Statistics Canada Statistique Canada

SURVEY METHODOLOGY

A JOURNAL
OF
STATISTICS CANADA

VOLUME 13, NUMBER 2
DECEMBER 1987

Canada

SURVEY METHODOLOGY

A JOURNAL OF STATISTICS CANADA

DECEMBER 1987

Published under the authority of
the Minister of Supply and
Services Canada

May 1988

Price: Canada, $20.00 a year
Other Countries, $23.00 a year

Payment to be made in Canadian funds or equivalent

Catalogue 12-001, Vol. 13, No. 2

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics.

EDITORIAL POLICY

The Survey Methodology Journal 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

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Two nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is $20.00 per year in Canada, $23.00 per year for other countries (payment to be made in Canadian funds or equivalent). Subscription order should be sent to: Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price of US $10.00 ($14.00 Can.) is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada. Please subscribe through your organization.

SURVEY METHODOLOGY

A Journal of Statistics Canada
Volume 13, Number 2, December 1987

CONTENTS

Survey Methodology, December 1987
Vol. 13, No. 2, pp. 125-126
Statistics Canada

In This Issue

Two new features appear for the first time in this issue of Survey Methodology. **"In This Issue"** summarizes papers appearing in the Journal and will appear regularly. The other new feature, a **"Short Communications"** section, will appear in the Journal from time to time.

This issue contains nine papers, four dealing with **estimation and weighting methods,** including two on family estimation. Fritz Scheuren's initiative and editorial assistance were instrumental in putting this special section together.

The first three papers in the special section deal (at least in part) with least-squares methods for weighting survey data. There is a certain historical irony in this. In their 1940 paper, Deming and Stephan introduced iterative proportional fitting as a quick practical way for approximating the estimates obtained by minimizing a squared function of the cells of a contingency table, subject to restrictions on the margins. The use of this technique has become fairly generalized in weighting survey data, where it is known as "raking ratio estimation".

In "An Alternative Method of Controlling Current Population Survey Estimates to Population Counts", Copeland, Peitzmeier and Hoy compare a raking ratio estimator to a generalized least-squares estimator under the same marginal restrictions. The comparison is carried out for estimates of individual characteristics obtained from the Current Population Survey, a household survey conducted by the United States Bureau of the Census. They note that the estimates produced by the two methods are very similar.

Most current methods of weighting data from household surveys produce weights that differ from person to person within the same household. A single weight per household, in addition to its conceptual appeal, would eliminate the recurrent and often awkward discrepancies between person-based and family-based estimates. Alexander, in "A Class of Methods for Using Person Controls in Household Weighting", considers a class of "constrained minimum distance" methods (including GLS) which actually yield a single weight per household yet respect person-level marginal totals. The properties of these methods in the presence of undercoverage are then studied through some simple coverage models.

Lemaître and Dufour, in "An Integrated Method for Weighting Persons and Families", propose a regression estimator that also yields a single weight per household and is equivalent to the GLS estimator under certain general conditions. Using Canadian Labour Force Survey data, they obtain large efficiency gains for estimates of families, and marginal gains for estimates of persons, relative to current methods.

In the last paper in this section, "Modified Raking Ratio Estimation", Oh and Scheuren describe an estimation procedure similar to the usual raking ratio. Their method can be used when population totals are available not only for the margins, but also for interior cells in a multi-way table. It combines conventional ratio estimation for cells with large sample sizes and raking ratio estimation for cells with sample sizes that are small (or zero). In an application involving sampling of corporate income tax returns, the Oh-Scheuren approach produced more efficient estimates relative to conventional ratio estimation. The authors stress that, before their method is offered for wide use, further work is needed including, among other things, comparison with conventional collapsing schemes.

The other four papers in this issue consider the development and application of methods and procedures with regard to probabilities of response in a survey context, rounding criteria for protection of confidentiality, data collection and analysis for retrospective type surveys, and variance estimation for the Canadian Labour Force Survey.

Every survey has some nonresponse problems. These are usually handled by imputation or adjustment procedures based on the assumption that nonresponse occurs at random within imputation or adjustment classes. The resulting estimates are generally biased whenever this assumption is not satisfied. Various methods of estimating response probabilities involving models have been proposed, notably by Cassel, Särndal and Wretman (CSW), but these methods are not effective when the assumed model is inadequate. In "Nonparametric Methods for Estimating Individual Response Probabilities", Giommi describes nonparametric procedures for estimating response probabilities using auxiliary information, providing an alternative to the CSW estimator that is robust against both population and response model breakdown. The resulting estimators perform well in Monte Carlo simulation studies.

Random rounding is used to ensure the confidentiality of information about individuals in statistical aggregates. In the context of the 1971 Canadian Census, Nargundkar and Saveland developed a rounding process that is unbiased in the sense that the expected value of the rounded data is the same as that of the unrounded data. Fellegi (SMJ, 1975) introduced controlled random rounding, a procedure that, in addition to being unbiased, also preserves additivity. Several other papers have since appeared, including the very recent work of Cox (JASA, 1987), generalizing and extending the applications to other fields. In "Estimates Based on Randomly Rounded Data", Withers develops an expression for the variance of unbiased estimates of cell probabilities and presents a comparison of efficiencies involving the rounding processes used in Australia, the United Kingdom, New Zealand and Canada. He also extends his results to any smooth function of the cell probabilities for applications in different areas of statistics.

In "Variance Estimation for the Canadian Labour Force Survey", Choudhry and Lee describe studies conducted to select a variance estimator for raking ratio estimates from the Canadian Labour Force Survey. Their paper reports on a comparison of three variance estimators for the random group sampling design: Keyfitz, Rao-Hartley-Cochran and Rao. In spite of its slight inferiority to the other two methods in terms of bias and stability, the Keyfitz method is suggested for actual use because of its operational simplicity.

In "The "AGEVEN" Record: A Tool for the Collection of Retrospective Data", Antoine, Bry and Diouf describe techniques used to collect data on natality and mortality of women in Pikine, a suburb of Dakar, Senegal. The retrospective procedure employed involved placing observed events (mainly births and deaths) in their socio-economic context and, according to the authors, made it possible to "better assess the relationship between urban insertion and changes in demographic behaviour". Analysis of data from the survey clearly indicates that child mortality rates are higher for children born in rural villages than for those born in Pikine.

It is well known that the Hansen-Hurwitz strategy is inferior to the Horvitz-Thompson strategy associated with a number of IPPS (inclusion probability proportional to size) sampling procedures. In the final piece in this issue, in the "Short Communications" section, Prabhu-Ajgaonkar presents proofs of these results that are much simpler than those already available in the literature.

The Editor

Survey Methodology, December 1987
Vol. 13, No. 2, pp. 127–134
Statistics Canada

Nonparametric Methods for Estimating Individual Response Probabilities

ANDREA GIOMMI[1]

ABSTRACT

This paper deals with the nonresponse problem in the estimation of the mean of a finite population, following an approach closely related to that of Cassel, Särndal and Wretman (1983). Two very simple methods are proposed for estimating the individual response probabilities; these are then used, in connection with a superpopulation model, to construct estimators for the population mean. A first evaluation of the properties of the proposed methods is given by a Monte Carlo experiment. The results shed some light on their effectiveness.

KEY WORDS: Nonresponse; Individual response probability; Nonparametric methods.

1. INTRODUCTION

Dealing with the estimation of finite population mean (or total, etc.) in the presence of nonresponse, Cassel, Särndal and Wretman (1983) introduced a very general estimation method based on the fundamental concept of individual response probability (IRP). The authors proposed estimators which are in part determined by a superpopulation model and in part by a response model, i.e., a model formalizing the response mechanism and by which IRP can be estimated from sample data. The estimation of IRP is the crucial point of their theory. In fact, if the superpopulation model is not correctly chosen, as is often the case, only a correct choice of the response model may guard the estimators from design bias. By a Monte Carlo experiment, Giommi (1985a) showed that a response model supplying a "good approximation" of the "true" response model can restore virtual unbiasedness; but little is known about the extent of a good approximation and in any case the choice of a response model may prove cumbersome besides being arbitrary. A natural way of avoiding these difficulties is to estimate the IRP by nonparametric procedures. In the present paper we propose two very simple methods to estimate IRP when available auxiliary information (which is assumed to be related to the response behaviour) is represented by a single continuous variable. The methods which make use of some tools of the kernel estimation theory may be viewed as an extension of the popular correction technique for nonresponse consisting in reweighting units by adjustment cells.

In this paper some empirical evaluations of these methods are described and the results regarding the bias and efficiency of the related estimators are presented.

2. ESTIMATION OF THE INDIVIDUAL RESPONSE PROBABILITIES

Let us consider a population of N units labelled k $(k=1, 2, \ldots, N)$, and let Y be a variable under study, of which we want to estimate the mean $\bar{Y}=\Sigma_k\, y_k/N$ from a sample s of n units, the selection being based on a given design $p(s)$. For the estimation, auxiliary information is available, represented by known values x_k $(k=1, \ldots, N)$, of a scalar continuous

[1] Andrea Giommi, Department of Statistics, University of Florence, Via Curtatone, 1, 50123 Florence, Italy.

variable X (the extension of the procedures proposed for the multidimensional case is, in principle, straightforward).

In the sample, Y is observable only in a subset r of n_r respondents and not on the $n - n_r$ nonrespondents. After the selection of the sample, the available information can be represented as follows:

$$(k,\ I_k,\ I_k y_k,\ x_k) \qquad k \in s;\ N,\ n,$$

where I_k is an indicator random variable such that $E(I_k) = q_k$ and q_k is the IRP.

To estimate q_k, a parametric model is generally assumed (Cassel *et al.* 1983) such that:

$$q_k = q(\Theta,\ x_k),$$

where Θ is an unknown parameter (or vector of parameters) and $q(\cdot,\cdot)$ is a functional form to be specified. Estimated q_k are then obtained replacing in the above parametric model estimated values $\hat{\Theta}$ of Θ.

In this paper the estimates of q_k $(k \in r)$ are obtained by avoiding any parametric specification of the function $q(\cdot,\cdot)$; nevertheless, maintaining the hypothesis that the IRPs depend on the values x_k. Two procedures (methods (1) and (2)) are proposed.

In the first, q_k $(k \in r)$ is estimated as the response rate (i.e. the proportion of respondents) in a group of units centered on the unit k, corresponding to an appropriate interval of x-values centered at x_k. Assuming that $2h_k$ is the length of such an interval, q_k is estimated by the following ratio:

$$\hat{q}_k = \sum_{j \in r} D(x_k - x_j) \Big/ \sum_{j \in s} D(x_k - x_j), \tag{1}$$

where

$$D(x_k - x_j) = \begin{cases} 1 \text{ if } |x_k - x_j| \leq h_k \\ 0 \text{ otherwise.} \end{cases}$$

It is evident that the estimate $\hat{q}_k$ depends on h_k or h if we adopt – as in this paper – a constant interval; the numerical specification of h is a main problem in applications.

In the second procedure, all the sample units, rather than a group, contribute to the estimation of q_k. By this method the possible limitation due to the classification of responding units in groups is removed. In other words, one might consider overly restrictive the fact that in the estimation of q_k some units contribute with weight 1 and some others with weight 0. With method (2), the estimate is given by:

$$\hat{q}_k = \sum_{j \in r} D^*(x_k - x_j) \Big/ \sum_{j \in s} D^*(x_k - x_j) \tag{2}$$

where D^* has to be specified. In this case, each value x_j contributes towards the estimate $\hat{q}_k$ through D^*, an amount inversely related to the difference $|x_k - x_j|$.

In (2), the problem is twofold: i) to specify the functional form D^* and ii) to define the values of its parameters. In this paper we adopt a function D^* of the normal type:

$$D^*(z) = (h^2 2\pi)^{-1/2} \exp(-z^2/2h^2); \qquad z = x_k - x_j, \tag{3}$$

in which the standard deviation, indicated by h, plays a role analogous to that of the parameter h in the expression (1). In both (1) and (2), when h increases, $\hat{q}_k$ approaches to the constant value n_r/n. In (1), it reaches n_r/n when h covers the whole range of the x-values.

An empirical study was designed to evaluate the properties of the proposed procedures, using a very wide range of h values. In the present paper we have limited ourselves to reporting results for only three (constant) values of h, equal to 1/10, 3/10 and 5/10 of the range of the x-sample values. Finally, we must observe that both expressions (1) and (2), apart from a normalizing factor, show themselves as the ratio of two probability density kernel estimators (in the approach of Rosenblatt (1956)) over different sets of x-values. Therefore, as suggested by Giommi (1985b), the value of h may be selected considering proposals put forward in that theory.

3. SUPERPOPULATION MODEL AND ESTIMATORS

For the choice of the estimator of $\bar{Y}$, we assume a superpopulation model Φ in which the population values y_k, $k=1, 2, \ldots, N$, are considered to be a random sample such that:

$$\begin{aligned} E_\Phi(Y_k) &= \mu_k = \beta x_k, \\ \mathrm{Var}_\Phi(Y_k) &= \sigma_k^2 = \sigma^2 x_k, \end{aligned} \tag{4}$$

where β and Φ unknown and x_k is the known value of the auxiliary variable X. It is apparent that the superpopulation model employed here is mainly applicable to quantitative rather than qualitative variables; other models should be employed in such cases. We further limit ourselves to the consideration of simple random samples. Providing the variance of Y may be specified as in (4), Cassel *et al.* (1983) have shown that the following estimator:

$$T = \bar{X}\left(\sum_r y_k/q_k\right) / \left(\sum_r x_k/q_k\right),$$

where Σ_r indicates the sum over the set r and $\bar{X}=\Sigma_k^N x_k/N$, is approximately unbiased, thanks to the q_k correction, even if the first equation in (4) fails to specify the true relationship between X and Y. This may happen, for example, when the "true" model has an intercept or has two regression coefficients (see (5) below), etc.

Unfortunately, in practice the estimator T cannot be used since q_k is unknown. The problem is, therefore, to evaluate its properties when q_k is replaced by its estimate derived either from method (1) or (2).

We shall examine such estimators, for the three chosen values of h. We denote the estimators by TD_i and TD_i^* where $i=1, 3, 5$ as in Table 1.

Table 1

Definition of Estimators

h	Estimators Method (1)	Method (2)
0.1	TD_1	TD_1^*
0.3	TD_3	TD_3^*
0.5	TD_5	TD_5^*

In addition, also the following estimators are considered in the Monte Carlo study:

$$TC = \bar{X}\left(\sum_s y_k / \sum_s x_k\right) \quad \text{and} \quad TI = \bar{X}\left(\sum_r y_k / \sum_r x_k\right).$$

TC is the full sample estimator, that is, the ratio estimator under the hypothesis of complete response and *TI* is the same estimator based on the set of respondents, on which no q_k-correction is made for nonresponse. Note that *TI* is also an estimator derived from a well known procedure of imputation (by regression) of missing values (Cassel *et al.* 1983) and equals *TD* when *h* covers the whole range of the *x*-values. *TI* is approximately unbiased only if (4) is true. The bias, as we shall see, depends on the divergence between the conditions in (4) and those of the population under study. As in the experiment of the next section model (4) will be a "false" model (that is, the study populations are specified by models different from (4)), the simulation also contributes to the knowledge of this very simple and widely used imputation method.

4. THE MONTE CARLO EXPERIMENT

In the Monte Carlo experiment two populations, POP1 and POP2, were generated following the same procedure as that of Särndal and Hui (1981). POP1 and POP2 are both composed of two strata, say *S*1 and *S*2, 500 units each and satisfy the following equations:

$$\begin{aligned} E_\Phi(Y_k) &= \beta_1 x_{k1} + \beta_2 x_{k2}, \\ \mathrm{Var}_\Phi(Y_k) &= \sigma_1^2 x_{k1} + \sigma_2^2 x_{k2}, \end{aligned} \tag{5}$$

where $x_{k1} = x_k \partial_k$ and $x_{k2} = x_k(1 - \partial_k)$, with $\partial_k = 1$ if $k \in S1$ and $\partial_k = 0$ if $k \in S2$. The difference between (4) and (5) simulates one of the many errors which one can incur in specifying the superpopulation model. The numerical characteristics of POP1 and POP2 are shown in Table 2.

The simulation procedure can briefly be described in the following steps:

1) A simple random sample *s* of *n* ($n = 50$, 100) units is selected from each population.

Table 2

Characteristics of Simulated Populations

Population and strata		POP1 Mean	POP1 SD	POP1 CV	POP1 SK	POP2 Mean	POP2 SD	POP2 CV	POP2 SK
Stratum 1	x	19.305	12.71	.66	1.30	20.037	14.50	.72	2.25
	y	7.612	5.38	.71	1.62	1.961	2.21	1.13	3.03
Stratum 2	x	50.325	21.32	.42	.77	49.775	23.28	.47	1.21
	y	30.325	13.38	.44	.72	44.862	21.31	.47	1.04
Total	x	34.815	23.42	.67	.90	34.906	24.44	.70	1.32
	y	18.969	15.26	.80	1.06	23.411	26.25	1.12	1.15

SD = population standard deviation; SK = skewness (3rd moment / (2nd moment)$^{3/2}$); CV = coefficient of variation.

2) The full sample values are recorded and nonresponse is then generated by each of the two following parametric models:

$$\text{Model A: } q_k = \exp(-\Theta x_k),$$

$$\text{Model B: } q_k = \Theta_1^{\partial_k}\,\Theta_2^{1-\partial_k}; \qquad \partial_k = 1\ (0) \text{ if } k \in S1\ (S2),$$

where the parameters Θ, Θ_1, Θ_2 are chosen in such a way that the average response rate $\bar{q}$ over the whole population is alternatively 0.6 and 0.7. In practice, sets of respondents are obtained by performing a Bernoulli trial for each unit $k \in s$, with probability q_k for "success" (response) and $1-q_k$ for "failure" (nonresponse).

3) The IRP is estimated by method (1) and (2) and, for each sample, the values of *TC*, *TI*, *TD*, *TD** are calculated.

4) Steps 1 to 3 are repeated 1000 times and at the end we calculate: bias, variance (VAR) and mean squared error (MSE) of the estimators for each sample size (50, 100), response model (A, B), average response rate (0.6, 0.7) and population (POP1, POP2).

The experimental results are reported in Tables 3 and 4.

5. RESULTS OF THE MONTE CARLO EXPERIMENT

Some interesting elements emerge from the examination of Tables 3 and 4.

1. As expected, *TC* is approximately unbiased in all of the experimental trials.
2. In this experiment the bias of *TI* is always larger than that of *TD* and *TD**. Therefore, at least in the situations of the experiment, the adjusted estimator is to be preferred over the non-adjusted one, which corresponds to a procedure of imputation by regression.
3. For the same h value, the bias of *TD* is always smaller than that of *TD**. The differences are negligible for $h=.1$. As h increases, *TD** tends toward *TI* faster than *TD*; for $h=.5$ the differences between *TD** and *TI* are irrelevant for practical purposes.
4. The reduction of the bias we are able to obtain using *TD* instead of *TI* is always significant, varying from 55% to 82% for model A, from 67% to 92% for model B. *TD** also experiences a notable reduction of the bias: from 51% to 68% for model A, from 61% to 84% for model B.
5. *TD* and *TD** are equivalent in terms of MSE for $h=.1$, even though TD_1^* is slightly more stable (i.e. has a lower variance). For $h=.3$ and $h=.5$, the lesser stability of *TD* in comparison with *TD** is generally compensated by the smaller bias, more than enough to make *TD* preferable to *TD** in terms of MSE.
6. The estimators adjusted by the estimated IRP are not very stable but, in terms of MSE, must be preferred to *TI*.
7. As expected, the bias is directly related to the increase of the nonresponse rate and to the divergence between the true superpopulation model and the one assumed (i.e. the false model on which the estimators are based). No relevant differences are revealed due to the response models considered in this paper (see Giommi (1984) for the effect of alternative models).
8. The increase of the sample size seems to reduce the bias slightly for all the estimators considered. TD_1 and TD_1^* are exceptions: in this case, the reduction of the bias cannot be attributed to experimental fluctuations but to the actual improvement of the estimate q_k when n increases.

In the end, we may conclude that, in situations similar to the ones considered in this paper, the two methods suggested can be used, with a certain preference for method (1) given its simpler application. The problem of determination of the best value for h (or h_k, in the general case) remains to be examined. We found that, within certain limits, small values for h reduce the bias but also reduce the stability of the adjusted estimator. We have found that, for our experimental examination, the optimum value of h is in the neighbourhood of 0.1. Results obtained from the same experiment but not reported in this paper indicate that a further reduction of h tends to increase the bias. This is to be expected since making h get closer to 0 results in a collection of estimates $\hat{q}_k$ $(k=1, \ldots, n)$, equal to 1 and 0 respectively for the respondents and nonrespondents.

6. ACKNOWLEDGEMENT

I am indebted to Prof. Luigi Biggeri for his support throughout the course of the study. I also wish to thank the referees for their helpful comments on the first draft of this paper.

