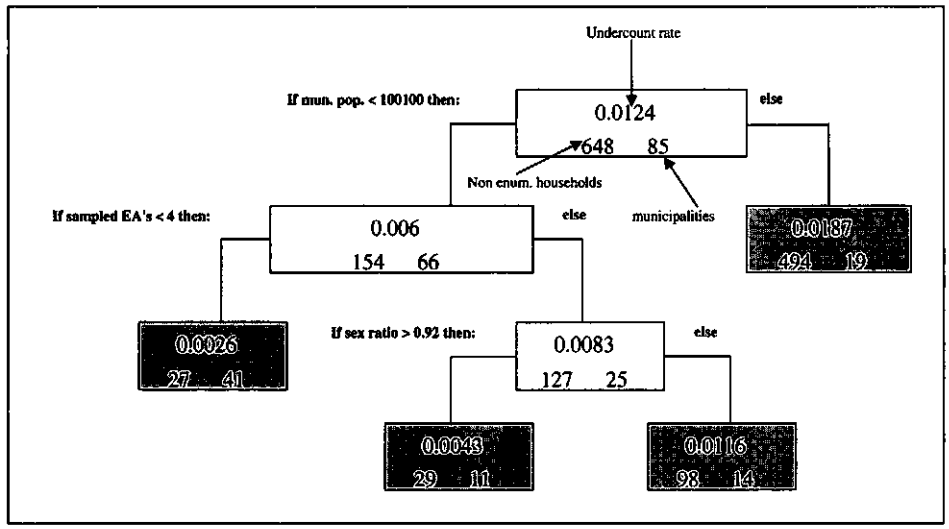

Figure 1. Tree 1 based on demographic variables.

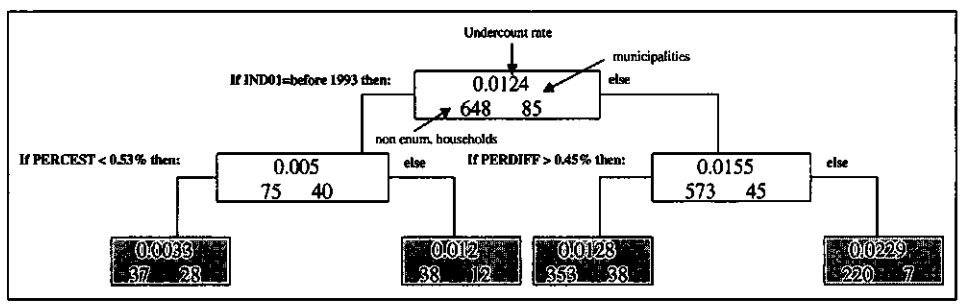

Figure 2. Tree 2 based on municipal statistical quality variables.

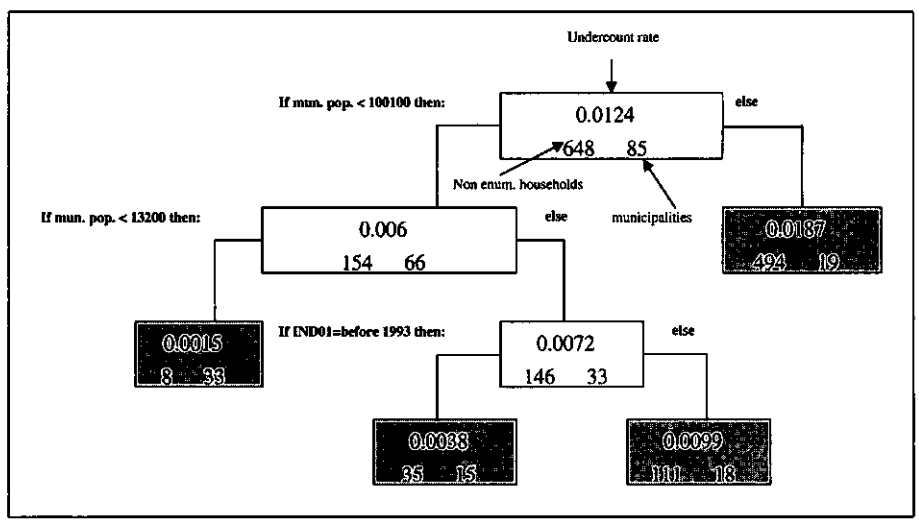

Figure 3. Tree 3 based on demographic and quality variables.

4. HIERARCHICAL POISSON-GAMMA MODELS

We denote the observed number of not enumerated households in each municipal sample with $y_i (i = 1, ..., 85)$. As an initial approximation, these counts can be modeled using a Poisson distribution:

$$y_i \mid \delta_i, e_i \sim \text{Pois}(\delta_i e_i) \tag{2}$$

where δ_i represents the rate of undercount to be estimated and e_i is given by the number of households in the sampled EAs within the municipality. Dependency on a set of explanatory variables is expressed by means of a canonical log-linear link:

$$\ln(\delta_i e_i) = X_i\beta + Z_i\xi \tag{3}$$

where Z_i is the i-th row of a categorical design matrix introduced for modelling group effects. Each X_i is a p-vector of explanatory variables associated with the i-th municipality and β and ξ are the regression parameters.

The occurrence of failure to enumerate is relatively rare when compared to the number of observed households. For this reason, the data may show strong overdispersion. Overdispersion can be managed by hierarchically modelling the parameters δ_i in (2). If the δ_i are Gamma(α, ν) distributed, the Negative Binomial distribution is marginally obtained for y_i by integrating out the parameters δ_i: *i.e.* $y_i \mid \alpha, \nu, e_i \sim \text{NegBin}(\alpha, \nu/(\nu + e_i))$ with moments:

$$E(y_i \mid \alpha, e_i, \nu) = \frac{\alpha e_i}{\nu}, \quad V(y_i \mid \alpha, e_i, \nu) = \frac{\alpha e_i(\nu + e_i)}{\nu^2}$$

(see Lawless 1987).

Instead of the parameterization above, we adopt the parameterization of the Gamma distribution at the second level of the hierarchy according to the proposal made by Christiansen and Morris (1997). When assuming

$$\delta_i \mid \lambda_i, \zeta \sim \text{Gamma}(\zeta, \zeta/\lambda_i) \tag{4}$$

with moments $E(\delta_i \mid \lambda_i, \zeta) = \lambda_i$ and $V(\delta_i \mid \lambda_i, \zeta) = \lambda_i^2/\zeta$, we have

$$y_i \mid e_i, \lambda_i, \zeta \sim \text{NegBin}\left(\zeta, \frac{\zeta/\lambda_i}{\zeta/\lambda_i + e_i}\right),$$

where $V(y_i \mid e_i, \lambda_i, \zeta) - E(y_i \mid e_i, \lambda_i, \zeta) = e_i^2\lambda_i^2/\zeta$. As ζ moves towards infinity, the variance of the Negative Binomial converges towards that of the Poisson (the variance of the Gamma in (4) tends towards 0), while small values of ζ point to high overdispersion.

From (4) it is immediate to see that:

$$E(\delta_i e_i \mid e_i, \lambda_i, \zeta) = \lambda_i e_i;$$

therefore the dependence assumption (3) is re-stated in terms of $\lambda_i e_i$:

$$\ln(\lambda_i e_i) = X_i\beta + Z_i\xi.$$

The prior (4) is conjugate to the likelihood defined by (2). Consequently one obtains

$$\delta_i \mid y_i, e_i, \lambda_i, \zeta \sim \text{Gamma}(y_i + \zeta, e_i + \zeta/\lambda_i)$$

from which it follows that

$$E(\delta_i \mid y_i, e_i, \lambda_i, \zeta) = (1 - B_i)\, r_i + B_i\lambda_i \tag{5}$$

where $r_i = y_i/e_i$ and $B_i = \zeta/(\zeta + e_i\lambda_i)$.

Each posterior mean (5) can be seen as a composite Small Area estimator where both the direct and the synthetic components are weighted according to the information available from the sample.

From (5) we note that the posterior mean of the distribution of the rate parameters δ_i is a linear combination of the observed undercount rate r_i and the prior mean λ_i. In other words, the model features posterior linearity. The two terms in (5) are weighted according to B_i, which varies between 0 and 1. The larger the B_i, the more the prior means λ_i (synthetic estimators) receive weight and the model estimates gain in importance compared with the observed rates. We note that each B_i is inversely proportional to the $e_i\lambda_i$, expressing the amount of information provided by the sample of each domain.

To complete the full Bayesian specification of the model we assign a distribution to the third level parameters ζ, β, ξ. According to an approximate non-informative criterion, we introduce proper, but flat, prior distributions. In particular we assume that:

$$\beta_j \overset{\text{iid}}{\sim} N(0, 100), \qquad j = 1, ..., p \tag{6}$$

$$\xi_k \overset{\text{ind}}{\sim} N\left(k\,\bar{u}_k, \frac{1}{\tau\,\bar{n}_k}\right), \qquad k = 1, ..., q \tag{7}$$

where $\bar{u}_k$ is the average undercount in the k-th group and $\bar{n}_k$ is the average number of sampled households in the municipalities of the same group. Priors (7), associated to group effects, are therefore centered on groups means and their precision is proportional to the group size. They are built to be weakly informative for improving the stability and convergence properties of the model. Priors for regression coefficients (6) associated to the remaining regressors are centered in 0. For the overdispersion parameter ζ we select the prior

$$\zeta \sim 1{,}000^{*}\ \text{Gamma}(0.001, 1) \tag{8}$$

following the suggestion given by Christiansen and Morris (1997). Note that the first two prior moments of (8) are $E(\zeta) = 1$ and $V(\zeta) = 1{,}000$; thus the prior is very diffuse and characterized by high positive skewness.

At the fourth level of the hierarchy we specify the following priors:

$$k \sim N(0, 100) \tag{9}$$

$$\tau \sim \text{Gamma}(0.001, 0.001). \quad (10)$$

which are both designed to have a very mild impact on posterior inferences.

We compute the posterior distributions of $(\delta_i \mid y_i, e_i)$ by using Markov chain Monte Carlo (McMC) sampling algorithms. For these calculations we use the software BUGS (Spiegelhalter, Thomas, Best and Gilks 1995), which is based on Gibbs sampling. Since the solution of models involving discrete distributions is computationally very demanding, we specify the prior distributions (6) – (10) by selecting simple well known functional forms, as Normal and Gamma, that facilitate fast computations. We examinated the sensitivity of the posterior means in (6) – (10), and we did not find any substantial changes in the posterior means. Hence, these priors can be considered noninformative. For the convergence assessment we consider the multiple chain approach suggested by Gelman and Rubin (1992), running three different chains with well separated starting points for each model. The visual inspection of the chains path and the modified Gelman and Rubin statistic (Brooks and Gelman 1998) are considered as basic convergence assessment tools. We run 10,000 iterations for each chain, discarding on average a conservative "burn in" of 3,000, thus yielding an approximate 20,000 draws from the posterior of each model.

5. MODEL COMPARISON AND DISCUSSION OF EMPIRICAL RESULTS

We estimated a variety of models for different definitions of the matrixes of regressors X and Z. As regards the design matrix Z we consider seven different cases, in which municipalities are grouped using either traditional stratification criteria (geographical area and demographic size) or the results of the partitioning techniques discussed in section 3. They are: a) geographical area (North, Center, South and Islands), b) demographic size classes only, c) demographic classes by geographical area, d) demographic size classes and geographical areas, e) Tree 1 (based on demographic variables), f) Tree 2 (based on quality variables), g) Tree 3 (based on both quality and demographic variables). Two kinds of variables may be proposed in matrix X: the quality variables of section 2.2 and the demographic variables of section 2.3. Matrix X has therefore three different possible compositions: I) quality variables only, II) demographic variables only, III) both quality and demographic variables. By matching the different definitions of X and Z, twenty-eight different models have been estimated. In this way we do not perform variable selection procedures, rather we introduce alternative blocks of variables.

The quantity commonly used for comparing models within the Bayesian framework is the Bayes factor (BF). A large sample approximation of $-2\ln(BF)$ is given by

$$\Delta\text{BIC} = -2\ln\left[\frac{\sup_{M_0} f(y \mid \theta_0)}{\sup_{M_k} f(y \mid \theta_k)}\right] - (p_k - p_0)\ln n \quad (11)$$

(see Schwarz 1978) which, moreover, makes no reference to the prior assumptions. We note that in (11) the $M_k (k = 1, ..., K)$ index the set of competing models and θ_k is the p_k dimensional parameter indexing the likelihood associated to each model. The null model against which all the others are compared is the one with the only intercept, and is denoted by M_0. Positive and large values of (11) support model M_k.

The complexity penalization in (11) depends on the size of the subset of third level parameters; that is, all models are compared as if they were non hierarchical. Since they share a similar hierarchical structure, this operational modification of the standard BIC criterion does not alter the results of model comparison summarized in Table 1.

We note that those models where group effects are based on geographical area perform very poorly (row 1), and the same happens when the geographical area is combined with the demographic size of the municipalities (rows 3 and 4). This is rather surprising, since geographical areas are employed in designing the stratification of the PES sample, and the efficiency of administrations, together with other social and economic indicators, are currently supposed to be clustered with respect to Italy's large geographical subdivisions (North, Center, South). This outcome may be ascribed to the predominant role that the specific organization of each municipality plays in determining the efficiency of Census operations within its territory.

Models with tree-based group effects (rows 5-7) clearly perform better than models with group effects based on ISTAT traditional stratification criteria (rows 1-4). The only exception to this behavior are those models relying on Tree 2 (row 5), which perform rather poorly when demographic size and other demographic variables are not included. In fact, the municipal population can be thought of as a proxy of municipal organizational complexity. It seems that quality variables are powerful in discriminating the level of undercount among municipalities with similar demographic features, but have little relevance when the effect of a different degree of organizational complexity is not accounted for by introducing a variable of demographic size. We point out that adding a design matrix Z based on Poisson regression trees grouping of municipalities allows us to model non linear relations between the undercount and the predictors.

Actually, the models based on Tree 3 provide the best performance. A number of comments about the model with maximum ΔBIC follow. This model uses demographic and quality variables as regressors. The adequacy of the selected model is assessed by means of posterior predictive checks. In particular the general purpose goodness-of-fit discrepancy measure proposed by Brooks, Catchpole and Morgan (2000) as a suitable tool for rare occurrencies as census undercounts:

$$D(y;\theta) = \sum_i \left(\sqrt{y_i} - \sqrt{\text{Exp}_i}\right)^2, \qquad (12)$$

where $\text{Exp}_i = e_i E(\delta_i \mid y_i, e_i)$, is adopted. The associated 0.46 tail area probability highlights a good fit for the selected model.

The set of models has been estimated again after eliminating the greatest municipality, which is potentially an influential case. Again, the model based on Tree 3 with demographic and quality variables as regressors has been selected using the criterion (11). This model shows a good fit (the Bayesian p-value associated to the discrepancy measure (12) is equal to 0.51). Moreover, composite estimates do not change much when compared with those obtained with the whole sample.

In order to check model fitting, in Figure 4, composite estimates against direct estimates of the number of not enumerated household in each municipality are plotted (the values of the largest 10 municipalities are reported with a different scale). The composite estimates are $w_i e_i E(\delta_i \mid y_i, e_i)$, while the direct estimates are $w_i y_i$, w_i being the expansion factor due to EA sampling in each municipality. Composite estimates are posterior expectations of first level parameters and, conditionally on the hyperparameters, are composite estimates in which the model predictions represented by the λ_i receive little weight when there is sound sampling evidence. From (5) we know that this weighting process is ruled by the municipal shrinkage factors B_i. They weight the direct estimates y_i/e_i in proportion to $e_i\lambda_i$, *i.e.* the number of not enumerated households within the municipal sample predicted by the model.

Table 1
ΔBIC of the estimated models compared with the reference model M_0

		Variables in the models			
		Only group effects	Group eff. + quality vars	Group eff. + demographic vars	Group eff. + quality and demographic vars
Group Effects	Area	-4.22	-0.39	18.52	23.32
	Classes of Mun. Pop.	15.34	17.87	17.32	20.09
	Area* Mun. Pop. Classes	2.08	6.13	4.91	8.45
	Area + Mun. Pop. Classes	9.68	13.20	13.74	17.83
	Tree2 (quality vars)	11.81	8.34	23.48	26.15
	Tree1 (demographic vars)	35.14	35.37	32.28	35.53
	Tree3 (quality + demographic vars)	38.89	35.76	41.12	41.45

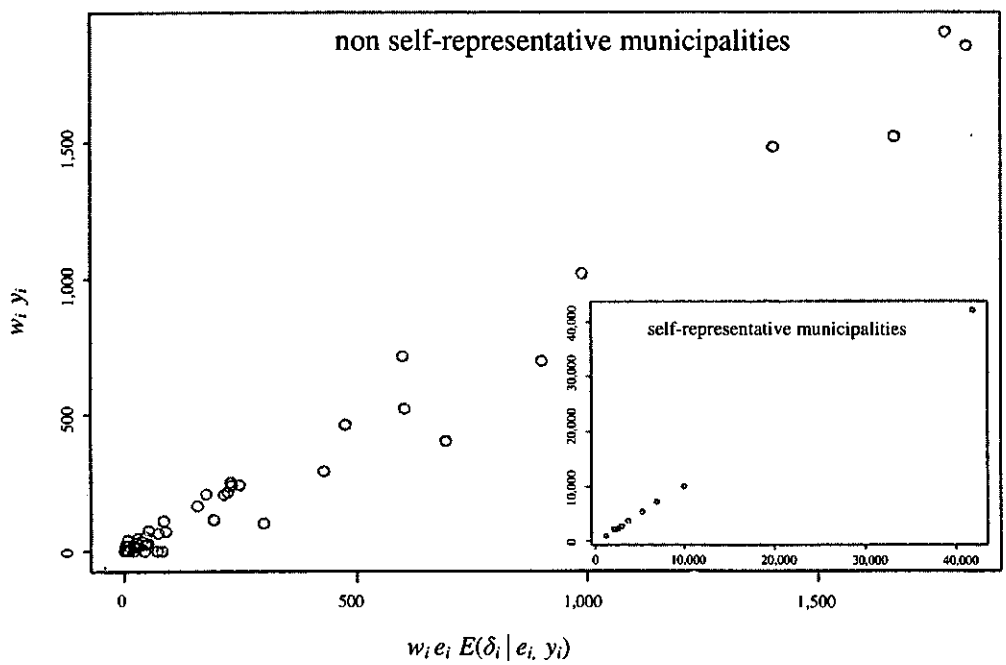

Figure 4. Composite estimates against direct estimates of the number of not enumerated households in each municipality.

For municipalities with resident population of up to 10,000 (this value is relatively close to the splitting value 13,200 of Tree 3) in almost all cases we have B_i values that are very close to 1; this means that, for small municipalities, the role of the model component in the determination of the composite estimate is overwhelming. In Figure 5 composite estimates (and their 95% credibility intervals) are plotted against direct estimates.

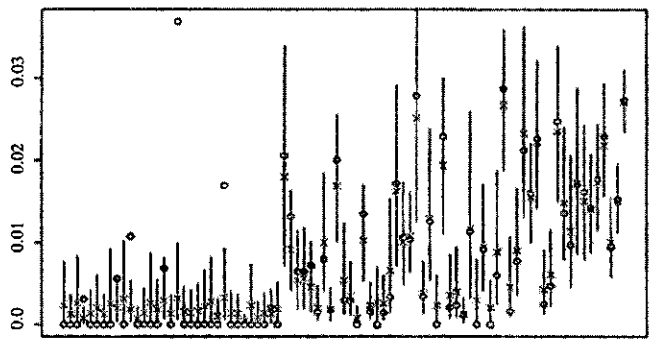

Figure 5. Composite estimates (×) and their 95% credibility intevals; (○) direct estimates. Municipalities are sorted by demographic size.

The width of the credibility intervals depends on the undercount level and, as should be expected, is large when the size of the sample within the municipality is small. Composite estimates associated with large credibility intervals are also characterized by large shrinkage factors, as a consequence of the scarce sample information. Large intervals for some middle-sized municipalities can be justified with the fact that they are under-sampled with respect to their size.

In small municipalities, where Census is conducted more easily, the undercount is generally very small. The undercount estimate is difficult since very few EAs are currently sampled from each of the small municipalities, often providing no evidence of undercount. In such cases, the composite estimate essentially consists in the model based component. Therefore, for the next PES, given the overall sample size, our suggestion is not to insist in sampling a great number of small municipalities, but to redirect sampling towards middle-sized municipalities, which are more etherogeneous. Moreover, the number of EAs to sample in the selected small municipalities ought to be increased.

The results of this work, which considers for the first time a criterion for grouping together municipalities according to their performance in statistical operations, confirm that an improvement may be reached for future similar surveys by modifying the stratified sampling design and by modelling undercount by means of the covariates mimicking the difficulties of the municipality behaviour in conducting censuses.

ACKNOWLEDGEMENTS

We would like to thank Angela Ferruzza, Marco Fortini, Aldo Orasi and Fernanda Panizon of the ISTAT team working on the 2001 Census and PES, together with Mariella Dimitri and Ersilia Di Pietro, of the ISTAT group working on surveys of statistical performance of municipalities, for their useful suggestions and continuous assistance.

The work has been partially funded by the (1999-2000) "Quality of total and partial surveys" Research Project grant from the University of Bologna (60%).

The PES data set and the archives containing the data on municipalities have been made available thanks to a special agreement between ISTAT and the Department of Statistics of the University of Bologna.

We would like to thank Francesca Bruno and Loredana Di Consiglio for their invaluable contribution in preparing the basic data sets, and Meri Raggi for her constant support and her discussion of the subjects of this research.