Table 3

Performance of Different Estimators under Response Model A

Estimators		TC	TI	TD_1	TD_3	TD_5	TD_1^*	TD_3^*	TD_5^*
				Average response rate $\bar{q}=.60$					
				POP1					
$n=50$	BIAS	.015	.861	.349	.420	.669	.380	.620	.765
	VAR	.405	.973	1.115	1.036	1.007	1.041	.995	.989
	MSE	.405	1.714	1.237	1.212	1.455	1.185	1.379	1.574
$n=100$	BIAS	.007	.805	.164	.323	.610	.227	.544	.686
	VAR	.186	.416	.443	.429	.412	.415	.404	.402
	MSE	.186	1.064	.470	.533	.784	.467	.700	.873
				POP2					
$n=50$	BIAS	.090	3.125	1.433	1.682	2.544	1.544	2.378	2.887
	VAR	3.952	8.744	9.821	9.823	9.743	9.390	9.233	9.118
	MSE	3.960	18.510	11.874	12.652	16.215	11.774	14.888	17.453
$n=100$	BIAS	.056	2.959	.749	1.387	2.337	1.004	2.104	2.566
	VAR	1.710	4.144	4.515	5.122	4.819	4.238	4.632	4.518
	MSE	1.713	12.900	5.076	7.046	10.281	5.246	9.059	11.102
				Average response rate $\bar{q}=.70$					
				POP1					
$n=50$	BIAS	.015	.581	.226	.271	.418	.249	.415	.439
	VAR	.405	.765	.794	.750	.738	.754	.752	.753
	MSE	.405	1.103	.845	.823	.913	.816	.924	.946
$n=100$	BIAS	.007	.531	.099	.205	.396	.143	.357	.457
	VAR	.186	.328	.323	.307	.327	.313	.327	.336
	MSE	.186	.610	.333	.349	.484	.333	.454	.545
				POP2					
$n=50$	BIAS	.090	2.130	.813	.939	1.542	.887	1.453	1.822
	VAR	3.952	6.996	7.122	6.827	6.991	6.708	6.753	6.871
	MSE	3.960	11.533	7.783	7.709	9.396	7.495	8.864	10.191
$n=100$	BIAS	.056	1.966	.473	.953	1.541	.658	1.406	1.732
	VAR	1.710	3.071	3.005	3.062	3.027	2.926	3.008	3.040
	MSE	1.713	6.937	3.229	3.970	5.402	3.359	4.985	6.040

Table 4

Performance of Different Estimators under Response Model B

Estimators		TC	TI	TD_1	TD_3	TD_5	TD_1^*	TD_3^*	TD_5^*
				Average response rate $\bar{q}=.60$					
				POP1					
$n=50$	BIAS	.015	1.086	.290	.383	.716	.323	.688	.992
	VAR	.405	.966	1.208	1.011	.937	1.050	.907	.928
	MSE	.405	2.145	1.29	1.158	1.450	1.154	1.380	1.912
$n=100$	BIAS	.007	1.079	.120	.349	.732	.196	.668	.902
	VAR	.186	.422	.513	.429	.420	.447	.401	.403
	MSE	.186	1.586	.527	.551	.956	.485	.847	1.217
				POP2					
$n=50$	BIAS	.090	4.046	1.362	1.757	2.826	1.562	2.749	3.562
	VAR	3.952	10.285	12.519	12.089	12.010	11.605	11.046	10.994
	MSE	3.960	26.655	14.374	15.176	19.996	14.045	18.603	23.682
$n=100$	BIAS	.056	3.897	.454	1.531	2.707	.853	2.521	3.284
	VAR	1.710	4.151	5.432	5.121	5.103	4.798	4.541	4.381
	MSE	1.713	19.338	5.638	7.465	12.431	5.525	10.896	15.166
				Average response rate $\bar{q}=.70$					
				POP1					
$n=50$	BIAS	.015	.584	.179	.221	.409	.196	.376	.499
	VAR	.405	.751	.826	.425	.716	.769	.723	.743
	MSE	.405	1.092	.858	.474	.883	.807	.864	.992
$n=100$	BIAS	.007	.536	.046	.173	.365	.087	.317	.436
	VAR	.186	.307.	318	.295	.295	.299	.295	.302
	MSE	.186	.594	.320	.325	.428	.307	.395	.492
				POP2					
$n=50$	BIAS	.090	2.057	.682	.891	1.477	.804	1.392	1.822
	VAR	3.952	6.199	6.788	6.165	6.232	6.340	6.093	6.270
	MSE	3.960	10.430	7.253	6.959	8.414	6.986	8.031	9.590
$n=100$	BIAS	.056	1.918	.157	.755	1.311	.374	1.175	1.562
	VAR	1.710	2.826	2.897	2.884	2.867	2.796	2.836	2.923
	MSE	1.713	6.506	2.922	3.454	4.586	2.936	4.217	5.363

REFERENCES

CASSEL, C.M., SÄRNDAL, C.E., and WRETMAN, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. In *Incomplete Data in Sample Surveys* (eds. W.G. Madow and I. Olkin), Vol. 3, New York: Academic Press, 143-160.

GIOMMI, A. (1984). On a simple method for estimating individual response probabilities in sampling from finite populations, *Metron*, 42, 185-200.

GIOMMI, A. (1985a). On estimation in nonresponse situations. *Statistica*, 1, 57-63.

GIOMMI, A. (1985b). On the estimation of the individual response probabilities. *Proceedings of the 45th Session of the International Statistical Institute*, Vol. 2 (Contributed Papers), 577-578.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates for the density function. *Annals of Mathematical Statistics*, 27, 832-837.

SÄRNDAL, C.E., and HUI, T.K., (1981). Estimation for nonresponse situations: to what extent must we rely on models? In *Current Topics in Survey Sampling*, (eds. D. Krewski, R. Platek and J.N.K. Rao), New York: Academic Press, 227-246.

Survey Methodology, December 1987
Vol. 13, No. 2, pp. 135–146
Statistics Canada

Estimates Based on Randomly Rounded Data

C.S. WITHERS[1]

ABSTRACT

Methods are given to estimate functions of the cell probabilities associated with a table of multinomial data that has been randomly rounded to multiples of a given number, say l. We show that: (i) random rounding causes only second order effects on bias and variance; (ii) the loss of efficiency in using the natural estimates of cell probability is negligible provided that the cell entry is large compared with $(l^2 - 1)/(6R)$ where R is the number of cells in the table; and (iii) estimates of apparently exponentially small bias are available for moments of these natural estimates and for polynomials in the cell probabilities.

KEY WORDS: Random rounding; Bias reduction; Efficiency.

1. INTRODUCTION AND SUMMARY

This paper gives methods of estimating a function of the cell probabilities associated with a table of multinomial data that has been randomly rounded. Random rounding is a widely used method for preserving confidentiality in situations where an entry of 1 in a table might identify an individual and so break a confidentiality requirement. Instead of tabling the value of a table entry, say N, one rounds N to the nearest multiple of a given number l above N with probability (w.p.) α or below N w.p. $1 - \alpha$, where α is chosen so that the rounded value M satisfies

$$E(M \mid N) = N.$$

That is, if for some integer j, $jl \leq N < (j + 1)l$, then

$$M = \begin{cases} jl \text{ w.p. } 1 - \alpha \\ (j + 1)l \text{ w.p. } \alpha \end{cases} \tag{1.1}$$

where $\alpha = r/l$ and $r = N - jl$.

The rounding base l used by the Department of Statistics in New Zealand is $l = 3$, while Statistics Canada reportedly uses $l = 5$. See Penny and Ryan (1986).

Random rounding should not be confused with grouping or non-random rounding of sample values to the nearest integral multiple of l (associated with Sheppard's corrections for moments). Nor should it be confused with intentional contamination, another method of preserving confidentiality where one simply adds to N an independent random variable with mean 0. (The main disadvantage of intentional contamination is the possibility of a negative cell entry). For some references on these methods see Gastwirth *et al.* (1978) and Kendall and Stuart (1977). Some references on random rounding for multivariate data and grouped data are also given in Gastwirth *et al.* (1978).

[1] C.S. Withers, Applied Mathematics Division, Department of Scientific and Industrial Research, Box 1335, Wellington, New Zealand.

In this paper we confine our attention to problems of estimating a function of the cell probabilities associated with a table of R values that have been randomly rounded. For convenience we label these cell probabilities as $p_1, \ldots, p_R$ rather than $\{p_{ij}, 1 \le i \le I, 1 \le j \le J\}$, as is more usual for an $I \times J$ table.

Thus, $1 = \Sigma_1^R p_i$ and $n = \Sigma_1^R N_i$ is the sum of the entries in the table. Let $\{M_i\}$ be the rounded values of $\{N_i\}$. Given n, we assume $\{N_i\}$ has the multinomial distribution with parameters n and $\{p_i\}$. This is true with $p_i = m_i/\Sigma_j m_j$ if, unconditionally, $\{N_i\}$ are independent Poisson variables with means $\{m_i\}$.

Two unbiased estimates of p_1 are

$$p_1^* = N_1/n \text{ and } \hat{p}_1 = M_1/n. \tag{1.2}$$

The first is not a true estimate since N_1 is not made available. The second is the natural estimate. (We assume n is reported. If it is not, there is negligible difference in replacing n by $\Sigma_1^R M_i$.) However, other unbiased estimates exist, namely the "complementary estimate"

$$\tilde{p}_1 = -\sum_{j \ne 1} M_j/n, \tag{1.3}$$

and hence

$$p_1(\lambda) = (1 - \lambda)\hat{p}_1 + \lambda\tilde{p}_1 \text{ for any given } \lambda. \tag{1.4}$$

This raises the issue of what is the best λ to use, and what loss of efficiency there is in sticking to the natural estimate — that is, using $\lambda = 0$. An answer requires the variances of these estimators. These are given by

Theorem 1.1.

$$\operatorname{var}(\hat{p}_1) = (p_1 - p_1^2)\, n^{-1} + \{(l^2 - 1)/6 + \Delta_n(p_1)\}n^{-2} = v_n(p_1), \tag{1.5}$$

where

$$\Delta_n(p_1) = \sum_{i=0}^{l-1} i(l - i)\,\{P(N_1 \bmod l = i) - l^{-1}\}. \tag{1.6}$$

Also,

$$\operatorname{var}(\tilde{p}_1) = (p_1 - p_1^2)n^{-1} + \{(R - 1)(l^2 - 1)/6 + \sum_{j \ne 1} \Delta_n(p_j)\}n^{-2}, \tag{1.7}$$

and

$$\operatorname{var}(p_1(\lambda)) = (p_1 - p_1^2)n^{-1} + \{\alpha(\lambda)(l^2 - 1)/6 + \nabla_n(p)\}n^{-2}, \tag{1.8}$$

where

$$\alpha(\lambda) = (1 - \lambda)^2 + (R - 1)\lambda^2 \tag{1.9}$$

and

$$\nabla_n(p) = (1 - \lambda)^2\, \Delta_n(p_1) + \lambda^2\, \Sigma_{i \ne 1}\, \Delta_n(p_i). \tag{1.10}$$

Proofs of the theorems in this paper are given in Section 2.

In Appendix A we give evidence that for $0 < p_1 < 1$, $P(N_1 \bmod l = i) - l^{-1} \to 0$ exponentially fast as $n \to \infty$, so that $\Delta_n(p_1) \to 0$ exponentially fast as $n \to \infty$, and hence $\nabla_n(p)$ also, provided $p_i \neq 0$ for all i.

Since $\alpha(\lambda)$ is minimised by $\lambda_R = R^{-1}$ and $\alpha(\lambda_R) = 1 - R^{-1}$ so, asymptotically, is $\mathrm{var}(p_1(\lambda))$. Hence the loss of efficiency in using the natural estimate $\hat{p}_1$ rather than the asymptotically optimal unbiased estimate $p_1(\lambda_R)$ when R is large, is

$$\{\mathrm{var}(\hat{p}_1) - \mathrm{var}(p_1(\lambda_R))\}/\mathrm{var}(p_1(\lambda_R)) \approx (l^2 - 1)/\{6Rn(p_1 - p_1^2)\} \qquad (1.11)$$

which is negligible provided $M_1(1 - M_1/n) \approx n(p_1 - p_1^2)$ is large compared with $(l^2 - 1)/\{6R\}$.

Generally $M_1(1 - M_1/n)$ can be approximated by M_1. This then gives a convenient rule of thumb as to when the natural estimates are efficient. (If one or more $\{p_i\}$ are zero, since $p_i = 0$ implies $N_i = M_i = 0$, $\Sigma_{i \neq 1}$ must be interpreted as excluding cells for which $p_i = 0$, and R as the number of cells in the table for which $p_i \neq 0$.)

Using (1.5) we can now make a brief comparison with the method of contamination. The Australian and U.K. statistics departments reportedly round by adding to each cell entry 1 w.p. 1/4, 0 w.p. 1/2 and -1 w.p. 1/4, so that

$$\mathrm{var}(\hat{p}_1) = (p_1 - p_1^2)n^{-1} + 1/2n^{-2}.$$

The factor 1/2 improves on 4/3 for the New Zealand system ($l = 3$) and 4 for the Canadian system ($l = 5$). The cost is less protection (a maximum change of 1 as opposed to 2 for the New Zealand system and 4 for the Canadian system), and a possibly negative cell entry if the procedure is applied to cells with zero entries.

Theorem 1.1 shows that random rounding has only a second order effect on the efficiency of estimating p_1 — the variance is only increased by a term of magnitude n^{-2}. The next result shows that this very important result is also true for estimating any smooth function of $\{p_i\}$. Set $r = R - 1$, $\mathbf{p} = (p_1, \ldots, p_r)$, $\mathbf{N} = (N_1, \ldots, N_r)$, $\mathbf{M} = (M_1, \ldots, M_r)$, $\mathbf{p}^* = \mathbf{N}/n$ and $\hat{\mathbf{p}} = \mathbf{M}/n$. Thus we have $\mathrm{cov}(\mathbf{p}^*) = V/n$ where $V = \mathrm{diag}(\mathbf{p} - \mathbf{p}\mathbf{p}')$. Suppose now we wish to estimate $f(\mathbf{p})$, a function with continuous second derivatives.

That is, $\dot{f}(\mathbf{p}) = \partial f(\mathbf{p})/\partial \mathbf{p}$ is a continuous $r \times 1$ function and $\ddot{f}(\mathbf{p}) = \partial^2 f(\mathbf{p})/\partial \mathbf{p} \partial \mathbf{p}'$ is a continuous $r \times r$ function.

Theorem 1.2. As $n \to \infty$ both $E(f(\mathbf{p}^*))$ and $E(f(\hat{\mathbf{p}}))$ equal

$$f(\mathbf{p}) + B(\mathbf{p})n^{-1} + O(n^{-2}) \text{ where } B(\mathbf{p}) = \mathrm{trace}\,(\ddot{f}(\mathbf{p})V/2). \qquad (1.12)$$

Also both $\mathrm{var}(f(\mathbf{p}^*))$ and $\mathrm{var}\,(f(\hat{\mathbf{p}}))$ equal

$$v(\mathbf{p})n^{-1} + O(n^{-2}) \text{ where } v(\mathbf{p}) = \dot{f}(\mathbf{p})' V \dot{f}(\mathbf{p}). \qquad (1.13)$$

This theorem shows that

(a) random-rounding increases the variance of the natural estimate for $f(\mathbf{p})$ by only $O(n^{-2})$; and

(b) random-rounding likewise has only a second order effect on the bias of the natural estimate for $f(\mathbf{p})$.

According to (1.12), the natural estimate of $f(\mathbf{p})$, $f(\hat{\mathbf{p}})$, has bias of magnitude n^{-1}. We now show how to reduce this to n^{-2}.

Corollary 1.1. If for some function $f_n(\mathbf{p})$, $E(f_n(\mathbf{p}^*)) = f(\mathbf{p}) + O(n^{-2})$ then $E(f_n(\hat{\mathbf{p}})) = f(\mathbf{p}) + O(n^{-2})$.

Two such choices for $f_n(\hat{\mathbf{p}})$ are the "delta-estimate" for which

$$f_n(\mathbf{p}) = f(\mathbf{p}) - \left\{ \sum_{i=1}^{r} f_{ii}(\mathbf{p})p_i - \mathbf{p}'\ddot{f}(\mathbf{p})\mathbf{p} \right\} / (2n), \tag{1.14}$$

where $f_{ii}(\mathbf{p}) = \partial^2 f(\mathbf{p}) / \partial p_i^2$, and the "jack-knife estimate" for which

$$f_n(\mathbf{p}) = nf(\mathbf{p}) - (n-1)\bar{f}, \tag{1.15}$$

where

$$\bar{f} = \Sigma_{i=1}^{r} p_i f([(n\mathbf{p} - e_i)/(n-1)]) + (1 - \Sigma_1^r p_i) f([n\mathbf{p}/(n-1)]),$$

$$e_i = \text{the } i\text{-th unit vector in } R^r,$$

$$\text{and } [x] : R^r \to R^r \text{ is defined by } [x]_i = \begin{cases} 0, & x_i < 0 \\ x_i, & 0 \le x_i \le 1. \\ 1, & x_i > 1 \end{cases}$$

These estimates were derived in Withers (1987a and 1987b). In particular, if $f(\mathbf{p})$ is only a function of p_1, say $f(\mathbf{p}) = g(p_1)$, then $f_n(\mathbf{p}) = g(p_1) - \ddot{g}(p_1)(p_1 - p_1^2)/(2n)$ and $\bar{f} = p_1 g([(np_1 - 1)/(n-1)]) + (1 - p_1) g([np_1/(n-1)])$. For example if $f(\mathbf{p}) = p_1^2$ then the delta-estimate uses $f_n(\mathbf{p}) = p_1^2 \{1 - (1 - p_1)/n\}$.

We now illustrate that if $f(\mathbf{p})$ is a polynomial we can in fact find an estimate of $f(\mathbf{p})$ based on the natural estimate with bias apparently exponentially small. We do this for the case $f(\mathbf{p}) = p_1^2$.

Theorem 1.3. $\hat{\lambda}_1 = \{\hat{p}_1^2 - n^{-1}\hat{p}_1 - n^{-2}(l^2 - 1)/6\}(1 - n^{-1})^{-1}$ estimates $\lambda_1 = p_1^2$ with bias $\Delta_n(p_1)(n^2 - n)^{-1}$.

Similarly if $f_n(\mathbf{p})$ is a moment of $\hat{\mathbf{p}}$ then we can also find an estimate of $f_n(\mathbf{p})$ with bias apparently exponentially small. We illustrate this for the case $f_n(\mathbf{p}) = \mathrm{var}(\hat{p}_1)$.

Theorem 1.4. $\hat{\lambda}_{2n} = n^{-1}(\hat{p}_1 - \hat{\lambda}_1) - n^{-2}(l^2 - 1)/6$ estimates $\lambda_{2n} = \mathrm{var}(\hat{p}_1)$ with bias $-\Delta_n(p_1)(n^2 - n)^{-1}$.

These results may be generalised to higher order polynomials and moments using the expression for moments and cumulants of $\hat{\mathbf{p}}$ given in Appendix B. We now show that for the special case of $f(\mathbf{p})$ collinear, an unbiased estimate exists.

Theorem 1.5. Set $f_I(\mathbf{p}) = \Pi_{i=1}^{I} p_i$ where $1 \le I \le R$ and

$$a_{nI} = n^{-I} n!/(n-I)! = (1 - n^{-1})(1 - 2n^{-1})\ldots(1 - \{I - 1\}n^{-1}). \tag{1.16}$$

Then

$$E(f_I(\hat{\mathbf{p}})) = E(f_I(\mathbf{p}^*)) = f_I(\mathbf{p})a_{nI}. \tag{1.17}$$

Hence an unbiased estimate of $f(\mathbf{p})$ is $f_I(\hat{\mathbf{p}})/a_{nI}$.

Corollary 1.2. $\mathrm{cov}(\hat{p}_1, \hat{p}_2) = -p_1 p_2 / n$. Its unbiased estimate is $-\hat{p}_1 \hat{p}_2 / (n-1)$. More generally for $1 \le I \le R$, $E(\Pi_{i=1}^{I} (\hat{p}_i - p_i)) = c_{nI} \Pi_{i=1}^{I} p_i$ with unbiased estimate $(\Pi_{i=1}^{I} \hat{p}_i)$ a_{nI} / c_{nI} where $c_{nI} = \Sigma_{j=0}^{I} (-1)^{I-j} \binom{I}{j} a_{nj}$. (The same result holds with $\hat{\mathbf{p}}$ replaced by $\mathbf{p}^*$.)

From (1.16) one may derive unbiased estimates for other special polynominals in $\mathbf{p}$ such as p_1^2, $p_1 p_2 (p_1 + p_2)$ and $\Sigma_{i=1}^{R} p_i^3$ - but not for $p_1^2 p_2$ or p_1^3.

Corollary 1.3. For $1 \le I < R$ an unbiased estimate of

$$f_I(\mathbf{p}) \sum_{1}^{I} p_i \text{ is } f_I(\hat{\mathbf{p}}) \left\{ 1 - In^{-1} - \sum_{I+1}^{R} \hat{p}_i \right\} / a_{n,I+1}. \tag{1.18}$$

In particular an unbiased estimate of p_1^2 is

$$\hat{p}_1 (\tilde{p}_1 - n^{-1})(1 - n^{-1})^{-1}. \tag{1.19}$$

We emphasize that the results of this paper are based on the assumption that table entries are independent Poisson's, or at least multinomial conditional on the total. The Poisson and multinomial models are appealing as they have a ready interpretation, and because sums of Poisson variables are Poisson. But sums of multinomials are multinomial only if they share the same cell probabilities $\mathbf{p}$. This suggests that conclusions drawn from such models may be less accurate if the populations modelled are composed of two or more inhomogeneous groups.

2. PROOFS

Proof of Theorem 2.1. Set $r = N_1 \bmod l$. Then (1.1) holds for $N = N_1$, $M = M_1$ with $jl = N - r$ and

$$E(M_1^2 \mid r) = (N_1 - r)^2 (1 - r/l) + (N_1 - r + l)^2 r/l = N_1^2 + lr - r^2.$$

Hence

$$E(\hat{p}_1^2) = E(p_1^{*2}) + n^{-2} A_n(p_1), \tag{2.1}$$

where

$$A_n(p_1) = E(M_1^2 - N_1^2) = E(lr - r^2) = \sum_{i=0}^{l-1} (li - i^2) P(N = i)$$

$$= (l^2 - 1)/6 + \Delta_n(p_1)$$

since

$$l^{-1} \sum_{i=0}^{l-1} i(l - i) = (l^2 - 1)/6. \tag{2.2}$$

But

$$E(p_1^{*2}) = p_1^2 + (p_1 - p_1^2) n^{-1}, \tag{2.3}$$

so (1.5) follows. Now $\tilde{p}_1 = \hat{p}_1 - \sum(M_j - N_j)/n$,

$$\text{so } E(\tilde{p}_1^2) = E(\hat{p}_1^2) - 2n^{-2}\sum E(M_1(M_j - N_j)) + n^{-2}\sum E((M_i - N_i)(M_j - N_j))$$

$$= E(\hat{p}_1^2) - 2n^{-2}A_n(p_1) + n^{-2}\sum A_n(p_i)$$

since $E(\Pi_i f_i(M_i)|\{N_i\}) = \Pi_i E(f_i(M_i)|N_i)$. (2.4)

Hence $\operatorname{var}(\tilde{p}_1) = (p_1 - p_1^2)n^{-1} + n^{-2}\Sigma_{i\neq 1} A_n(p_i)$ so (1.7) holds.

Also,

$$E(\hat{p}_1\tilde{p}_1) = p_1 - n^{-2}\sum_{i\neq 1} E(M_1M_i) = p_1 - \sum_{i\neq 1} E(p_1^*p_i^*)$$

$$= p_1 - \sum_{i\neq 1} p_1p_i(1 - n^{-1}) = p_1 - p_1(1 - p_1)\ (1 - n^{-1}),$$

so

$$\operatorname{cov}(\hat{p}_1,\tilde{p}_1) = (p_1 - p_1^2)n^{-1}. \tag{2.5}$$

Hence $\operatorname{var}(p_1(\lambda)) = (p_1 - p_1^2)n^{-1} + \{(1 - \lambda)^2A_n(p_1) + \lambda^2\,\Sigma_{i\neq 1} A_n(p_i)\}n^{-2}$ and (1.8) holds.

Proof of Theorem 1.2. This was proved for $\mathbf{p}^*$ in Withers (1987a). Also since $\ddot{f}$ is finite in a neighborhood of $\mathbf{p}$,

$$f(\hat{\mathbf{p}}) = f(\mathbf{p}^*) + (\hat{\mathbf{p}} - \mathbf{p}^*)'\dot{f}(\mathbf{p}^*) + O(|\hat{\mathbf{p}} - \mathbf{p}^*|^2).$$

$$E((\hat{\mathbf{p}} - \mathbf{p}^*)|N) = 0,\ E((\hat{p}_1 - p_i^*)^2|N) = 2n^{-2}I(N_l \bmod l \neq 0),$$

where $I(A) = 1$ or 0 for A true or false, that is, $I(\cdot)$ is the indicator function Hence $E(f(\hat{\mathbf{p}})) = E(f(\mathbf{p}^*)) + O(n^{-2})$ and $\operatorname{var}(f(\hat{\mathbf{p}})) = \operatorname{var}(f(\mathbf{p}^*)) + O(n^{-2})$.

Proof of Theorem 1.3. This follows directly from (2.1) and (2.3).

Proof of Theorem 1.4. This follows from (2.1) and (1.5).

Proof of Theorem 1.5. The first equality in (1.16) follows from (2.4), and the second from the multinomial theorem. Corollary 1.2 follows immediately.

Proof of Corollary 1.3. From (1.16), for $1 \leq I < i \leq R$ we have

$$E(f_I(\hat{\mathbf{p}})\hat{p}_i) = f_I(\mathbf{p})p_ia_{n,I+1}$$

so

$$E(f_l(\hat{\mathbf{p}}) \sum_{l+1}^{R} \hat{p}_i / a_{n,l+1}) = f_l(\mathbf{p})(1 - \Sigma_1^l p_i)$$

$$= E(f_l(\hat{\mathbf{p}}) / a_{nl}) - f_l(\mathbf{p}) \Sigma_1^l p_i.$$

ACKNOWLEDGEMENT

I wish to thank Peter McGavin for doing the computing in Appendix A.

APPENDIX A

One expects that for f a smooth function

$$E(f(\hat{\mathbf{p}})I(N_1 = j_1 \bmod l, \ldots, N_s = j_s \bmod l)) \to f(\mathbf{p})l^{-s} \tag{A.1}$$

as $n \to \infty$ provided $0 < p_i < 1$ for $1 \leq i \leq s \leq R$.

If $E(f(\hat{\mathbf{p}})) = f(\mathbf{p})$, one expects the rate of convergence to be exponential, $O(e^{-\lambda n})$ for some $\lambda > 0$. If $f(\hat{\mathbf{p}})$ is biased, then its bias is $O(n^{-1})$, so that one would expect this rate also to apply to (A.1). Convergence will in general break down as $\mathbf{p}$ approaches the boundary of $[0,1]^r$, since

$$E(f(\hat{\mathbf{p}})I(N_1 = j_1 \bmod l, \ldots, N_s = j_s \bmod l))$$

$$= \begin{cases} f(\mathbf{p})I(j_1 = j_2 = \ldots = j_s = 0) & \text{if } \mathbf{p} = 0 \\ f(\mathbf{p})I(j_1 = n \bmod l) & \text{if } p_1 = 1. \end{cases}$$

To test these expectations we considered the case $s = 1$, $l = 3$, $j = 0$ and the functions (a) $f(\mathbf{p}) = 1$, (b) $f(\mathbf{p}) = p_1$, and (c) $f(\mathbf{p}) = \exp(p_1)$. Computations were done in quadruple precision on a VAX11/780, giving a precision for

$$\Delta = E(f(\hat{\mathbf{p}})I(N_1 = j_1 \bmod l, \ldots, N_s = j_s \bmod l)) - f(\mathbf{p})l^{-s}$$

of 112 bits - nearly 34 decimal places. Figures 1a, 1b and 1c plot Δ versus p_1 for $n = 6$, 18, 54. Since $n \bmod 3 = 0$, Δ is symmetric about $p_1 = 1/2$ for (a).

Since $\Delta = 2/3f(0)$ at $p_1 = 0$, and is equal to $2/3$, 0 and $2/3$ for (a), (b) and (c) respectively, convergence breaks down at $p_1 = 0$ for (a) and (c), but not for (b). At $n = 18$, Δ is already negligibly different from 0 for p_1 in (.2, .8) for (a) and for p_1 in (.1, .8) for (b) and (c). At $n = 54$, these ranges have grown to cover (.1, .9) for (a), (.02, .95) for (b), and (.07, .95) for (c).

Figures 2a and 2b plot $Y = \log(-\log|\Delta|)$ versus $X = \log(n)$ for (a) $f(\mathbf{p}) = 1$ and (b) $f(\mathbf{p}) = p_1$. As expected, except for small n, the curves are roughly parallel to $Y = X$ (except for (b) with $p_1 = .01$), consistent with $\Delta = O(e^{-\lambda n})$ for some $\lambda > 0$. The curves are not smooth, as Δ has only been calculated at n a power of 2 ($n = 2^i$ for $0 \leq i \leq 7$).

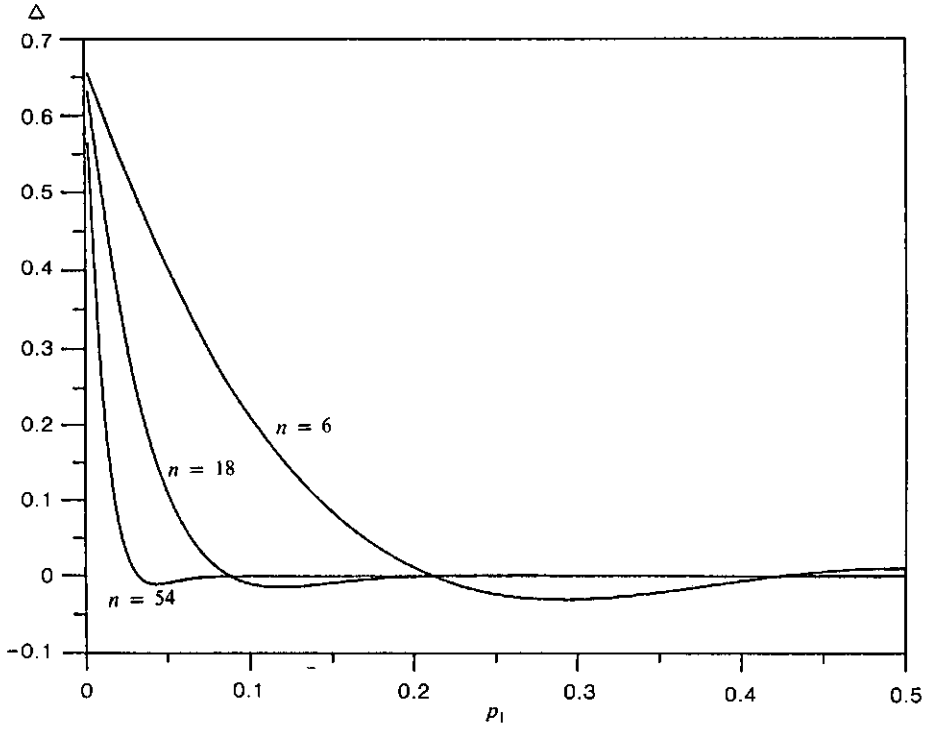

Figure 1a. Evidence for (A.1) When $f(\mathbf{p}) = 1$.

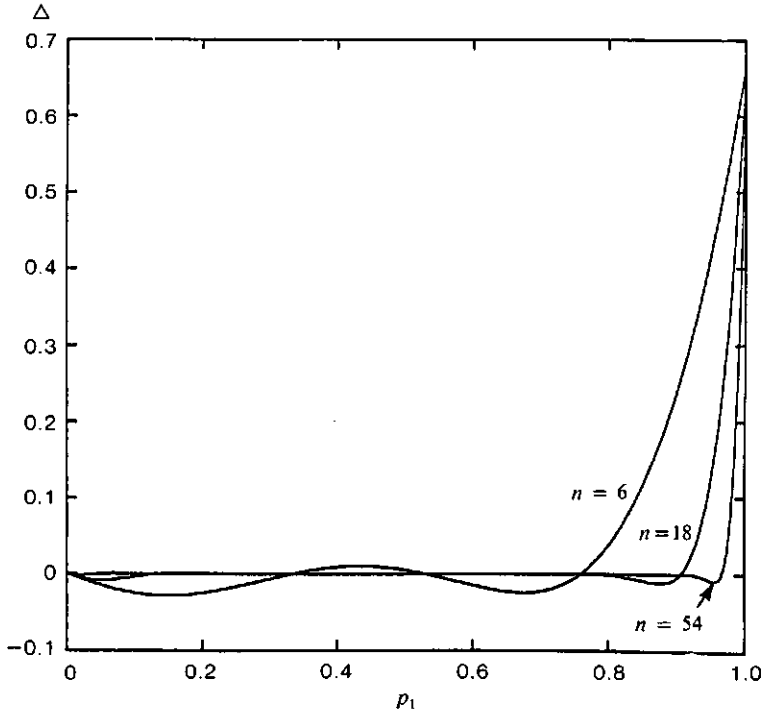

Figure 1b. Evidence for (A.1) When $f(\mathbf{p}) = p_1$.

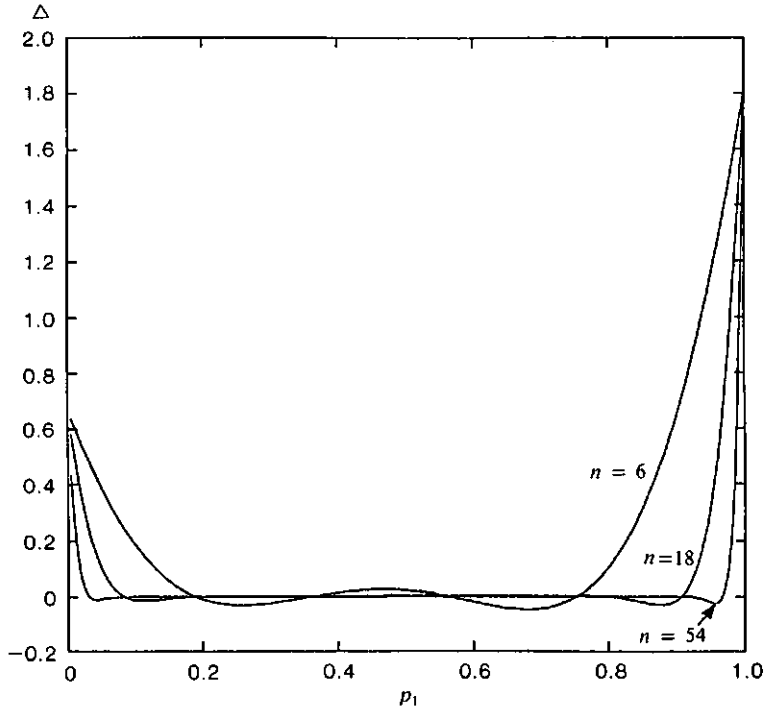

Figure 1c. Evidence for (A.1) When $f(\mathbf{p}) = \exp(p_1)$.

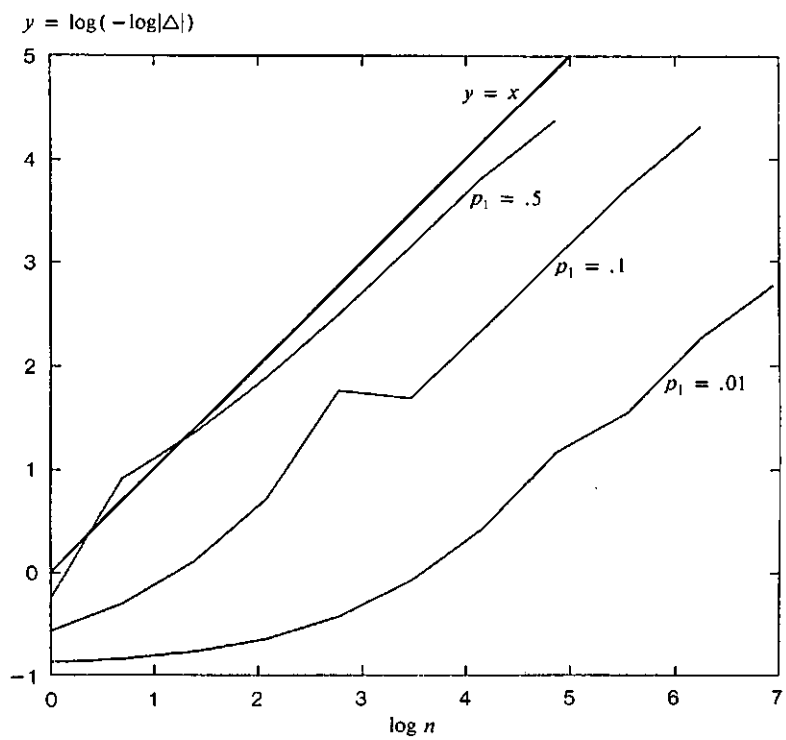

Figure 2a. Evidence for Exponential Convergence in (A.1) for $f(\mathbf{p}) = 1$.

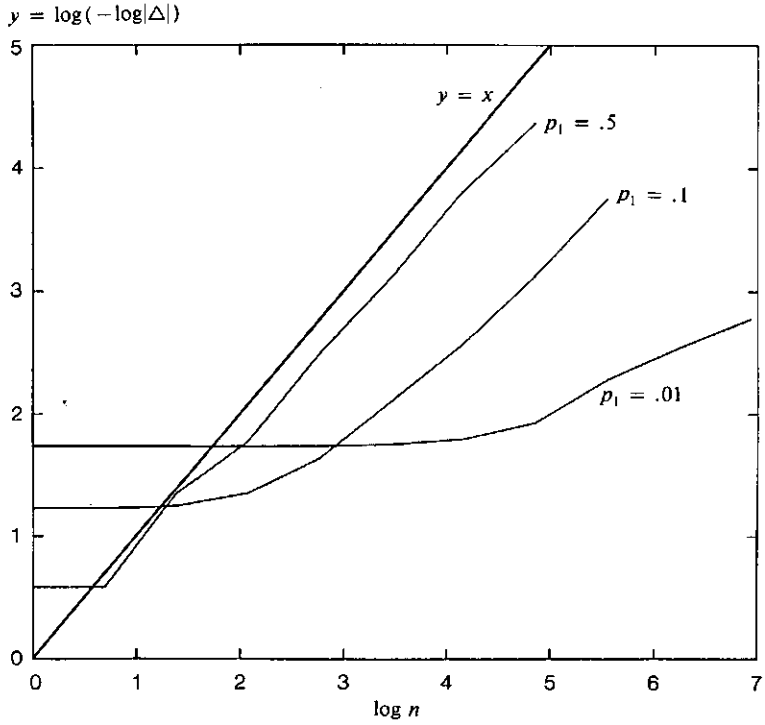

Figure 2b. Evidence for Exponential Convergence in (A.1) for $f(\mathbf{p}) = p_1$.

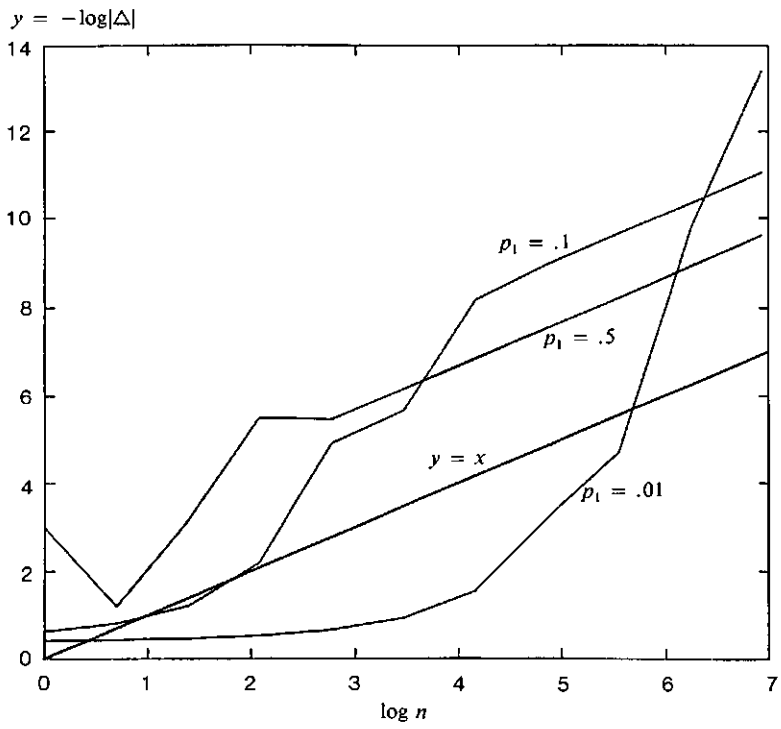

Figure 3. Evidence for Convergence at Rate $\sim n^{-1}$ in (A.1) for $f(\mathbf{p}) = \exp(p_1)$.

Figure 3 plots $Y = -\log|\Delta|$ versus $X = \log(n)$ for (c) $f(\mathbf{p}) = \exp(p_1)$. For n large the curves are parallel to $Y = X$ for $p_1 = .5$ and .1 consistent with $\Delta = O(n^{-1})$, but for $p_1 = 0.1$ the increase is much faster than linear. The graphs generally confirm our expectations on the rate of convergence in (A.1). To obtain analytic proofs would appear to require some sophisticated number theory.

APPENDIX B

Here we compare the moments and cumulants of $\mathbf{p}^* = N/n$ and $\hat{\mathbf{p}} = M/n$. Set $q_1 = 1 - p_1$, $n_i = N_i \bmod l$, and $m_i(j) = E(p_1^{*j} I(n_1 = j)) \to p_1^j/l$ as $n \to \infty$, assuming $p_1 \neq 0$ or 1. Elementary calculations yield

$$\mu(\hat{\mathbf{p}}) = \mu(\mathbf{p}^*) = \mathbf{p},$$

$$\mu_2(\hat{p}_1) = \mu_2(p_1^*) + M_{22}n^{-2} = p_1q_1n^{-1} + O(n^{-2}),$$

where

$$M_{22} = A_n(p_1) = \sum_{i=0}^{l-1} i(l-i)m_0(i) \to (l^2-1)/6$$

as $n \to \infty$,

$$\mu_3(\hat{p}_1) = \mu_3(p_1^*) + 3n^{-2}\sum_{j=0}^{l-1}(lj - j^2)\{m_2(j) - 2p_1m_1(j) + p_1^2m_0(j)\}$$

$$+ n^{-3}\sum_{j=0}^{l-1} a_{jl}m_0(j)$$

$$= \mu_3(p_1^*) + o(n^{-2}) = p_1q_1(1-2p_1)n^{-2} + o(n^{-2}),$$

and

$$a_{jl} = -j^3(1 - j/l) + (l-j)^3 j/l.$$

Similarly $\mu_4(\hat{p}_1)$ has the form $\mu_4(p_1^*) + \Sigma_2^4 M_{4i}n^{-i} = O(n^{-2})$ and $\kappa_4(\hat{p}_1)$ has the form $\Sigma_2^4 k_{4i}n^{-i}$ where $k_{42} = M_{42}$ does not converge to 0 as $n \to \infty$. Hence $\kappa_4(\hat{p}_1) \sim n^{-2}$, not n^{-3}. Hence $\hat{\mathbf{p}}$ does not satisfy the Cornish-Fisher assumption that $\kappa_r(\hat{\mathbf{p}}) = O(n^{1-r})$ for $r \geq 1$: see for example Kendall and Stuart (1977).

Moments and cumulants may also be obtained from the m.g.f. (moment generating function), which we now obtain.

$$E(\exp(t_1M_1/n) \mid N_1) = \exp(t_1N_1/n)S(t_1,n_1)$$

where

$$S(t_1,n_1) = (1 - n_1/l)\exp(-n_1t_1/n) + (n_1/l)\exp(l - n_1)t_1/n.$$

Hence by (2.4), the m.g.f. is

$$E(\exp(t'\hat{\mathbf{p}})) = E(\exp(t'\mathbf{N}/n))S(t) \text{ where } S(t) = \Pi_1^r S(t_i, n_i).$$

Also at $t = 0$, $S_1 = 0$ and so $S_{ij\ldots} = 0$ if a subscript occurs exactly once. For example, setting

$$S = S(t),\ \partial_i = \partial/\partial t_i,\ S_i = \partial_i S,\ S_{ij} = \partial_i \partial_j S,\ \ldots$$

gives

$$E(\hat{p}_1^2 \exp(t'\hat{\mathbf{p}})) = E(\exp(t'\mathbf{N}/n)\{p_1^{*2} S + 2p_1^* S_1 + S_{11}\}),$$

$$E(\hat{p}_1^2 \hat{p}_2^2 \exp(t'\hat{\mathbf{p}})) = E(\exp(t'\mathbf{N}/n)\{p_1^{*2}(p_2^{*2} S + 2p_2^* S_2 + S_{22}) + 2p_1^*(p_2^{*2} S_1 + 2p_2^* S_{12} + S_{122}) + (p_2^{*2} S_{11} + 2p_2^* S_{112} + S_{1122})\}).$$

Hence $E(\hat{p}_1^2) = E\{p_1^{*2} + S_{11}(0)\}$ and

$$E(\hat{p}_1^2 \hat{p}_2^2) = E\{p_1^{*2} p_2^{*2} + p_1^{*2} S_{22}(0) + p_2^{*2} S_{11}(0) + S_{1122}(0)\}.$$

where $S_{ii}(0) = S_{11}(0, n_i) = n^{-2}(l - n_i)n_i = n^{-2} \Sigma_{k=0}^{l-1} (l - k)kI(n_i = k)$ and $S_{1122}(0) = S_{11}(0)\, S_{22}(0)$. Some further simplifications can be obtained using $N_2 \mid N_1 \sim Bi(\theta, n - N_1)$ where $\theta = p_2/(1 - p_1)$. From the multinomial m.g.f. one obtains

$$E(p_1^{*2} p_2^{*2}) = n^{-4} p_1 p_2 \{(n)_4 p_1 p_2 + (n)_3 (p_1 + p_2) + (n)_2\}$$

where $(n)_i = n!/(n - i)! = n(n - i) \ldots (n - i + 1)$.

REFERENCES

GASTWIRTH, J.L., KRIEGE, A.M., and RUBIN, D.B. (1978). Statistical analyses from summary data and their impact on the issue of confidentiality. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 183-188.

KENDALL, M.G., and STUART, A. (1977). *The Advanced Theory of Statistics, Volume* 7. London: Griffin.

NARGUNDAR, M.S., and SAVELAND, W. (1972). Random rounding to prevent statistical disclosures. *Proceedings of the Social Statistics Section, American Statistical Association,* 382-385.

PENNY, R., and RYAN, M. (1986) A problem associated with random-rounding. *New Zealand Statistician,* 21, 43-52.

WITHERS, C.S. (1987a). Bias reduction by Taylor series. *Communications in Statistics - Theory and Methods* (forthcoming).