We thank the Editor, an Associate Editor and two anonymous referees for comments and suggestions which helped us in revising and improving the manuscript.

REFERENCES

ABBATE, C., MASSELLI, M. and SIGNORE M. (1993). A combined post-enumeration survey for the 1991 Italian population and industrial censuses. *Bulletin of the International Statistical Institute, Firenze, 48th Session.* Tome LV, 2, 159-173.

ALHO, J.M., MULRY, M.H., WURDEMAN, K. and KIM, J. (1993). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association.* 88, 1130-1136.

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A. and STONE, C.J., (1984). *Classification and Regression Trees.* Wadsworth, California.

BROOKS, S.P., CATCHPOLE, E.A. and MORGAN, B.J.T. (2000). Bayesian animal survival estimation. *Statistical Science.* 15, 357-276.

BROOKS, S.P., and GELMAN, A. (1998). Alternative methods for monitoring convergence of iterative simulation. *Journal of Computational and Graphical Statistics.* 7, 434-455.

CHRISTIANSEN, C.L., and MORRIS, C. (1997). Hierarchical Poisson regression models. *Journal of the American Statistical Association.* 92, 618-632.

DI PIETRO, E. (1998). Anagrafi comunali: funzione statistica e livello di informatizzazione. *Atti Della Quarta Conferenza Nazionale di Statistica.* Tomo 1 - Sessioni Plenarie, Workshop: Il progetto anagrafi. Roma 11-13 novembre.

DI PIETRO, E. (1999). Anagrafe informatizzata e Censimenti demografici: dal censimento tradizionale al censimento basato sugli Archivi. *Società Italiana di Statistica: Atti Del Convegno "Verso i Censimenti del 2000.* Udine 7-9 giugno. 169-182.

FORTINI, M. (1994). Un'applicazione del modello a classi latenti per l'analisi dell'errore di copertura del XIII censimento della popolazione. *Atti della XXXVII Riunione Scientifica della Società Italiana di Statistica*. San Remo 6-8 Aprile. 2, 423-430.

GELMAN, A., and RUBIN, D.B. (1992). Inference from iterative simulation using multiple sequence. *Statistical Science*. 7, 457-72.

LAWLESS, J.F. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*. 15, 209-225.

MOURA, F.A.S., and HOLT, D. (1999). Small area estimation using multilevel models. *Survey Methodology*. 25, 73-80.

SCHWARTZ, G. (1978). Estimating the dimension of a model. *Annals of Statistics*. 6, 461-464.

SPIEGELHALTER, D.J., THOMAS, A., BEST, N. and GILKS, W.R. (1995). *BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.50*. Technical Report, Medical Research Council biostatistics Unit, Institute of Public Health, Cambridge University.

THERNEAU, T.M., and ATKINSON, E.J. (1997). *An Introduction to Recursive Partitioning Using the RPART Routines*. Technical report, Mayo Foundation.

WOLTER, K. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*. 81, 338-346.

Survey Methodology, December 2003
Vol. 29, No. 2, pp. 177-180
Statistics Canada

Estimation of a Measure of Disclosure Risk for Survey Microdata Under Unequal Probability Sampling

C.J. SKINNER and R.G. CARTER[1]

ABSTRACT

Skinner and Elliot (2002) proposed a simple measure of disclosure risk for survey microdata and showed how to estimate this measure under sampling with equal probabilities. In this paper we show how their results on point estimation and variance estimation may be extended to handle unequal probability sampling. Our approach assumes a Poisson sampling design. Comments are made about the possible impact of departures from this assumption.

KEY WORDS: Confidentiality protection; Finite population inference; Poisson sampling; Statistical disclosure control; Uniqueness.

1. INTRODUCTION

Microdata files of survey data may be of great analytic value to researchers. When deciding whether and how to make such files available, agencies conducting surveys need to protect against risks of possible statistical disclosure (Willenborg and de Waal 2001). Skinner and Elliot (2002, abbreviated henceforth to SE) proposed a simple measure of statistical disclosure risk for survey microdata, for use as evidence to inform such decisions. They showed that this measure may be estimated simply under sampling with equal probabilities. In this paper we show how their results may be extended to handle unequal probability sampling.

The measure is introduced in section 2. Point estimation and variance estimation for the measure are considered in sections 3 and 4 respectively. See SE for the relation of this measure to the literature on statistical disclosure risk.

2. THE MEASURE OF DISCLOSURE RISK

We consider the possible release of a microdata file consisting of a set of records for units (*e.g.*, individuals or households) in a sample s, selected by probability sampling from a population U. Each record consists of a vector of values of a specified set of variables for the given unit. Following a standard approach to disclosure risk assessment (*e.g.*, Bethlehem, Keller and Pannekoek 1990), we suppose that an intruder attempts to match the microdata records to known population units using a specified subset of variables. We assume that these 'identifying variables' are categorical and that the possible combinations of their values define the categories $1, ..., J$ of a variable X. (J will usually be very large).

We suppose that the intruder is able to determine the value of X for a population unit with known identity and that the intruder 'claims' that a microdata record has been identified if and only if this value matches the value of X recorded in the microdata for *just one* microdata record. Assuming (a) that the population unit with known identity is randomly drawn from U with equal probabilities and (b) that the value of X for this unit is measured in the same way that X is measured in the microdata, the probability that the intrude's claim is correct is:

$$\theta = \Pr(\text{correct match} \mid \text{unique match})$$

$$= \sum_{j=1}^{J} I(f_j = 1) \Big/ \sum_{j=1}^{J} F_j I(f_j = 1),$$

where f_j and F_j are the frequencies of units in s and U respectively, for which $X = j$ and where $I(.)$ is the indicator function ($I(A) = 1$ if A is true, $I(A) = 0$ otherwise). The numerator of θ is the number of microdata records which are unique in the microdata with respect to X and the denominator of θ is the number of units in the population which share the same value of X with any of these records.

The quantity, θ, is the measure of disclosure risk considered in this paper. To protect against disclosure, θ might be estimated under alternative forms of microdata release (implying alternative specifications of X) and a form of release chosen so that θ is inferred to be acceptably small. A sensitivity analysis will usually be required in which the specification of X is varied not only according to the form of release but also to allow for alternative plausible forms of external information which an intruder might hold about known population units. For example, one might consider both an intruder with access only to publicly available information, such as the visible characteristics of an individual, and an intruder with access to a private database held by an organisation.

[1] C.J. Skinner, University of Southampton, Southampton, United Kingdom, S017 1BJ and R.G. Carter, Statistics Canada, B-2 Jean Talon Building, Ottawa, Ontario, K1A 0T6.

3. ESTIMATION OF θ

We suppose that the data consist of the values of X for the sample units. Hence, the sample frequencies f_j are known but the population frequencies F_j are unknown ($j = 1, ..., J$). The 'parameter' of interest, θ, is also unknown and must be estimated. We adopt a design-based approach to inference in which the f_j are random and the F_j are fixed. As discussed by SE, the 'parameter', θ, therefore depends on s, unlike standard finite population parameters considered in survey sampling.

SE motivate a point estimator of θ by a resampling argument, which may be generalised to the case of unequal probability sampling, as follows.

Repeat the following steps K times.

Step 1: remove a single unit i from the microdata sample s with probability

$$\alpha_i = \pi_i^{-1} / \sum_s \pi_i^{-1},$$

where π_i is the (first-order) inclusion probability of unit i;

Step 2: copy the removed unit back into the sample with probability π_i;

Step 3: record whether the removed unit matches a unique record in the microdata and whether this match is correct.

The idea is that Step 1 mimics the intruder's (equal probability) selection of a unit from U (using the inverse sampling idea of Hinkins, Oh and Scheuren 1997). Step 2 mimics the inclusion of that unit in s. The estimator of θ is the empirical proportion of unique matches which are correct. Following the argument of SE, this estimator converges almost surely as $K \to \infty$ to

$$\hat{\theta} = \frac{\sum_{s^{(1)}} \Pr(\text{unit } i \text{ removed and then copied back})}{\left[\sum_{s^{(1)}} \Pr(\text{unit } i \text{ removed and then copied back}) + \sum_{s^{(2)}} \Pr(\text{unit } i \text{ removed and then not copied back})\right]}$$

$$= \sum_{s^{(1)}} \alpha_i \pi_i \Big/ \left[\sum_{s^{(1)}} \alpha_i \pi_i + \sum_{s^{(2)}} \alpha_i (1 - \pi_i)\right]$$

$$= n^{(1)} \Big/ \left[n^{(1)} + \sum_{s^{(2)}} (\pi_i^{-1} - 1)\right], \qquad (1)$$

where $s^{(1)}$ is the subsample of unique units in s, $s^{(2)}$ is the subsample of units which occur in pairs and $n^{(1)} = \sum_j I(f_j = 1)$ is the size of $s^{(1)}$. In the case of equal probability sampling with $\pi_i = \pi$, $\hat{\theta}$ reduces to $n^{(1)} / [n^{(1)} + 2n^{(2)} (\pi^{-1} - 1)]$, where $2n^{(2)} = 2\sum_j I(f_j = 2)$ is the size of $s^{(2)}$, as in SE.

We are interested in $\hat{\theta}$, defined in (1), as an estimator of θ. SE show that $\hat{\theta}$ is consistent for θ in the equal probability sampling case. The basic steps of their argument may be generalised to the case of unequal probability sampling as follows. We may write

$$\theta = n^{(1)} \Big/ \left[n^{(1)} + \sum_j (F_j - 1) I(f_j = 1)\right]. \qquad (2)$$

Hence, by comparing (1) and (2), $\hat{\theta}$ will be a 'good' estimator of θ if $\sum_{s^{(2)}} (\pi_i^{-1} - 1)$ is a 'good' estimator of $\sum_j (F_j - 1) I(f_j = 1)$. In fact, we prove in Appendix 1 that the latter estimator is unbiased, that is

$$E\left[\sum_{s^{(2)}} (\pi_i^{-1} - 1)\right] = E\left[\sum_j (F_j - 1) I(f_j = 1)\right], \qquad (3)$$

under the assumption of Poisson sampling, that is where population units are sampled independently. Equation (3) generalizes Proposition 2 of SE. In the equal probability sampling case SE show how the result in equation (3) may be extended to prove consistency of $\hat{\theta}$ as an estimator of θ, using an asymptotic framework where $J \to \infty$ and under some regularity conditions, in particular that the F_j are bounded.

Having established the main unbiasedness result in (3), we conjecture that this consistency result will generalise to the case of unequal probability Poisson sampling, subject to additional weak conditions on the π_i, for example that the π_i are bounded above by a positive constant.

The Poisson sampling assumption generalises the Bernoulli sampling assumption in SE. They conclude that in practice $\hat{\theta}$ will remain approximately unbiased for θ under a number of other equal probability sampling designs including simple random sampling, (equal probability) systematic sampling or proportionate stratified simple random sampling. We suggest that in a similar way $\hat{\theta}$ will remain approximately unbiased for θ under corresponding unequal probability designs, *i.e.*, disproportionate stratified simple random sampling and unequal probability systematic sampling. We also suggest that it may be reasonable to allow for nonresponse in $\hat{\theta}$ if s is the set of respondents and if π_i^{-1} consists of a weight which may be interpreted as the reciprocal of the estimated probability of both being sampled and responding.

As discussed in SE, the form of sampling which seems to have the potential to lead to most bias in $\hat{\theta}$ as an estimator of θ in practice is multistage sampling, where the multistage units are strongly related with respect to X. For example, bias might be non-negligible when households form clusters within which all adult individuals are sampled, where the microdata includes individual-level records and where X is primarily determined by household-level variables. This might lead to a higher value of $n^{(2)}/n^{(1)}$ than expected under Poisson sampling and hence to underestimation of θ. Such an example is somewhat contrived, however, and we suspect that the bias of $\hat{\theta}$ as an estimator of θ will be modest in most typical social surveys.

4. VARIANCE ESTIMATION

SE present a linearization estimator of $\text{var}(\hat{\theta}-\theta)$, which depends on $n^{(1)}$ and $n^{(2)}$, like $\hat{\theta}$, as well as on $n^{(3)}=\sum_j I(f_j=3)$, the number of values of X for which there are exactly three microdata records. We show in Appendix 2 that this variance estimator may be generalised, in the case of unequal probability Poisson sampling, to

$$\hat{v} = \hat{\theta}^2 \frac{\sum_{j=1}^{J}\left\{I(f_j=3)(\gamma_{1j}^2-\gamma_{2j})+I(f_j=2)(\gamma_{1j}^2+\gamma_{1j})\right\}}{\left[n^{(1)}+\sum_j I(f_j=2)\gamma_{1j}\right]^2} \quad (4)$$

where $\gamma_{1j}=\sum_{s_j}\beta_i$, $\gamma_{2j}=\sum_{s_j}\beta_i^2$, $\beta_i=\pi_i^{-1}-1$ and $s_j=\{i\in s; X_i=j\}$, where X_i is the value of X for unit i.

Note that, in this notation, we may write

$$\hat{\theta} = n^{(1)} \Big/ \left[n^{(1)}+\sum_j I(f_j=2)\gamma_{1j}\right].$$

As in the equal probability case, both $\hat{\theta}$ and $\hat{v}$ can be computed straightforwardly from the values X_i and π_i for $i\in s$. The expression given above for $\hat{v}$ reduces to the expression given in Proposition 3 of SE when $\pi_i=\pi$ for all $i\in s$.

The linearisation argument which gives $\hat{v}$ assumes J is large. This seems a weak condition relative to the assumption of Poisson sampling. The linearisation variance estimator does not appear to generalise straightforwardly to other complex sampling designs. This is because the linearised form of $\hat{\theta}-\theta$ depends on the F_j and these cannot simply be replaced by consistent estimators. It also does not appear to be straightforward to apply replication methods to estimate the variance of $\hat{\theta}-\theta$, since θ is unknown and, as indicated by the simulation study in SE, the variance of θ may not be negligible in practice relative to the variance of $\hat{\theta}$.

5. CONCLUDING REMARKS

The estimated measure $\hat{\theta}$ considered in this paper may be used as evidence in assessing whether or not a proposed microdata file has an acceptable level of disclosure risk. The aim may be to ensure that $\hat{\theta}$ does not exceed some specified probability. To allow for sampling variation in $\hat{\theta}$ a more conservative procedure would be to require that the upper bound of a confidence interval for θ, say $\hat{\theta}+2\hat{v}^{1/2}$, does not exceed the specified probability.

As well, $\hat{\theta}$ may be used to compare alternative strategies to control disclosure risk. For example, variables may be included in microdata files with more or less classification detail. Greater detail may enhance the value of the file for analysis but may also increase disclosure risk if the variable could be used to match against external information. The estimated measure $\hat{\theta}$ could, therefore, be used to assess the relative risk resulting from different ways of collapsing the level of classification in specific identifying variables, including geography.

The measure may be estimated not only for the population as a whole, but also for subpopulations. Such a breakdown of the measure permits a more realistic assessment of the risk posed by intruders who target specific subpopulations. Such a targeted threat invalidates the basic assumption underlying the definition of whole population measure, θ, that the population unit with known identity is randomly drawn from U with equal probabilities. Separate estimation of the measure in different strata with different sampling fractions also provides a simple method of handling unequal probabilities of selection. This paper has shown how to allow for more general sources of unequal probability sampling in $\hat{\theta}$ and $\hat{v}$. More research is required to assess the robustness of these estimators to departures from Poisson sampling, especially multi-stage sampling.

A potential problem with estimating the measure separately by subpopulations is the impact of the reduction in sample size. SE found $\hat{\theta}$ to be stable in their numerical investigations, with a coefficient of variation never exceeding 6%. Their minimum sample size was, however, about 9,000 so further numerical work is needed to assess the stability of $\hat{\theta}$ for smaller sample sizes. The proposed variance estimation method provides some guidance for any specific case. Stability could, in principle, be improved by the use of model assumptions and one of us (CJS) is conducting further research on the limiting case of a small subpopulation, a single unit, extending θ to a record-level measure of risk analogous to that considered by Skinner and Holmes (1998).

APPENDIX 1

Proof of Equation (3)

Let $\beta_i=\pi_i^{-1}-1$ and $U_j=\{i\in U; X_i=j\}, j=1,...,J$, where X_i denotes the value of X for unit i. The size of U_j is F_j. Instead of labelling units in U by the single index i, consider the double index (jk), $j=1,...,J$, $k=1,...,F_j$, so that, for example, $\pi_{(jk)}$ denotes the inclusion probability for the k-th unit in U_j and $\beta_{(jk)}=\pi_{(jk)}^{-1}-1$. Under Poisson sampling the right side of (3) is

$$E\left[\sum_j (F_j-1)I(f_j=1)\right]$$

$$= \sum_{j=1}^{J}(F_j-1)\sum_{k=1}^{F_j}\pi_{(jk)}\prod_{\substack{\ell=1\\ \ell\neq k}}^{F_j}(1-\pi_{(j\ell)}) \quad (A.1)$$

and the left side of (3) is

$$E\left[\sum_{i\in s^{(2)}}\beta_i\right]$$

$$=\sum_{j=1}^{J}\sum_{k=1}^{F_j}\sum_{\substack{\ell=1\\k<\ell}}^{F_j}\pi_{(jk)}\pi_{(j\ell)}\left[\prod_{\substack{m=1\\m\neq k,\ell}}^{F_j}(1-\pi_{(jm)})\right][\beta_{(jk)}+\beta_{(j\ell)}]$$

$$=\sum_{j=1}^{J}\sum_{k=1}^{F_j}\sum_{\substack{\ell=1\\k\neq\ell}}^{F_j}\pi_{(jk)}\pi_{(j\ell)}\left[\prod_{\substack{m=1\\m\neq k,\ell}}^{F_j}(1-\pi_{(jm)})\right]\beta_{(jk)}$$

$$=\sum_{j=1}^{J}\sum_{k=1}^{F_j}\sum_{\substack{l=1\\k\neq\ell}}^{F_j}\pi_{(jl)}\left[\prod_{\substack{m=1\\m\neq\ell}}^{F_j}(1-\pi_{(jm)})\right]$$

$$=\sum_{j=1}^{J}(F_j-1)\sum_{\ell=1}^{F_j}\pi_{(j\ell)}\left[\prod_{\substack{m=1\\m\neq\ell}}^{F_j}(1-\pi_{(jm)})\right]$$

which is identical to (A.1) so (3) follows.

APPENDIX 2

Derivation of Linearisation Variance Estimator

Write $\hat{\theta}-\theta=\tau_1/(\tau_1+\tau_2)-\tau_1/\tau_3$, where

$$\tau_1=\sum_j I(f_j=1),\ \tau_2=\sum_j I(f_j=2)\gamma_{1j},\ \tau_3=\sum_j F_j I(f_j=1).$$

Let $\mu_t=E(\tau_t)$, t = 1,2,3, and note that $\mu_1+\mu_2=\mu_3$ from (3). A linearised expression for $\hat{\theta}-\theta$ is $\mu_1(-\tau_1-\tau_2+\tau_3)/\mu_3^2$, the variance of which may be expressed as

$$\begin{aligned}&\mathrm{var}(\hat{\theta}-\theta)\\ &\approx \mathrm{var}\left[(\mu_1/\mu_3^2)\sum_{j=1}^{J}\{(F_j-1)I(f_j=1)-\gamma_{1j}I(f_j=2)\}\right]\\ &=(\mu_1/\mu_3^2)^2\sum_{j=1}^{J}\left[(F_j-1)^2\Pr(f_j=1)+E\{\gamma_{1j}^2I(f_j=2)\}\right]. \quad (A.2)\end{aligned}$$

This generalises the expression for the variance in Proposition 3 of SE. The expression for $\hat{v}$ in (4) is obtained by replacing terms in (A.2) by their unbiased estimators. First, μ_1 and μ_3 are estimated by τ_1 and $\tau_1+\tau_2$ respectively so that μ_1/μ_3^2 is estimated by $\hat{\theta}/(\tau_1+\tau_2)$. Next note that

$$E\left[I(f_j=3)(\gamma_{1j}^2-\gamma_{2j})\right]$$

$$=\sum_{k=1}^{F_j}\sum_{\substack{\ell=1\\k\neq\ell\neq m}}^{F_j}\sum_{m=1}^{F_j}\pi_{(jk)}\pi_{(j\ell)}\pi_{(jm)}\left[\prod_{\substack{n=1\\n\neq k,\ell,m}}^{F_j}(1-\pi_{(jn)})\right]\beta_{(j\ell)}\beta_{(jm)})$$

$$=\sum_{k=1}^{F_j}\sum_{\substack{\ell=1\\k\neq\ell\neq m}}^{F_j}\sum_{m=1}^{F_j}\pi_{(jk)}\left[\prod_{\substack{n=1\\n\neq k}}^{F_j}(1-\pi_{(jn)})\right]$$

$$=(F_j-1)\ (F_j-2)\Pr(f_j=1),$$

using the notation of Appendix 1. We may also show that

$$E\left[I(f_j=2)\gamma_{1j}\right]=(F_j-1)\Pr(f_j=1) \qquad (A.3)$$

by following the proof of (3) in Appendix 1, but omitting the summation over j. (Note that the sides of (3) are equal to the corresponding sides of (A.3) summed over j). Hence, an unbiased estimator of $(F_j-1)^2\Pr(f_j=1)$ is

$$I(f_j=3)(\gamma_{1j}^2-\gamma_{2j})+I(f_j=2)\gamma_{1j}.$$

It follows that the numerator of the expression for $\hat{v}/\hat{\theta}^2$ in (4) is unbiased for the second part of the expression on the right side of (A.2) (omitting $(\mu_1/\mu_3^2)^2$) as required.