WITHERS, C.S. (1987b). Jackknifing binomials and multinomials. Unpublished manuscript, Department of Scientific and Industrial Research.

Survey Methodology, December 1987
Vol. 13, No. 2, pp. 147–161
Statistics Canada

Variance Estimation for the Canadian Labour Force Survey

G.H. CHOUDHRY and H. LEE[1]

ABSTRACT

The biases and stabilities of alternative variance estimators for the two stage random group design (Rao et al. 1962) are evaluated in a Monte Carlo study in the context of Canadian Labour Force Survey. The variance formula for raking ratio estimation procedure is derived using Taylor linearization method. The properties of the variance formula are investigated by a Monte Carlo simulation.

KEY WORDS: Keyfitz's variance estimator; Raking ratio estimator; Taylor linearization; Monte Carlo simulation.

1. INTRODUCTION

The Canadian Labour Force Survey (LFS) is the largest monthly household survey conducted by Statistics Canada and is used to produce estimates of various labour force characteristics at national, provincial and sub-provincial levels. It follows a stratified multistage rotating sample design with six rotation panels (Platek and Singh 1976).

Following each decennial census of population, the LFS has undergone a sample redesign. As part of the 1981 post-censal redesign, an extensive program of research was undertaken in the areas of sampling, data collection, and estimation methodologies (Singh and Drew 1981). The post-stratified ratio estimation procedure used in the old design was replaced by a raking ratio estimation procedure to improve the reliability of subprovincial data. This paper presents the results related to variance estimation methodology.

The methodology for variance estimation for the old LFS was based on Woodruff's generalization (Woodruff 1971) of the Keyfitz procedure (Keyfitz 1957) using Taylor linearization applied to the post-stratified ratio estimates (Platek and Singh 1976). This method will be called the Keyfitz method as in Platek and Singh (1976).

There are three area types identified in the LFS design, i.e., self-representing (SR) areas consisting of major cities, non-self-representing (NSR) areas which are smaller urbans and rural areas, and special areas composed of military, institutions and remote areas. For the NSR and special areas it was decided to use the Keyfitz method with modification to incorporate the raking ratio estimation procedure.

However, for the two-stage random group design in SR areas, two alternative variance estimators given by Rao, Hartley, and Cochran (1962) and by Rao (1975) were evaluated and compared with Keyfitz's method using Monte Carlo simulation. The alternative variance estimators of estimates with and without ratio adjustment were compared with respect to their biases and stabilities. The impact on the Keyfitz variance estimator due to increase of the number of replicates was also examined. Details are reported in Section 2. Based on the results of the evaluation, the Keyfitz method was adopted for SR areas as well.

[1] G.H. Choudhry and H. Lee, Social Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

The Keyfitz variance formula for raking ratio estimates used for all area types in the LFS is derived in Section 3 and evaluated by Monte Carlo study. Finally in Section 4, some concluding remarks are given.

2. VARIANCE ESTIMATION FOR THE SR DESIGN

2.1 SR Design

The LFS design in the SR areas is a two-stage random group design (Rao et al. 1962) with probability proportional to size (PPS) selection of primary sampling units (PSU's) and systematic selection of dwellings at the second stage such that the design becomes self-weighting. Suppose that there are N PSU's in a given stratum and let x_j and M_j, $j = 1, 2, \ldots, N$, respectively be the size measure and dwelling count for the j-th PSU in the stratum. Let $1/W$ be the sampling rate in the stratum, where W is an integer, and n be the number of PSU's to be selected from the stratum. The N PSU's in the stratum are randomly partitioned into n groups so that the i-th random group contains N_i PSU's, and $\Sigma_{i=1}^{n} N_i = N$.

Define

$$p_j = \frac{x_j}{\sum_{t=1}^{N} x_t}, \quad j = 1, 2, \ldots, N,$$

and

$$\delta_{ij} = 1 \text{ if the } j\text{-th PSU is in the } i\text{-th group}$$
$$= 0 \text{ otherwise.}$$

Then $\pi_i = \Sigma_{j=1}^{N} \delta_{ij} p_j$ is the relative size of the i-th random group.

Now define W_{ij}, the sampling interval for systematic sampling, as follows: Let $a_{ij} = \delta_{ij} W p_j / \pi_i$ and $r_{ij} = a_{ij} - [a_{ij}]$ where $[a]$ is the greatest integer less than or equal to a. Without loss of generality, we assume that the set $\{r_{ij}, j = 1, 2, \ldots, N\}$ is in descending order. Then, W_{ij} is defined as

$$W_{ij} = [a_{ij}] + 1, \quad j = 1, 2, \ldots, R$$
$$= [a_{ij}], \quad j = R + 1, \ldots, N$$

where $R = \Sigma_{j=1}^{N} r_{ij}$. Then, by definition $\Sigma_{j=1}^{N} W_{ij} = W$ for the i-th random group, $i = 1, 2, \ldots, n$.

Since W_{ij} is the sampling interval for systematic sampling from the selected cluster in the i-th random group, it is defined as an integer for operational simplicity.

One PSU is selected with probability proportional to W_{ij}'s from each of the n random groups independently. The selected PSU j from the i-th random group is sub-sampled systematically at the rate $1/W_{ij}$. Then the overall sampling rate in each of the n random groups is $1/W$ so that the design becomes self-weighting with a design weight equal to W. Each random group is assigned a panel number from 1 to 6. The number of random

groups n is usually a multiple of six so that each panel has the same number of random groups.

Since only one PSU is selected from each random group, we denote by $1/W_i$ the subsampling rate in the selected PSU from the i-th random group and by m_i the number of selected dwellings from the random group i.

2.2 Alternative Variance Estimators

Suppose that we are interested in the total of a characteristic y for the stratum. Let y_{jk} be the y-value for the k-th dwelling in the j-th PSU where $k = 1, 2, \ldots, M_j$. Then the total $Y = \Sigma_{j=1}^{N} \Sigma_{k=1}^{M_j} y_{jk}$ can be estimated by $\hat{Y} = W \Sigma_{i=1}^{n} y_i$, where y_i is the sum of y-values for the m_i sampled dwellings from the PSU selected from the i-th group, $i = 1, 2, \ldots, n$. We consider the following variance estimators for estimating the variance of the estimated total $\hat{Y}$:

(1) Keyfitz's (1957) Variance Estimator

This estimator was used in the old design with two pseudo-replicates formed by collapsing the odd numbered panels into one replicate and the even into the other. Ignoring the finite population correction (fpc), the variance is obtained by

$$\hat{V}_1(\hat{Y}) = W^2 \left(\sum_o y_i - \sum_e y_i \right)^2 \tag{2.1}$$

where $\sum_o$ is the summation over all the odd numbered panels and $\sum_e$ is the summation over all the even numbered panels. Alternatively, the generalized Keyfitz variance estimator for $n(\geq 2)$ replicates which is given by

$$\hat{V}_2(\hat{Y}) = W^2 \frac{n}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{2.2}$$

where $\bar{y} = (1/n) \Sigma_{i=1}^{n} y_i$, can be used. In this case each PSU or panel is taken as a replicate. $\hat{V}_2$ was considered because it was thought that this variance estimator might have better efficiency (stability) than $\hat{V}_1$ due to its larger number of degrees of freedom.

(2) Rao, Hartley, and Cochran's (1962) Variance Estimator

This variance formula is derived under the assumption that the number of secondaries m_i to be selected from the i-th group is fixed for $i = 1, 2, \ldots, n$, and simple random sampling (SRS) is also assumed at the second stage. The variance estimator is given by:

$$\hat{V}_3(\hat{Y}) = A \sum_1^n \pi_i \left(\frac{M_i y_i}{m_i p_i} - \hat{Y} \right)^2 + \sum_1^n \frac{\pi_i}{p_i} M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i} \right) s_i^2 \tag{2.3}$$

where

$$A = \frac{\sum_1^n N_i^2 - N}{N^2 - \sum_1^n N_i^2}, \tag{2.4}$$

$$s_i^2 = \frac{1}{m_i - 1} \sum_{k=1}^{m_i} (y_{ik} - \bar{y}_i)^2. \tag{2.5}$$

M_i is the number of dwellings in the selected PSU from the i-th group and m_i out of M_i dwellings are selected with systematic sampling but the variance estimate is obtained under the assumption of SRS. The y-value for the k-th selected dwelling from the selected PSU in the i-th group is y_{ik} and $\bar{y}_i = y_i / m_i$.

Since $\pi_i / p_i = W / W_i$ and $M_i / m_i = W_i$, (these equalities are not strict due to the use of integer values for W_i), the variance formula (2.2) can be written as:

$$\hat{V}_3(\hat{Y}) = A \sum_1^n \pi_i \left(W \frac{y_i}{\pi_i} - \hat{Y} \right)^2 + W \sum_1^n \left(1 - \frac{m_i}{M_i} \right) M_i s_i^2. \tag{2.6}$$

(3) Rao's (1975) Variance Estimator

In this case it is assumed that m_i secondaries are selected with SRS but, since the design is self-weighting, the sample size m_i at the second stage is treated as a random variable. The variance estimator is given by:

$$\hat{V}_4(\hat{Y}) = A \sum_1^n \pi_i \left(W \frac{y_i}{\pi_i} - \hat{Y} \right)^2 + \sum_1^n \left\{ \frac{\pi_i^2}{p_i^2} - A \left(\frac{\pi_i}{p_i^2} - \frac{\pi_i^2}{p_i^2} \right) \right\} \frac{M_i^2 s_i^2}{m_i} - \sum_1^n \frac{\pi_i}{p_i} M_i s_i^2. \tag{2.7}$$

where A is defined by (2.4) and s_i^2 by (2.5). After some simplification (2.7) can be written as:

$$\hat{V}_4(\hat{Y}) = \hat{V}_3(\hat{Y}) + W^2 \sum_1^n m_i s_i^2 \left\{ \left(1 - \frac{W_i}{W} \right) - A \left(\frac{1}{\pi_i} - 1 \right) \right\}. \tag{2.8}$$

We note that there is an additional term, which could be positive or negative, in the variance formula when random sample size is assumed at the second stage.

2.3 Monte Carlo Study

In order to evaluate the biases of the four variance estimators and their relative stabilities, a Monte Carlo study was carried out with 19 Labour Force strata from the Census Metropolitan Area (CMA) of Halifax using data from the 1981 census. The census data for the purpose of this study was the census sample given the long questionnaire which is 20% systematic sample of dwellings within Enumeration Areas. The sampling rate $1 / W$ was taken to be 0.04 to obtain the same expected sample size as in the actual redesigned LFS. The number of random groups within each stratum was even and was determined so that the expected sample size within random groups would be as close to 4.5 as possible to correspond to the actual LFS. The 19 strata chosen for the study are shown in Table 1 with the number of PSU's, the number of selected PSU's, the number of dwellings, and the expected sample sizes along with the corresponding totals for all the strata. Within each of the 19 strata, 1,000

Table 1

Strata Used for the Monte Carlo Study

Stratum	No. of Dwellings	No. of PSU's	No. of Selected PSU's	Expected Sample Size
1	737	49	6	29.5
2	490	33	4	19.6
3	745	45	6	29.8
4	720	34	6	28.8
5	621	37	6	24.8
6	630	38	6	25.2
7	503	31	4	20.1
8	340	23	4	13.6
9	472	33	4	18.9
10	468	33	4	18.7
11	367	28	4	14.7
12	390	23	4	15.6
13	626	36	6	25.0
14	650	39	6	26.0
15	350	22	4	14.0
16	736	46	6	29.4
17	573	35	6	22.9
18	773	48	6	30.9
19	866	64	8	34.6
Total	11,057	697	100	442.3

samples were generated independently using a Monte Carlo technique, employing the random group design described in Subsection 2.1.

Let $\hat{Y}_{ht}$ be the estimate of the total Y_h for stratum h from the t-th Monte Carlo draw, $h=1, 2, \ldots, 19$, and $t=1, 2, \ldots, 1{,}000$. Similarly $\hat{V}_{jht}$, $j=1, 2, 3, 4$ are the four variance estimators of $\hat{Y}_{ht}$.

Now define

$$Y = \sum_{h=1}^{19} Y_h,$$

$$\hat{Y}_t = \sum_{h=1}^{19} \hat{Y}_{ht},$$

$$\hat{V}_{jt} = \sum_{h=1}^{19} \hat{V}_{jht}, \quad j = 1, 2, 3, 4,$$

where $t = 1, 2, \ldots, 1000$.

$\hat{Y}_t$ is the estimate of the total Y obtained from the t-th Monte Carlo draw and $\hat{V}_{jt}$, $j = 1, 2, 3, 4$ are the corresponding variance estimates.

The Monte Carlo expectation and variance denoted by E^* and V^* respectively are defined for T Monte Carlo draws as follows:

$$E^*(\hat{\theta}) = \frac{1}{T}\sum_{t=1}^{T} \hat{\theta}_t,$$

$$V^*(\hat{\theta}) = \frac{1}{T}\sum_{t=1}^{T} [\hat{\theta}_t - E^*(\hat{\theta})]^2,$$

where $\hat{\theta}$ is an estimator of the unknown parameter θ and $\hat{\theta}_t$ is the estimate obtained from the t-th draw. Using these definitions, we obtain the Monte Carlo variance of the estimator $\hat{Y}$, $V^*(\hat{Y})$, and the Monte Carlo expectations and variances of the variance estimators $\hat{V}_j$, $E^*(\hat{V}_j)$ and $V^*(\hat{V}_j)$ respectively for $j = 1, 2, 3, 4$.

Now define the bias of the variance estimator $\hat{V}_j$ by:

$$B_j = E^*(\hat{V}_j) - V^*(\hat{Y}),$$

and percent bias as:

$$PB_j = 100\,\frac{B_j}{V^*(\hat{Y})},\ j = 1, 2, 3, 4.$$

Then the Mean Square Error (MSE) of $\hat{V}_j$ is given by:

$$MSE_j = V^*(\hat{V}_j) + B_j^2,\ j = 1, 2, 3, 4.$$

We define the efficiency of $\hat{V}_j$, relative to the Keyfitz variance estimator with two replicates (i.e., $\hat{V}_1$) as:

$$\text{Rel. Eff}(\hat{V}_j \text{ vs. } \hat{V}_1) = (MSE_1/MSE_j)^{1/2},\ j = 2, 3, 4.$$

In this study, we consider three labour force characteristics: Employed, Unemployed, and In Labour Force. The relative biases and efficiencies of the variance estimators are reported in Tables 2A and 3A respectively for the three characteristics. We observe that, with respect to bias, the variance estimators 1 and 2 are similar and so are 3 and 4. The variance estimators 1 and 2 have very large positive biases notably for Employed and In Labour Force while 3 and 4 have relatively small biases. In efficiency comparison, the variance estimators 3 and 4 are much superior to 1 and 2 and very similar to each other. Moreover, the variance estimator 2 also performed better than 1.

The four variance estimators were also evaluated for ratio estimates by total population at the level of aggregation of all the strata. The corresponding variance estimators denoted by $\hat{V}_j^{(R)}$, $j = 1, 2, 3, 4$ were also obtained from each Monte Carlo draw by the Taylor linearization method. Then we obtained ratio adjusted version of percent biases of the four variance estimators (Table 2B) and relative efficiencies of the latter three variance estimators with respect to the first one (Table 3B).

We note that the biases of the variance estimators 1 and 2 were substantially reduced for ratio adjusted estimates especially for Employed and In Labour Force. For the variance estimators 3 and 4, the biases were also reduced for Employed and In Labour Force but there was very little change for Unemployed. Although the biases of the four variance estimators are small, the only nonsignificant bias at 5% level was that of the variance estimator 3 for In Labour Force. All the observed differences between biases were significant at 5% level except those of the variance estimators 1 and 2 for the three characteristics.

Table 2A

Percent Biases of the Variance Estimators of the Estimates of LF Characteristic Totals without Ratio Adjustment

Characteristic	Percent Bias			
	$\hat{V}_1$	$\hat{V}_2$	$\hat{V}_3$	$\hat{V}_4$
Employed	23.4	24.5	−4.7	−6.3
Unemployed	6.3	6.6	3.7	1.2
In Labour Force	24.2	25.2	−5.1	−6.7

Table 2B

Percent Biases of the Variance Estimators of the Estimates of LF Characteristic Totals with Ratio Adjustment

Characteristic	Percent Bias			
	$\hat{V}_1^{(R)}$	$\hat{V}_2^{(R)}$	$\hat{V}_3^{(R)}$	$\hat{V}_4^{(R)}$
Employed	3.7	4.3	−1.1	−3.1
Unemployed	5.3	5.5	4.0	1.4
In Labour Force	4.5	5.0	−0.5	−2.5

Table 3A

Relative Efficiencies of $\hat{V}_2$, $\hat{V}_3$, and $\hat{V}_4$ with Respect to $\hat{V}_1$
(Rel. Eff. of $\hat{V}_j = [MSE(\hat{V}_1)/MSE(\hat{V}_j)]^{1/2}$, $j = 2,3,4$

Characteristic	Relative Efficiency		
	$\hat{V}_2$	$\hat{V}_3$	$\hat{V}_4$
Employed	1.51	3.22	3.11
Unemployed	1.52	1.71	1.76
In Labour Force	1.49	3.24	3.12

Table 3B

Relative Efficiencies of $\hat{V}_2^{(R)}$, $\hat{V}_3^{(R)}$, and $\hat{V}_4^{(R)}$ with Respect to $\hat{V}_1^{(R)}$
(Rel. Eff. of $\hat{V}_j^{(R)} = [MSE(\hat{V}_1^{(R)})/MSE(\hat{V}_j^{(R)})]^{1/2}$, $j = 2,3,4$

Characteristic	Relative Efficiency		
	$\hat{V}_2^{(R)}$	$\hat{V}_3^{(R)}$	$\hat{V}_4^{(R)}$
Employed	2.13	2.59	2.52
Unemployed	1.57	1.71	1.76
In Labour Force	2.08	2.56	2.51

Table 4

Coverage Rates of 95% Confidence Intervals for the Estimates of LF Characteristic Totals with Ratio Adjustment

Characteristic	Coverage Rate			
	$\hat{V}_1^{(R)}$	$\hat{V}_2^{(R)}$	$\hat{V}_3^{(R)}$	$\hat{V}_4^{(R)}$
Employed	93.6	95.4	94.6	94.2
Unemployed	94.3	95.1	95.3	95.0
In Labour Force	93.2	95.3	94.6	94.2

We also computed the 95% confidence intervals (CI's) for the ratio-adjusted estimates from each Monte Carlo draw using the four variance estimators. The coverage rates were obtained as the proportion of CI's which include the true value of characteristic total. The results are given in Table 4 and show that the performances of all the 4 variance estimators are very good for all the characteristics. Since the variance estimators of ratio- adjusted estimates provide confidence intervals which have coverage rates very close to the nominal value, the small biases of the variance estimators are of no practical consequence. Thus, from the bias point of view, all four variance estimators for the ratio-adjusted estimates are not much different from each other. The relative efficiencies of the variance estimators 3 and 4 are now only marginally better than 2 regardless of characteristic. The relative efficiencies of the 3 alternatives in this case are over 2 for Employed and In Labour Force. For unemployed they are somewhat lower and lie between 1.5 and 1.8, which are almost the same as those for the unadjusted case. We should note here that the variance estimator 1 is computed with 19 degrees of freedom (1 per stratum). On the other hand, in the case of the 3 alternatives we have 81 degrees of freedom since each PSU is a replicate. Hence, we conclude that the stability of the Keyfitz variance estimator for the ratio-adjusted estimates is significantly improved by increasing the number of replicates and becomes comparable with the other two alternatives (see Table 3B).

2.4 Keyfitz's Variance Estimators with 2 vs. 6 Replicates for the LFS

The results of the Monte Carlo study reported in the previous sub-section have shown that the Keyfitz variance estimator compares well with the alternate methods for the variances of the ratio-adjusted estimates both from the bias and efficiency point of view when each method uses the same number of replicates. In addition, Keyfitz's method has the advantage of simplicity and estimating the variances of changes and averages under the alternative methods involves many complications. Therefore, the Keyfitz method was retained for the SR areas as well. In order to improve the efficiency of Keyfitz's method, 6 rotation panels were adopted as replicates as opposed to 2 replicates in the old design. One major concern with using the rotation panels as replicates was whether there would be any serious inflation of the variance estimate due to panel bias.

This aspect was investigated for the three LF characteristics by computing the variance estimates using the variance formula developed in Section 3 with 2 and 6 replicates from the actual LFS data for 24 months (March '85 - February '87). From the 24 estimated variances for each of the LF characteristics, the means and standard deviations (SD's) of the variances were obtained. The ratios of the means and SD's of the variances under the two alternatives (2 vs. 6 replicates) are averaged over 24 Census Metropolitan Areas (CMA's) and given in Table 5. The following observations can be made from the table:

Table 5

Comparison of SR Variance Estimates with 2 vs. 6 Replicates per Stratum Based on CMA Data of the LFS Mar '85 - Feb '87

Characteristic	Average Ratio of Means of Variances (2 vs. 6)	Average Ratio of SD's of Variances (2 vs. 6)
Employed	0.997	1.813
Unemployed	0.995	1.515
In Labour Force	1.003	1.833

Note: For each CMA, means and standard deviations of variance estimates were obtained from 24 months data for 2 and 6 replicates. Then the ratios (2 rep. vs. 6 rep.) of means of variances and of standard deviations (SD's) of variances were calculated for each CMA. The average ratios in the table are the averages over 24 CMA's.

(i) The effect on the levels of the variances due to using 6 replicates as compared to 2 is very minimal, which means that adopting rotation panels as replicates has little impact on the bias of the variance estimates.

(ii) As expected, the variances are more stable with 6 replicates than with 2 and the results are not much different from those of the Monte Carlo study (see the first column in Table 3B)

From the above observations, we conclude that the efficiency of the Keyfitz method is improved substantially without having serious impact on the bias by adopting the 6 rotation panels as replicates as opposed to using only 2 replicates.

3. VARIANCE ESTIMATION FOR RAKING RATIO ESTIMATES

3.1 Raking Ratio Estimation for the LFS

In the old LFS, post-stratified ratio estimation was used. The subweight, which is the design weight adjusted for non-response, was ratio-adjusted to external estimates of the LFS target population for 38 post-strata defined by age and sex at provincial level. The LFS target population is the population 15 years of age and over excluding armed forces, inmates of institutions, and population living on Indian reserves.

This ratio estimation enhanced the quality of provincial data substantially but subprovincial data still had somewhat poor reliability. In order to improve subprovincial data especially for Economic Regions (ER's) and Census Metropolitan Areas (CMA's), a raking ratio estimation procedure was adopted, through which simultaneous ratio adjustment at provincial and subprovincial levels is achieved.

The raking procedure is carried out in a sequence of adjustments: first, the subweight is adjusted to the subprovincial (CMA's and Non-CMA parts of ER's) population and then the provincial level adjustment by age / sex (the number of age / sex groups were reduced from 38 to 24 in the redesigned sample) is applied to the resulting weight. This procedure is repeated once more to obtain a second pair of weights. Note that for the ER's containing CMA(s), the CMA part is excluded when defining adjustment cells for the ER's so that the subprovincial adjustment cells are mutually exclusive. Let W_0 be the subweight and let (W_1, W_2) and (W_3, W_4) be the two pairs of weights resulting from the first and second iteration respectively. Labour force characteristics are estimated using W_4. Due to the order of adjustments, the marginal totals of W_4 at provincial age / sex groups are exactly the same as the external population estimates of the corresponding groups but the marginal totals of W_4 at

subprovincial level (ER and CMA) are not quite equal to the corresponding external population estimates. However, the differences are very small.

The special area frames, which are composed of military establishments, institutions, and remote areas, in general, do not respect the ER and CMA boundaries and hence, are treated differently during the raking procedure. Each special area type forms a stratum at the provincial level. The only exceptions are remote areas in the provinces of Quebec and Alberta where further stratification is carried out. Those ER's and CMA's which contribute to the special area frame will be called "contributing" ER's and CMA's. The special area records on the sample file are copied to each of the contributing ER's or CMA's with deflated subweights in proportion to the population of that particular type of special area in the contributing ER or CMA. The raking procedure is then carried out in the usual manner as described earlier.

3.2 Variance Formula for One-Iteration Raking Ratio Estimates

The variance formula for one-iteration raking ratio estimates is derived here. The basic methodology employed here is successive application of Taylor series approximation to the raking ratio estimates until we obtain a linear form of subweights. Then the replication formula is applied as in Woodruff (1971). The successive application of the Taylor series approximation was also used by Arora and Brackstone (1977a,b) and Brackstone and Rao (1979) to obtain variance formula of raking ratio estimates for simple random sampling of units or clusters. We have adopted this method for the stratified multi-stage PPS sampling design following Woodruff's approach.

Let $Y^{(0)}$, $Y^{(1)}$, $Y^{(2)}$ be the estimates of a labour force characteristic y in a province based on W_0, W_1, and W_2, respectively. The superscripts in parentheses correspond to the subscripts of W's.

Then $Y^{(2)}$ can be expressed as follows:

$$Y^{(2)} = \sum_a \frac{Y_a^{(1)}}{P_a^{(1)}} P_a \tag{3.1}$$

where $Y_a^{(1)}$ = W_1-weighted estimate of characteristic y for the age/sex group a in the province,

$P_a^{(1)}$ = W_1-weighted estimate of population for the age/sex group a in the province,

P_a = External estimate of population for the age/sex group a in the province.

Let $F_a = Y_a^{(1)}/P_a^{(1)}$. The first order Taylor approximation to F_a at $(E(Y_a^{(1)}), E(P_a^{(1)}))$ is

$$F_a \doteq \frac{E(Y_a^{(1)})}{E(P_a^{(1)})} + \frac{1}{E(P_a^{(1)})}\left\{Y_a^{(1)} - E(Y_a^{(1)})\right\} - \frac{E(Y_a^{(1)})}{\{E(P_a^{(1)})\}^2}\left\{P_a^{(1)} - E(P_a^{(1)})\right\}$$

where E denotes expectation.

Then a Taylor approximation to the variance of $Y^{(2)}$ can be written as

$$V(Y^{(2)}) = V\left(\sum_a F_a P_a\right) \doteq V\left\{\sum_a \frac{P_a}{E(P_a^{(1)})}(Y_a^{(1)} - R_{Ya}^{(1)} P_a^{(1)})\right\} \tag{3.2}$$

where

$$R_{Ya}^{(1)} = \frac{E(Y_a^{(1)})}{E(P_a^{(1)})}.$$

Now the W_1-weighted estimates $Y_a^{(1)}$ and $P_a^{(1)}$ can be expressed in terms of W_0-weighted estimates as follows:

$$Y_a^{(1)} = \sum_s \frac{Y_{sa}^{(0)}}{P_s^{(0)}} P_s,$$
$$P_a^{(1)} = \sum_s \frac{P_{sa}^{(0)}}{P_s^{(0)}} P_s, \tag{3.3}$$

where s denotes a CMA or an ER or the complementary part of an ER after removing the CMA part and P_s is population of the subprovincial area s. Substituting the expressions for $Y_a^{(1)}$ and $P_a^{(1)}$ from (3.3) into (3.2) and applying the first order Taylor approximation to the ratios of W_0-weighted estimates, we obtain

$$V(Y^{(2)}) \doteq V\left[\sum_a \frac{P_a}{E(P_a^{(1)})} \sum_s \frac{P_s}{E(P_s^{(0)})} \left\{\left(Y_{sa}^{(0)} - R_{Ysa}^{(0)} P_s^{(0)}\right)\right.\right.$$
$$\left.\left. - R_{Ya}^{(1)} \left(P_{sa}^{(0)} - R_{Psa}^{(0)} P_s^{(0)}\right)\right\}\right], \tag{3.4}$$

where

$$R_{Ysa}^{(0)} = \frac{E(Y_{sa}^{(0)})}{E(P_s^{(0)})} \text{ and } R_{Psa}^{(0)} = \frac{E(P_{sa}^{(0)})}{E(P_s^{(0)})}.$$

The expression in (3.4) can be written in terms of replicate level estimates. Define

$$Z_{Yshia}^{(0)} = \frac{P_a}{E(P_a^{(1)})} \frac{P_s}{E(P_s^{(0)})} (Y_{shia}^{(0)} - R_{Ysa}^{(0)} P_{shi}^{(0)}),$$
$$Z_{Pshia}^{(0)} = \frac{P_a}{E(P_a^{(1)})} \frac{P_s}{E(P_s^{(0)})} (P_{shia}^{(0)} - R_{Psa}^{(0)} P_{shi}^{(0)}), \tag{3.5}$$

where h denotes a stratum belonging to s and i denotes a replicate in h.

Then (3.4) can be rewritten by rearranging the order of summations as follows:

$$V(Y^{(2)}) \doteq V\left\{\sum_s \sum_{h \in s} \sum_{i=1}^{n_h} \sum_a \left(Z_{Yshia}^{(0)} - R_{Ya}^{(1)} Z_{Pshia}^{(0)}\right)\right\}$$
$$= V\left(\sum_s \sum_{h \in s} \sum_{i=1}^{n_h} D_{shi}^{(0)}\right) \tag{3.6}$$

where

$$D_{shi}^{(0)} = \sum_a \left(Z_{Yshia}^{(0)} - R_{Ya}^{(1)} Z_{Pshia}^{(0)}\right).$$

Apart from special area strata, $(\Sigma_{i=1}^{n_h} D_{pshi}^{(0)})$'s are independent because they are based on subweights. However, for the special area strata they are highly correlated because the same records are attributed to the contributing subprovincial areas.

We can rewrite (3.6) as

$$V(Y^{(2)}) \doteq V\left(\sum_{h\epsilon s}\sum_{h}\sum_{i=1}^{n_h} D_{shi}^{(0)}\right)$$

$$= V\left(\sum_{h}\sum_{i=1}^{n_h}\sum_{s\ni h} D_{shi}^{(0)}\right) \tag{3.7}$$

where $\Sigma_{s\ni h}$ is summation over all the subprovincial areas containing the stratum h. For a non-special stratum, the stratum appears only in one subprovincial area, and the summation $(\Sigma_{s\ni h})$ is redundant. However, a special area stratum could appear in several subprovincial areas and the summation $(\Sigma_{s\ni h})$ sums up all D-values $(D_{shi}^{(0)})$. belonging to the special area stratum.

Define

$$D_{hi}^{(0)} = \sum_{s\ni h} D_{shi}^{(0)}.$$

Then (3.7) becomes

$$V(Y^{(2)}) \doteq V\left(\sum_{h}\sum_{i=1}^{n_h} D_{hi}^{(0)}\right). \tag{3.8}$$

The variables, $\Sigma_i D_{hi}^{(0)}$, are independent since they are based on subweights. Then, ignoring the fpc, the variance can be estimated by

$$\hat{V}(Y^{(2)}) \doteq \sum_{h} \frac{n_h}{n_h-1}\sum_{i=1}^{n_h} (D_{hi}^{(0)} - \bar{D}_h^{(0)})^2 \tag{3.9}$$

where

$$\bar{D}_h^{(0)} = \frac{1}{n_h}\sum_{i=1}^{n_h} D_{hi}^{(0)}.$$

In this expression, however, expected values are involved and these are unknown. The variance can be approximated reasonably well by substituting expected values with their estimates and hence, from (3.9), we obtain the final form of $\hat{V}$ as follows:

$$\hat{V} \doteq \sum_{h} \frac{n_h}{n_h-1}\sum_{i=1}^{n_h} (D_{hi}^{(2)} - \bar{D}_h^{(2)})^2 \tag{3.10}$$

where

$$D_{hi}^{(2)} = \sum_{s \ni h} D_{shi}^{(2)},$$

$$\bar{D}_h^{(2)} = \frac{1}{n_h} \sum_{i=1}^{n_h} D_{hi}^{(2)},$$

$$D_{shi}^{(2)} = \sum_a \left(Z_{Yshia}^{(2)} - R_{Ya}^{(2)} Z_{Pshia}^{(2)} \right),$$

$$Z_{Yshia}^{(2)} = \frac{P_a}{P_a^{(1)}} \frac{P_s}{P_s^{(0)}} \left(Y_{shia}^{(0)} - \frac{Y_{sa}^{(0)}}{P_s^{(0)}} P_{shi}^{(0)} \right)$$

$$= Y_{shia}^{(2)} - \frac{P_{shi}^{(0)}}{P_s^{(0)}} Y_{sa}^{(2)},$$

$$Z_{Pshia}^{(2)} = \frac{P_a}{P_a^{(1)}} \frac{P_s}{P_s^{(0)}} \left(P_{shia}^{(0)} - \frac{P_{sa}^{(0)}}{P_s^{(0)}} P_{shi}^{(0)} \right)$$

$$= P_{shia}^{(2)} - \frac{P_{shi}^{(0)}}{P_s^{(0)}} P_{sa}^{(2)},$$

and

$$R_{Ya}^{(2)} = \frac{Y_a^{(1)}}{P_a^{(1)}} = \frac{P_a}{P_a^{(1)}} \frac{Y_a^{(1)}}{P_a} = \frac{Y_a^{(2)}}{P_a}.$$

The formula (3.10) gives the variance for W_2-weighted estimates of LF characteristics and requires two weights W_0 and W_2.

3.3 Application of the One-Iteration Variance Formula to Two-Iteration Raking Ratio Estimates

The variance formula for the two-iteration raking ratio estimates can be obtained by successive application of the Taylor linearization as described in the previous section. However, the formula thus obtained is very complex. It was conjectured that the variance formula for one-iteration would be a reasonably good approximation for estimating the variance of the two-iteration raking ratio estimates. The rationale behind this conjecture was that there were only small perturbations in the weights after the first iteration. Now, the one-iteration variance formula uses the pair of weights (W_0, W_2). However, it was decided to use (W_0, W_4) instead of (W_0, W_2) since it was found that the use of W_4 instead of W_2 does not have any impact on the CV's of LF estimates which are based on W_4. The one-iteration variance

formula using the pair of weights (W_0, W_4) will be referred to as the one-iteration variance estimator.

To verify our conjecture, a Monte Carlo simulation study was carried out using the 1981 Census data from the province of Nova Scotia. In each Monte Carlo sample, the LFS design was simulated through all stages of sampling and a total of 1,000 Monte Carlo samples were selected independently. For each Monte Carlo sample, the following statistics were calculated for the three labour force characteristics at subprovincial and provincial levels;

1. Two-iteration raking ratio estimate, $Y^{(4)}$.
2. Variance estimate $\hat{V}(Y^{(4)})$ using the one-iteration variance estimator and the corresponding estimate of CV.
3. 95% confidence interval (i.e., $Y^{(4)} \pm 1.96\sqrt{\hat{V}(Y^{(4)})}$).

At the end of simulation, the average of 1,000 CV's was computed and compared with the Monte Carlo CV which is very close to the true value. The results are given in Table 6A. In all 21 cases (3 characteristics for each of 7 areas) the differences are less than 8% and in 13 cases less than 4%.

Also, the proportion of confidence intervals which cover the true characteristic value was obtained. The results are shown in Table 6B. Coverage rates for Employed and In Labour Force are very close to the nominal value in general, whereas those for Unemployed are somewhat lower but still acceptable.

It was also found that the two-iteration raking ratio estimate is nearly unbiased with a maximum of 0.35 percent bias in all 21 cases.

Table 6A

Average CV's Obtained by the One-Iteration Variance Estimator and the Monte Carlo CV's

Characteristic	ER 210	ER 220	ER 230	ER 240	ER 250	CMA Halifax	Province (Nova Scotia)
				Average CV's			
Employed	3.52	3.46	3.14	3.05	1.96	2.01	1.08
Unemployed	10.36	12.28	13.13	13.43	10.35	10.55	5.27
In Labour Force	2.98	3.17	2.85	2.73	1.77	1.83	0.91
				Monte Carlo CV's			
Employed	3.48	3.35	2.95	2.86	1.97	1.99	1.11
Unemployed	10.90	12.71	13.28	13.37	11.12	11.31	5.59
In Labour Force	2.76	3.08	2.76	2.53	1.72	1.74	0.92

Table 6B

Coverage Rates of 95% Confidence Intervals Constructed by the One-Iteration Variance Estimator

Characteristic	ER 210	ER 220	ER 230	ER 240	ER 250	CMA Halifax	Province (Nova Scotia)
Employed	94.5	92.8	94.0	94.7	94.7	94.9	92.5
Unemployed	92.1	90.7	91.4	91.8	92.7	92.7	93.1
In Labour Force	96.2	93.0	93.6	95.2	95.2	96.0	94.0

4. CONCLUSIONS

It has been shown that the Keyfitz variance estimation method for estimates without ratio adjustment (in this case it becomes just a replication method) has very large positive biases and low efficiencies while the alternatives have negligible biases and higher efficiencies for the labour force characteristics considered in this study.

However, for the ratio-adjusted estimates, all the methods considered here have negligibly small biases. It has also been shown that the efficiency of the Keyfitz method can be improved substantially and made comparable to the alternatives by increasing the number of replicates. It was demonstrated using actual LFS data that using 6 rotation panels as replicates in the Keyfitz variance estimator as opposed to 2 pseudo replicates does not introduce bias due to the phenomenon of rotation panel bias. As shown by Monte Carlo results, the one-iteration variance formula derived by the Keyfitz method using Taylor linearization gives reasonably good variance estimates for the two-iteration raking ratio estimates and has good coverage properties.

ACKNOWLEDGEMENT

The authors are grateful to the two referees, an Associate Editor and the Editor for their useful comments on the earlier version of the paper.

REFERENCES

ARORA, H.R., and BRACKSTONE, G.J. (1977a). An investigation of the properties of raking ratio estimators: I with simple random sampling. *Survey Methodology,* 3, 62-83.

ARORA, H.R., and BRACKSTONE, G.J. (1977b). An investigation of the properties of raking ratio estimators: II with cluster sampling. *Survey Methodology,* 3, 232-252.

BRACKSTONE, G.J., and RAO, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyā,* Series C, 41, 97-114.

KEYFITZ, N. (1957). Estimates of sampling variance where two units are selected from each stratum. *Journal of the American Statistical Association,* 52, 503-510.

PLATEK, R., and SINGH, M.P. (1976). Methodology of the Canadian Labour Force Survey. Catalogue No. 71-526, Statistics Canada.

RAO, J.N.K. (1975). Unbiased variance estimation for multi-stage designs. *Sankhyā,* Series C, 37, 133-139.

RAO, J.N.K., HARTLEY, H.O., and COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society,* 24, 482-490.

SINGH, M.P., and DREW, J.D. (1981). Redesigning continuous surveys in a changing environment. *Survey Methodology,* 7, 44-73.

WOODRUFF, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association,* 66, 411-414.

Survey Methodology, December 1987
Vol. 13, No. 2, pp. 163–171
Statistics Canada

The "AGEVEN" Record: A Tool for the Collection of Retrospective Data

PHILIPPE ANTOINE, XAVIER BRY and PAP DEMBA DIOUF[1]

ABSTRACT

Because it is easy to use, the "AGEVEN" record makes it possible to date events more precisely and to classify retrospectively demographic events (births and deaths), changes in marital status and changes in place of residence. The data collected are used to accurately recreate the socio-economic conditions that were present when the demographic events being studied took place.

KEY WORDS: Retrospective survey; Biographies; Demographic survey.

1. INTRODUCTION

Two major data collection methods are available to demographers to collect data on natural movement (natality and mortality): longitudinal observation and retrospective questionnaires. The longitudinal observation method (following a population sample over a relatively long period of time) is, in theory, the method which provides the most accurate results. It does, however, have its drawbacks. It is expensive because of the amount of travel required for observation, and a relatively lengthy period of time is needed to obtain results. Finally, in urban areas, the method is difficult to apply because of the high degree of mobility of the population, which leads to a significant deterioration of the sample, such as that encountered in IFORD's infant and child mortality surveys (Scott 1985; Fargues 1985).

The retrospective method gives less reliable results because it depends more on the memory of the respondents. However, the total observation period is generally longer than that of the longitudinal surveys introduced in recent years in African countries. The risk of omitting events remains high and dating them is inaccurate. Finally, in urban areas there is a tendency when reconstituting the past to mix events which took place in the city being surveyed with other, earlier events, which took place in other places of residence (urban or rural).

Since we wished to determine mortality and fertility differences in Pikine, a suburb of Dakar, and also wished to obtain fairly reliable results quickly, we selected a data collection method that would enable us to recreate accurately the infant and child mortality risk factors at the time of death of each of the children of the women surveyed. The survey was conducted jointly by the Senegal Statistics Branch and Orstom (Antoine et Diouf 1986). The field work was carried out between March and May 1986. The first results were available in September 1986. The method we selected is different from the retrospective method most frequently used, which takes into account only the socio-economic and cultural characteristics of the women at the time of the survey. These characteristics could, in fact, have changed considerably during the women's child-bearing years (improvement or deterioration of living conditions, change of marital status, change of activity, and so forth). Our method makes it possible to better assess the relationship between urban insertion and changes in demographic behaviour. The following objectives determined our collection strategy:

- to obtain a complete list of the events observed (mainly births and deaths);

[1] Philippe Antoine, demographer, and Xavier Bry, statistician, ORSTOM, P.O. Box 1386, Dakar, Senegal; Pap Demba Diouf, demographer, Statistics Branch, P.O. Box 116, Dakar, Senegal.

- to date these events as accurately as possible;
- to place the events in their socio-economic context (marital status, professional status of the husband and wife, living conditions).

2. COLLECTION AND DATING OF DEMOGRAPHIC EVENTS

To conduct a successful retrospective survey means, in particular, establishing as accurate a biography as possible (in relation to the field studied) for each person surveyed. A method has to be found, therefore, to situate past events chronologically.

A number of methodological improvements have been proposed in the past. Ferry (1977) used an "event file", which involved assigning a record to each event. According to the author, the originality of this method lay in placing the events in order together with the person surveyed (pregnancies, marriages and divorces, places of residence and so forth) and situating them in relationship with each other. The technique consisted in recreating, with the person surveyed, the succession, logic, interferences and, finally, the individual biography. However, it is a relatively complex method and involves handling numerous records in the field and during processing.

Another method of classifying and dating events was used in the Senegalese survey on fertility in 1978: the "AGEVEN" graph. There were two reasons for using the "AGEVEN" graph in the Senegalese survey:
- to make it possible to better estimate the age of the women and their children with the help of relatively precise dating;
- to make it possible to accurately estimate fertility by preparing the pregnancy histories of all the women.

The "AGEVEN" graph used in the Senegalese fertility survey (Figure 1) plots two curves. The righthand curve describing the lifeline of the woman (LL curve) is graduated in intervals of three months, making it possible to plot inside a year the events affecting the woman. The lefthand curve, called the AE (age of events) curve, indicates the time which has passed between the event and the date of the survey. Thus, an age on the AE curve corresponds to each year on the LL curve, and vice versa. This graph, which was also used in the Ivory Coast fertility survey, seems to be mainly an instrument for dating events.

3. USE OF THE "AGEVEN" RECORD IN THE PIKINE SURVEY

We tried to combine some of the advantages of each of these collection methods: the "AGEVEN" graph, which is easy to use to date events, and the event file, which makes it possible to take various kinds of events and to classify them in relation to each other. We systematized the "AGEVEN" record by distinguishing between demographic events (births, deaths), changes in marital status and changes in place of residence. For convenience, we retained the name given the graph used in the Senegalese fertility survey for our record, but while the name is the same, the uses which can be made of it are different. The "AGEVEN" record (see Figure 2) contains three columns:
- the first covers demographic events (births (B); deaths (DT); abortions (A); miscarriages (MC); stillbirths (SB)). Each event (birth or death) must be followed by its chronological ranking, the first and last names of the child and, possibly, the exact date;
- the second column covers matrimonial events and the chronological ranking of each of the spouses or partners (marriages (M); divorces (D); widowhood (W), the rank of the various fathers (indicated as F1, F2, ... Fn).