REFERENCES

BETHLEHEM, J.G., KELLER, W.J. and PANNEKOEK, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*. 85, 38-45.

HINKINS, S., OH, H.L. and SCHEUREN, F. (1997). Inverse sampling design algorithms. *Survey Methodology*. 23, 11-21.

SKINNER, C.J., and ELLIOT, M.J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society, Series B*. 64, 855-867.

SKINNER, C.J., and HOLMES, D.J. (1998). Estimating the re-identification risk per record for microdata. *Journal of Offical Statistics*. 14, 361-372.

WILLENBORG, L., and DE WAAL, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer.

Survey Methodology, December 2003
Vol. 29, No. 2, pp. 181-188
Statistics Canada

Inference for Partially Synthetic, Public Use Microdata Sets

J.P. REITER[1]

ABSTRACT

To avoid disclosures, one approach is to release partially synthetic, public use microdata sets. These comprise the units originally surveyed, but some collected values,.for example sensitive values at high risk of disclosure or values of key identifiers, are replaced with multiple imputations. Although partially synthetic approaches are currently used to protect public use data, valid methods of inference have not been developed for them. This article presents such methods. They are based on the concepts of multiple imputation for missing data but use different rules for combining point and variance estimates. The combining rules also differ from those for fully synthetic data sets developed by Raghunathan, Reiter and Rubin (2003). The validity of these new rules is illustrated in simulation studies.

KEY WORDS: Confidentiality; Disclosure; Multiple imputation; Synthetic data.

1. INTRODUCTION

When releasing data to the public, statistical agencies seek to provide detailed data without disclosing respondents' sensitive information. To reduce the risk of disclosures, agencies typically alter the original data for public release, for example by recoding variables, swapping data, or adding random noise to data values (Willenborg and de Waal 2001). However, these methods can distort relationships among variables in the data set. They also complicate analyses for users: to analyze properly perturbed data, users should apply the likelihood-based methods described by Little (1993) or the measurement error models described by Fuller (1993). These are difficult to use for non-standard estimands and may require analysts to learn new statistical methods and specialized software programs.

An alternative approach was proposed by Rubin (1993): release fully synthetic data sets comprised entirely of multiply-imputed data rather than actual values. This can protect confidentiality, since identification of units and their sensitive data can be difficult when the released data are not actual, collected values. And, with appropriate imputation and estimation methods based on the concepts of multiple imputation (Rubin 1987), the approach can allow data users to obtain valid inferences using standard, complete-data statistical methods and software. Such inferences can be made using the methods developed by Raghunathan *et al.* (2003), whose rules for combining point and variance estimates differ from those of Rubin (1987). Other discussions and variants of synthetic data approaches appear in Little (1993); Fienberg, Steele and Makov (1996); Fienberg, Makov and Steele (1998); Dandekar, Cohen and Kirkendall (2002a); Dandekar, Domingo-Ferrer and Sebe (2002b); Franconi and Stander (2002, 2003); Polettini, Franconi and Stander (2002); Polettini (2003) and Reiter (2002, 2003).

Although no data producers have adopted the fully synthetic approach on a production basis yet, some have adopted a variant of the approach: release partially synthetic data sets comprising a mix of actual and multiply-imputed values. For example, to protect data in the U.S. Survey of Consumer Finances, the U.S. Federal Reserve Board replaces monetary values at high disclosure risk with multiple imputations, then releases a mixture of these imputed values and the unreplaced, collected values (Kennickel 1997). Another partially synthetic approach has been implemented by Abowd and Woodcock (2001) to protect data in longitudinal, linked data sets. They replace all values of some sensitive variables with multiple imputations, but leave other variables at their actual values. A third approach has been implemented by Liu and Little (2002), who develop an algorithm for simulating multiple values of key identifiers for selected units. All these partially synthetic approaches are appealing because they promise to maintain many of the benefits of fully synthetic data – protecting confidentiality while allowing users to make inferences without learning complicated statistical methods or software – with decreased sensitivity to the specification of imputation models.

Even though partially synthetic data sets are being publicly released, the literature does not contain technical results on how to obtain inferences from them. At first glance, it may appear appropriate to use the inferential methods for multiple imputation of missing data in Rubin (1987). Unfortunately, as shown in this article, these methods can result in biased variance estimates. Furthermore, and also as shown, the methods developed by Raghunathan *et al.* (2003) for analyzing fully synthetic data are not valid when applied on partially synthetic data. New methods of inference are required.

This paper describes methods for obtaining inferences from multiply-imputed, partially synthetic data sets. The derivation of these methods also provides prescriptions for

[1] J.P. Reiter, Institute of Statistics and Decision Sciences, Box 90251, Duke University, Durham, NC 27708-0251. E-mail: jerry@stat.duke.edu

generating partially synthetic data. The paper is organized as follows. Section 2 presents the new methods of inference. Section 3 shows a derivation of these methods from a Bayesian perspective, and it discusses conditions under which the resulting inferences should be valid from a frequentist perspective. Section 4 describes simulation studies that illustrate the validity of these methods, as well as the ineffectiveness of competing rules for combining multiple point and variance estimates. Section 5 concludes with suggestions of future areas of research.

2. INFERENCES FROM MULTIPLY-IMPUTED, PARTIALLY SYNTHETIC DATA SETS

Let $I_j = 1$ if unit j is selected in the original survey, and $I_j = 0$ otherwise. Let $I = (I_1, ..., I_N)$. Let Y_{obs} be the $n \times p$ matrix of collected (real) survey data for the units with $I_j = 1$; let Y_{nobs} be the $(N-n) \times p$ matrix of unobserved survey data for the units with $I_j = 0$; and, let $Y = (Y_{obs}, Y_{nobs})$. For simplicity, we assume that all sampled units fully respond to the survey. Let X 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 approximately for all population units. It may come, for example, from census records or the sampling frame(s).

The agency releasing synthetic data, henceforth abbreviated as the imputer, constructs synthetic data sets based on the observed data, $D = (X, Y_{obs}, I)$, in a two-part process. First, the imputer selects the values from the observed data that will be replaced with imputations. Second, the imputer imputes new values to replace those selected values. Let $Z_j = 1$ if unit j is selected to have any of its observed data replaced with synthetic values, and let $Z_j = 0$ for those units with all data left unchanged. Let $Z = (Z_1, ..., Z_n)$. Let $Y_{rep,i}$ be all the imputed (replaced) values in the i-th synthetic data set, and let Y_{nrep} be all unchanged (unreplaced) values of Y_{obs}. The $Y_{rep,i}$ are assumed to be generated from the Bayesian posterior predictive distribution of $(Y_{rep,i} \mid D, Z)$. The values in Y_{nrep} are the same in all synthetic data sets. Each synthetic data set, d_i, then comprises $(X, Y_{rep,i}, Y_{nrep}, I, Z)$. Imputations are made independently for $i = 1, ..., m$ times to yield m different synthetic data sets. These synthetic data sets are released to the public.

The values in Z can and frequently will depend on the values in D. For example, the imputer may choose to simulate sensitive variables or identifiers only for units in the sample with rare combinations of identifiers; or, the imputer may replace only those incomes above $100,000 with imputed values. To avoid bias, imputers should account for such selections by imputing from the posterior predictive distribution of Y for those units with $Z_j = 1$. In practice, this can be done by using only the units with $Z_j = 1$ as the data when finding the posterior distributions for imputations. Using all units with $I_j = 1$ can result in biased estimates or wider confidence intervals with overly conservative coverage rates, as illustrated in the simulations of section 4.

From these synthetic data sets, some user of the publicly released data, henceforth abbreviated as the analyst, seeks inferences about some estimand $Q = Q(X, Y)$, where the notation $Q(X, Y)$ means that Q is a function of (X, Y). For example, Q could be the population mean of Y or the population regression coefficients of Y on X. In each synthetic data set d_i, the analyst estimates Q with some point estimator q and estimates the variance of q with some estimator v. It is assumed that the analyst determines the q and v as if the synthetic data were in fact collected data from a random sample of (X, Y) based on the actual survey design used to generate I.

For $i = 1, ..., m$, let q_i and v_i be respectively the values of q and v in synthetic data set d_i. Under certain conditions to be described in section 3, the analyst can obtain valid inferences for scalar Q by combining the q_i and v_i. Specifically, the following quantities are needed for inferences:

$$\bar{q}_m = \sum_{i=1}^{m} q_i / m \tag{1}$$

$$b_m = \sum_{i=1}^{m} (q_i - \bar{q}_m)^2 / (m-1) \tag{2}$$

$$\bar{v}_m = \sum_{i=1}^{m} v_i / m. \tag{3}$$

The analyst then can use $\bar{q}_m$ to estimate Q and

$$T_p = b_m / m + \bar{v}_m \tag{4}$$

to estimate the variance of $\bar{q}_m$. When q is a function of only (X, Y_{nrep}, I) and not any imputed values, the synthetic data inferences are identical to the observed data inferences; that is, the $q_i = q_{obs}$ and $v_i = v_{obs}$ for all i, and the $b_m = 0$. When n is large, inferences for scalar Q can be based on t-distributions with degrees of freedom $v_p = (m-1)(1 + r_m^{-1})^2$, where $r_m = (m^{-1} b_m / \bar{v}_m)$. In many cases, r_m^{-1} and hence v_p will be large enough that a normal distribution provides an adequate approximation to the t-distribution. Extensions for multivariate Q are not presented here.

T_p differs from the variance estimator for multiple imputation of missing data, $T_m = (1 + 1/m) b_m + \bar{v}_m$ (Rubin 1987). In the partially synthetic data context, the $\bar{v}_m$ estimates $\mathrm{Var}(q_{obs})$ and the b_m / m estimates the additional variance due to using a finite number of imputations. In the missing data context, the $\bar{v}_m$ and b_m / m have the same interpretations, but an additional b_m is needed to average over the nonresponse mechanism (Rubin 1987, Chapter 4). This additional averaging is unnecessary in partially synthetic data settings, since the selection mechanism Z, which is set by the imputer, is not treated as stochastic.

T_p also differs from the variance estimator for analyzing fully synthetic data, $T_s = (1 + 1/m)\, b_m - \bar{v}_m$ (Raghunathan *et al.* 2003). To generate fully synthetic data, new units are sampled off the frame(s) for each synthetic data set, and their data are imputed. As shown by Raghunathan *et al.* (2003), this re-sampling and imputation process results in $b_m - \bar{v}_m$ as an appropriate estimate of $\mathrm{Var}(q_{\mathrm{obs}})$. For partially synthetic data, the original units are released for each data set, so that $\bar{v}_m$ is an appropriate estimate of $\mathrm{Var}(q_{\mathrm{obs}})$.

3. JUSTIFICATION OF NEW COMBINING RULES

This section shows a Bayesian derivation of the inferences described in section 2 and conditions under which these inferences are valid from a frequentist perspective. These results are based on, and closely follow, the theory developed in Raghunathan *et al.* (2003).

3.1 Bayesian Derivation

For this derivation, we assume that the analyst and imputer use the same Bayesian model. The posterior distribution for $(Q \mid d^m)$, where $d^m = \{d_1, d_2, ..., d_m\}$, can be decomposed as

$$f(Q \mid d^m) = \int f(Q \mid d^m, D, B) f(D \mid d^m, B) f(B \mid d^m)\, dD\, dB \quad (5)$$

where $B = \mathrm{Var}(q_i \mid D, Z)$. The integration with respect to $f(D \mid d^m, B)dD$ is only over the values of Y_{obs} that are replaced with imputations; the $(X, Y_{\mathrm{nrep}}, I)$ components of D remain fixed. Given D, the synthetic data are irrelevant, so that $f(Q \mid d^m, D, B) = f(Q \mid D)$. We assume standard Bayesian asymptotics hold, so that $f(Q \mid D) \sim N(q_{\mathrm{obs}}, v_{\mathrm{obs}})$, where q_{obs} and v_{obs} are the posterior mean and variance of Q determined using D.

Integrating (5) over D, we obtain $f(Q \mid d^m, B)$. Since only q_{obs} and v_{obs} are needed for inferences about $(Q \mid D)$, for $f(D \mid d^m, B)$ it is sufficient to determine $f(q_{\mathrm{obs}}, v_{\mathrm{obs}} \mid d^m, B)$. We assume imputations are made so that, for all i, $(q_i \mid D, B) \sim N(q_{\mathrm{obs}}, B)$ and $(v_i \mid D, B) \sim (v_{\mathrm{obs}}, << B)$. Here, the notation $F \sim (G, << H)$ means that the random variable F has a distribution with expectation of G and variability much less than H. In actuality, v_i is typically centered at a value larger than v_{obs}, since synthetic data incorporate uncertainty due to drawing values of the parameters. For large sample sizes n, this bias should be minimal. The assumption that $E(q_i \mid D, B) = q_{\mathrm{obs}}$ should be reasonable when the imputations are drawn from the correct posterior distribution of Y for those units with $Z_j = 1$.

Assuming flat priors for q_{obs} and v_{obs}, standard Bayesian theory implies that $(q_{\mathrm{obs}} \mid d^m, B) \sim N(\bar{q}_m, B/m)$ and $(v_{\mathrm{obs}} \mid d^m, B) \sim (\bar{v}_m, << B/m)$. Hence, the posterior mean and variance of $(Q \mid d^m, B)$ are

$$\begin{aligned} E(Q \mid d^m, B) &= E(E(Q \mid D, d^m, B) \mid d^m, B) \\ &= E(q_{\mathrm{obs}} \mid d^m, B) = \bar{q}_m \end{aligned} \quad (6)$$

$$\begin{aligned} \mathrm{Var}(Q \mid d^m, B) &= E(\mathrm{Var}(Q \mid D, d^m, B) \mid d^m, B) \\ &\quad + \mathrm{Var}(E(Q \mid D, d^m, B) \mid d^m, B) \\ &= \bar{v}_m + B/m. \end{aligned} \quad (7)$$

Since all the convolutions involve normal distributions, $f(Q \mid d^m, B) \sim N(\bar{q}_m, \bar{v}_m + B/m)$.

To integrate this distribution over $f(B \mid d^m)$, we use the fact that $((m-1) b_m B^{-1} \mid d^m) \sim \chi^2_{m-1}$ and, following the approximation in Rubin (1987), fit the first two moments of $\bar{v}_m + B/m$ to a mean-square random variable. The resulting approximation to the posterior distribution of Q is $(Q \mid d^m) \sim t_{v_p}(\bar{q}_m, T_p)$, where v_p is as defined in section 2.

3.2 Randomization Validity

For inferences based on (1) - (4) to have valid frequentist properties, we require two conditions. First, the analyst must use randomization valid estimators, q and v. That is, when q and v are applied on D to get q_{obs} and v_{obs}, the $(q_{\mathrm{obs}} \mid X, Y) \sim N(Q, U)$ and $(v_{\mathrm{obs}} \mid X, Y) \sim (U, << U)$, where the relevant distribution is that of I. Second, the synthetic data generation methods must be proper in a sense similar to Rubin (1987). Specifically, the data generation methods should satisfy the following conditions:

C1: Averaging over imputations of $Y_{\mathrm{rep},i}$, it is required that

(i) $(q_i \mid X, Y, I, Z) \sim N(q_{\mathrm{obs}}, B)$;
(ii) $(b_m \mid X, Y, I, Z) \sim (B, << B)$; and,
(iii) $(\bar{v}_m \mid X, Y, I, Z) \sim (v_{\mathrm{obs}}, << B/m)$, where $B = \mathrm{Var}(q_i \mid X, Y, I, Z)$.

C2: Averaging over the sampling and replacement mechanisms $(I, Z \mid X, Y)$, it is required that $(B \mid X, Y) \sim (B_0 << U)$ where $B_0 = E(b_m \mid X, Y)$.

Essentially, these conditions require the synthetic data be generated so that the q_i are unbiased for q_{obs}, the b_m is unbiased for B_0, and the $\bar{v}_m$ is unbiased for v_{obs}. Further discussion of proper imputation can be found in Rubin (1987).

Using these assumptions, it follows that

$$\begin{aligned} E(\bar{q}_m \mid X, Y) &= E(E(\bar{q}_m \mid X, Y, I, Z) \mid X, Y) \\ &= E(q_{\mathrm{obs}} \mid X, Y) = Q \end{aligned} \quad (8)$$

$$\begin{aligned} \mathrm{Var}(\bar{q}_m \mid X, Y) &= E(\mathrm{Var}(\bar{q}_m \mid X, Y, I, Z) \mid X, Y) \\ &\quad + \mathrm{Var}(E(\bar{q}_m \mid X, Y, I, Z) \mid X, Y) \\ &= E(B \mid X, Y)/m + \mathrm{Var}(q_{\mathrm{obs}} \mid X, Y) = B_0/m + U. \end{aligned} \quad (9)$$

Since $(q_{obs}|X, Y)$ and the $(q_i|X, Y, I, Z)$ are assumed to have normal distributions, it follows that $(\bar{q}_m|X, Y) \sim N(Q, B_0/m + U)$.

When C1 and C2 hold, T_p is an unbiased estimator of $B_0/m + U$. The t-approximation is justified using the method outlined in Rubin (1987). Specifically, the t-approximation follows since $((m-1)b_m B_0^{-1}|X, Y) \sim \chi^2_{m-1}$, and the degrees of freedom of a chi-squared random variable equals two times the square of its expectation over its variance.

4. SIMULATION STUDIES

This section illustrates the validity of these new combining rules, as well as the ineffectiveness of T_m and T_s as variance estimators, using simulation studies of partially synthetic strategies. Section 4.1 describes two studies in which the imputer generates synthetic data only for selected units. Section 4.2 describes a study in which the imputer generates synthetic data for all values of one survey variable, leaving the others at their observed values. For illustrations, the simulations use artificial data and correct posterior distributions for imputations. Of course, in real settings the correct imputation model typically is not known and must be estimated using the observed data and subject-matter expertise. For all simulations, the population sizes are considered infinite so that finite population correction factors are ignored.

4.1 Imputation for Selected Units

Imputers may decide to replace the observed values for some units in the collected data, then release a mixture of the imputed and observed values. This strategy is employed in two simplistic although illustrative simulations, the first involving a single variable and the second four variables.

4.1.1 Simulations Using a Single Variable

Each observed dataset, D, comprises $n = 100$ values drawn randomly from $Y \sim N(0, 10^2)$. Two different schemes are used to specify the units with $Z_j = 1$, so that two sets of partially synthetic data sets are generated for each D. The first scheme, labelled "Random", replaces Y for 20 units randomly sampled from D. The second scheme, labelled "Big Y", replaces Y only for units with $Y_j > 10$.

For each D, and for each scheme, there are $m = 5$ synthetic data sets $d_i = (Y_{rep,i}, Y_{nrep}, I, Z)$, for $i = 1, ..., 5$. The $Y_{rep,i}$ are generated by using a Bayesian bootstrap (Rubin 1987, pages 123-124), which draws values of Y from a donor pool of selected values of Y_{obs}. Let Y_{elig} be the $n_0 \times 1$ vector of values of Y_{obs} that make up the donor pool. Let $n_{rep} = \sum_{j=1}^{100} Z_j$. The Bayesian bootstrap proceeds as follows:

1. Draw $(n_0 - 1)$ uniform random numbers. Sort these numbers in ascending order. Label these ordered numbers as $a_0 = 0, a_1, a_2, ..., a_{n_0-1}, a_{n_0} = 1$.
2. Draw n_{rep} uniform random numbers, $u_1, u_2, ..., u_j, ..., u_{n_{rep}}$. For each of these u, impute $Y_{elig,j}$ when $a_{j-1} < u \leq a_j$.

This Bayesian bootstrap is not likely to be used to impute data in real settings, since data sets contain more than one variable. It is used here because it provides straightforward, proper imputations for this illustration.

As mentioned in section 2, the correct posterior predictive distribution is $f(Y|D, Z)$, not $f(Y|D)$. This implies that the donor pool, Y_{elig}, should equal the set $\{Y_j: Z_j = 1\}$. This set is labelled "SELECT." For comparisons, synthetic values also are imputed using the donor set $\{Y_j: I_j = 1\}$. This set is labelled "ALL". Imputations based on ALL donors do not meet condition C1 in section 3.2, since $E(q_i|X, Y, I, Z) = \left(\sum_{j=1}^{100-n_{rep}} y_{nrep,j} + n_{rep}\bar{y}_{obs}\right)/n \neq \bar{y}_{obs}$, whereas imputations based on SELECT donors are proper.

Table 1 summarizes the results from 5,000 runs of this simulation. For both the Random and Big Y schemes, the averages of the $\bar{q}_5$ based on the SELECT donors approximately equal the average of q_{obs}. In the Random scheme, the $\bar{q}_5$ based on ALL donors is also unbiased, because $E(\bar{y}_{nrep}|X, Y, I) = q_{obs}$ when averaged over Z (which is in fact stochastic in this scheme). However, when using ALL donors in the Big Y scheme, $\bar{q}_5$ has a large, negative bias. This results because imputed values are not restricted to be greater than 10 when using ALL donors.

In both the Random and Big Y schemes, 94.5% of the 5,000 synthetic 95% confidence intervals based on T_p and the SELECT donors cover zero. This rate is identical to the 94.5% coverage rate for the confidence intervals based on the observed data $(q_{obs} \pm 1.96\sqrt{v_{obs}})$. The nominal rates are less than 95% due to simulation error. The 2-3% difference between the averages of the T_p and the $\text{Var}(\bar{q}_5)$ roughly equals the difference between the average v_{obs} and $\text{Var}(q_{obs})$. The usual multiple imputation variance estimator, T_m, tends to overestimate the $\text{Var}(\bar{q}_5)$, leading to overly conservative confidence interval coverage rates, showing that T_m is not the correct variance estimator when analyzing properly imputed, partially synthetic data.