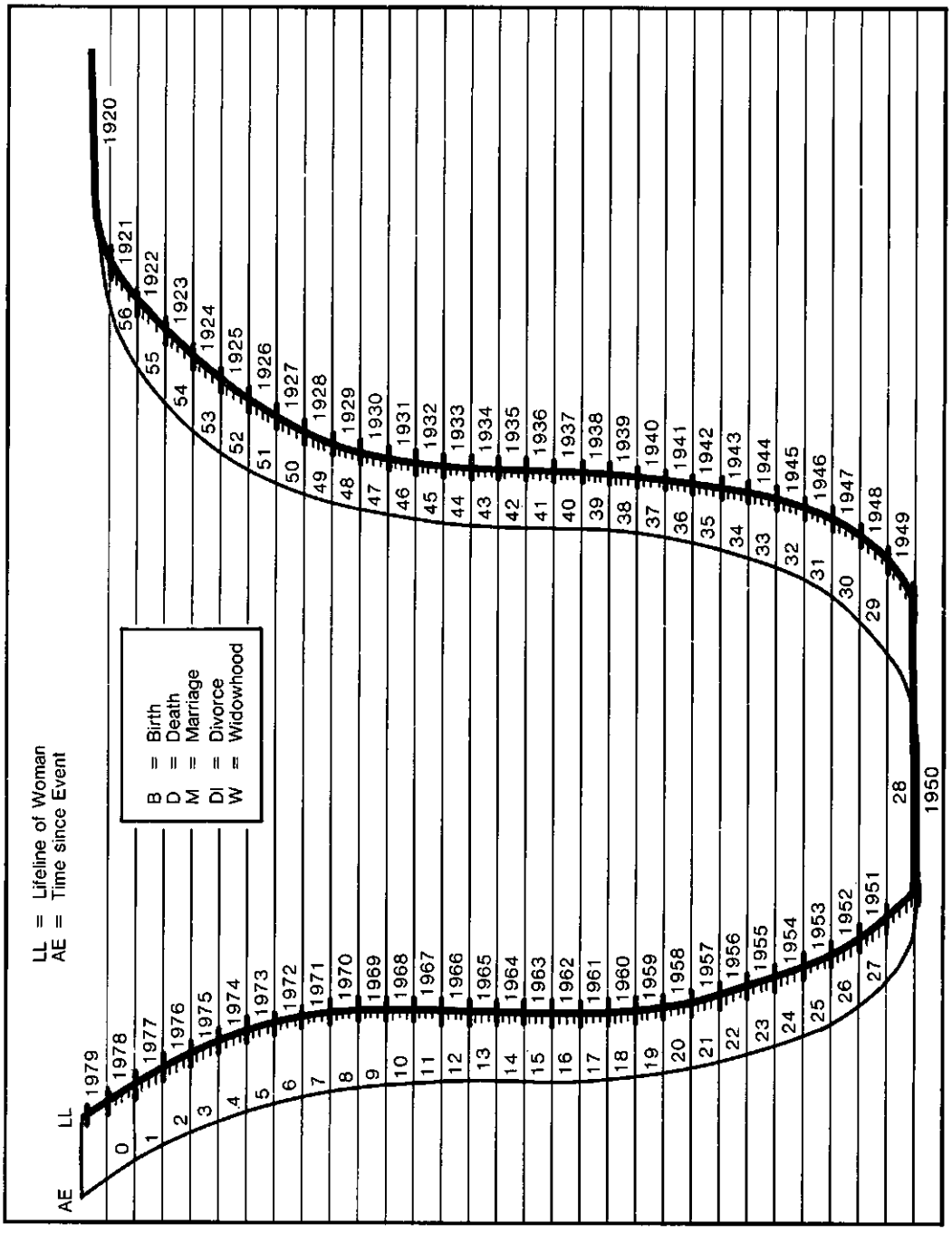

Figure 1. AGEVEN Graph Used in the Senegalese Fertility Survey

Year	Age	Demographic Events	Matrimonial Events	Places of Residence
1986	March			
1985	0			
1984	1	B_3 Aminata 18-12-84	F_2	Pikine
1983	2			
1982	3		M_2	Pikine
1981	4			
1980	5			
1979	6	D_1 Ibrahima at age 4	DI_1	Pikine
1978	7	B_2 Abdoul 5-01-78	F_1	Dakar
1977	8			
1976	9			
1975	10	B_1 Ibrahima	F_1	Dakar
1974	11			
1973	12		M_1	Thiès
1972	13			
1971	14			
1970	15			
1969	16			
1968	17			
1967	18			
1966	19			
1965	20			
1964	21			
1963	22			
1962	23			Kaolack
1961	24			
1960	25			
1959	26			
1958	27			

AGEVEN

BNR/ORSTOM

CODES
B_i + date + name + place = birth child no. i
D_i + date + name + place = death child no. i
M_i + date + name + place = marriage no. i
DI_i + date + name = divorce spouse no. i
V_i + date + name = widowhood spouse no. i
MC = miscarriage A = abortion SB = stillbirth
F_i = father no. i

IDENTIFIER
Name:
Block:
Concession:
Household:
Woman No.:

Year	Age	Demographic Events	Matrimonial Events	Places of Residence
1936	49			
1937	48			
1938	47			
1939	46			
1940	45			
1941	44			
1942	43			
1943	42			
1944	41			
1945	40			
1946	39			
1947	38			
1948	37			
1949	36			
1950	35			
1951	34			
1952	33			
1953	32			
1954	31			
1955	30			
1956	29	BW Awa		
1957	28			Kaolack

Figure 2. Example of use of the "AGEVEN" record.

- the third column indicates the place of residence at the time of each of these demographic and matrimonial events. This column makes it possible to follow the migratory paths of the women and to determine the date of their arrival in Pikine.

The "AGEVEN" record is a methodological tool that serves various purposes:
- situating events chronologically;
- helping the woman situate chronologically events for which she has forgotten the date;
- ensuring that all the demographic events lived by the woman surveyed are recorded;
- identifying changes of residence and the location where events took place;
- checking the consistency of events among themselves.

The interview consists of two phases: one involving the household and the other involving the women between the ages of 15 and 49. The "household" questionnaire, which lists all members of the household, whether currently residing in the household or not, deals in particular with the filiation of the persons surveyed, their blood relationship with the head of the household or "nucleus," their sex, their marital status, and their date of birth or age. The "women's" questionnaire concerns all the women, resident and present in the household, between the ages of 15 and 49. The "AGEVEN" record is used to complete this questionnaire.

To transcribe the data collected on this record, the investigator can take various points of reference (the date of birth of the woman, the date of birth of her first child, and so forth) and, with the help of the respondent, reconstitute her entire lifeline, namely all the other events which have taken place during her life, such as marriage, divorce, and various pregnancies. This operation may be broken down as follows:

1. After recording the first live birth, the investigator asks the respondent to state all subsequent live births, in chronological order, indicating whether or not the child is still alive and whether or not he or she is still living in the household.
2. The investigator then records these births on the record, using the official documents shown to him. In our case, official documents were available mainly for children born in the Dakar area. For the age of the women, however, as well as for the birthdates of some children, the investigator has to rely on elements in the historical calendar to determine the dates (month and year).

The "AGEVEN" record makes it possible to situate events according to the age of the woman at the time of the event, the time which has passed since the event took place, or the date of the event. Any large gap between two births or other inconsistency between two events is easily detected during the interview with the woman.

It is also possible to use the "AGEVEN" record to check the consistency of events. For example, two children cannot be born within nine months of each other; a woman cannot say that she was married at age 12 and had her first child in 1970 at age 14, and then go on to say that she was born in 1950. In the latter case, there is likely an error in the date of birth of the woman and it should be corrected.

The record makes it possible to record both events for which an exact date is given and events for which only an age is given (such and such a child is now ten years old; I was married 15 years ago). Finally, with the help of this record, events for which the date is not clear can be situated. For example, such and such a child was born between the one born on 10-2-74 and the one born in 1978. It is highly likely that this child was born in 1976. To use this record successfully, the investigator must take a critical look at the chain of events and must try to make it as complete as possible, taking care to check the reliability and consistency of the responses provided. This is possible only if confidence is established in the dialogue with the respondent.

After having recorded all the live births declared by the respondent, the investigator turns to the intervals between successive births. All events are not always reported in the initial responses, but by using the "AGEVEN" record, the investigator can track down the

omitted events. The investigator thus asks himself what happened each time an interval of more than two years is recorded between two live births. The responses provided by the respondent may reveal abortions, stillbirths, death soon after birth, information obtained on contraceptives, and so forth. Although this was not an objective of the Pikine survey, the dialogue that is established can make it possible to delve deeper into matters relating to family planning.

Each of the events is linked to the location, marital status and partner of the woman at the time of the event. After recording all the events affecting the woman, the investigator then has to estimate more accurately the date of birth of the mother. The investigator has in fact already recorded the date of birth of the mother, as indicated either by the woman or the head of the household, when completing the "household" questionnaire. Now, in a one-on-one interview with the respondent and having recorded the events which affected her, he can provide the best possible estimate of the respondent's age.

For example. Awa was born in 1956 in Kaolack. She says that she has had three children: Ibrahima, who would now be 10 years old, born in Dakar, died at age 4 in Pikine; Abdoul, born on January 5, 1978 in Dakar; and Aminata, born on December 18, 1984 in Pikine. Awa was married for the first time at age 17 in Thies. She was divorced in 1979 (while living in Pikine). She remarried in 1982, at which time she was still living in Pikine (see Figure 2). During the interview, the investigator will notice a gap of almost 7 years between Abdoul and Aminata. He should ask whether there were other births or pregnancies during this period. In the case of Awa, the divorce and subsequent remarriage three years later may explain the gap. However, the investigator must check with the woman to ensure that the gap does not hide other demographic events.

The interactive form of the interview seems to encourage dialogue with the respondent and improves contact between the investigator and respondent, which is unfortunately only too often clouded by doubt on the part of the investigator and mistrust on the part of the respondent Bonnet (1984). As the investigator continues his or her investigation, new events are mentioned. When he or she asks whether there was another event between two births separated by more than two years, the respondent is often surprised and responds in one of two ways. If no event has occurred, she asks, "Why do you ask that?" If, however, an event has indeed occurred, she often asks, "Who told you that?" since she has the impression that the investigator already knows something. The "AGEVEN" record becomes a kind of crystal ball, like the cowry shell. Sometimes the interview becomes a game, and the respondent is pleased to place past events in order. A woman with a complicated marital and reproductive history may even want a copy of her "AGEVEN" record. As in any survey, there are problems with the use of this record. Sometimes it is difficult or awkward to be alone with the respondent, and often women are embarrassed if the record brings up events concerning a partner preceding the current husband.

In practice, the record is incomplete because there is no question which eliminates possible confusion between stillbirths and infants who die shortly after birth. This kind of confusion often arises in responses given in the Wolof language, in which it is difficult to distinguish between miscarriages and abortions and between stillbirths and deaths immediately after birth. Some French terms or words cannot be translated directly into Wolof. For stillbirths, for example, there is no single question that elicits the desired response. At least two questions are therefore required. When confronted with an interval between successive births, the investigators asks the following question, for example: "Lou am dikhane té Moussa ak Ali?" (what happened between Moussa and Ali?). This question correctly leads the women to stillbirths, abortions, miscarriages and so forth. To elicit a satisfactory response, clarifications are needed: "Dikhane té Moussa ak Ali, amo fi dom diou dé guinaw bou mou indé bakhane?" (did you have a child who died after giving some sign of life between Moussa and Ali?). The confusion results mainly from the fact that the distinction between a miscarriage and stillbirth is not always clear and from the fact that a child is not given a name

until he or she is a week old. Also, for certain ethnic groups, it is not until the child has a name that he or she is really taken into account. A column indicating whether or not the infant cried at birth would therefore have been very useful.

The "AGEVEN" record used in the Pikine survey did indeed provide more satisfactory data than the graph used in the Senegalese fertility survey, in terms of both the nature and quantity of data collected. However, it did not eliminate the tendency to round off the intervals between successive births in years (approximately 37% of the intervals), particularly in intervals of two years, which account for approximately 20% of the intervals observed between successive births. In addition, it was not possible, using this technique, to list all the issue of young girls who had been pregnant but who had had no live births. Some biases, which are certainly classic in demography, do persist therefore, and this method does not eliminate the need to take extreme care in the field.

4. TRANSCRIPTION FROM THE "AGEVEN" RECORD TO THE QUESTIONNAIRE AND ELECTRONIC DATA PROCESSING

The questionnaire regarding the reproductive history of the women was designed in such a way as to permit the best possible transcription of the data collected using the "AGEVEN" record. First, the characteristics of each of the children are noted in chronological order by birth, along with the date of death, if appropriate. The investigator then records the marital status at the time of each of these events in order to note any possible change in spouse. Then, changes in the socio-economic situation of the father and mother are taken into account, as well as changes in living conditions and in place of residence. The survey also included other questionnaires regarding the characteristics of the household, individuals and women observed.

The data collection method allows for two kinds of analysis. The first involves a classical analysis of mortality by generation and sub-population (according to neighbourhood, type of housing and so forth). However, what is especially interesting about this study is that it allows for analysis of mortality (and fertility) taking into account migratory behaviour and changes in the socio-economic conditions of the women surveyed. When this method is used, mortality is no longer interpreted solely according to the socio-economic conditions at the time of the survey. Rather, it is related to the conditions which really existed at the time of the event, and it is therefore possible to better understand the differences relating specifically to living conditions in urban areas (Pikine in this case).

Depending on the place of birth of the child, different morality rates were recorded. Many of the respondents are migrant women from other cities or from villages in the interior of the country. Children born to them in rural areas suffered a significantly higher risk of mortality than those born in the Dakar area.

The child mortality rate (between 1 and 4 years) clearly reveals the risks resulting from socio-economic differences. The risk of dying between the ages of 1 and 4 is 2.84 times higher for children born in villages than for those born in Pikine. The z-test shows that the difference between the two rates (Pikine mortality rate and rural mortality rate) is significant. We tested the hypothesis that the mortality rate for children born in Pikine is the same as that for children born in rural areas. Since the sample sizes are relatively large, approximation using the normal distribution is justified. Under the hypothesis that the mortality rates

Table 1

Mortality by place of birth (in thousands)

	Pikine	Dakar	Other Cities	Rural	Total	Pkn-Rural Test
Infants	52	57	45	114	58	-6,586**
Children	55	62	90	156	68	-10,093**
Population	5155	1513	644	704	8016	

are equal, the z-statistic is distributed as a standard normal variable. The symbol "**" indicates a significant difference at the $\alpha = 0.05$ level. Classic restrospective data collection without distinction as to the place of birth of the child would have led us to class births outside Pikine with those inside Pikine and would have resulted in a higher mortality rate (child mortality rate of 68 per thousand rather than 55 per thousand).

Moreover, a second analysis can be made for each of the women observed. A simplified biographical file can be created in which the successive stages are defined in terms of births. A relationship is thus established between matrimonial events, changes in residence and reproductive data. The principal stages in the migratory path followed since the birth of the first child, or since marriage, can also be reconstructed. Longitudinal data gathered in this way lend themselves very well to recent methods for the analysis of interference between phenomena (Courgeau and Lelievre (1986); Cox and Oakes (1984)).

5. CONCLUSION

The data collected for each of the variables are very brief, but they should make it possible to detect some significant differences and to determine the living conditions at the time of birth and death. The collection methodology used is adapted to the collection of data on the reproductive histories of the women and the destiny of their children. The main advantage of the "AGEVEN" record is its facility in pinpointing various events chronologically and in classifying these events in relationship with each other, without eliminating the possibility of inserting events omitted as the interview proceeds. The flexibility of the "AGEVEN" record leads us to suggest that it could be used in other fields, for professional biographies or migratory routes, for example, by establishing a parallel between place of residence, profession, marital status, family situation, living conditions and so forth. A great deal of methodological research has been conducted in the analysis of demographic biographies (Courgeau 1984; Haeringer 1972; Riandey 1985). Our method is intended merely as a simple and reliable tool for the collection of data. It is up to each user to determine which variables he or she wishes to arrange chronologically using the "AGEVEN" record and, once the biographical framework has been collected, to obtain more data on the field(s) he or she is studying, using the questionnaire.

ACKNOWLEDGMENTS

The authors would like to thank the referees for their helpful comments.

REFERENCES

ANTOINE, Ph., and DIOUF, P.D. (1986). Changements démographiques en milieu urbain. Paper presented at Séminaire sur la mortalité au Sénégal. Dakar.

BONNET, D. (1984). Occultation, omissions. Quelques problèmes soulevés par l'enquête quantitative en matière de santé. *Medicus Mundi,* 11.

COURGEAU D. (1984). Relations entre cycle de vie et migrations. *Population,* 39, 483-513.

COURGEAU, D., and LELIEVRE, E. (1986). Nuptialité et agriculture. *Population*, 41, 303-326.

COX, R., and OAKES, D. (1984). *Analysis of Survival Data.* London: Chapman and Hall.

DIRECTION DE LA STATISTIQUE (1981). *Enquête Sénégalaise sur la Fécondité, 1978 - Rapport National d'Analyse,* 1.

FARGUES, Ph. (1985). L'évaluation du niveau de la mortalité à partir des données des enquêtes EMIJ. *Les enquêtes sur la mortalité infantile et juvénile (EMIJ).* 1, 60-84.

FERRY, B. (1977). Le fichier événement. Une nouvelle méthode d'observation rétrospective. In *l'Observation démographique dans les pays à statistiques déficientes.* Liege, Belgium: Ordina Editions, 137-150.

HAERINGER, Ph. (1972). Méthodes de recherche sur les migrations africaines. Un modèle d'interview biographique et sa transcription synoptique. *Cahiers ORSTOM,* 9, 439-453.

SCOTT, Ch. (1985). Les problèmes de déperdition dans les enquêtes suivies. In *Les enquêtes sur la mortalité infantile et juvénile (EMIJ),* 1, 44-47.

RIANDEY, B. (1985). L'enquête "biographie familiale professionnelle et migratoire" (INED, 1981). Le bilan de la collecte. In *Migrations internes, collecte des données et méthode d'analyse. Département de démographie.* Université de Louvain, 117-134.

Survey Methodology, December 1987
Vol. 13, No. 2, pp. 173–181
Statistics Canada

An Alternative Method of Controlling Current Population Survey Estimates to Population Counts

K.R. COPELAND, F.K. PEITZMEIER, and C.E. HOY[1]

ABSTRACT

The CPS uses raking ratio estimation in post-stratification estimation to adjust sample estimates of population to census-based estimates of the population. An alternative procedure, using generalized least squares, is compared to the current procedure.

KEY WORDS: Generalized least squares; Post-stratification; Raking ratio estimation.

1. INTRODUCTION

The Current Population Survey (CPS) produces labor force estimates for the total U.S. working-age civilian noninstitutional population, based on a monthly multi-stage probability sample of approximately 60,000 housing units in the U.S. Each month a rotating sample comprised of 8 panels (called rotation groups) of housing units is interviewed, with demographic and labor force data being collected for all civilian adult occupants of the sample housing units.

Monthly estimates are published, subaggregated by demographic characteristics. Estimates for other subaggregates of the population (states, families, veterans, wage and salary earners, persons not in the labor force, etc.) are also produced on a monthly, quarterly, and/or annual basis.

Sample person weights are derived through the application of probability of selection, adjustment for nonresponse, and ratio adjustment to reduce the contribution to the variance due to the sampling of primary sampling units. A post-stratification estimation procedure adjusts the sample person weights so as to control the survey estimates of population to independently derived estimates of the population. The resultant weights are used in a composite estimation procedure and then seasonally adjusted to produce national estimates (Hanson 1978).

Detailed estimates for certain population subdomains (families, wage and salary earners, persons not in the labor force, family earnings, and veterans) make use of sample weights derived from adjustment procedures built on top of the post-stratification estimation.

The use of a generalized least squares (GLS) approach could potentially be used in place of post-stratification estimation or to integrate the various CPS adjustment procedures. The use of GLS has been proposed and investigated for use in the Consumer Expenditure Survey (Zieschang 1986).

This article discusses and compares the current CPS post-stratification estimation (which uses raking ratio estimation) and the GLS procedure, based on two months' CPS data (July 1983 and July 1984). Both macro and micro level data were examined to evaluate differences, if any, in the two procedures in this application.

[1] K.R. Copeland, F.K. Peitzmeier, and C.E. Hoy, Division of Statistical Methods, Office of Employment and Unemployment Statistics, Bureau of Labor Statistics, Washington, D.C. 20212 U.S.A.

2. CURRENT CPS POST-STRATIFICATION ESTIMATION

The CPS post-stratification estimation uses raking ratio estimation (RRE) to adjust the sample weights within a rotation group so as to control the sample estimates for the population to independently derived estimates of the population in each of three categories (state, age/sex/ethnicity, age/sex/race).

The methodology for RRE was first proposed by Deming and Stephan (1940) as an iterative alternative to least squares adjustment of table data. The RRE procedure has been shown to produce best asymptotically normal (BAN) estimates under simple random sampling, and to minimize the adjustments made to the sample weights based on one measure of closeness, as discussed in subsection 4.2 (Ireland and Kullback 1968). In addition, RRE, although producing biased estimates, can sometimes be effective in reducing the mean square error of survey estimates. This is believed to be the case in the application of RRE for CPS (Hanson 1978).

For the CPS, the RRE procedure attempts to adjust the sample counts $\{n_{ijk}\}$ obtained from previous stages of weighting to adjusted sample counts $\{\tilde{n}_{ijk}\}$ under the condition that:

$$\text{(A)} \qquad \sum_{j,k} \tilde{n}_{ijk} = m_{i..}$$

$$\text{(B)} \qquad \sum_{i,k} \tilde{n}_{ijk} = m_{.j.}$$

$$\text{(C)} \qquad \sum_{i,j} \tilde{n}_{ijk} = m_{..k}$$

be satisfied simultaneously,

where i = state ($i = 1, \ldots, 51$),
j = age/sex/ethnicity ($j = 1, \ldots, 16$),
k = age/sex/race ($k = 1, \ldots, 70$),
$m_{i..}$ = independent state estimate,
$m_{.j.}$ = independent age/sex/ethnicity estimate,
$m_{..k}$ = independent age/sex/race estimate.

The RRE procedure proportionately ratio adjusts the sample data each way (i.e., state, age/sex/ethnicity, and age/sex/race) of the table in successive steps, as follows.

(1) Ratio adjustment by state:

$$n_{ijk}^{(1,1)} = (m_{i..}/n_{i..})\, n_{ijk} = a_i^{(1)}\, n_{ijk}.$$

(2) Ratio adjustment by age/sex/etchnicity:

$$n_{ijk}^{(1,2)} = (m_{.j.}/n_{.j.}^{(1,1)})\, n_{ijk}^{(1,1)} = b_j^{(1)}\, n_{ijk}^{(1,1)}$$
$$= a_i^{(1)}\, b_j^{(1)}\, n_{ijk}.$$

(3) Ratio adjustment by age/sex/race:

$$n_{ijk}^{(1,3)} = (m_{..k}/n_{..k}^{(1,2)})\, n_{ijk}^{(1,2)} = d_k^{(1)}\, n_{ijk}^{(1,2)}$$
$$= a_i^{(1)}\, b_j^{(1)}\, d_k^{(1)}\, n_{ijk},$$

where $n_{i..}$ = sample row total
$n_{.j.}$ = sample column total
$n_{..k}$ = sample layer total.

The completion of the three adjustment steps constitutes one iteration of the raking process. The three steps are repeated substituting the current value of $n_{ijk}^{(h,3)}$ (adjusted sample count following the third way rake of the h-th iteration) for n_{ijk} in step (1) each time until 6 iterations are completed. (The number of iterations used in CPS was determined based on the convergence properties of the RRE for CPS and the relative gains achieved by number of iterations.) The final $\{n_{ijk}^{(6,3)}\}$ is taken as $\{\tilde{n}_{ijk}\}$.

In order to adjust the sample weights, the adjustment factor for sample records in cell $\{ijk\}$ is

$$F_{ijk} = n_{ijk}^{(6,3)} / n_{ijk}$$

$$= \prod_{h=1}^{6} a_i^{(h)} \, b_j^{(h)} \, d_k^{(h)}.$$

The sample weights prior to RRE are multiplied by the appropriate F_{ijk} to obtain the adjusted weights.

3. APPLICATION OF THE GLS IN THE CPS

The generalized least squares (GLS) procedure adjusts the sample weights from prior stages of weighting by minimizing the weighted squared adjustments, subject to a set of linear 'control' constraints the adjusted weights must satisfy. This is the problem which Deming and Stephan attempted to address in developing the RRE. The GLS procedure, like RRE, produces BAN estimates under certain conditions, in this case when all the cells are nonempty (Neyman 1949). GLS, by definition, minimizes the adjustments to the sample weights based on one measure of closeness (see subsection 4.2).

For the CPS, each dimension that defines a set of controls in the current post-stratification will define a set of linear constraints for the GLS procedure. The function to be minimized is

$$f(\underline{F}) = (\underline{F} - \underline{P})' \, P_0^{-1} \, (\underline{F} - \underline{P})$$

$$= \sum_i (W_{2i} - W_{1i})^2 / W_{1i},$$

subject to $X'\underline{F} = \underline{N}$,

where $\underline{F}$ = $(n \times 1)$ vector of derived final weights (W_{2i}) for each of the n sample persons,
$\underline{P}$ = $(n \times 1)$ vector of sample person weights prior to post-stratification (W_{1i}),
$\underline{P}_0$ = $(n \times \mathrm{n})$ diagonal matrix with the W_{1i} on the diagonal,
X = $(n \times \mathrm{k})$ design matrix whose rows correspond to sample persons, and whose columns correspond to control cells. The entries of the matrix (x_{ij}) are 0's or 1's, indicating the appropriate control categories for each of the n sample persons.
$\underline{N}$ = $(k \times 1)$ vector of independent population estimates, corresponding to the columns of X. These estimates are the same as those used in the CPS RRE.

The columns of X are required to be linearly independent so that an inverse of the matrix $(X' P_0 X)$ is achievable. In setting up matrices X and $\underline{N}$ for CPS, the 137 control cells used in the RRE (state, age/sex/ethnicity, age/sex/race) were reduced to a set of $k = 132$ linearly independent cells.

The unique solution to $X'\underline{F} = \underline{N}$ that minimizes $f(\underline{F})$ is, as shown in Luery (1986)

$$\underline{F} = \underline{P} + P_0 X (X' P_0 X)^{-1} (\underline{N} - X' \underline{P})$$

Although the elements of $\underline{F}$ are not constrained to be positive, in this application of GLS for CPS, the elements of $\underline{F}$ were all positive without the need for additional constraints. Methodology for providing non-negative weights in this context is discussed in Huang and Fuller (1978) and Zieschang (1986), among others.

4. RESULTS

4.1 Macro-Level

a. Estimates

Labor force estimates were tabulated for several demographic groups for July 1983 and July 1984, using the final weights derived from RRE and GLS. Standard errors for both RRE and GLS were calculated using a random group estimator of the form Wolter (1985)

$$\sum_{k=1}^{8} (8Y_k - \hat{Y})^2 / 56,$$

where Y_k = sum of the weights for sample records from the k-th rotation group with the characteristic Y,
$\hat{Y}$ = sum of the Y_k.

This variance estimator, while not accounting for the multi-stage design of the CPS, was used due to the unavailability of design information on the CPS public use microdata file.

Relative differences were calculated for both estimates of level and estimates of standard error. The relative difference was defined as:

$$(Y_{GLS} - Y_{RRE}) / Y_{RRE},$$

where Y_{RRE} = estimate of Y based on the weights derived through the use of RRE,
Y_{GLS} = estimate of Y based on the weights derived through the use of GLS.

As the data in Table 1 indicate, neither weighted labor force estimates nor estimates of standard error based on the current CPS RRE procedure and the GLS procedure showed any noticeable differences or trends when subaggregated to the sex by race/ethnicity level.

For labor force estimates by sex by race/ethnicity the estimated absolute relative differences between the CPS RRE and GLS estimates were all less than 0.3% (well below the estimated CVs of each estimate). For the majority of these estimates, in particular for total and whites, the absolute relative difference was less than 0.1%.

For many of the characteristics the sign of the relative difference changed from 1983 to 1984; thus there does not appear to be a pattern to the differences in the estimates obtained from the two procedures.

Table 1

Labor Force Estimates by Sex/Race or Ethnicity

		1983				1984			
		GLS		(GLS-RRE)/ RRE		GLS		(GLS-RRE)/ RRE	
		Total (000)	S.E. (000)	Total (%)	S.E. (%)	Total (000)	S.E. (000)	Total (%)	S.E. (%)
Total									
Total	Emp	103516	403	0.00	−0.14	107535	352	−0.01	1.12
	UE	10669	221	−0.04	−0.75	8765	118	−0.06	−0.21
	Rate	*9.34%*	*0.19%*	*−0.04*	*−0.56*	*7.54%*	*0.09%*	*−0.05*	*0.27*
	NILF	59938	373	0.01	−0.68	60080	419	0.02	0.41
White	Emp	91338	344	0.00	−0.33	94417	274	0.00	0.70
	UE	7928	236	0.00	−0.27	6282	120	0.00	−0.14
	Rate	*7.99%*	*0.23%*	*0.00*	*−0.26*	*6.24%*	*0.10%*	*0.00*	*−0.16*
	NILF	51915	340	0.00	−0.36	51700	358	0.00	0.39
Black	Emp	9871	69	0.06	−3.44	10371	98	0.02	0.17
	UE	2434	68	−0.12	−1.07	2202	60	−0.03	1.41
	Rate	*19.78%*	*0.55%*	*−0.14*	*−1.60*	*17.51%*	*0.42%*	*−0.04*	*1.49*
	NILF	6628	26	−0.04	−1.47	6765	109	−0.02	0.09
Hispanic	Emp	6132	73	−0.03	−0.59	6607	102	−0.03	1.90
	UE	920	79	−0.05	−0.29	786	70	−0.08	−0.03
	Rate	*13.04%*	*1.10%*	*−0.02*	*−0.33*	*10.63%*	*0.96%*	*−0.05*	*0.35*
	NILF	3760	31	0.05	−0.39	3786	73	0.04	1.02
Male									
Total	Emp	58985	147	0.00	−1.58	61045	188	0.00	1.74
	UE	5980	134	−0.05	−0.88	4682	79	−0.02	0.77
	Rate	*9.20%*	*0.19%*	*−0.05*	*−0.79*	*7.12%*	*0.11%*	*−0.02*	*1.30*
	NILF	17495	178	0.01	−1.81	17840	214	0.02	0.64
White	Emp	52674	482	0.00	0.42	54261	111	0.00	0.34
	UE	4484	131	0.01	−0.49	3394	93	0.01	−0.12
	Rate	*7.84%*	*0.21%*	*0.00*	*−0.47*	*5.89%*	*0.15%*	*0.01*	*−0.13*
	NILF	14985	160	−0.02	−0.40	15077	150	0.00	0.16
Black	Emp	5047	56	0.07	−1.70	5263	84	0.01	−0.50
	UE	1300	45	−0.20	−1.87	1137	33	0.08	1.12
	Rate	*20.49%*	*0.71%*	*−0.21*	*−2.02*	*17.76%*	*0.51%*	*0.05*	*0.94*
	NILF	2097	40	−0.04	−0.13	2236	88	−0.07	−0.48
Hispanic	Emp	3781	48	0.01	−0.86	4064	79	−0.02	1.29
	UE	534	45	−0.16	−0.83	451	41	−0.05	0.51
	Rate	*12.38%*	*0.99%*	*−0.15*	*−0.89*	*9.99%*	*0.95%*	*−0.03*	*0.66*
	NILF	981	42	0.00	−0.42	964	57	0.07	1.40
Female									
Total	Emp	44531	320	−0.01	−0.01	46490	194	−0.01	1.48
	UE	4689	107	−0.04	−0.19	4083	88	−0.10	−1.22
	Rate	*9.53%*	*0.23%*	*−0.03*	*−0.02*	*8.07%*	*0.16%*	*−0.09*	*−0.80*
	NILF	42443	287	0.01	−0.26	42240	217	0.02	0.34
White	Emp	38664	315	0.00	−0.29	40156	191	0.00	0.66
	UE	3444	115	−0.01	0.16	2888	68	0.00	−0.32
	Rate	*8.18%*	*0.28%*	*−0.01*	*0.11*	*6.71%*	*0.15%*	*0.00*	*−0.34*
	NILF	36929	283	0.01	−0.32	36623	214	0.00	0.53
Black	Emp	4824	57	0.05	0.56	5108	50	0.02	1.69
	UE	1134	46	−0.02	0.07	1065	46	−0.14	−0.62
	Rate	*19.03%*	*0.80%*	*−0.06*	*0.08*	*17.25%*	*0.67%*	*−0.13*	*−0.63*
	NILF	4531	24	−0.04	2.99	4529	59	0.01	1.49
Hispanic	Emp	2350	44	−0.08	−0.46	2543	38	−0.05	3.04
	UE	385	41	0.10	0.51	335	34	−0.13	−0.62
	Rate	*14.08%*	*1.46%*	*0.16*	*0.57*	*11.64%*	*1.18%*	*−0.07*	*−0.11*
	NILF	2778	33	0.07	−0.87	2822	27	0.03	0.13

The absolute relative differences between the CPS RRE and GLS estimates of standard errors for national labor force estimates were all less than: 1.9% for total population; 0.7% for whites; 3.5% for blacks; and 3.1% for Hispanics.

b. Month-in-Sample Indexes

It is a well-documented fact that the estimates produced from the CPS final weights have certain patterns of relative bias based upon the time the rotation group has been in sample (Bailar 1975). Month-in-sample indexes

$$I_k = (8Y_k / \hat{Y}) \times 100,$$

were calculated for both July 1983 and July 1984 based upon both the RRE estimates and the GLS estimates.

Month-in-sample indexes for labor force by race, labor force by sex, and labor force by ethnicity were virtually identical for estimates based upon the CPS RRE and GLS procedures.

4.2 Micro-Level

a. Adjustments to Sample Weights

Both RRE and GLS minimize some measure of closeness between the pre- and post- adjustment sample weights. For RRE the measure is (Ireland and Kullback 1968)

$$M_A = \sum_i W_{2i} \ln (W_{2i} / W_{1i}).$$

For GLS, the measure is (Luery 1986)

$$M_B = \sum_i (W_{2i} - W_{1i})^2 / W_{1i},$$

where W_{1i} = weight for sample record i prior to adjustment,
W_{2i} = weight for sample record i following adjustment.

Tabulation of the measures of closeness (summarized in Table 2) provided some interesting and, in some cases, puzzling results. The CPS RRE yielded smaller values for both measures. The GLS procedure did tend to produce smaller values for the measures for certain subgroups, most notably for blacks and Hispanics. It should be noted that the differences between the values for the measures for RRE and GLS were almost always less than 1%.

Although M_B should be minimized through the use of the GLS procedure, the value of M_B based upon the GLS weights for the total sample was greater than the value of M_B for the CPS RRE weights for 11 of the 16 rotation groups.

In seeking a reason for this apparent contradiction, it was noted that the CPS RRE had yet to converge to the age/sex/ethnicity controls after six iterations. The extent of this non/convergence is *very small;* less than 1.0% for all control categories. However, given the difference in M_B between the RRE and GLS, a change in the RRE sample weights of only 0.1%-0.2% could reverse the results. Rerunning RRE using 15 iterations, although still not achieving convergence did provide indications that the slight lack of convergence of the RRE is the reason for the results for M_B. (It should be noted that the GLS procedure minimizes M_B among the class of adjustment procedures yielding estimates that meet the population controls. Since the CPS RRE did not converge to the population controls, it is not a member of this class.)

Table 2

Comparison of measures of closeness based on 8 RGs for each year (# of RGs with RRE < GLS)

	M_A		M_B	
	1983	1984	1983	1984
Total	8	8	4	7
White	7	7	3	4
Black	3	3	1	1
Hispanic	0	0	0	0
Male	2	7	1	5
Female	8	8	8	8

Although an adjustment procedure such as RRE or GLS may minimize some measure of closeness for the total sample, it does not necessarily minimize that measure of closeness for subaggregates of the sample which were controlled for (e.g., blacks, Hispanics, males). Given the use of controls, and the fact that the overall measure of closeness is being minimized, it would seem desirable to have an adjustment procedure produce small measures of closeness at the subaggregate level also. The GLS procedure yielded smaller measures in almost every rotation group for Hispanics, in many rotation groups for blacks, and in several rotation groups for whites and males.

b. Comparison of Adjustments

Both RRE and GLS determine adjustment factors within cells defined by the intersection of the marginal constraints. Each sample record within a cell receives the same factor. To compare the adjustments made by the two procedures, the factors determined for each sample record by each procedure were compared using the following ratio

$$RRE/GLS = [(W_{2i}/W_{1i})_{RRE}] / [(W_{2i}/W_{1i})_{GLS}].$$

This ratio indicates the relationship between the adjustments made to a sample person weight by the RRE and GLS procedures. For comparison purposes, values of *RRE/GLS* less than 0.95 or greater than 1.05 were used to denote differences in the adjustments made by RRE and GLS.

For each set of independent population controls, ratios E/C (i.e., coverage rates), where E is the sample estimate based on the sample person weights prior to post-stratification and C is the independent control, were derived.

Within each set of controls (state, age/sex/ethnicity, age/sex/race) sample records were categorized by their coverage rates. Table 3 provides the sample distribution by coverage rate categories and by the *RRE/GLS* values, as well as the proportion of records within each coverage rate category that have the *RRE/GLS* values.

The data in Table 3 indicate that, for each set of controls, sample records from population groups which were over- or under-covered to some extent by the survey (i.e., for which the coverage rate is not near 1) were more likely to be adjusted differently by RRE and GLS than were sample records in population groups adequately covered by the survey.

Table 3

Comparison of RRE and GLS adjustments, 1984

Control Marginal	Coverage Rate Category	Proportion of Total Sample	Proportion of Sample with RRE/GLS <0.95 or >1.05	Proportion of Category with RRE/GLS <0.95 or >1.05
Age/Sex/ Race	<0.7	0.007	0.057	0.219
	0.7-0.8	0.022	0.116	0.136
	0.8-0.9	0.241	0.147	0.019
	0.9-1.1	0.699	0.504	0.019
	1.1-1.2	0.021	0.069	0.084
	>1.2	0.010	0.106	0.275
Age/Sex/ Ethnicity	<0.7	0.010	0.078	0.198
	0.7-0.8	0.014	0.032	0.058
	0.8-0.9	0.106	0.135	0.033
	0.9-1.1	0.869	0.741	0.022
	1.1-1.2	0.001	0.007	0.202
	>1.2	0.001	0.007	0.373
State	<0.7	0.056	0.068	0.031
	0.7-0.8	0.111	0.180	0.042
	0.8-0.9	0.278	0.325	0.030
	0.9-1.1	0.479	0.342	0.018
	1.1-1.2	0.026	0.009	0.009
	<1.2	0.049	0.077	0.040

4.3 Computer Resources

The CPS RRE and GLS procedures were run on an IBM System 370 at the National Institutes of Health using PROC MATRIX in the SAS System. The CPU time to prepare the files and perform the weighting was approximately three times as much for the GLS procedure than it was for the RRE procedure. There was also more storage of files involved with the GLS procedure. (The size of the matrices involved for CPS are quite large, with the number of rows for $\underline{P}$, P_0, X, and $\underline{N}$ being around 14,000 for each rotation group.)

5. SUMMARY AND CONCLUSIONS

This investigation was intended to provide a comparison of RRE and GLS as applied to the CPS, at both the macro and micro level.

The results obtained at the macro level do not indicate any difference in the estimates obtained from the RRE and GLS procedures.

The measures of closeness indicated that the CPS RRE made slightly smaller changes overall to the sample weights to meet the control constraints than did the GLS. The CPS RRE tended to produce slightly larger measures of closeness for subaggregates of minority populations. The two procedures differ most notably in the adjustments made to portions of the population which are either over- or under-covered.

Based on the work done in this investigation, it does appear that the RRE takes less computer time to run for the CPS second-stage adjustment than the GLS.

ACKNOWLEDGEMENT

The authors are grateful to Fritz Scheuren for his review of the original version of this paper, and to the referees and the Associate Editor for their very useful comments, incorporation of which resulted in the improvement of the paper.

REFERENCES

BAILAR, B. (1975). The Effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association,* 70, 23-30.

DEMING, W.E., and STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics,* 11, 427-444.

HANSON, R.H. (1978). The Current Population Survey design and methodology. Technical Paper 40, U.S. Bureau of the Census.

HUANG, E.T., and FULLER, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Section on Social Statistics, American Statistical Association,* 300-305.

IRELAND, C.T., and KULLBACK, S. (1968). Contingency tables with given marginals. *Biometrika,* 55, 179-188.

LUERY, D. (1986). Weighting sample survey data under linear constraints on the weights. *Proceedings of the Section on Social Statistics, American Statistical Association,* 325-350.

NEYMAN, J. (1949). Contribution to the Theory of the X^2 Test. In *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability,* (Ed. J. Neyman), Berkeley: University of California Press, 239-273.

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

ZIESCHANG, K.D. (1986). A Generalized least squares weighting system for the Consumer Expenditure Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 64-71.

Survey Methodology, December 1987
Vol. 13, No. 2, pp. 183–198
Statistics Canada

A Class of Methods for Using Person Controls in Household Weighting

CHARLES H. ALEXANDER[1]

ABSTRACT

A class of "constrained minimum distance" methods is considered for constraining household weights to be consistent with auxiliary information on the number of persons in various age × race × sex cells. The constrained weights are as close as possible to the initial weights based on the inverse probability of selection. This class of methods includes raking and generalized least square methods, as well as multinomial maximum likelihood, (where the cells of the distribution are household types.) The properties of the methods in the presence of systematic undercoverage of the household types are studied through some simple models for coverage. Comparisons with the principal person method are made and the paper concludes with the observation that it is necessary to know more about the nature of survey undercoverage before deciding on which of the constrained minimum distance or principal person methods is to be preferred in applications.

KEY WORDS: Weighting; Auxiliary information; Raking ratio estimation; Principal person method; Survey coverage.

1. INTRODUCTION

Post-stratification is commonly used to adjust survey weights to take into account independent information about the number of units of certain kinds in the population. For example, independent estimates of the population in various age × race × sex post-stratification cells may be available from adjusting census counts for known changes in the number of persons since the census. These independent estimates are often referred to as "control counts". Prior to post-stratification, each sample person (or household) has an initial weight, typically corresponding to the inverse of the selection probability. A post-stratification ratio adjustment factor is applied to the weights of all sample persons in each cell, so that the sum of the adjusted person weights equals the independent control count for the cell. This adjustment is especially important when there is systematic undercoverage of households or persons within households.

For most U.S. Census Bureau demographic surveys, post-stratification is used in assigning weights to sample persons, but is not used directly in assigning weights to sample households. This is due to the greater difficulty of obtaining independent estimates for households. Instead, household weights for these surveys are assigned using some version of the "principal person" method. In the basic principal person method, the household weight is set equal to the final post-stratified person weight of the "principal" person in the household. The rule for identifying this person will be described in Section 2. By using the post-stratified person weight, the principal person method does incorporate the independent estimates of persons into the weights assigned to households.

The most obvious problem with the principal person method is that when the resulting household weights are used to calculate weighted estimates of the number of persons in each post-stratification cell, with each person being given his or her household's weight, these

[1] Charles H. Alexander, Statistical Methods Division, U.S. Census Bureau, Washington, D.C. 20233 U.S.A.

estimates do not agree with the control counts used in the post-stratification. Consequently, there has been interest in methods of assigning weights to households which are constrained to produce person estimates which agree with the independent control counts.

This paper considers a class of methods for assigning survey weights to households, constrained to be consistent with the "known" control counts in various person cells. The general idea is to find household weights which satisfy the constraints and are as close as possible to the initial vector of weights assigned to the households. The different methods within the class correspond to different ways of measuring the distance between the initial vector of weights and the adjusted vector of weights.

Section 2 describes six "constrained minimum distance" weighting methods of this type plus a version of the principal person method. Three of the six methods have been investigated previously, and the others are added in this paper to round out the picture. Section 3 describes the computation of the weights. Section 4 discusses how the adjusted weight depends on the composition of the household. Section 5 discusses results and examples which may help in understanding what these methods do. Section 6 describes areas for further research.

This work has numerous antecedents. The general class of constrained minimum distance methods is suggested for household weighting by Luery (1986). Extending Luery's work, Zieschang (1986a) proposes using one of these methods, generalized least squares, for weighting the U.S. Consumer Expenditure Surveys. Another member of the class is the "minimum discriminant information method", otherwise known as raking ratio estimation or, simply, raking. Oh and Scheuren (1978a) specifically discuss the raking approach to the household weighting problem, and give additional references to a rich literature on raking and related methods. The idea of viewing raking as a constrained minimum distance problem dates back at least to Deming and Stephan (1940). The fundamental principles of this approach are explored in Ireland and Kullback (1968). Applications to survey weight adjustment are well covered in Brackstone and Rao (1979). The class of methods also includes two criterion functions related to multinomial maximum likelihood. The relationship of this to raking has been extensively studied; see, for example, Bishop, Fienberg, and Holland (1976). Fienberg (1986) points out that the distance criteria considered in this paper may be viewed as special cases of a parametric family of functions considered in Cressie and Read (1984).

2. CONSTRAINED MINIMUM DISTANCE METHODS

2.1 Methods Based on Household Weights

Consider a sample of K households, whose initial weights are given by the vector $\underline{S} = (S_1 \ldots, S_K)'$. In this paper, S_k will be the inverse of the probability of selection of the k-th household; in some applications other adjustments such as nonresponse factors may be included in the initial weight.

Suppose that there are J post-stratification cells, and that the number of persons in the population (N_j) is known for each cell. For example, for the U.S. Consumer Expenditure Survey, there are $J = 48$ cells corresponding to combinations of the two sexes, two races (black, nonblack), and twelve age categories. In that survey, persons younger than 14 are not included. The control counts for these cells will be treated as a vector $\underline{N} = (N_1, \ldots, N_J)'$.

The composition of the sample households will be described by a matrix $A = (a_{kj})$, where a_{kj} is equal to the number of persons in the k-th sample household who are in the j-th post-stratification cell. Summing over the post-stratification cells for the k-th household gives $a_{k.}$, the total number of persons in the k-th household. For household k, the vector

$(a_{k1}, \ldots, a_{kJ})$ describes the composition of the household. For example, if the vector is (2,1,0,0, . . . , 0), then the household contains exactly two persons in the first cell and one in the second.

Using the initial weights $\underline{S}$, the weighted sample estimate of the number of persons in cell j would be $\hat{N}_j = \Sigma_k a_{kj} S_k$ or in general $\underline{\hat{N}} = A'\underline{S}$.