When imputations are based on ALL donors – an improper imputation method – in the Random scheme, T_p is negatively biased, and only 92.6% of the synthetic 95% confidence intervals cover zero. Using T_m increases the coverage rate to 95%, suggesting that it is safer to use T_m instead of T_p when ALL units are used for imputations. The confidence intervals based on ALL and T_m are on average wider than those based on SELECT and T_p. This illustrates the advantage of conditioning on Z to obtain proper imputations, even when the scheme used to set the $Z_j = 1$ does not depend on the values of Y.

Table 1
Simulation Results when Imputing Single Variable

Scheme and Imputation Method	Avg. $\bar{q}_5$	Var $\bar{q}_5$	Avg. T_p	Avg. T_m	Coverage of 95% CIs Using T_p	Coverage of 95% CIs Using T_m
$Z_j = 1$ for 20 randomly selected units						
SELECT	0.024	1.097	1.067	1.420	94.5%	96.7%
ALL	0.020	1.233	1.044	1.281	92.6%	94.9%
$Z_j = 1$ for units with $Y_j > 10$						
SELECT	0.016	1.031	1.011	1.068	94.5%	95.0%
ALL	-2.383	0.796	0.736	0.921	20.7%	28.8%
Observed data results*	0.016	1.021	1.000		94.5%	

* The column labels do not apply for this row. The average of the q_{obs} = 0.016, the variance of the q_{obs} = 1.021, the average of the v_{obs} = 1.000, and 94.5% of the five thousand 95% observed-data confidence intervals cover zero.

Although not shown in Table 1, the variance estimator for fully synthetic data, T_s, is negative in every one of the 5,000 simulations for both schemes and both imputation methods. Clearly, although valid for fully synthetic data (Raghunathan *et al.* 2003), T_s is not generally appropriate for partially synthetic data.

4.1.2 Simulations Using Four Variables

Each observed dataset, D, comprises n = 200 values of four variables, (Y_1, Y_2, Y_3, Y_4), generated as follows: $(y_1, y_2, y_3) \sim MVN(\mathbf{0}, \Sigma)$, where Σ has all variances equal to one and all covariances equal to 0.5; and, $(y_4 \mid y_1, y_2, y_3) \sim N(10y_1 + 7y_2 + 4y_3, 25^2)$. To fix ideas, the variable Y_1 can be considered a key identifier and Y_4 the sensitive variable. The plan is to simulate values of the sensitive Y_4 for all units with "unusual" values of the key identifier, defined as $Y_1 > 1$. Hence, Y_{nrep} comprises sampled values of (Y_1, Y_2, Y_3) and values of Y_4 for those units with $Y_1 \leq 1$. Typically, around 30 units per observed data set have $Y_1 > 1$.

As before, we examine two schemes for determining the posterior predictive distribution for imputations. SELECT uses only the units with $Z_j = 1$ as the data for the posteriors, and ALL uses all observed units. Imputations under each scheme are made by (i) drawing values of the parameters of the regression of Y_4 on (Y_1, Y_2, Y_3) from their posterior distribution, which is estimated using either the SELECT or ALL units, and (ii) drawing values of Y_4 for units with $Z_j = 1$ using the drawn values of parameters. There are m = 5 synthetic data sets generated for each observed data set D.

The estimands of interest include β, the regression coefficient of Y_1 in the linear regression of Y_4 on (Y_1, Y_2, Y_3); α, the regression coefficient of Y_4 in the regression of Y_1 on (Y_2, Y_3, Y_4); and $\bar{Y}_4$, the population average of Y_4. For inferences about β and α, q is the usual ordinary least squares estimator and v its variance estimator. For inferences about $\bar{Y}_4$, q is the sample average and v its standard error.

Table 2 summarizes results from 5,000 runs of this simulation. When imputations are based on the SELECT units, the averages of the $\bar{q}_5$ and T_p are within simulation errors of the averages of the q_{obs} and Var$(\bar{q}_5)$. Additionally, the coverage rates for the synthetic 95% confidence intervals are similar to the coverage rates for the observed data 95% confidence intervals. The T_m are substantially larger than the Var$(\bar{q}_5)$, resulting in coverage rates around 97%. Although not shown in Table 2, T_s is negative in all 5,000 simulation runs. Taken together, these results are consistent with the findings in section 4.1.1: when imputations are drawn from a posterior distribution that conditions on Z, point and interval estimates based on T_p are more accurate than those based on T_m and T_s.

Although imputations based on ALL units are not proper, it is informative to examine the performances of T_p and T_m for such imputations. Imputers might base imputations on all observed units for practical reasons, for example because the units with $Z_j = 1$ do not provide sufficient data to fit the imputation models. The results mirror those in section 4.1.1: the T_p underestimate the Var$(\bar{q}_5)$, leading to coverage rates around 94%, whereas using T_m increases coverage rates to around 96%, primarily due to the positive bias in T_m. This again suggests that, when imputers do in fact base imputations on all observed units even though only some $Z_j = 1$, analysts are safer using T_m as the variance estimator rather than T_p. Just as seen in section 4.1.1, the intervals based on ALL units are typically wider than those based on SELECT units, suggesting that, when possible, imputers are better off basing imputations only on the units with $Z_j = 1$.

4.2 Imputation of all Values of Y for one Variable

Each observed data set comprises n = 200 values of four variables generated as follows: $(y_1, y_2, y_3) \sim MVN(\mathbf{0}, \mathbf{I})$ where $\mathbf{I}$ is the identity matrix; and, $(y_4 \mid y_1, y_2, y_3) \sim N(10y_1 + 10y_2 + 10y_3, 25^2)$. Hence, the $Y_{nrep} = (Y_1, Y_2, Y_3)$. Values of Y_4 are imputed from the Bayesian posterior predictive distribution of $(Y_4 \mid Y_{obs})$, derived by fitting the

regression of Y_4 on (Y_1, Y_2, Y_3). All units have $Z_j = 1$ and are used as data for the posterior distributions. The estimands are the same as those described in section 4.1.2.

Table 3 summarizes the results from 5,000 simulation runs using $m = 5$ partially synthetic data sets. For all estimands, the averages of the $\bar{q}_5$ are practically identical to those of the q_{obs}. Additionally, the estimated variances based on T_p are close to the actual variances of the $\bar{q}_5$. The slight upward bias results because $\bar{v}_m$ tends to overestimate v_{obs}, as explained in section 3.1. The T_m on average overestimate the $\text{Var}(\bar{q}_5)$ by factors of more than two, and the T_s severely underestimate the $\text{Var}(\bar{q}_5)$ for α and $\bar{Y}_4$. These problems are not due to small m; in simulations with large m they persist. Although errors of these magnitudes may not occur in other settings, the results in this simple setting again indicate that T_m and T_s are not appropriate in general for analyzing partially synthetic data, especially when synthesizing entire variables.

Imputers have incentive to release small numbers of synthetic data sets. Each additional data set requires extra storage, and more importantly, releasing too many data sets might jeopardize confidentiality if intruders somehow combine the imputed values to learn about the actual values. Table 4 displays results of independent replications of 5,000 simulation runs using different values of m. Point estimates are unbiased for all three estimands and so are not displayed in the table. The 95% confidence interval coverage rates are close to 95% for all values of m greater than two. The inflations in the T_p are again due to positive biases in the $\bar{v}_m$.

Table 4 illustrates that, when imputing entire variables, substantial efficiency gains can be made by increasing m beyond five. The amount of efficiency gain depends on the magnitude of b_m. When b_m is small relative to $\bar{v}_m$, for example when imputing values only for a small number of selected units, efficiency gains from increasing m will not be large. For any partially synthetic strategy, imputers can compare gains in efficiency with potential tradeoffs in confidentiality by simulation studies of intruder behavior on different numbers of released synthetic data sets.

Table 2
Simulation Results when Imputing Y_4 for Units with $Y_1 > 1$

Type of Inference	Avg. $\bar{q}_5$	Var $\bar{q}_5$	Avg. T_p	Avg. T_m	Coverage of 95% CIs Using T_p	Coverage of 95% CIs Using T_m
Estimand is β						
SELECT	10.02	5.45	5.68	8.97	95.3%	98.2%
ALL	10.04	5.89	5.28	7.57	93.7%	96.9%
Observed data*	10.00	4.70			95.5%	
Estimand is α						
SELECT	9.25×10^{-3}	4.49×10^{-6}	4.76×10^{-6}	6.97×10^{-6}	95.4%	97.9%
ALL	9.59×10^{-3}	5.03×10^{-6}	4.75×10^{-6}	6.31×10^{-6}	94.1%	96.5%
Observed data*	9.66×10^{-3}	4.26×10^{-6}			95.4%	
Estimand is $\bar{Y}_4$						
SELECT	-1.45×10^{-2}	4.97	5.01	6.09	95.0%	96.6%
ALL	-1.24×10^{-3}	5.19	4.82	5.59	93.8%	95.4%
Observed data*	-2.34×10^{-3}	4.76			94.5%	

* The column labels do not apply for this row. These are the averages of the q_{obs}, the variance of the q_{obs}, and the percentage of 95% observed-data confidence intervals that cover their Q.

Table 3
Simulation Results when Imputing an Entire Variable

Estimand	Avg. q_{obs}	Avg. $\bar{q}_5$	Var q_{obs}	Var $\bar{q}_5$	Avg. T_p	Avg. T_m	Avg. T_s
β	9.95	9.94	3.19	4.46	4.54	11.10	4.63
α	0.0137	0.0135	6.12	7.69	7.94	17.30	5.17
$\bar{Y}_4$	0.00	0.00	4.55	5.83	6.00	12.30	2.87

Table 4
Sensitivity of Partially Synthetic Inferences to Value of m

Setting	Var $\bar{q}_m$	Avg. T_p	95% CI cov.
Inference for β			
$m = 2$	6.52	6.50	92.7
$m = 3$	5.38	5.38	94.4
$m = 4$	4.64	4.89	95.4
$m = 5$	4.46	4.54	95.1
$m = 10$	3.87	3.88	94.4
$m = 50$	3.30	3.37	95.1
Inference for α			
$m = 2$	10.62	10.89	93.4
$m = 3$	8.92	9.15	94.9
$m = 4$	8.41	8.45	94.9
$m = 5$	7.69	7.94	95.4
$m = 10$	6.99	7.02	94.8
$m = 50$	6.05	6.28	95.5
Inference for $\bar{Y}_4$			
$m = 2$	8.13	7.96	93.4
$m = 3$	6.51	6.86	95.5
$m = 4$	6.11	6.33	95.6
$m = 5$	5.83	6.00	95.3
$m = 10$	5.13	5.38	95.4
$m = 50$	4.66	4.87	95.5

Variances associated with α are multiplied by 10^6.

5. CONCLUDING REMARKS

The simulations in this article illustrate that the usual rules for combining multiply-imputed data sets can result in positively biased variance estimates when applied on partially synthetic data. The new rules presented here appear to remedy this problem, thereby leading to more reliable inferences. Further research is needed to assess the performance of these new rules when using partially synthetic strategies for genuine data, for which the correct imputation models are unlikely to be known. Additionally, evaluations of the new rules are needed when the released data sets also contain multiple imputations of missing data, for example imputations for item nonresponse. As conjectured by a referee of this article, when significant fractions of imputations are for missing data, T_m may not perform so unfavorably relative to T_p.

The simulations and theory also suggest that, when possible, imputers should use only units with values selected for replacement as the data when estimating posterior predictive distributions for imputations. Further examination of this prescription when simulating more than one variable in genuine data sets would be valuable.

Lastly, this article does not examine the implications of various partially synthetic data strategies for protecting confidentiality, nor does it compare partially synthetic approaches to alternative techniques for disclosure control. Such comparisons would help imputers determine whether partially synthetic approaches are appropriate for their public use microdata releases.

ACKNOWLEDGEMENTS

This work was supported by the United States Bureau of the Census through a contract with Datametrics Research. The author thanks Trivellore Raghunathan, Donald Rubin, and Laura Zayatz for providing statistical support and general motivation for this research, and two referees and an associate editor for their valuable comments and suggestions.

REFERENCES

ABOWD, J.M., and WOODCOCK, S.D. (2001). Disclosure limitation in longitudinal linked data. In P. Doyle, J. Lane, L. Zayatz and J. Theeuves (Eds.). *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*. Amsterdam: North-Holland. 215-277.

DANDEKAR, R.A., COHEN, M. and KIRKENDALL, N. (2002a). Sensitive micro data protection using Latin hypercube sampling technique. In J. Domingo-Ferrer (Ed.). *Inference Control in Statistical Databases*. Berlin: Springer-Verlag. 117-125.

DANDEKAR, R.A., DOMINGO-FERRER, J. and SEBE, F. (2002b). LHS-based hybrid microdata versus rank swapping and microaggregation for numeric microdata protection. In J. Domingo-Ferrer (Ed.). *Inference Control in Statistical Databases*. Berlin: Springer-Verlag. 153-162.

FIENBERG, S.E., MAKOV, U.E. and STEELE, R.J. (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics*. 14, 485-502.

FIENBERG, S.E., STEELE, R.J. and MAKOV, U.E. (1996). Statistical notions of data disclosure avoidance and their relationship to traditional statistical methodology: Data swapping and log-linear models. In *Proceedings of Bureau of Census 1996 Annual Research Conference*. 87-105.

FRANCONI, L., and STANDER, J. (2002). A model based method for disclosure limitation of business microdata. *The Statistician*. 51, 1-11.

FRANCONI, L., and STANDER, J. (2003). Spatial and non-spatial model-based protection procedures for the release of business microdata. *Statistics and Computing*. Forthcoming.

FULLER, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*. 9, 383-406.

KENNICKELL, A.B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. In W. Alvey and B. Jamerson (Eds.). *Record Linkage Techniques, 1997*. Washington, D.C.: National Academy Press. 248-267.

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

LIU, F., and LITTLE, R.J.A. (2002). Selective multiple imputation of keys for statistical disclosure control in microdata. In *Proceedings of the Survey Research Methods Section*, American Statistical Association. 2133-2138.

POLETTINI, S. (2003). Maximum entropy simulation for microdata protection. *Statistics and Computing*. Forthcoming.

POLETTINI, S., FRANCONI, L. and STANDER, J. (2002). Model-based disclosure protection. In J. Domingo-Ferrer (Ed). *Inference Control in Statistical Databases*. Berlin: Springer-Verlag. 83-96.

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

REITER, J.P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*. 18, 531-544.

REITER, J.P. (2003). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. Tech. Rep., Institute of Statistics and Decision Sciences, Duke University.

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

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

WILLENBORG, L., and DE WAAL, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.

Survey Methodology, December 2003
Vol. 29, No. 2, pp. 189-196
Statistics Canada

The High Entropy Variance of the Horvitz-Thompson Estimator

K.R.W. BREWER and MARTIN E. DONADIO[1]

ABSTRACT

Using both purely design-based and model-assisted arguments, it is shown that, under conditions of high entropy, the variance of the Horvitz-Thompson (HT) estimator depends almost entirely on first order inclusion probabilities. Approximate expressions and estimators are derived for this "high entropy" variance of the HT estimator. Monte Carlo simulation studies are conducted to examine the statistical properties of the proposed variance estimators.

KEY WORDS: Horvitz-Thompson estimator; Model assisted survey sampling; Monte Carlo simulation; Variance estimation.

1. INTRODUCTION

Let U denote a finite population of N units labelled $i = 1, ..., N$, and let Y_i denote the value for the i-th unit of a certain characteristic y. Consider the problem of estimating the population total $Y_\bullet = \sum_{i=1}^{N} Y_i$. If a sample, s, of n units is drawn without replacement from U with first order inclusion probabilities $\pi_i, i \in U$, the Horvitz-Thompson (HT) (1952) estimator of the total is $\hat{Y}_{\bullet HT} = \sum_{i \in s} Y_i \pi_i^{-1}$. In this paper, we confine consideration to fixed size sampling designs. For this important special case, Sen (1953) and Yates and Grundy (1953) showed independently that $\hat{Y}_{\bullet HT}$ has the variance

$$V(\hat{Y}_{\bullet HT}) = (1/2)\sum_{i \in U} \sum_{j(\neq i) \in U} (\pi_i \pi_j - \pi_{ij})(Y_i \pi_i^{-1} - Y_j \pi_j^{-1})^2, \quad (1)$$

where π_{ij} is the second order or joint inclusion probability of the i-th and j-th population units together in the same sample. They therefore suggested the variance estimator

$$\hat{V}_{SYG}(\hat{Y}_{\bullet HT}) = (1/2)\sum_{i \in s} \sum_{j(\neq i) \in s} \pi_{ij}^{-1}(\pi_i \pi_j - \pi_{ij})(Y_i \pi_i^{-1} - Y_j \pi_j^{-1})^2. \quad (2)$$

This is known to perform better than the variance estimator proposed by Horvitz and Thompson (1952) (the latter, however, usually being unbiased for random n), but the critical dependence of (2) on π_{ij} has proved problematical (Brewer 1999). If one or more of the $N(N-1)/2$ distinct values of π_{ij} are zero, the estimator (2) is biased downwards. And if any of them should be very small compared with their corresponding values of $\pi_i \pi_j$, (2) will be unstable (that is, it will itself be subject to high variance). In addition, the double sum feature of (2) is quite inconvenient, especially for large sample sizes. Not only are there many more π_{ij}'s than there are π_i's; it is also frequently the case that the individual π_{ij}'s are problematic to evaluate. In view of these difficulties, the aim of this paper is to provide alternative variance estimators, which do not depend on the π_{ij}'s and are simple to compute.

In the next section, a new expression for the design-variance of the HT estimator is presented. This new expression leads, under high entropy conditions, to the derivation of an approximate formula for $V(\hat{Y}_{\bullet HT})$, which is π_{ij}–free. In section 3, we check the usefulness of our approximate formulae using a model assisted approach. An estimator of our approximate variance is proposed in section 4; this variance estimator is expected to perform well under conditions of high entropy (meaning the absence of any detectable pattern or ordering in the selected sample units). Most sample selection schemes though, result in the selection of high entropy samples. With the aim of testing the usefulness of the variance estimator presented in section 4, some empirical studies were conducted. The main findings from these studies are reported in section 5. Some concluding remarks are provided in section 6.

2. SOME APPROXIMATE FORMULAE FOR THE DESIGN-VARIANCE OF THE HT ESTIMATOR

We begin this section by presenting an alternative formulation for the variance of the HT estimator, valid only when the sampling design is of fixed size. Before proceeding, we state the following relations, which will be useful later:

$$\sum_{j(\neq i) \in U} \pi_{ij} = (n-1)\pi_i, \quad i \in U \quad (3)$$

$$\sum_{j(\neq i) \in U} \pi_i \pi_j = (n-\pi_i)\pi_i, \quad i \in U \quad (4)$$

$$\sum_{i \in U} \sum_{j(\neq i) \in U} \pi_{ij} = n(n-1) \quad (5)$$

[1] Ken Brewer, School of Finance and Applied Statistics, Australian National University, ACT 0200, Australia. E-mail: Ken.Brewer@anu.edu.au and Martin E. Donadio, Department of Mathematics and Statistics, The University of Melbourne, VIC 3010, Australia. E-mail: M.Donadio@abs.gov.au.

$$\sum_{i\in U}\sum_{j(\neq i)\in U}\pi_i\pi_j = n^2 - \sum_{i\in U}\pi_i^2. \qquad (6)$$

The alternative formulation is obtained as follows. We start with a trivial modification of (1),

$$\begin{aligned}
&V(\hat{Y}_{\bullet HT})\\
&= (1/2)\sum_{i\in U}\sum_{j(\neq i)\in U}(\pi_i\pi_j - \pi_{ij})\left\{(Y_i\pi_i^{-1} - Y_\bullet n^{-1})\right.\\
&\quad \left.-(Y_j\pi_j^{-1} - Y_\bullet n^{-1})\right\}^2\\
&= (1/2)\sum_{i\in U}\sum_{j(\neq i)\in U}(\pi_i\pi_j - \pi_{ij})\left\{(Y_i\pi_i^{-1} - Y_\bullet n^{-1})^2\right.\\
&\quad \left.+(Y_j\pi_j^{-1} - Y_\bullet n^{-1})^2 - 2(Y_i\pi_i^{-1} - Y_\bullet n^{-1})(Y_j\pi_j^{-1} - Y_\bullet n^{-1})\right\}.
\end{aligned}$$

Using the relations (3) and (4), the above equation may be shown to be identical to

$$\begin{aligned}
&V(\hat{Y}_{\bullet HT})\\
&= \sum_{i\in U}\pi_i(Y_i\pi_i^{-1} - Y_\bullet n^{-1})^2 - \sum_{i\in U}\pi_i^2(Y_i\pi_i^{-1} - Y_\bullet n^{-1})^2\\
&\quad - \sum_{i\in U}\sum_{j(\neq i)\in U}(\pi_i\pi_j - \pi_{ij})(Y_i\pi_i^{-1} - Y_\bullet n^{-1})(Y_j\pi_j^{-1} - Y_\bullet^{-1}). \qquad (7)
\end{aligned}$$

The first term in (7) is virtually the same as the variance of the corresponding Hansen-Hurwitz (1943) estimator of total for sampling at n draws with replacement, the probability of selecting unit i at each draw being $p_i = \pi_i/n, i\in U$. The second term can be viewed as a finite population correction. Consequently, these two terms together plausibly constitute a first approximation to the entire variance of the HT estimator and, importantly, neither of them depends on the π_{ij}'s.

The magnitude of the third term depends mostly on the sampling design $p(s)$. Thus, if $p(s)$ is such that $\pi_{ij} \approx \pi_i\pi_j$, for all $i \neq j \in U$, then we can expect a very small third term in (7) (compared with the other two). This condition seems to be satisfied by high entropy sampling designs. For example, in simple random sampling without replacement (*srswor*), which maximizes the entropy among all fixed sized designs (see Hájek 1981), the second order inclusion probabilities can be written as $\pi_{ij} = \pi_i\pi_j [N(n-1)/\{n(N-1)\}]$. The factor $N(n-1)/\{n(N-1)\}$ is less than 1, and tends to 1 for large population and sample sizes. For this design, the third term in (7) accounts for only $1/N$ of the entire variance of the HT estimator. Furthermore, for several probability proportional-to-size designs, such as rejective sampling (Hájek 1964) and randomized systematic πps sampling (Hartley and Rao 1962), the condition $\pi_{ij} \approx \pi_i\pi_j$ also holds, provided N and n are large enough.

There are some exceptions, however, in which the third term in (7) can be important. The most important of these exceptions is systematic sampling from a population in which the units are arranged in a meaningful order prior to the selection. In such a case, a number of second order inclusion probabilities can even be equal to zero. This and other special cases need to be dealt with separately, and are not discussed further in this paper.

The rest of this section is devoted to deriving an approximation to $V(\hat{Y}_{\bullet HT})$ that uses first order inclusion probabilities only. We start by proposing a simple approximation to the π_{ij} of the form

$$\pi_{ij} \cong \tilde{\pi}_{ij} = \pi_i\pi_j(c_i + c_j)/2, \quad i \neq j \in U. \qquad (8)$$

Three possible choices for $c_i, i\in U$, are then:

$$c_i = (n-1)/(n-\pi_i), \qquad (9)$$

$$c_i = c = (n-1)\Big/\left(n - n^{-1}\sum_{k\in U}\pi_k^2\right) \text{ and} \qquad (10)$$

$$c_i = (n-1)\Big/\left(n - 2\pi_i + n^{-1}\sum_{k\in U}\pi_k^2\right). \qquad (11)$$

The first two choices of c_i are prompted by ratios of sums of π_{ij} to the corresponding sums of $\pi_i\pi_j$. Thus, on the one hand, formula (9) is obtained by comparing (3) with (4). On the other hand, formula (10) is suggested by the comparison of (5) and (6). Finally, formula (11) is based on the asymptotic expressions for π_{ij} obtained by Hartley and Rao (1962) and by Asok and Sukhatme (1976) for randomized systematic πps sampling and for Sampford's (1967) procedure respectively. To order $O(n^3N^{-3})$, both these asymptotic expressions simplify to

$$\tilde{\pi}_{ij} = \pi_i\pi_j\{(n-1)/n\}\left\{1 + n^{-1}(\pi_i + \pi_j) - n^{-2}\sum_{k\in U}\pi_k^2\right\},$$

which in turn implies $c_i = \{(n-1)/n\}(1 - 2n^{-1}\pi_i - n^{-2}\sum_{k\in U}\pi_k^2)$. Under *srswor*, however, this choice of c_i does not yield the exact formula for the π_{ij}'s. For this reason, the slightly different expression given by (11) is used here, $(1 - 2n^{-1}\pi_i + n^{-2}\sum_{k\in U}\pi_k^2)$ being the first two terms in the Taylor expansion of the reciprocal of $(1 + 2n^{-1}\pi_i - n^{-2}\sum_{k\in U}\pi_k^2)$ and *vice versa*.

The next step consists of replacing the π_{ij}'s in the third term of (7) by the approximation (8). This replacement yields

$$\begin{aligned}
&-\sum_{i\in U}\sum_{j(\neq i)\in U}(\pi_i\pi_j - \pi_{ij})(Y_i\pi_i^{-1} - Y_\bullet n^{-1})(Y_j\pi_j^{-1} - Y_\bullet n^{-1})\\
&\cong -\sum_{i\in U}\sum_{j(\neq i)\in U}\pi_i\pi_j[1 - (c_i + c_j)/2]\\
&\qquad (Y_i\pi_i^{-1} - Y_\bullet n^{-1})(Y_j\pi_j^{-1} - Y_\bullet n^{-1})\\
&= \sum_{i\in U}(1 - c_i)\pi_i^2(Y_i\pi_i^{-1} - Y_\bullet n^{-1})^2,
\end{aligned}$$

and thus the variance of the HT estimator may be approximated by

$$\tilde{V}(\hat{Y}_{\bullet HT}) = \sum_{i\in U} [\pi_i - \pi_i^2 + (1-c_i)\pi_i^2](Y_i\pi_i^{-1} - Y_\bullet n^{-1})^2$$

$$= \sum_{i\in U} \pi_i(1-c_i\pi_i)(Y_i\pi_i^{-1} - Y_\bullet n^{-1})^2. \quad (12)$$

This approximate variance has a very simple form. It is also without error under *srswor* for all the three choices of c_i presented above.

3. A MODEL ASSISTED CHECK ON THE USEFULNESS OF THE APPROXIMATE VARIANCE FORMULAE

Consider the following ratio model as a possible description of the population being sampled:

$$\xi:\ Y_i = \beta\pi_i + \varepsilon_i;\ E_\xi\varepsilon_i = 0;\ E_\xi\varepsilon_i^2 = \sigma_i^2;$$

$$E_\xi(\varepsilon_i\varepsilon_j) = 0,\ i \neq j;\ i, j \in U. \quad (13)$$

This is a shorthand model. It is intended to reflect the situation where the expected values of the Y_i are *intrinsically* proportional to the values X_i of an auxiliary variable x, and the inclusion probabilities π_i are *chosen* to be proportional to the X_i. It is of course impossible for the Y_i to be directly dependent on the inclusion probabilities as such, since those probabilities may be set quite arbitrarily by the person designing the sample.

The prediction or model expectation under ξ of the approximate variance expression (12) is

$$E_\xi\tilde{V}(\hat{Y}_{\bullet HT}) = E_\xi \sum_{i\in U} \pi_i(1-c_i\pi_i)(Y_i\pi_i^{-1} - Y_\bullet n^{-1})^2$$

$$= E_\xi \sum_{i\in U} \pi_i(1-c_i\pi_i)(\varepsilon_i\pi_i^{-1} - \varepsilon_\bullet n^{-1})^2$$

$$= \sum_{i\in U} \sigma_i^2\left\{\pi_i^{-1} - n^{-1} - c_i(1-2n^{-1}\pi_i) - n^{-2}\sum_{k\in U} c_k\pi_k^2\right\}, \quad (14)$$

where $\varepsilon_\bullet = \sum_{i\in U}\varepsilon_i$. Ideally, expression (14) should be equal to $E_\xi V(\hat{Y}_{\bullet HT})$, namely $\sum_{i\in U}\sigma_i^2(\pi_i^{-1} - 1)$ (Godambe 1955; Godambe and Joshi 1965). This condition leads to the implicit formula

$$c_i = \left(1 - n^{-1} - n^{-2}\sum_{k\in U} c_k\pi_k^2\right)\Big/(1-2n^{-1}\pi_i),$$

which can be solved for c_i iteratively, starting with the trial value $c_i^{[1]} = (n-1)/n$. To $O(N^{-1})$, this iterative solution is identical to (11). Alternatively, a closed expression can be derived by putting (14) equal to $\sum_{i\in U}\sigma_i^2(\pi_i^{-1} - 1)$ and then requiring that $c_i = c$ for all $i \in U$, in which case we obtain

$$c = (1-n^{-1})\sum_{i\in U}\sigma_i^2\Big/\sum_{i\in U}\sigma_i^2\left(1 - 2n^{-1}\pi_i - n^{-2}\sum_{k\in U}\pi_k^2\right). \quad (15)$$

Under *srswor*, (15) becomes $c = N(n-1)/\{n(N-1)\}$, which yields the exact expression for $V(\hat{Y}_{\bullet HT})$. Even without *srswor*, replacing σ_i^2 by $\sigma^2\pi_i$ in (15) returns (10) for c. It is reassuring that the purely design-based analysis and the model-assisted one produce results in such close agreement.

4. ESTIMATING THE DESIGN-VARIANCE OF THE HT ESTIMATOR

The aim of this section is to propose a plausible sample estimator for the approximate design-variance of the HT estimator given in (12). One such estimator is

$$\hat{\tilde{V}}(\hat{Y}_{\bullet HT}) = \sum_{i\in s}(c_i^{-1} - \pi_i)(Y_i\pi_i^{-1} - \hat{Y}_{\bullet HT}n^{-1})^2, \quad (16)$$

which is arrived at by replacing each population sum in (12) by the corresponding HT estimator, and adjusting by the factor c_i^{-1}. This estimator has some attractive properties: (i) For all three choices of c_i, it reduces to the standard variance estimator in the case of *srswor*; (ii) it is simple to compute, since no double sums are involved; and (iii) using Taylor linearization technique, it can be shown that (16) is approximately design-unbiased for (12).

A further attractive property of the estimator (16) is the following. When c_i is specified by (9), we have

$$c_i^{-1} - \pi_i = (n-\pi_i)/(n-1) - \pi_i = \{n/(n-1)\}(1-\pi_i). \quad (17)$$

The factor $(1-\pi_i)$ is easily interpretable as a finite population correction, while the factor $n/(n-1)$ has an entirely different role, which can be explained as follows. It is easy to see that $\hat{\beta} = \hat{Y}_{\bullet HT}n^{-1}$ is a model unbiased estimator of β in model (13). Let us write $\hat{\sigma}_i^2 = (Y_i - \hat{\beta}\pi_i)^2$, for all i. Then $(Y_i\pi_i^{-1} - \hat{Y}_{\bullet HT}n^{-1})^2 = (Y_i - \hat{\beta}\pi_i)^2\pi_i^{-2} = \hat{\sigma}_i^2\pi_i^{-2}, i \in U$. It is not difficult to show that the factor $n/(n-1)$ removes the (model) bias from $\sum_{i\in s}(Y_i\pi_i^{-1} - \hat{Y}_{\bullet HT}n^{-1})^2 = \sum_{i\in s}\hat{\sigma}_i^2\pi_i^{-2}$ as an estimator of $\sum_{i\in s}\sigma_i^2\pi_i^{-2}$.

The choice of (9) to specify the value of c_i also renders particularly simple the calculation both of the HT estimate itself and of its estimated variance; for substituting (17) into (16) and expanding that expression into individual terms we obtain:

$$\hat{\tilde{V}}(\hat{Y}_{\bullet HT}) = \{n/(n-1)\}\left\{\sum_{i\in s} Y_i^2\pi_i^{-2} - n^{-1}\hat{Y}_{\bullet HT}^2 - \sum_{i\in s} Y_i^2\pi_i^{-1} + 2n^{-1}\hat{Y}_{\bullet HT}\sum_{i\in s} Y_i - n^{-2}\hat{Y}_{\bullet HT}^2\sum_{i\in s}\pi_i\right\}.$$

This formula involves six expressions, namely n, $\hat{Y}_{\bullet HT}$, $\sum_{i\in s} Y_i^2 \pi_i^{-2}$, $\sum_{i\in s} Y_i^2 \pi_i^{-1}$, $\sum_{i\in s} Y_i$, and $\sum_{i\in s} \pi_i$, which are the sample sums of 1 (unity), $Y_i \pi_i^{-1}$, $Y_i^2 \pi_i^{-2}$, $Y_i^2 \pi_i^{-1}$, Y_i, and π_i respectively. If these individual terms are cumulated over every sample unit, then $\hat{Y}_{\bullet HT}$ and $\hat{\hat{V}}(\hat{Y}_{\bullet HT})$ can be evaluated together, using only a single pass of the sample data.

Note that, if non-response is present, a first order correction for it may be obtained by conditioning the sample on the achieved sample size, which we may denote here by n'. That would involve replacing the original first order inclusion probabilities, π_i, by the "adjusted inclusion probabilities", $\pi_i' = \pi_i n'/n$. (This terminology has been taken from Furnival, Gregoire and Grosenbaugh (1987), where the same type of adjustment was used in a different context). The summations over the achieved sample, s', would then be n', $\sum_{i\in s'} Y_i \pi_i'^{-1}$, $\sum_{i\in s'} Y_i^2 \pi_i'^{-2}$, $\sum_{i\in s'} Y_i^2 \pi_i'^{-1}$, $\sum_{i\in s'} Y_i$, and $\sum_{i\in s'} \pi_i'$ respectively.

Beyond the properties listed above, a further study of (16) is possible with the aid of the model ξ of (13). The most desirable expression for the ξ-expectation of an estimator of $V(\hat{Y}_{\bullet HT})$ is $\sum_{i\in s} \sigma_i^2 \pi_i^{-1}(\pi_i^{-1} - 1)$, because this in turn has design-expectation $\sum_{i\in U} \sigma_i^2 (\pi_i^{-1} - 1)$, which is the lower bound for the anticipated variance of any unbiased estimator (Godambe 1955; Godambe and Joshi 1965). For all the three definitions of c_i, the ξ-expectation of (16) differs from $\sum_{i\in s} \sigma_i^2 \pi_i^{-1}(\pi_i^{-1} - 1)$ by terms of order $O(Nn^{-1})$. Although these "unwanted" terms have opposite signs and therefore tend to cancel, they are not entirely negligible, being only $O(N^{-1})$ smaller than the variance itself.

In view of this, a new version of c_i, which retained the (design) properties (i)-(iii) for (16) and provided a closer expression to $\sum_{i\in s} \sigma_i^2 \pi_i^{-1}(\pi_i^{-1} - 1)$ for the ξ-expectation of (16), was desirable. These requirements are satisfied by a c_i defined as follows:

$$c_i = (n-1)/\left\{n - (2n-1)(n-1)^{-1}\pi_i + (n-1)^{-1}\sum_{k\in U}\pi_k^2\right\}, \quad (18)$$

for all $i \in U$. With this definition of c_i, the ξ-expectation of (16) still contains some "unwanted" terms, but they now consist only of a single term of order $O(Nn^{-2})$ – which is therefore smaller than $V(\hat{Y}_{\bullet HT})$ by a factor of order $O(N^{-1}n^{-1})$ – and other terms of smaller magnitude still.

5. SOME EMPIRICAL RESULTS

With the aim of evaluating the performance of the variance estimator proposed in section 4, some empirical studies were conducted. Three other variance estimators were also included in these studies: (i) the SYG estimator, given in (2); (ii) the variance estimator suggested by Hájek (1964, page 1520),

$$\hat{V}_{HAJ}(\hat{Y}_{\bullet HT}) = \{n/(n-1)\}\sum_{i\in s}(1-\pi_i)(Y_i\pi_i^{-1} - A_s)^2, \quad (19)$$

where $A_s = \sum_{i\in s} a_i Y_i \pi_i^{-1}$, $a_i = (1-\pi_i)/\sum_{k\in s}(1-\pi_k)$; and (iii) a slight modification of (19) proposed by Deville (1999),

$$\hat{V}_{DEV}(\hat{Y}_{\bullet HT}) = \frac{1}{1-\sum_{i\in s} a_i^2}\sum_{i\in s}(1-\pi_i)(Y_i\pi_i^{-1} - A_s)^2. \quad (20)$$

It is worth mentioning that the estimator (19) was originally intended only for a particular high entropy design, namely rejective sampling, and not for all the high entropy ones. Later on, however, this estimator was proposed for its use with some other high entropy designs. For example, Rosén (1997) suggested the use of (19) in the context of Pareto sampling.

The inclusion of the estimators (2), (19) and (20) in our empirical studies deserves a brief explanation. The SYG variance estimator would usually be the preferred choice if the π_{ij} were known and were neither zero nor very small compared with the corresponding $\pi_i\pi_j$. Under these conditions, it would then be natural to ask: Is there a significant difference, in terms of performance, between (2) and the simpler estimator (16)? On the other hand, a comparison with (19) and (20) is of interest because these two estimators share with (16) the simplicity and π_{ij}-free features. Thus, they are "competitors" in the same class.

The performance of a variance estimator can be assessed in different ways; here we will focus on *bias* and *stability*. The main findings from our studies are reported in the remainder of this section. We will consider two cases separately, namely $n = 2$ and $n > 2$.

5.1 Case $n = 2$

With the aim of testing the variance estimators under different situations, nine small populations were used in this study, most of which were also included in the stability studies carried out by Rao and Bayless (1969). Table 1 summarizes the main features of each population, including the coefficients of variation (CV) of y and x, and the correlation coefficient, ρ, between y and x. Here, y is the variable for which total estimates are sought, and x is an auxiliary variable that may be used for sample selection. Note that N varies from 10 to 20, CV(x) from 0.14 to 0.73, and ρ from 0.49 to 0.94. This provides a good mixture of populations with different characteristics.

The inclusion probabilities are chosen to be proportional to x, *i.e.*, $\pi_i = 2X_i/X_\bullet$, for all i. Two sampling designs are considered here, namely Brewer's (1963) procedure (BREWER) and Tillé's (1996) elimination procedure (TILLÉ). For both procedures, the π_{ij} are simple to compute and, for these nine populations, they are strictly positive (this condition is not always satisfied by TILLÉ). Moreover, since $n = 2$, for any sample $s = \{i, j\}$ we have $p(s) = \pi_{ij}$. Hence we can obtain the exact statistical properties of any given variance estimator $\hat{V}$.

To this end, let S denote the set of all possible samples of size $n = 2$ from a population U. The expectation of $\hat{V}$ is then defined as

$$E(\hat{V}) = \sum_{s \in S} p(s)\,\hat{V}(s),$$

and its standard error (SE) as

$$SE(\hat{V}) = \left\{\sum_{s \in S} p(s)\,[\hat{V}(s) - E(\hat{V})]^2\right\}^{1/2}.$$

For each of the two sampling designs mentioned above, Table 2 displays the *relative bias* $RB(\hat{V}) = E(\hat{V})/V(\hat{Y}_{\bullet HT}) - 1$, expressed as a percentage, of the six π_{ij}-free variance estimators. The first two of these estimators need no explanation; the other four correspond to (16) coupled with (9), (10), (11), and (18) respectively. Since for $n = 2$ (only), $\hat{V}_{DEV}$ and $\hat{V}_{16.9}$ are identical, they both appear in the same row. In order to simplify the reading of the table, the smallest RB (in absolute terms) in each population and sampling design has been highlighted.

The results in Table 2 prompt the following comments: (i) the performance of the π_{ij}-free variance estimators is reasonably good for all populations, with the possible exception of Population 4. An examination of the relationship between x and y for this population reveals the presence of some curvature, with larger cities growing at a higher rate. There is also an outlier – city 26 – for which the number of people almost tripled in the 10-year period between 1920 and 1930. Another interesting case is given by Populations 5 and 6. These two populations have identical definitions, thus one would expect to obtain similar results for them. However, the RB figures for Population 5 are considerably worse than those for Population 6, specially for BREWER. The only noticeable difference between these two populations is that Population 5 contains an outlier (Farm 14 in the reference provided). It would appear then that the presence of outliers may result in some additional bias in these variance estimators. (ii) The estimator $\hat{V}_{16.18}$ seems to be the best of the class, performing remarkably well in all situations, and showing the smallest bias figures (in absolute values) in most cases; (iii) The estimator $\hat{V}_{16.10}$ tends to exhibit the largest bias figures.

Regarding stability, Table 3 reports the *coefficient of variation* $CV(\hat{V}) = SE(\hat{V})/E(\hat{V})$, expressed as a percentage, of all the seven variance estimators. It can be seen that the π_{ij}-free variance estimators tend to be more efficient (lower CVs) than $\hat{V}_{SYG}$, although the gains are small. Otherwise, there is little to choose from among these variance estimators, even though $\hat{V}_{16.10}$ is the best performer in all but the last population.

Table 1
Description of the Nine Small Populations

Pop.	*Source*	*y*	*x*	*N*	CV(*y*)	CV(*x*)	ρ
1	Cochran (1963, page 325)	No. of persons per block	No. of rooms per block	10	0.15	0.14	0.65
2	Yates (1981, page 150) Kraals 26-38	No. of persons absent	Total no. of persons	13	0.67	0.47	0.72
3	Rao (1963, page 207)	Corn acreage in 1960	Corn acreage in 1958	14	0.39	0.43	0.93
4	Cochran (1963, page 156) Cities 19-33	No. of people in 1930	No. of people in 1920	15	0.67	0.69	0.94
5	Sampford (1962, page 61) Even units	Oat acreage in 1957	Total acreage in 1947	17	0.61	0.71	0.80
6	Sampford (1962, page 61) Odd units	Oat acreage in 1957	Total acreage in 1947	18	0.75	0.73	0.91
7	Yates (1981, page 153)	Vol. of timber	Eye-estimated vol. of timber	20	0.51	0.48	0.49
8	Sukhatme (1954, page 279) Circles 1-20	Wheat acreage	No. of villages	20	0.63	0.50	0.59
9	Horvitz and Thompson (1952, page 682)	No. of households	Eye-estimated no. of households	20	0.44	0.40	0.87

Table 2
RB (%) of Variance Estimators for $n = 2$

		Pop1	Pop2	Pop3	Pop4	Pop5	Pop6	Pop7	Pop8	Pop9
B	$\hat{V}_{HAJ}$	-1.04	-2.97	-2.60	-6.05	-3.64	0.08	-0.81	-1.48	1.13
R	$\hat{V}_{DEV}, \hat{V}_{16.9}$	-0.98	-2.52	-2.29	-5.21	-3.00	0.54	-0.63	-1.33	1.24
E	$\hat{V}_{16.10}$	-1.37	-3.55	-3.21	-7.16	-4.31	0.82	-0.94	-1.89	1.80
W	$\hat{V}_{16.11}$	-0.59	-1.49	-1.37	-3.26	-1.69	0.26	-0.31	-0.76	0.68
.	$\hat{V}_{16.18}$	**-0.20**	**-0.46**	**-0.46**	**-1.31**	**-0.38**	**-0.01**	**0.00**	**-0.19**	**0.13**
T	$\hat{V}_{HAJ}$	-1.06	-4.40	-1.07	-5.90	-1.86	-0.41	0.32	-1.10	0.82
I	$\hat{V}_{DEV}, \hat{V}_{16.9}$	-1.00	-3.94	-0.75	-5.03	-1.19	**0.07**	0.51	-0.95	0.93
L	$\hat{V}_{16.10}$	-1.39	-4.91	-1.68	-6.91	-2.47	0.33	**0.19**	-1.50	1.48
L	$\hat{V}_{16.11}$	-0.62	-2.98	**0.17**	-3.14	**0.09**	-0.20	0.83	-0.39	0.38
É	$\hat{V}_{16.18}$	**-0.23**	**-2.02**	1.10	**-1.25**	1.37	-0.46	1.15	**0.17**	**-0.17**

Table 3
CV (%) of Variance Estimators for $n = 2$

		Pop1	Pop2	Pop3	Pop4	Pop5	Pop6	Pop7	Pop8	Pop9
B	$\hat{V}_{SYG}$	123	126	118	245	138	127	158	127	**133**
R	$\hat{V}_{HAJ}$	121	119	115	238	131	125	155	124	134
E	$\hat{V}_{DEV}, \hat{V}_{16.9}$	121	119	115	238	131	125	155	124	134
W	$\hat{V}_{16.10}$	**120**	**117**	**114**	**236**	**128**	**124**	**153**	**123**	135
E	$\hat{V}_{16.11}$	122	122	116	241	133	126	157	125	133
R	$\hat{V}_{16.18}$	122	125	117	243	136	127	158	126	133
T	$\hat{V}_{SYG}$	123	143	118	248	147	131	164	131	134
I	$\hat{V}_{HAJ}$	121	118	115	238	128	125	155	124	134
L	$\hat{V}_{DEV}, \hat{V}_{16.9}$	121	118	115	238	128	125	155	124	134
L	$\hat{V}_{16.10}$	**121**	**116**	**114**	**235**	**125**	**124**	**154**	**123**	135
É	$\hat{V}_{16.11}$	122	121	115	240	130	125	157	125	133
	$\hat{V}_{16.18}$	122	123	116	243	133	126	159	126	**133**

5.2 Case $n > 2$

In this section, we adopt a standard Monte Carlo simulation approach to examine the performance of the variance estimators. Two real populations are used in this study. The first one is a population of 220 blocks (BL220) taken from Appendix E in Kish (1965). The dataset contains two variables: Y_i = no. of dwellings occupied by renters in block i, and X_i = total no. of dwellings in block i. Some features of this population are: CV (y) = 1.05, CV(x) = 0.85, and $\rho = 0.97$.

The second population comprises 281 municipalities (MU281), and is given in Särndal, Swensson, and Wretman (1992). The role of the study variable, y, is played by RMT85, revenues from the 1985 municipal taxation, while P75, the municipality population in 1975, is used as a measure of size. The main characteristics of this population are: CV(y) = 1.06, CV(x) = 0.96, and $\rho = 0.99$.

Samples of sizes $n = 10, 20$ and 40 with $\pi_i \propto X_i$, $i \in U$, are drawn from BL220 and MU281 by means of randomized systematic πps sampling (RANSYS) and TILLÉ. For each sample, we compute a total estimate using the HT estimator, and variance estimates using the seven variance estimators mentioned in the previous section (for RANSYS, however, the Hartley and Rao (1962) approximation to the π_{ij}, instead of the exact π_{ij}, is used in formula (2)). This sampling-estimation process is repeated R=50,000 times.

Table 4 shows the observed Monte Carlo relative biases of the variance estimators for RANSYS and TILLE. Note that, for TILLE, no values have been provided in the row corresponding to the SYG variance estimator. This is because, given the populations, measures of size, and sample sizes employed here, TILLE produces strictly positive π_{ij}, which means that the SYG variance estimator is design unbiased. All the figures in this table are reasonably small, which seems to support our belief that, under conditions of high entropy, the calculation of the π_{ij} is not essential for obtaining nearly unbiased variance estimates.

Within the group of π_{ij}-free estimators, there are no noticeable differences among them so far as RANSYS is concerned, but $\hat{V}_{HAJ}$ and its relative, $\hat{V}_{DEV}$, seem to perform somewhat better than the $\hat{V}_{16.*}$ family so far as TILLE is concerned, especially for $n = 40$. However, all the observed TILLE biases are positive and tend to increase as the sample size increases. This seems to indicate that TILLE is slightly lower in entropy than RANSYS, in which case the higher observed biases for the $\hat{V}_{16.*}$ family are reflecting the actual facts quite accurately.

Table 4
RB (%) of Variance Estimators for n > 2

Variance estimators	RANSYS			TILLE		
	$n = 10$	$n = 20$	$n = 40$	$n = 10$	$n = 20$	$n = 40$
			BL220			
$\hat{V}_{SYG}$	0.13	1.02	**-0.27**	–	–	–
$\hat{V}_{HAJ}$	-0.14	**0.47**	-2.35	1.49	**2.18**	**3.27**
$\hat{V}_{DEV}$	-0.12	0.54	-2.15	1.52	2.25	3.48
$\hat{V}_{16.9}$	**-0.06**	0.83	-0.52	1.58	2.54	5.21
$\hat{V}_{16.10}$	-0.23	0.64	-0.75	**1.41**	2.34	4.97
$\hat{V}_{16.11}$	0.11	1.02	-0.30	1.75	2.73	5.45
$\hat{V}_{16.18}$	0.13	1.03	-0.29	1.77	2.74	5.45
			MU281			
$\hat{V}_{SYG}$	-0.27	-0.43	0.77	–	–	–
$\hat{V}_{HAJ}$	-0.40	-0.75	-0.59	0.64	**1.01**	**1.93**
$\hat{V}_{DEV}$	-0.37	-0.68	**-0.39**	0.67	1.09	2.14
$\hat{V}_{16.9}$	-0.34	-0.51	0.67	0.70	1.26	3.22
$\hat{V}_{16.10}$	-0.40	-0.58	0.58	**0.63**	1.19	3.13
$\hat{V}_{16.11}$	**-0.27**	-0.43	0.76	0.77	1.34	3.31
$\hat{V}_{16.18}$	-0.27	**-0.43**	0.76	0.78	1.34	3.32

In order to test whether TILLÉ is of slightly lower entropy than RANSYS or not, we compared their Monte

Carlo variances (MCV) with formula (12), the high entropy approximation to the HT variance. The most accurate version of c_i, that is (18), was used to compute (12). The comparison is presented in Table 5. It is seen that the TILLE variances are somewhat smaller than the corresponding RANSYS variances. Moreover, the approximate variances provided by (12) are in closer agreement with the RANSYS variances. These findings support our previous conjecture that the entropy for TILLE is slightly lower than that for RANSYS, particularly when the finite population correction is appreciable.

Table 5
Comparison of Variances (all values in 10^4)

	BL220			MU281		
	$n = 10$	$n = 20$	$n = 40$	$n = 10$	$n = 20$	$n = 40$
(12)+(18)	14.06	6.572	2.830	565.5	264.3	113.7
MCV-RANSYS	14.07	6.520	2.841	566.2	265.3	112.8
MCV-TILLE	13.87	6.404	2.691	560.0	257.6	108.9

Next we focus on stability. Table 6 reports the observed Monte Carlo SE of the variance estimators. Clearly, there are no differences worth mentioning among the variance estimators. The same is true for a comparison of the two sampling procedures. It seems that stability does not constitute a relevant factor when choosing between these variance estimators.

Table 6
CV (%) of Variance Estimators for n > 2

Variance estimators	RANSYS			TILLE		
	$n = 10$	$n = 20$	$n = 40$	$n = 10$	$n = 20$	$n = 40$
			BL220			
$\hat{V}_{SYG}$	58.31	41.16	30.70	57.43	40.41	29.54
$\hat{V}_{HAJ}$	57.90	40.49	**29.48**	57.39	40.24	**29.08**
$\hat{V}_{DEV}$	57.90	40.49	29.48	57.39	40.24	29.08
$\hat{V}_{16.9}$	57.02	40.54	29.64	57.41	40.29	29.24
$\hat{V}_{16.10}$	**57.79**	**40.45**	29.56	**57.29**	**40.19**	29.16
$\hat{V}_{16.11}$	58.04	40.64	29.73	57.53	40.39	29.32
$\hat{V}_{16.18}$	58.05	40.65	29.73	57.55	40.39	29.32
			MU281			
$\hat{V}_{SYG}$	54.90	37.29	25.33	55.07	37.50	25.45
$\hat{V}_{HAJ}$	54.69	36.98	24.96	54.79	37.07	24.78
$\hat{V}_{DEV}$	54.68	36.98	24.95	54.79	37.07	24.77
$\hat{V}_{16.9}$	54.67	36.92	24.70	54.77	37.01	24.52
$\hat{V}_{16.10}$	**54.63**	**36.89**	**24.66**	**54.74**	**36.98**	**24.48**
$\hat{V}_{16.11}$	54.70	36.95	24.74	54.81	37.04	24.56
$\hat{V}_{16.18}$	54.71	36.96	24.74	54.81	37.04	24.56

6. SUMMARY

Estimators are derived for what, in the context of any high entropy selection procedure, is a close approximation to the design variance of the HT estimator of a total.

These estimators resemble, but are not identical to other variance estimators suggested for certain particular high entropy selection procedures by Hájek (1964), Rosen (1997), and Deville (1999). All these estimators have the important advantage over the standard SYG variance estimator that their formulae do not involve the second order inclusion probabilities, π_{ij}.

Empirical investigations indicate that these estimators all behave acceptably well, both for the important special case $n = 2$ and when n takes larger values. The estimator given by (16) with c_i defined by (18), which has certain near-optimal theoretical properties, appears to be noticeably less biased than the others for $n = 2$, but not for larger values of n.

For the case $n > 2$, two high entropy procedures were used, namely systematic sampling from a randomly ordered population (RANSYS) and the procedure proposed by Tillé (1996) (TILLÉ). The biases in all the variance estimators were consistently higher (meaning algebraically larger) for TILLÉ than for RANSYS, and particularly so when n took its largest value of 40. The differences between the TILLE biases and the RANSYS biases were also positive for all values of n, and again particularly so when $n = 40$. We conjecture that these differences may indicate that TILLÉ is a slightly lower entropy (and typically lower variance) selection procedure than RANSYS.

ACKNOWLEDGEMENTS

The authors wish to thank Dr. P.S. Kott for suggesting equation (10) in a private communication, and an anonymous referee for three other sugestions that have added value to this paper.

REFERENCES

ASOK, C., and SUKHATME, B.V. (1976). On sampford's procedure of unequal probability sampling without replacement. *Journal of the American Statistical Association*.71, 912-918.

BREWER, K.R.W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*. 5, 5-13.

BREWER, K.R.W. (1999). Cosmetic calibration for unequal probability samples. *Survey Methodology*. 25, 205-212.

COCHRAN, W.G. (1963). *Sampling Techniques*. 2nd Ed. New York: John Wiley & Sons, Inc.

DEVILLE, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*. 25, 193-203.

FURNIVAL, G.M., GREGOIRE, T.G. and GROSENBAUGH, L.R. (1987). Adjusted inclusion probabilities with 3P sampling. *Forest Science*. 33, 617-631.

GODAMBE, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society B*. 17, 269-278.

GODAMBE, V.P., and JOSHI, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations I, II, and III. *Annals of Mathematical Statistics*. 36, 1707-1742.

HÁJEK, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*. 35, 1491-1523.

HÁJEK, J. (1981). *Sampling from a finite population*. New York: Marcel Dekker.

HANSEN, M.H., and HURWITZ, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*. 14, 333-362.

HARTLEY, H.O., and RAO, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*. 33, 350-374.

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

KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

RAO, J.N.K. (1963). On three procedures of unequal probability sampling without replacement. *Journal of the American Statistical Association*. 58, 202-215.

RAO, J.N.K., and BAYLESS, D.L. (1969). An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling of two units per stratum. *Journal of the American Statistical Association*. 64, 540-559.

ROSÉN, B. (1997). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*. 62, 159-191.

SAMPFORD, M.R. (1962). *An Introduction to Sampling Theory*. Edinburgh and London: Oliver and Boyd Ltd.

SAMPFORD, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*. 54, 499-513.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SEN, A.R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*. 5, 119-127.

SUKHATME, P.V. (1954). *Sampling Theory of Surveys with Applications*. Ames, Iowa: Iowa State College Press.

TILLÉ, Y. (1996). An elimination procedure for unequal probability sampling without replacement. *Biometrika*. 83, 238-241.

YATES, F. (1981). *Sampling Methods for Censuses and Surveys*. 4th Ed. London: Charles Griffin and Co.

YATES, F., and GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*. 15, 235-261.

Survey Methodology, December 2003
Vol. 29, No. 2, pp. 197-205
Statistics Canada

Estimation with Link – Tracing Sampling Designs – A Bayesian Approach

MOSUK CHOW and STEVEN K. THOMPSON[1]

ABSTRACT

In link-tracing designs, social links are followed from one respondent to another to obtain the sample. For hidden and hard-to-access human populations, such sampling designs are often the only practical way to obtain a sample large enough for an effective study. In this paper, we propose a Bayesian approach for the estimation problem. For studies using link-tracing designs, prior information may be available on the characteristics that one wants to estimate. Using this information effectively via a Bayesian approach should yield better estimators. When the available information is vague, one can use noninformative priors and conduct a sensitivity analysis. In our example we found that the estimators were not sensitive to the specified priors. It is important to note that, under the Bayesian setup, obtaining interval estimates to assess the accuracy of the estimators can be done without much added difficulty. By contrast, such tasks are difficult to perform using the classical approach. In general, a Bayesian analysis yields one distribution (the posterior distribution) for the unknown parameters, and from this a vast number of questions can be answered simultaneously.

KEY WORDS: Link-tracing designs; Snowball samples; Adaptive sampling; Graph sampling; Network sampling; Beta prior.

1. INTRODUCTION

Social network data include measurements on the relationships between people or other social entities as well as measurements on entities themselves. Collecting network data on entire networks requires a great deal of time and effort, especially when networks are large. It is thus important to be able to estimate network properties from samples. In link-tracing sampling designs, social links are followed from one respondent to another to obtain the sample. For hidden and hard-to-access human populations, such sampling designs are often the only practical way to obtain a sample large enough for an effective study. For example, in a study of injection drug use in relation to the spread of the HIV infection, social leads from initial respondents may be traced and the linked individuals added to the sample. (*e.g.*, see Neaigus, Friedman, Goldstein, Ildefonso, Curtis and Jose 1995; Neaigus, Friedman, Jose, Goldstein, Curtis, Ildefonso and Des Jarlais 1996 and Thompson and Collins 2002). Similarly, for studies of homeless people, respondents may be asked about other homeless people who will then be sampled.

Populations with social structure are often modeled as graphs, with the nodes of the graph representing populations and the arcs of the graph representing social links, relationships, or transactions. In the graph setting, the variables of interest include both those associated with nodes and those associated with pairs of nodes. The population graph itself can be viewed either as a fixed structure or as a realization of a stochastic graph model. Samples are taken to obtain information about the population graph. Usually, the sampling method will take advantage of the arcs or links from one entity to another.

There is a large literature on network sampling, both applied and theoretical. Frank (1977a, 1977b, 1977c, 1978, 1979, 1980, 1997) has many important results in sampling for social networks. His classic work (Frank 1971) presents basic solutions for estimating graph quantities from the sample data. Snijders and Nowicki (1997) propose various statistical approaches, including a Bayesian approach, for estimation and prediction with stochastic blockmodels for graphs in which the node values are not observed.

Snowball sampling (Goodman 1961) is one type of link-tracing sampling design in which individuals in an initial sample are asked to identify acquaintances, who in turn were asked to identify acquaintances, and so on for a fixed number of stages or waves. Erickson (1978) and Frank (1979) review snowball sampling designs with the goal of understanding how other "chain methods" (methods designed to trace ties through a network from a source to an end) can be used in practice. Snijders (1992) used the same term "snowball sampling" to include designs in which only a subsample of links from each node is traced. Frank and Snijders (1994) consider model and design-based estimation of a hidden population size, that is, the number of nodes in the graph, based on snowball samples. Another link-tracing procedure for which design-based estimators are available is adaptive cluster sampling Thompson and Seber 1996), which has been formulated in the graph setting as well as the spatial setting.

With a fixed-population, design-based approach in the graph setting, both the characteristics of the people and the social network structure of the population are viewed as fixed, unknown values. The properties such as design-unbiasedness do not depend on any assumptions about the

[1] Mosuk Chow and Steven K. Thompson, Department of Statistics, The Pennsylvania State University, 326 Thomas Building, University Park, PA 16802, U.S.A.

population itself but they do depend on the sampling design being carried out as specified. In this paper, we consider the model-based methods since they can be applied to a wide range of sample selection procedures. In many studies of hidden and hard-to-reach populations, the sample selection procedures, including link-tracing, are not readily analyzed based on idealized design induced probabilities, but results from the model-based methods can be applied for the cases.

Thompson and Frank (2000) used a model-based approach to inference with link-tracing designs. In their paper, maximum likelihood estimators of population graph parameters and predictors of realized population graph quantities were described. In this paper, we adopt a Bayesian approach for the graph estimation problem. For real problems with sampling designs that follow social links from one person to another, prior information may be available on the characteristics that one wants to estimate. Using this information effectively via a Bayesian approach should yield improved estimators. Moreover, when the available information is vague, we can use noninformative priors and conduct a sensitivity analysis. It is important to note that under the Bayesian setup, obtaining interval estimates to assess the accuracy of the estimators can be done without much added diffculty whereas such tasks would be difficult to perform using the maximum likelihood approach. We deal with inferences for both the characteristic of nodes and also of arcs such as the prevalence of disease in a certain community and also the transmission rate of that disease between two subjects.

Notation for a full graph model with links related to node values and its likelihood function will be given in section 2. In section 3, the likelihood function for the sample obtained from a link-tracing design will be presented and a Bayesian inference method will be introduced. In section 4, an illustrative example will be given. The paper will be concluded by an empirical example and a discussion in section 5.

2. THE MODEL

Using notation similar to Frank (1971) and Thompson and Frank (2000), we denote the full set of node labels by $U = \{1, 2, ..., N\}$ which form the population of N units. A variable of interest associated with an individual node u will be denoted Y_u while a variable of interest associated with pair of nodes u and v will be denoted A_{uv}. The sequence of node variables of interest is denoted by $\mathbf{Y} = (Y_1, ..., Y_N)$. Here we consider the variable of interest A_{uv} as an indicator variable which equals one if there is an arc (directional link) from u to v and zero otherwise for two distinct nodes u and v. The matrix of arc indicators, having A_{uv} as the element in the u-th row and v-th column, is the graph adjacency matrix, denoted **A**. For convenience we will assume that the diagonal elements A_{uu} are zero. The ordered pair (u,v) is referred to as a dyad of type $(Y_u, Y_v; A_{uv}, A_{vu})$. In the following assumed model the node variables $Y_1, ..., Y_N$ are independent, identically distributed (i.i.d.) Bernoulli random variables with probabilities $P(Y_u = i) = \theta_i$, for $i = 0,1$, and $\theta_0 + \theta_1 = 1$. Conditional on the node values $Y_1, ..., Y_N$, the dyads (A_{uv}, A_{vu}) are independent, for $1 \le u < v \le N$, with conditional distribution given by $P[(A_{uv}, A_{vu}) = (k, l) \mid Y_u = i, Y_v = j] = \lambda_{ijkl}$ for all combinations of $i = 0,1$; $j = 0,1$; $k = 0,1$; and $l = 0,1$. For all combinations of i and j, the sums over k and l are denoted $\lambda_{ij..} = \sum_k \sum_l \lambda_{ijkl}$ and equal 1. In order to get graph probabilities not depending on node identities, the following natural symmetry conditions are assumed: $\lambda_{1110} = \lambda_{1101}$, $\lambda_{1011} = \lambda_{0111}$, $\lambda_{1010} = \lambda_{0101}$, $\lambda_{1001} = \lambda_{0110}$, $\lambda_{0010} = \lambda_{0001}$ and $\lambda_{1000} = \lambda_{0100}$. For example, the first and the fifth conditions say that between two nodes having the same value, the probability of an arc in either direction is the same. Let N_i denote the total number of nodes with value i in the graph so that $N_0 + N_1 = N$. Let further M_{ijkl} denote the total number of dyads of type $(ijkl)$, that is, the total number of ordered node pairs (u, v) such that $(Y_u, Y_v; A_{uv}, A_{vu}) = (ijkl)$. The likelihood for the full graph under the model with parameters (θ, λ) is $L(\theta, \lambda; Y, A) = (\Pi_{i=0}^{1} \theta_i^{N_i})(\Pi_{i=0}^{1} \Pi_{j=0}^{1} \Pi_{k=0}^{1} \Pi_{l=0}^{1} \lambda_{ijkl}^{M_{ijkl}})$.

3. BAYESIAN INFERENCE FROM LINK-TRACING DESIGNS

3.1 Likelihood Function given the Sample Data

A sample s from the graph is a subset of nodes from U and a subset of node pairs from U^2. The sample data $d = (s, y_s, a_s)$ are a function of the sample selected and of the graph values y and a. For any design in which the selection of the sample depends on graph y and a values only through those values y_s and a_s included in the data, the design does not affect the value of estimators or predictors based on direct likelihood methods such as maximum likelihood or Bayes estimators (Rubin 1976, Thompson and Frank 2000). For example, many of the snowball and other link-tracing designs are ignorable for likelihood-based inference provided the selection procedure for the initial sample is ignorable. Any carefully implemented conventional or adaptive survey design would be ignorable in this sense. Nonignorable initial samples can occur when the selection is uncontrolled and selection probabilities are related to unobserved node and link values, as when people with risk-averse behaviors and low numbers of relationships are less conspicuous to investigators, thereby influencing what units are missed and hence influencing sample selection probabilities in ways that are not measured.

Consider the link-tracing design in which an initial sample s_0 is selected and all links out from nodes in s_0 are followed to add the set s_1 of nodes not in s_0 that are adjacent to nodes in s_0. The whole sample is $s = s_0 \cup s_1$. The entire set of labels in the population can be written as

the union of three disjoint sets, $U = s_0 \cup s_1 \cup \bar{s}$ where $\bar{s}$ denotes the nonsampled nodes. Here, we consider a design in which the decision to follow the links from node u depends on the node value y_u. For example, in a study on injection drug use, the initial sample may contain both users and nonusers. If the investigators choose to follow social links only from users, then the design depends adaptively on the node y-values as well as the links. The design then can be written $P(s \mid y_s, a_{s_{0U}})$, since the selection procedure depends on both node and link values. The data are $d = (s, y_s, a_{s_{0U}})$. Since the decision depends on y and a values only through the observed data, the design factors out of the likelihood function and divides out of the Bayes posterior, so that likelihood or Bayes inference depends only on the assumed model.

With the graph model described in the previous section, it then follows (Thompson and Frank 2000) that the likelihood with the sample data is:

$$L(\theta, \lambda; d) = P(s \mid y_s, a_{s_{0U}}) \sum \left(\prod_{u=1}^{N} \theta_{y_u} \right) \left(\prod_{u<v} \lambda_{y_u y_v a_{uv}, a_{vu}} \right)$$

where the sum is over all values of y_u and a_{uv} that are not fixed by the sample data.

For link-tracing designs in which all links, rather than a subsample, from the initial sample nodes are traced, all of the elements in the submatrix $a_{s_0\bar{s}}$ are zero. It has been shown by Thompson and Frank (2000) that the likelihood function can then be written as:

$$L(\theta, \lambda; Y, A) = P(s \mid y_s, a_{s_{0U}}) \left(\prod_i \theta_i^{n_i(s)} \right) \left(\prod_{ijkl} \lambda_{ijkl}^{m_{ijkl}(s_0, s_0)} \right) \left(\prod_{ijk} \lambda_{ijk.}^{m_{ijk.}(s_0, s_1)} \right) \times \left[\sum_j \theta_j \prod_i \lambda_{ij0.}^{n_i(s_0)} \right]^{n(\bar{s})} \quad (1)$$

where $n_i(s), n_i(s_0)$, and $n_i(\bar{s})$ denote the numbers of nodes of type i in the full sample s, the initial sample s_0, and the nonsampled nodes $\bar{s}$, respectively, and $m_{ijkl}(s_0, s_0), m_{ijkl}(s_0, s_1)$ are the counts of node pairs in $s_0 \times s_0$ and $s_0 \times s_1$.

For a symmetric model, $\lambda_{ijkl} = 0$ for $k \neq l$ so that arcs are always two-way or, equivalently, they can be considered as undirected edges. The full symmetric model has parameters $\lambda_{ijkk} = \lambda_{jikk}$ for $i, j, k = 0,1$, with $\lambda_{ij00} + \lambda_{ij11} = 1$. To simplify notation for this model, let $\beta_{i+j} = \lambda_{ij11}$ and thus β_k denotes the probability of a mutual link between two nodes having total value k, for $k = 0,1$ or 2. The above likelihood simplifies to

$$L(\theta, \beta; d) = P(s \mid y_s, a_{s_{0U}}) \left(\prod_i \theta_i^{n_i(s)} \right) \left(\prod_{i,j} \beta_{i+j}^{m_{ij11}(s_0, s)} (1 - \beta_{i+j})^{m_{ij00}(s_0, s)} \right) \times \left[\sum_j \theta_j \prod_i (1 - \beta_{i+j})^{n_i(s_0)} \right]^{n(\bar{s})}. \quad (2)$$

Now define $r_{0,0} = m_{0000}(s_0, s), r_{0,2} = m_{0011}(s_0, s), r_{1,0} = m_{0100}(s_0, s) + m_{1000}(s_0, s), r_{1,2} = m_{0111}(s_0, s) + m_{1011}(s_0, s), r_{2,0} = m_{1100}(s_0, s), r_{2,2} = m_{1111}(s_0, s)$. Note that the r's are dyad counts where the first index represents the sum of the node values and the second index represents the sum of the link values. The above expression can be rewritten as:

$$L(\theta, \beta; d) = P(s \mid y_s, a_{s_{0U}}) \theta_0^{n_0(s)} (1 - \theta_0)^{n_1(s)} \beta_0^{r_{0,2}} (1 - \beta_0)^{r_{0,0}} \beta_1^{r_{1,2}} (1 - \beta_1)^{r_{1,0}} \beta_2^{r_{2,2}} (1 - \beta_2)^{r_{2,0}} \left[\theta_0 (1 - \beta_0)^{n_0(s_0)} (1 - \beta_1)^{n_1(s_0)} + (1 - \theta_0) (1 - \beta_1)^{n_0(s_0)} (1 - \beta_2)^{n_1(s_0)} \right]^{n(\bar{s})}. \quad (3)$$

In the remainder of this paper, we focus on the full symmetric model to illustrate the proposed Bayesian methodology for simplicity of presentation. The same method can be applied to the general model with the likelihood function given in (1).

3.2 Choice of Prior Distributions

Since there are no specific constraints on $\theta_0, \beta_0, \beta_1, \beta_2$, we may assume independent priors on $\theta_0, \beta_0, \beta_1, \beta_2$, all of which take values in the interval [0, 1]. It is quite common to put a beta prior on a parameter that takes values in [0, 1] because most smooth unimodal distributions on [0, 1] can be well approximated by some beta distributions and the class of beta distributions is reasonably rich to model the uncertainty about the parameter. Also, the expression in (3) is in general quite complex but beta priors can yield a tractable posterior distribution (to be shown later). Using beta priors, we obtain an analytic formula for the Bayes estimates and the marginal posterior distribution.

In this paper we consider independent beta priors for the parameters:

$$\pi(\theta_0, \beta_0, \beta_1, \beta_2) \propto \theta_0^{a-1} (1 - \theta_0)^{b-1} \beta_0^{c-1} (1 - \beta_0)^{d-1} \beta_1^{e-1} (1 - \beta_1)^{f-1} \beta_2^{g-1} (1 - \beta_2)^{h-1}. \quad (4)$$

When determining the constants a and b it is often useful to equate the mean $E[\theta_0] = a/(a + b)$ of Beta(a, b) to a value which represents your belief about the location of θ_0 and the variance $\mathrm{Var}[\theta_0] = ab/(a + b)^2 (a + b + 1)$ of Beta (a, b) to a value which represents the uncertainty put on the specified θ_0 value. Similarly, the values of c, d, e, f, g and

h can be determined. For example, if one is interested in the prevalence of injection drug use in a certain community, one may take an initial sample and trace links by asking the injection drug user in the sample to name the people with whom they share injection equipment. If the value $y_u = 1$ represents injection drug use, then θ_0 is the percentage of non-users in that community. Quite often an estimate for the central location and the spread of θ_0 may be provided.

In the case of complete ignorance, we will consider three commonly used noninformative priors and provide a comparison of the resulting Bayes estimates in our illustrative example in section 4. (For a fuller discussion of the noninformative priors, see Berger 1985, pages 89-90). The first one is the uniform prior, which corresponds to Beta(1,1). The second one, Beta(0, 0), suggested by Haldane (1931), has an improper density. It is equivalent to a prior uniform in the log-odds $\log\{\theta_0/(1-\theta_0)\}$. A possible compromise between Beta(1,1) and Beta(0,0) is Beta(1/2, 1/2), which has a proper density. This prior implies a uniform prior for $\sin^{-1}\sqrt{\theta_0}$.

3.3 Posterior Distribution and Bayes estimates

In our problem, the posterior distribution $\pi(\theta_0, \beta_0, \beta_1, \beta_2 \mid d)$ corresponding to the beta priors is given by:

$$\begin{aligned}\pi(\theta_0,\beta_0,\beta_1,\beta_2 \mid d) \propto\ & \theta_0^{n_0(s)+a-1}(1-\theta_0)^{n_1(s)+b-1} \\ & \beta_0^{r_{0,2}+c-1}(1-\beta_0)^{r_{0,0}+d-1} \\ & \beta_1^{r_{1,2}+e-1}(1-\beta_1)^{r_{1,0}+f-1} \\ & \beta_2^{r_{2,2}+g-1}(1-\beta_2)^{r_{2,0}+h-1} \\ & \Big[\theta_0(1-\beta_0)^{n_0(s_0)}(1-\beta_1)^{n_1(s_0)} \\ & \quad +(1-\theta_0)(1-\beta_1)^{n_0(s_0)} \\ & \quad (1-\beta_2)^{n_1(s_0)}\Big]^{n(\bar{s})}. \end{aligned} \tag{5}$$

To find the posterior mean (Bayes estimate) of θ_0, let

$$\begin{aligned} q(\theta_0,\beta_0,\beta_1,\beta_2) = & \theta_0^{n_0(s)+a-1}(1-\theta_0)^{n_1(s)+b-1} \\ & \beta_0^{r_{0,2}+c-1}(1-\beta_0)^{r_{0,0}+d-1} \\ & \beta_1^{r_{1,2}+e-1}(1-\beta_1)^{r_{1,0}+f-1} \\ & \beta_2^{r_{2,2}+g-1}(1-\beta_2)^{r_{2,0}+h-1} \\ & \Big[\theta_0(1-\beta_0)^{n_0(s_0)}(1-\beta_1)^{n_1(s_0)} \\ & \quad +(1-\theta_0)(1-\beta_1)^{n_0(s_0)} \\ & \quad (1-\beta_2)^{n_1(s_0)}\Big]^{n(\bar{s})}. \end{aligned}$$

Since $\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx = B(\alpha,\beta)$ is the beta function, we have the following two results:

$$\begin{aligned} M_1 = & \int_0^1\int_0^1\int_0^1\int_0^1 q(\theta_0,\beta_0,\beta_1,\beta_2)\,d\theta_0\,d\beta_0\,d\beta_1\,d\beta_2 \\ = & \sum_{i=0}^{n(\bar{s})}\binom{n(\bar{s})}{i} B(n_0(s)+a+i,\, n(\bar{s})+n_1(s)+b-i) \\ & B(r_{0,2}+c,\, i\,n_0(s_0)+r_{0,0}+d)\,B(r_{1,2} \\ & +e,\, i\,n_1(s_0)+(n(\bar{s})-i)\,n_0(s_0)+r_{1,0}+f) \\ & B(r_{2,2}+g,\,(n(\bar{s})-i)\,n_1(s_0)+r_{2,0}+h). \end{aligned}$$

$$\begin{aligned} M_2 = & \int_0^1\int_0^1\int_0^1\int_0^1 \theta_0\, q(\theta_0,\beta_0,\beta_1,\beta_2)\,d\theta_0\,d\beta_0\,d\beta_1\,d\beta_2 \\ = & \sum_{i=0}^{n(\bar{s})}\binom{n(\bar{s})}{i} B(n_0(s)+a+1+i,\, n(\bar{s})+n_1(s)+b-i) \\ & B(r_{0,2}+c,\, i\,n_0(s_0)+r_{0,0}+d)\,B(r_{1,2} \\ & +e,\, i\,n_1(s_0)+(n(\bar{s})-i)\,n_0(s_0)+r_{1,0}+f) \\ & B(r_{2,2}+g,\,(n(\bar{s})-i)\,n_1(s_0)+r_{2,0}+h). \end{aligned}$$

The Bayes estimate for θ_0 can thus be evaluated by the quotient of the righthand side of the above two equations since:

$$\begin{aligned} E(\theta_0 \mid d) &= \frac{\int_0^1\int_0^1\int_0^1\int_0^1 \theta_0\, q(\theta_0,\beta_0,\beta_1,\beta_2)\,d\theta_0\,d\beta_0\,d\beta_1\,d\beta_2}{\int_0^1\int_0^1\int_0^1\int_0^1 q(\theta_0,\beta_0,\beta_1,\beta_2)\,d\theta_0\,d\beta_0\,d\beta_1\,d\beta_2} \\ &= \frac{M_2}{M_1}. \end{aligned}$$

Similarly, the Bayes estimates for $\beta_0, \beta_1, \beta_2$ can be computed.

3.4 Prediction of Realized Graph Quantities

Consider the problem of estimating or predicting, from the sample data, the realized value of some graph quantity $Z = Z(\mathbf{Y}, \mathbf{A})$, an observable but unobserved finite-population quantity. Denoting the unknown parameters collectively by ψ, the relevant posterior predictive density is

$$\begin{aligned} f(z \mid d) &= \int f(z \mid d, \psi)\pi(\psi \mid d)\,d\psi \\ &\propto \int f(d, z \mid \psi)\pi(\psi)\,d\psi \end{aligned} \tag{6}$$

where the constant of proportionality is, as usual, $f(d)$.

For example, suppose the objective is to predict the proportion of nodes in the population that have value $y = 1$. Let $n_1(s)$ denote the number of nodes for which $y = 1$ in the sample, and let $n_1(\bar{s})$ denote the number of nodes with value 1 among the nodes not in the sample. Note that

$n_1(s)$ is observed and $n_1(\bar{s})$ is an unobserved quantity to be estimated or predicted. The realized proportion of value-1 nodes in the population is denoted $Z = (n_1(s) + n_1(\bar{s}))/N$, where N is the total number of nodes in the population.

For a one-wave snowball design with an ignorable initial sample from which all links are traced and with the nondirected stochastic block model, the joint predictive likelihood is

$$f(d, n_1(\bar{s}) \mid \theta_0, \beta_0, \beta_1, \beta_2) =$$

$$p(s \mid y_s, a_{s_0 U}) \binom{n(\bar{s})}{n_1(\bar{s})}$$

$$\theta_0^{n_0(s)+n_0(\bar{s})}(1-\theta_0)^{n_1(s)+n_1(\bar{s})}$$

$$\beta_0^{r_{02}(s_0,s)}(1-\beta_0)^{r_{00}(s_0,s)+n_0(s_0)n_0(\bar{s})}$$

$$\beta_1^{r_{12}(s_0,s)}(1-\beta_1)^{r_{10}(s_0,s)+n_0(s_0)n_1(\bar{s})+n_1(s_0)n_0(\bar{s})}$$

$$\beta_2^{r_{22}(s_0,s)}(1-\beta_2)^{r_{20}(s_0,s)+n_1(s_0)n_1(\bar{s})}. \qquad (7)$$

With joint likelihood (7) and independent beta priors and carrying out the integration, the posterior predictive density for the finite-population proportion Z becomes

$$f(n_1(\bar{s}) \mid d) \propto \binom{n(\bar{s})}{n_1(\bar{s})} B\left[n_0(s)+n_0(\bar{s})+a, n_1(s)+n_1(\bar{s})+b\right]$$

$$B\left[r_{02}+c, r_{00}+n_0(s_0)n_0(\bar{s})+d\right]$$

$$B\left[r_{12}+e, r_{10}+n_0(s_0)n_1(\bar{s})+n_1(s_0)n_0(\bar{s})+f\right]$$

$$B\left[r_{22}+g, r_{20}+n_1(s_0)n_1(\bar{s})+h\right].$$

The Bayes predictor of $n_1(\bar{s})$ is

$$E\left[n_1(\bar{s}) \mid d\right] = \sum_{n_1(\bar{s})=0}^{n(\bar{s})} n_1(\bar{s}) f(n_1(\bar{s}) \mid d).$$

Since $i\binom{n}{i} = n\binom{n-1}{i-1}$,

$$E[n_1(\bar{s}) \mid d] \propto n(\bar{s}) \sum_{i=1}^{n(\bar{s})} \binom{n(\bar{s})-1}{i-1} B\left[n_0(s)+n(\bar{s})-i\right.$$

$$\left.+a, n_1(s)+i+b\right]$$

$$B\left[r_{02}+c, r_{00}+n_0(s_0)(n(\bar{s})-i)+d\right]$$

$$B\left[r_{12}+e, r_{10}+n_0(s_0)i+n_1(s_0)(n(\bar{s})-i)+f\right]$$

$$B\left[r_{22}+g, r_{20}+n_1(s_0)i+h\right]$$

$$= M_3.$$

in which M_3 is defined to be the right hand side. Thus, since $M_1 = f(d)$ defined earlier is the proportionality constant, $E[n_1(\bar{s}) \mid d] = M_3/M_1$.

Therefore, the Bayes predictor $\hat{Z}$ of the realized proportion Z of positive nodes in the population is

$$\hat{Z} = E(Z \mid d) = E[(n_1(s) + n_1(\bar{s}))/N \mid d]$$

$$= \frac{n_1(s) + (M_3/M_1)}{N}. \qquad (8)$$

4. AN ILLUSTRATIVE EXAMPLE

Here, we consider an example which concerns estimating the percentages of injection drug users and nonusers among a certain target population. Let θ_0 represent the proportion of non injection drug users in the target population. Then $1 - \theta_0$ is the proportion of injection drug users. Suppose that there are 200 people in that population. In the first wave sample, 22 people are sampled randomly without replacement and 5 of those sampled are injection drug users whereas 17 are not. The injection drug users are asked to name their injection partners. Note that links are only possible between users and tracing these links can only add users to the sample. The initial users give 12 referrals, of which 10 are distinct users not in the initial sample. The statistics are:

$$n_1(s_0) = 5, n_0(s_0) = 17, n_1(s) = 15, n_0(s) = 17,$$

$$n(\bar{s}) = 168, r_{22} = 12, r_{20} = 93.$$

In terms of the notation of section 3, $\beta_0 = \lambda_{0011}$ is the probability of a mutual link between two non injection drug users. $\beta_1 = \lambda_{1011} = \lambda_{0111}$ is the probability of a mutual link between injection drug user and non injection drug user (it is natural that the two different orders of node values have the same probability). $\beta_2 = \lambda_{1111}$ is the probability of a mutual link between two injection drug users. Since non injection drug users will by definition not have injection partners, $\beta_0 = \beta_1 = 0$ for this example.

The Bayes estimates for θ_0 and β_2 corresponding to different noninformative priors are given in table 1.

Note that the three noninformative priors are very different from each other. For example, the improper non-informative prior corresponding to $a = b = g = h = 0$ place a lot of its weight on both 0 and 1. This would arise in practice when people in a certain neighbourhood are either all injection drug users or are all non injection drug users, but we just do not know which one. On the other hand, the prior corresponding to $a = b = g = h = 1$ place a flat weight to values between 0 and 1. Even though the three priors are very different, the posterior distributions corresponding to these three non-informative priors nearly coincide with each other. Figure 1 shows the posterior distribution of θ_0 and β_2 corresponding to the three non-informative priors. One can conclude that the Bayes estimates here are not sensitive to the specification of the three priors.

Table 1
Bayes estimates for noninformative priors corresponding to the specified values of a, b, g, h
(The values in the brackets are the 95% HPD regions)

Bayes estimate	$a=b=g=h=0$	$a=b=g=h=.5$	$a=b=g=h=1$
$\hat{\theta}_0$	.7273 (.5706, .8713)	.7285 (.5747, .8670)	.7295 (.5786, .8686)
$\hat{\beta}_2$	.0420 (.0153, .0738)	.0439 (.0164, .0766)	.0458 (.0175, .0791)

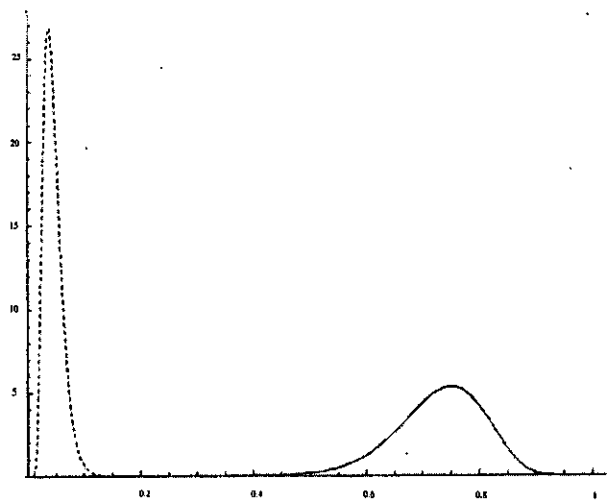

Figure 1. Marginal Posterior distributions: solid line for θ_0 and dashed line for β_2. (The posterior distributions corresponding to the three non-informative priors are given here and they nearly coincide)

For comparison purposes, it is of interest to note that the maximum likelihood estimates obtained using the likelihood function given in (3) are calculated to be: $\hat{\theta}_0 = .7604, \hat{\beta}_2 = .0501$, not far from the Bayes estimates. However, it is not easy to compute confidence intervals for the maximum likelihood estimate whereas one can obtain the posterior intervals for the Bayes estimates without any additional diffculty. For example, a $(1-\alpha)$ highest posterior density (HPD) region can be obtained for the specified α value for each parameter $\theta_0, \beta_0, \beta_1, \beta_2$, where HPD is the region of values that contains $(1-\alpha)$ of the posterior probability for that parameter with the characteristic that the density within the region is never lower than that outside. It is worthwhile to note that the posterior intervals can be directly regarded as having the stated probability of containing the unknown quantity in contrast to the repeated sampling property of frequentist confidence interval. See Gelman, Carlin, Stern and Rubin (1995, pages 104-106) for a discussion on the frequency property of some Bayesian procedures.

From Table 1, we can see that even though the width of the HPD interval of β_2 is large compared to the magnitude of its Bayes estimate, it gives us a rough order-of-magnitude estimate of β_2 and provides useful information to the subject matter specialists.

5. AN EMPIRICAL EXAMPLE AND DISCUSSION

To examine the properties of estimators and predictors under repeated sampling, socially-networked data from the Colorado Springs study on the heterosexual transmission of HIV/AIDS was used as an empirical population from which to repeatedly sample. The Colorado Springs study, which is described in Potterat, Woodhouse, Rothenberg, Muth, Darrow, Muth and Reynolds (1993); Rothenberg, Woodhouse, Potterat, Muth, Darrow and Klovdahl (1995), and Darrow, Potterat, Rothenberg, Woodhouse, Muth and Klovdahl (1999), involved a very thorough investigation of a population of people thought to be at high risk for infection with the human immunodeficiency virus. In the study, data were obtained not only on the risk-related behaviors of individuals, but also on their social relationships with other individuals. Risk-related behaviors included various sexual and drug-use behaviors, and the social links examined included sexual and drug-use relationships. Over the course of the study, data were obtained on several thousand people.

For our empirical population we have used the 595 individuals in the study for which the data on both individual risk-related behaviors and relationships to other people in the study are complete. For the node variable of interest we chose a high-risk sexual behavior (commercial sex work) and sexual relationship for the link variable of interest. Figure 2 shows a graphical representation of the empirical population, in which the nodes or circles represent people in the study and the lines represent sexual relationships between pairs of individuals. Presence of the high-risk sexual behavior ($y = 1$) is indicated by a dark colored circle, while presence of a sexual relationship between two individuals is indicated by a line between the two circles. The positioning of the nodes in the graph is arbitrary, but has been arranged to separate connected components. The largest connected component contains 219 of the 595 people in the population. The next largest connected component contains 12 people, followed by several components of 4, 3 and 2 people. There are 267 people without sexual relationships to others among the 595 in the empirical population. The extremely uneven distribution of connected component sizes exemplified by this population presents one of the challenges to sampling design and inference in such populations.

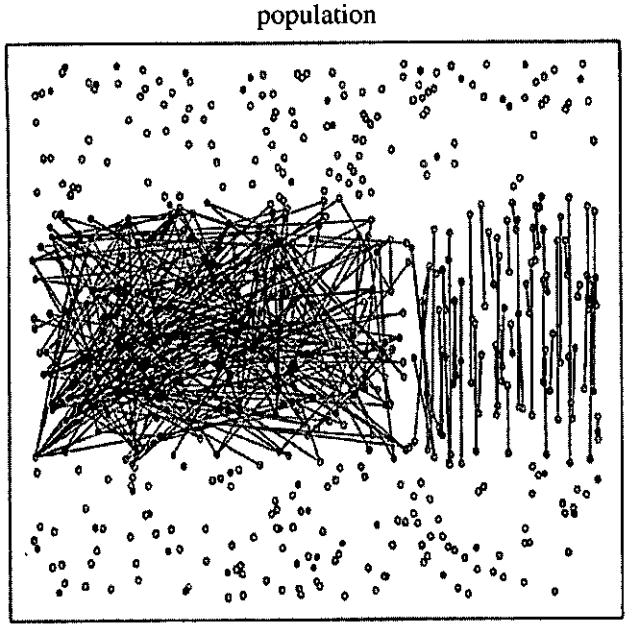

Figure 2. Colorado Springs study on the heterosexual transmission of HIV/AIDS (Potterat *et al.* 1991; Rothenberg *et al.* 1993; Darrow *et al.* 1999): The 595 people in the empirical population. Dark circles represent individuals with high-risk sexual behavior (sex work). Links between circles indicate sexual relationships.

Figure 3. shows a one-wave snowball sample from this population. First, a simple random sample of 40 nodes (circled in the figure) is selected. All links from these initial nodes are traced to add the additional nodes to the sample.

Repeated sampling of the empirical population was carried out using the one-wave snowball design with initial simple random sample of 40 individuals. The addition of a wave of new nodes brought the total sample size to 85, on average. For each sample, various estimators of the proportion of high-risk individuals $(y = 1)$ in the population were computed, and this procedure was repeated 1,000 times. The undirected stochastic block graph model was used for the maximum likelihood and Bayes estimators of θ and the Bayes predictor of the finite-population proportion z. A uniform prior was used for the Bayes procedures. Table 2 and Figure 4 summarize the properties under the repeated sampling of the different estimators. The actual proportion of nodes having value $(y = 1)$ in the empirical population is 0.2235. The sample proportion overestimates relative to the actual proportion because the linktracing has a tendency to enrich the sample with high-risk nodes. Each of the model-based estimators has relatively little bias with the link-tracing design.

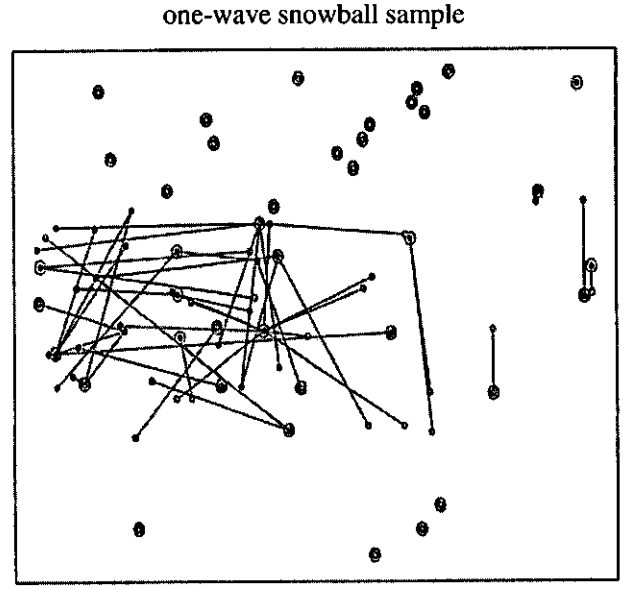

Figure 3. A one-wave snowball sample selected from the Colorado Springs empirical population. From an initial random sample of 40 individuals (circled), links are traced to add one wave of new individuals to the sample.

Table 2

Means and mean square errors of estimators of the population mean of the node values, for the Colorado Springs empirical population. The actual mean of node values in the population is 0.2235294. The design is a one-wave snowball sample with an initial random sample of 40 nodes. The average final sample size was 82.65. The number of simulation runs is 1,000.

Type of estimator:	sample proportion	m.l.e.	Bayes estimator	Bayes predictor
mean:	0.3147	0.2155	0.251	0.2142
m.s.e.:	0.011391	0.003279	0.003261	0.003275

In this paper, we employ a Bayesian approach to the estimation problem with link-tracing design and show that, corresponding to the independent beta priors, the posterior distribution can be evaluated analytically. If a more general prior is desired then one can use the Markov Chain Monte Carlo (MCMC) method to evaluate the posterior for that general prior. References for using MCMC techniques in Bayesian computations include Gilks, Richardson and Spiegelhalter (1996) and Gelman, Carlin, Stern and Rubin (1995) . The approach used in Gelfand and Smith (1990) can be adapted for the implementation of the MCMC simulations here.

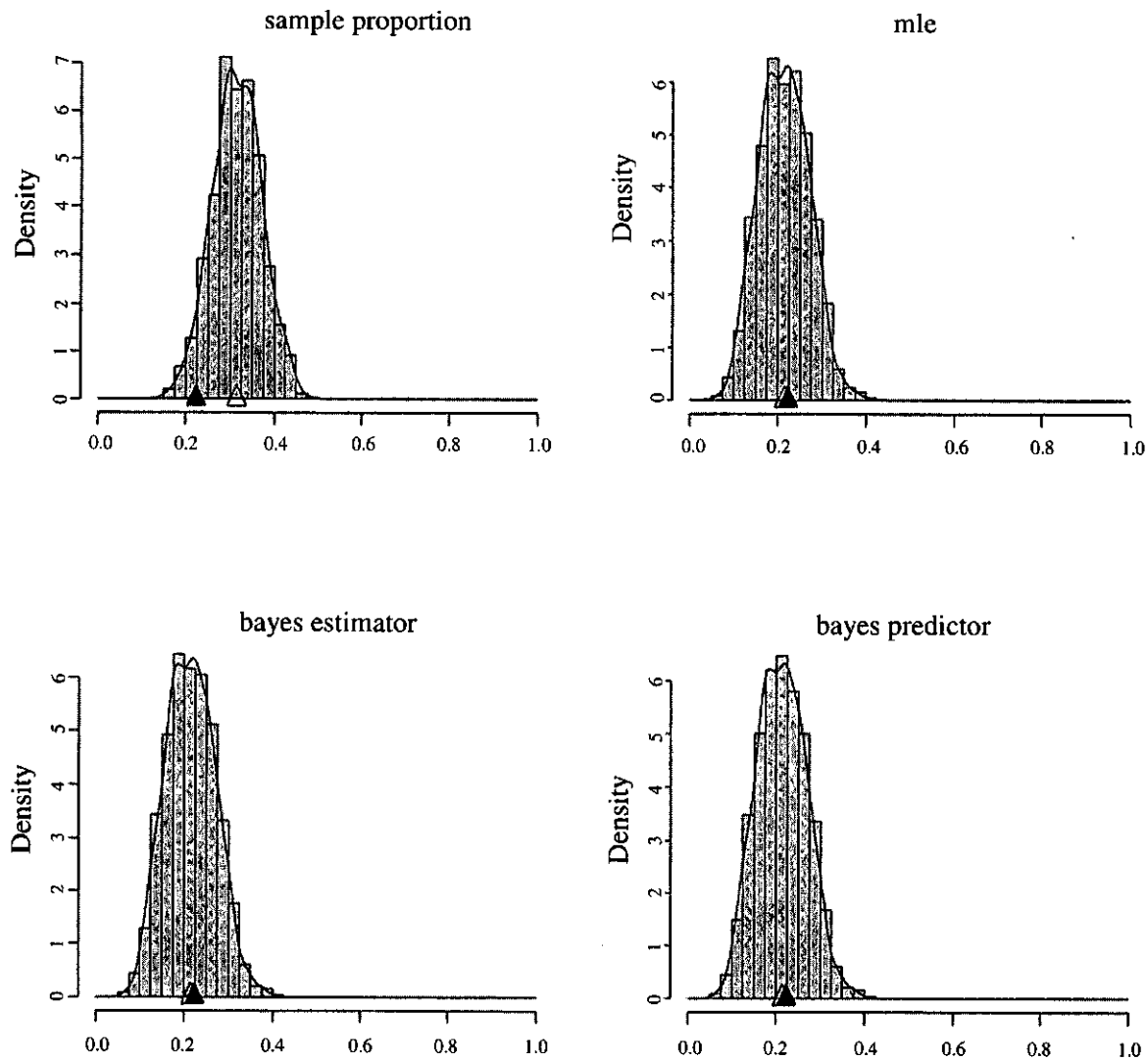

Figure 4. Distributions of estimators of the proportion of individuals in the high-risk category in the Colorado Springs empirical population, with the one-wave snowball design using an initial sample of 40. Solid triangle is the actual proportion in the population. Hollow triangle is the mean of the distribution of the estimator. The number of simulations was 1000.

ACKNOWLEDGEMENTS

Support for this work was provided by funding from the National Center for Health Statistics, the National Science Foundation (DMS-9626102), and the National Institutes of Health (R01-DA09872). The authors would like to thank John Potterat and Steve Muth for advice and use of the data from the Colorado Springs study. We would also like to thank the Associated Editor and the referees for their insightful comments and suggestions.

REFERENCES

BERGER, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, (2nd ed.). Berlin: Springer-Verlag.

DARROW, W.W., POTTERAT, J.J., ROTHENBERG, R.B., WOODHOUSE, D.E., MUTH, S.Q. and KLOVDAHL, A.S. (1999). Using knowledge of social networks to prevent human immunodeficiency virus infections: The Colorado Springs Study. *Sociological Focus*. 32, 143-158.

ERICKSON, B. (1978). Some problems of inference from chain data. In *Sociological Methodology*, 1979, K.F. Schuessler (Ed.) San Francisco: Jossey-Bass. 276-302.

FRANK, O. (1971). *Statistical Inference in Graphs*. Stockholm: Försvarets forskningsanstalt.

FRANK, O. (1977a). Survey sampling in graphs. *Journal of Statistical Planning and Inference*. 1, 235-246.

FRANK, O. (1977b). A note on Bernoulli sampling in graphs and Horvitz-Thompson estimation. *Scandinavian Journal of Statististics*. 4, 178-180.

FRANK, O. (1977c). Estimation of graph totals. *Scandinavian Journal of Statististics*. 4, 81-89.

FRANK, O. (1978). Estimating the number of connected components in a graph by using a sampled subgraph. *Scandinavian Journal of Statististics*. 5, 177-188.

FRANK, O. (1979). Estimation of population totals by use of snowball samples. In *Perspectives on Social Network Research*. (P.W. Holland, and S. Leinhardt, Eds.). New York: Academic Press. 319-348.

FRANK, O. (1980). Sampling and inference in a population graph. *International Statistical Review*. 48, 33-41.

FRANK, O. (1997). Composition and structure of social networks. *Mathematiques, Informatique et Sciences humaines*. 35, 11-23.

FRANK, O., and SNIJDERS, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*. 10, 53-67.

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.

GELMAN, A., CARLIN, J., STERN, H. and RUBIN, D (1995). *Bayesian Data Analysis*. London: Chapman & Hall.

GILK, W.R, RICHARDSON, S. and SPIEGELHALTER (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

GOODMAN, L.A. (1961). Snowball sampling. *Annals of Mathematical Statistics*. 20, 572-579.

HALDANE, J.B.S. (1931). A note on inverse probability. *Proc Cambridge Philos. Soc*. 28, 55-61.

NEAIGUS, A., FRIDEMAN, S.R., GOLDSTEIN, M.F., ILDEFONSO, G., CURTIS, R. and JOSE, B. (1995). Using dyadic data for a network analysis of HIV infection and risk behaviors among injection drug users. In *Social Networks, Drug Abuse, and HIV Transmission*. (R.H. Needle, S.G. Genser and R.T. II Trotter, Eds.) NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse. 20-37.

NEAIGUS, A., FRIEDMAN, S.R., JOSE, B., GOLDSTEIN, M.F., CURTIS, R., ILDEFONSO, G. and DES JARLAIS, D.C. (1996). High-risk personal networks and syringe sharing as risk factors for HIV infection among new drug injectors. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*. 11, 499-509.

POTTERAT, J.J., WOODHOUSE, D.E., ROTHENBERG, R.B., MUTH, S.Q., DARROW, W.W., MUTH, J.B. and REYNOLDS, J.U. (1993). AIDS in Colorado Springs: Is there an epidemic? *AIDS*. 7, 1517-1521.

ROTHENBERG, R.B., WOODHOUSE, D.E., POTTERAT, J.J., MUTH, S.Q., DARROW, W.W. and KLOVDAHL, A.S. (1995). Social networks in disease transmission: The Colorado Springs study. In *Social Networks, Drug Abuse, and HIV Transmission*. (R.H. Needle, S.G. Genser and R.T. II Trotter, Eds.) NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse. 3-19.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*. 63, 581-592.

SNIJDERS, T.A.B. (1992). Estimation on the basis of snowball samples: how to weight. *Bulletin de Methodologie Sociogique*. 36, 59-70.

SNIJDERS, T.A.B., and NOWICKI, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*. 14, 75-100.

THOMPSON, S.K., and COLLINS, L.M. (2002). Adaptive sampling in research on risk-related behaviors. *Drug and Alcohol Dependence*. 68, S57-S67.

THOMPSON, S.K., and FRANK, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology*. 26, 87-98.

THOMPSON, S.K., and SEBER, G.A.F. (1996). *Adaptive Sampling*. New York: John Wiley & Sons, Inc.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees during 2002. An asterisk indicates that the person served more than once.

M.Z. Anis, *Indian Statistical Institute*
J. Bethel, *Westat, Inc.*
J.-F. Beaumont, *Statistics Canada*
D.R. Bellhouse, *University of Western Ontario*
M. Bellow, *NASS*
Y. Berger, *University of Southampton*
D. Binder, *Statistics Canada*
E. Blair, *University of Houston*
J. Breidt, *Iowa State University*
J.M. Brick, *Westat, Inc.*
R. Chambers, *University of Southampton*
J. Chen, *University of Waterloo*
M.J. Cho, *Bureau of Labor Statistics*
J. Choi, *National Center for Health Statistics*
J. Church, *Worsey House*
C. Clark, *U.S. Bureau of the Census*
P. Clarke, *Office for National Statistics*
R. Clark, *Australian Bureau of Statistics*
M.P. Cohen, *U.S. Bureau of Transportation Statistics*
F. Conrad, *University of Michigan*
M.P. Couper, *University of Michigan*
F.A. Cowell, *London School of Economics and Political Science*
J. Dalen, *Eurostat*
J. de Haan, *Statistics Netherlands*
P. Dick, *Statistics Canada*
A. Dorfman, *U.S. Bureau of Labour Statistics*
P. Duchesne, *Université de Montréal*
M.R. Elliott, *University of Pensylvania*
J. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
M. Ghosh, *University of Florida*
A. Gower, *Statistics Canada*
B. Graubard, *National Cancer Institute*
S. Hawala, *U.S. Census Bureau*
D. Haziza, *Statistics Canada*
D. Hedeker, *University of Illinois*
D. Hedlin, *University of Southampton*
D.F. Heitjan, *University of Pennsylvania*
M.A. Hidiroglou, *Statistics Canada*
S. Hinkins, *National Opinion Research Centre*
T. Holt, *University of Southampton*
B. Hulliger, *Swiss Federal Statistical Office*
D. Judkins, *Westat, Inc.*
G. Kalton, *Westat Inc.*
A. Kennickell, *Federal Reserve System*
J.-K. Kim, *Hankuk University of Foreign Studies*
* P. Kott, *National Agricultural Statistics Service*
V. Kuusela, *Statistics Finland*
P.A. Lachenbruch, *U.S. Food and Drug Administration*
P. Lahiri, *JPSM, University of Maryland*
N. Laniel, *Statistics Canada*
* P. Lavallée, *Statistics Canada*
* H. Lee, *Westat, Inc.*
R. Lehtonen, *University of Jyvaskyla*
J. Lent, *U.S. Bureau of Transportation Statistics*
J. Lepkowski, *University of Michigan*
R.J.A. Little, *University of Michigan*
S. Linacre, *Australian Bureau of Statistics*
S. Lohr, *Arizona State University*
T. Maiti, *Iowa State*
H. Mantel, *Statistics Canada*
S. Matthews, *Statistics Canada*
X.-L. Meng, *Harvard University*
S.M. Miller, *U.S. Bureau of Labour Statistics*
S.R. Mohen, *ISI*
J.M. Montaquila, *Westat, Inc.*
G. Nathan, *The Hebrew University of Jerusalem*
D. Norris, *Statistics Canada*
* J. Opsomer, *Iowa State University*
D. Pfeffermann, *The Hebrew University of Jerusalem*
N.G.N. Prasad, *University of Alberta*
* T.E. Raghunathan, *University of Michigan*
J.N.K. Rao, *Carleton University*
P.S.R.S. Rao, *University Rochester*
T.J. Rao, *Indian Statistical Institute*
J. Reiter, *Duke University*
L.-P. Rivest, *Université Laval*
P. Saavedra, *ORC Macro*
* S. Sae-Ung, *U.S. Census Bureau*
C.-E. Särndal, *University of Montreal*
J. Schafer, *Pennsylvania State University*
N. Schenker, *National Center for Health Statistics*
F.J. Scheuren, *National Opinion Research Center*
M.D. Sinclair, *Mathematica Policy research*
R. Sitter, *Simon Fraser University*
* C. Skinner, *University of Southampton*
K.P. Srinath, *ABT Associates*
E. Stasny, *Ohio State University*
J.-L. Tambay, *Statistics Canada*
Y. Thibaudeau, *U.S. Bureau of the Census*
Y. Tillé, *Université de Neuchâtel*
R. Valliant, *Westat, Inc.*
J. van der Brakel, *Statistics Netherlands*
V. Vehovar, *University of Ljubljana*
J. Waksberg, *Westat, Inc.*
W.E. Winkler, *U.S. Bureau of the Census*

K.M. Wolter, *Iowa State University*
C. Wu, *University of Waterloo*
W. Yung, *Statistics Canada*
A. Zaslavsky, *Harvard University*
H. Zheng, *Harvard Medical School*
K. Zieschang, *International Monetary Fund*

Acknowledgements are also due to those who assisted during the production of the 2003 issues: H. Laplante, F. Pilon-Renaud and R. Guido (Dissemination Division) and L. Perreault (Official Languages and Translation Division). Finally we wish to acknowledge C. Cousineau, C. Ethier, and D. Lemire of Household Survey Methods Division, for their support with coordination, typing and copy editing.

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 19, No. 2, 2003

Contents
Volume 19, No. 3, 2003

All inquires about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

CONTENTS | TABLE DES MATIÈRES

Volume 31, No. 2, June/juin 2003, 115-238

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. Accepted articles must be submitted in machine-readable form, preferably in Word or WordPerfect. 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.

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, p. 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.