Typically $\underline{\hat{N}} \neq \underline{N}$, i.e., the initial weighted estimate of persons in the post-stratification cells may not equal the known population of the cell.

The goal is to define a new vector of weights $\underline{W} = (W_1, \ldots, W_K)'$ for the sample households, so that $\underline{N} = A'\underline{W}$ or

$$\sum_k a_{kj} W_k = N_j \text{ for } j = 1, \ldots, J. \tag{1}$$

The solution to (1) is not necessarily unique. The idea of the constrained minimum distance methods is to chose $\underline{W}$ so as to minimize some measure $D(\underline{W},\underline{S})$ of the distance between the vectors $\underline{W}$ and $\underline{S}$, subject to (1). In this way, the initial weights $\underline{S}$ are changed as little as possible in meeting the constraint that the adjusted weights should agree with the known control totals. Note that, for certain possible values $N_1, \ldots, N_J$, it may be impossible for any vector of weights $\underline{W}$ to satisfy the constraints (1). Practically speaking, this possible infeasibility does not seem to be a problem, provided the sample is large enough to include a good representation of different types of households, since the controls $\underline{N}$ are generated from the actual population and therefore can be expected to be "feasible".

There are numerous ways of measuring the difference between two vectors. Three distance criteria $D(\underline{W},\underline{S})$ will be considered, corresponding to a household-level generalized least squares (GLS-H) objective function, a minimum discriminant information (MDI-H) function, and a maximum likelihood estimation (MLE-H) criterion. The criteria are:

$$\text{GLS - H:} \quad \sum_k (W_k - S_k)^2 / S_k, \tag{2a}$$

$$\text{MDI - H:} \quad (S_. - W_.) + \sum_k W_k \ln(W_k / S_k), \tag{2b}$$

$$\text{MLE - H:} \quad (W_. - S_.) - \sum_k S_k \ln(W_k / S_k). \tag{2c}$$

Throughout the paper, the dot notation is used to denote summation over a subscript.

In each case $D(\underline{W},\underline{S})$ is nonnegative and is equal to zero if and only if $\underline{W} = \underline{S}$. This can be shown, in the usual way, by examining the first and second partial derivatives of each expression with respect to the W_k.

Algorithms for calculating $\underline{W}$ to minimize these three criteria, while meeting the constraint (1) to the degree of approximation desired, will be discussed in Section 3.

2.2 Methods Derived from Person Weights

An alternative approach to this problem leads to a slight but important modification of the three distance criteria. These modified criteria are given by (5a), (5b), and (5c) below. Although these criteria lead to weights for households, they are generated by an approach which starts out by trying to define weights for persons. Accordingly, first consider the problem as one of defining person weights as close as possible to their original household weights, subject to the constraint that the weighted estimate of persons in each post-stratification cell

equals the known control. Let the persons in the k-th household be numbered $i = 1, \ldots, a_{k.}$ and let S_{ki} be the initial weight of the i-th person in the k-th household; note that $S_{ki} = S_{k.}$

Let b_{kij} be a zero-one indicator variable showing whether the i-th person in the k-th household is in the j-th post-stratification cell. Then the condition for consistency with the controls is

$$\sum_k \sum_i b_{kij} W_{ki} = N_j. \tag{3}$$

The three criteria for the person weighting problem would be

$$\sum_k \sum_i (W_{ki} - S_{ki})^2 / S_{ki}, \tag{4a}$$

$$S_{..} - W_{..} + \sum_k \sum_i W_{ki} \ln(W_{ki}/S_{ki}), \tag{4b}$$

$$W_{..} - S_{..} - \sum_k \sum_i S_{ki} \ln(W_{ki}/S_{ki}). \tag{4c}$$

These criteria could be used for defining person weights. In fact the criterion (4c) would lead to the post-stratification weights which are used in person weighting for the Consumer Expenditure Survey, as described in Alexander (1986). However, our problem is to define weights for households. Household weights may be obtained from these criterion functions by imposing upon the person problem the additional constraint that all persons in the same household must have the same weight. Therefore, let $W_{ki} = W_k$ for $i = 1, \ldots, a_{k.}$ Under this constraint, (3) becomes

$$N_j = \sum_k \left(\sum_i b_{kij} \right) W_{ki} = \sum_k a_{kj} W_k,$$

which is the same as the constraint (1) in Section 2.1. The distance criteria (4a), (4b), and (4c) now become:

GLS-P: $$\sum_k a_{k.} (W_k - S_k)^2 / S_k, \tag{5a}$$

MDI-P: $$\sum_k a_{k.} S_k - \sum_k a_{k.} W_k + \sum_k a_{k.} W_k \ln (W_k/S_k), \tag{5b}$$

MLE-P: $$\sum_k a_{k.} W_k - \sum_k a_{k.} S_k - \sum_k a_{k.} S_k \ln (W_k/S_k). \tag{5c}$$

The criteria are now summations at the household level, but the household size $a_{k.}$ has been brought into the criterion for measuring the distance between the initial and adjusted vector of weights. These criteria will be seen to have advantages over the more direct approach which led to (2a), (2b), and (2c).

2.3 The Principal Person Method

In the basic principal person method, the post-stratified person weight of the household's "principal person" is used as the household's weight. To determine the principal person, it is first necessary to determine the household's "reference person". The reference person is identified by the interviewer as the first person mentioned in response to the instruction "start by giving me the name of someone who owns or rents this house." Household relationships are defined in terms of the other members' relationship to this reference person. "Reference person" has replaced the "head of household" concept for this purpose.

The principal person is the wife of the reference person if the reference person is a married male with spouse present. Otherwise, the principal person is the reference person himself or herself. The rationale for this choice is that the principal person should be a person who is not likely to be missed due to within-household undercoverage. In general, women have better coverage than men. Further, the principal owners or renters of the house or apartment seem unlikely to be overlooked.

The basic idea of the principal person method is that there is exactly one principal person in each household. Consequently, the number of households may be estimated by estimating the number of principal persons. This basic method is used for the U.S. National Crime Survey. Other surveys such as the U.S. Consumer Expenditure Surveys or Current Population Survey, make additional adjustments based on assumptions about within-household undercoverage of principal persons, as compared to other persons in the same post-stratification cell (Alexander 1986.)

The principal person method is difficult to model theoretically because the designation of the reference person is somewhat arbitrary. In the hypothetical examples of Section 5, a simplified version of the principal person method will be used, in which the principal person is the household member whose post-stratification cell has the best coverage, i.e., whose post-stratification factor is closest to one. A similar idea is used in Scheuren (1981).

This simplified principal person method will be represented symbolically as follows. For the k-th sample household, let $j(k)$ be the post-stratification cell of the household's principal person. Then the household's principal person weight is

$$W_k = S_k(N_{j(k)}/\hat{N}_{j(k)}).$$

3. COMPUTATION OF THE WEIGHTS

The two least squares methods, GLS-H and GLS-P, have closed-form expressions for $\underline{W}$, providing that there exists some solution to the constraints (1). For the GLS-H weights, the adjusted weights are given by

$$\underline{W} = \underline{S} + MA(A'MA)^{-1}(\underline{N} - A'\underline{S}) \tag{6}$$

where $\underline{S} = (S_1, \ldots, S_K)$, $\underline{N} = (N_1, \ldots, N_J)$, A is the matrix (a_{kj}) and M is the $K \times K$ diagonal matrix with the elements of $\underline{S}$ on the main diagonal. The weights $\underline{W}$ for the GLS-P method are also given by (6), except that M is the $K \times K$ diagonal matrix with the values $S_1/a_{1.}, \ldots, S_K/a_{K.}$ on the main diagonal.

A disadvantage of (6) for either method GLS-H or GLS-P is that the solution $\underline{W}$ may include negative weights. Conceptually this is unsettling, and for practical users negative weights are unacceptable. It is usually possible to incorporate additional constraints that the

weights must be positive. Ways of doing this are given by Zieschang (1986a) and Huang and Fuller (1978). However, the advantage of a simple closed-form solution is lost with these additional constraints.

The raking method (MDI-P) has been used before for household weighting, e.g., by Oh and Scheuren (1978a). A related method which has been extensively tested is described in Pugh, Tyler, and George (1976), based on the approach of Stephan (1942). Luery (1986) gives an iterative algorithm based on Darroch and Ratcliff (1972), which is proved to converge whenever there is a solution to (1). This method is presented here, since the iterative step has a simple interpretation. The iteration starts with "step 0" weights

$$W_k(0) = S_k(N_{.}/\hat{N}_{.})$$

In other words, the initial weight S_k is adjusted by an overall inflation factor equal to the known population $N_{.}$ divided by the initial weighted total population. At subsequent iterative steps, the adjustment is

$$W_k(i) = W_k(i-1) \prod_j \left(N_j / \sum_s a_{sj} W_s(i-1) \right)^{a_{kj}/a_{k.}}$$

Note that $W_k(i-1)$ is multiplied by the geometric mean of the post-stratification factors for the persons in the k-th household, where the post-stratification factors are calculated using the weights after iteration $i-1$.

The other three methods, MDI-H, MLE-H, and MLE-P, have not been extensively studied. The following iterative algorithms have worked successfully in small hypothetical examples such as those given in Section 5. In each case, a system of equations, which the weights must satisfy in order to minimize the distance criterion subject to the constraints, can be found by the use of Lagrange multipliers. The equations cannot be solved directly, but if an iterative method produces solutions of the proper form, then the solution minimizes the criterion. If the algorithms converge, the solutions will satisfy the equations. However, the author has no general proof of convergence. A possible alternative approach for the "maximum likelihood" criteria would be to apply the approach of Haber and Brown (1986). Other related work is Fagan and Greenberg (1985).

3.1 Method for MDI-H

The equation for the weights is

$$W_k = S_k \prod_j \gamma_j a_{kj} \tag{7}$$

subject to (1). If values $\gamma_1, \ldots, \gamma_J$ can be found so that the weights calculated according to (7) satisfy (1), then those weights minimize (2b) subject to (1). An iterative algorithm for generating such a vector $\underline{W}$ is as follows.

Initialize $W_k(0) = S_k$ and $\gamma_j(0) = 1$. Then at the i-th iteration let

$$\gamma_j(i) = \gamma_j(i-1) \left[1 - (\hat{N}_j(i-1) - N_j) / \sum_s a_{sj}^2 W_s(i-1) \right],$$

where $\hat{N}_j(i-1) = \sum_s a_{kj} W_s(i-1)$. Then let $W_k(i) = S_k \prod_i (\gamma_j(i))^{a_{kj}}$.

3.2 Method for MLE-H

The solution is of the form:

$$W_k = S_k / \left(1 + \sum_j \gamma_j a_{kj}\right).$$

subject to (1).

An iterative solution is

$$W_k(0) = S_k \quad \text{and} \quad \gamma_j(0) = 0,$$

$$\gamma_j(i) = \gamma_j(i-1) + (\hat{N}_j(i-1) - N_j) / \left(\sum_s (a_{sj} W_s(i-1))^2 / S_k\right),$$

$$W_k(i) = S_k / \left(1 + \sum_j \gamma_j(i) a_{kj}\right).$$

3.3 Method for MLE-P

The solution is of the form:

$$W_k = S_k / \left(\sum_j \gamma_{kj} a_{kj} / a_{k.}\right).$$

subject to (1).

An iterative solution is

$$W_k(0) = S_k \quad \text{and} \quad \gamma_j(0) = 1,$$

$$\gamma_j(i) = \gamma_j(i-1) \hat{N}_j(i-1) / N_j,$$

$$W_k(i) = S_k / \left(\sum_j \gamma_j(i) a_{kj} / a_{k.}\right).$$

4. THE ROLE OF A HOUSEHOLD'S "COMPOSITION TYPE"

For the six constrained minimum distance methods, the ratio of a household's initial weight to its adjusted weight depends on the number of people in the household in the different post-stratification cells. To discuss this further, the notion of a household's "composition type" will be introduced. Two sample households, say k and m will be said to "have the same type" if they have exactly the same number of people in each of the post-stratification cells, i.e., if

$$a_{kj} = a_{mj} \text{ for } j = 1, \ldots, J. \tag{8}$$

As an example, one household type would be a "household consisting of a white male 35-39 and a white female 30-34." Note that the composition type does not depend on family relationships.

The ratio of the adjusted weight to the initial weight, W_k / S_k, is the same for all households with the same type. In other words, if k and m satisfy (8), then $W_k / S_k = W_m / S_m$. This fact was used in Ireland and Scheuren (1975). A formal proof is given in Alexander and Roebuck (1986).

A useful consequence of this fact is that, in calculating the weights for the constrained minimum distance methods, the calculations may be done using the household type as the unit of analysis rather than the individual household. A simple example may make the implications of these results clearer. Suppose that there are two post-stratification cells, $j = 1$ for females and $j = 2$ for males. The sample consists of K households. For household k, the vector (a_{k1}, a_{k2}) describes how many females and males are in the household; a household with vector (2,1) has two females and one male.

Practically speaking, there is some upper limit on the size of a household, and there are only finitely many household types. For the example, assume that no household has more than three people. Then there are $T = 9$ household types corresponding to the vectors: (1,0), (0,1), (2,0), (1,1), (0,2), (2,1), (1,2), (3,0), (0,3). These types will be numbered consecutively $t = 1, \ldots, 9$. The types will also be labelled mnemonically, F, M, FF, FM, MM, FFM, FMM, FFF, MMM. Hypothetical sample data and control totals are given in Table 1. Note that S_t is the total initial weight given to households of type t.

The constrained minimum distance adjustments effectively may be calculated from the total weights for the household composition types, $S_1, \ldots, S_9$, without actually looking at the individual household weights. Adjusted weights $W_1, \ldots, W_9$ may be calculated using the algorithms from Section 3 replacing summation over k by summation over t. Then for any type t household, the adjusted weight given by the method is W_t / S_t times the initial weight for the household. (The potentially confusing notation of using S_k for the household weight and S_t for the total weight for a t household type is adopted to emphasize that the formulas of Sections 2 and 3 apply equally well to households or household types. In doing calculations, the meaning will be clear from the context.)

The reduction of the problem from individual households to household types is extremely convenient for presenting small examples. Even when applied to the full 48 post-stratification cells, the household-type approach may still be practical: despite the astronomical number of possible household types, the actual number of types in the sample can never be larger than the sample size and often is substantially smaller. This was found to be the case for related cells of households in Ireland and Scheuren (1975). Simply reducing the size of the computational task by combining the weights for single-person households of the same type may be useful; this has been done at the U.S. Bureau of Labor Statistics in applying the generalized least squares method to the Consumer Expenditure Surveys.

The simplified version of the principal person method also depends only on the household type. If two households have the same composition, then their principal persons will be in the same post-stratification cell, the one with the post-stratification factor closest to one. Consequently, the same ratio adjustment factor would be used for both households. In the actual principal person method, the principal person depends in part on who happens to be designated as reference person, so the adjustment factor is not completely determined by the household's composition type.

Note that the MLE-H method corresponds to calculating multinomial maximum likelihood estimates (subject to the constraint (1)) of p_t, $t = 1, \ldots, T$, where p_t is the population proportion of households with type t. The MLE-P method has a related interpretation. Neither of these models, which also pertain to the corresponding GLS and MDI methods, allows for systematic undercoverage.

5. DISCUSSION OF THE METHODS

This Section begins with some speculations about properties of the constrained minimum distance methods, based on the results of Section 4, and follows with some simple hypothetical examples, which generally appear to support the speculations.

The first conjecture is that MLE-H, GLS-H, and MDI-H will tend to give similar results, and also that MLE-P, GLS-P, and MDI-P will tend to be similar to one another, at least for large samples. This is based on the observation that these are all best asymptotic normal estimators under the relevant multinomial sampling model, where the cells are the household types. For small or moderate sample sizes, greater differences between the methods might be anticipated, especially if there are a large number of household composition types, so that the sample in individual "cells" of the multinomial may be small.

The examples given below tend to support this conjecture; the "household" methods all give very similar results, as do the "person" methods. This is true even in some cases when the hypothetical data do not fit the model very well. However, these examples involve only a small number of household types and post-stratification cells, and so are illustrative rather than conclusive.

The second conjecture is based on considering the nature of the sampling models under which the constrained minimum distance methods may be viewed as maximum likelihood estimates, or asymptotic approximations thereto. In these models, perfect coverage is assumed. The models assume a distribution corresponding to probabilities which are the actual proportions in the population, and these probabilities are consistent with the "true" control totals used in the constraints (1). According to these models, for sufficiently large samples, the initial sample estimates would approach agreement with the control totals. This would not be true when there is substantial undercoverage in the sampling frame. Such undercoverage is an important reason for using post-stratification. Coverage considerations may be especially important for telephone surveys where there is no supplemental frame to include households without telephones. If there is no special adjustment for noninterview "nonresponse", such as refusal or inability to provide the requested information, then nonresponse may be a further departure.

Based on these remarks, the second conjecture is that without adjustment the constrained minimum distance methods may not perform well in adjusting for systematic undercoverage, even for large samples. The methods are optimal under models which assume perfect coverage; one would expect that they might be less than optimal when this assumption is violated.

The examples given below partly support this conjecture. The constrained distance methods do not do as well as the simplified principal person method under certain assumptions about undercoverage. Under other assumptions, some of the methods may do quite well. The author concludes that it is necessary to know more about the nature of survey undercoverage before judging that any of these methods is superior to the principal person method. Oh and Scheuren (1978b) raise some related issues about mean square error of the raking estimator when there is undercoverage.

Two examples will be presented, representing two extreme forms of undercoverage. The first ("household undercoverage example") will assume that there is a uniform 10% undercoverage of all households, but that there is no within-household undercoverage. The second example ("within-household undercoverage example") assumes a 10% undercoverage of males due to within-household undercoverage in households where there are both males and females, and undercoverage of all-male households. For single-person households, any "within-household undercoverage" means that the whole household is missed.

In example 1, there is a 10% under-representation of all types of households in the sample. For a sufficiently large sample, this would obviously be due to systematic undercoverage, rather than sampling error. Applying the constrained minimum distance methods and the principal person method to this example gives the total adjusted weights for each household type shown in the last four columns of Table 1.

Note that the GLS-P, MDI-P, and MLE-P methods all bring the adjusted weight up to the actual population value. Thus, these methods give "unbiased" weights. Since all persons have a second-stage factor of 1/.9, the principal person method also achieves this result.

Table 1

Household Undercoverage Example:
Description of Population and Sample

Type & description	Actual Population	Total Initial Weights	Total Weight (W_i) for Methods: GLS-H	MDI-H	MLE-H	GLS-P MDI-P MLE-P Prin. Pers.
1: F	25,000	22,500	23,785	23,745	23,704	25,000
2: M	15,000	13,500	14,120	14,097	14,075	15,000
3: FF	7,000	6,300	7,020	7,016	7,013	7,000
4: FM	40,000	36,000	39,708	39,672	39,632	40,000
5: MM	5,000	4,500	4,913	4,906	4,900	5,000
6: FFM	12,000	10,800	12,529	12,506	12,594	12,000
7: FMM	12,000	10,800	12,408	12,428	12,449	12,000
8: FFF	0	0	0	0	0	0
9: MMM	0	0	0	0	0	0
Total	116,000	104,400	114,483	114,370	114,367	116,000

Control Totals:	Number of Females	=	115,000
	Number of Males	=	101,000
Initial Weighted Person Counts:	Females	=	103,500
	Males	=	90,900

The other methods, GLS-H, MDI-H, and MLE-H, all give substantially too little weight to one-person households and too much to the three-person households. Intuitively, this makes sense; since these methods do not allow for systematic undercoverage and must explain the shortage of sample persons as sampling error, the obvious explanation is that the sample has a below-average number of large households, due to chance. The better performance of MLE-P makes some sense, since it starts out with a multinomial sampling model which allows sampling of persons without regard to households.

Practically speaking, this example reflects very poorly on the GLS-H, MDI-H, and MLE-H methods. Even uniform undercoverage would cause these methods to distort the distribution of household sizes. Worse, the distortion goes opposite from what is commonly assumed about differential household coverage, namely that small households are more likely to be missed than large ones, so that small households need relatively higher weights, not relatively lower weights.

The second example will emphasize within-household undercoverage of males. The situation is more complicated than in the previous example, because a household may have an apparent composition type different than its actual type. For example, a household which actually consists of a male and a female may appear to be a single-person household. The actual and apparent type will be indicated by modifying our previous notation. For example, a FM household in which the male is missed will be denoted F[M]. A [M] household or [MM] household is missed entirely. Table 2 describes the hypothetical data. The actual population is the same as in the previous example.

Table 2

Within-household Undercoverage Example:
Description of Population and Sample

Actual Household Type	Apparent Type	Actual Number	Total Initial Weights
1: F	F	25,000	25,000
2: M	M	13,500	13,500
	[M]	1,500	0
3: FF	FF	7,000	7,000
4: FM	FM	36,000	36,000
	F[M]	4,000	4,000
5: MM	MM	4,500	4,500
	[MM]	500	0
6: FMM	FFM	10,800	10,800
	FF[M]	1,200	1,200
7: FMM	FMM	10,800	10,800
	FM[M]	1,200	1,200
8: FFF	FFF	0	0
9: MMM	MMM	0	0
		116,000	114,000
Control Counts:	Number of Females	115,000	
	Number of Males	101,000	
Initial Weighted Person Counts:	Females	115,000	
	Males	90,900	

Note that there is a 10% undercoverage of males, due to missing males within households, or missing all-male households. Each male has a 10% chance of being missed.

Neither column of numbers in table 2 is observed, since there are no household controls. Also the actual household type is not known for the sample units. Thus, the [FM] households appear to be the same as the F households. The data which would be observed are given in Table 3, along with the total initial weight for households which appear to have a given type. The adjusted weights are given for three methods, MLE-H, MLE-P, and principal person. The results for GLS-H and MDI-H are fairly close to MLE-H, and GLS-P and MDI-P are similar to MLE-P, so these other methods are omitted.

The last three columns of Table 3 show the total adjusted weight assigned to each actual household type by the MLE-H, MLE-P, and principal person methods. The principal person weights for each actual household type agree with the population counts for the actual types, shown in the third column of Table 1. In this sense, the principal person weights are unbiased.

This example corresponds to assumptions upon which the simplified principal person is based. The principal person adjusted weights for each actual type of household coincide with the population counts. The one difference is that totally missing [M] or [MM] households are given no weight; however, the weight of the non-missing M or MM households is increased accordingly. The total weighted number of households for the principal person method is equal to the number in the population.

Table 3

Within-household Undercoverage Example: Observed Types and Weights, with Adjusted Weights from Three Methods

Household Type	Total Initial Weight	Weight Assigned to Apparent Type			Weight Assigned to Actual Type		
		MLE-H	MLE-P	Principal Person	MLE-H	MLE-P	Principal Person
F	29,000	27,450	26,973	29,000	23,664	23,253	25,000
M	13,500	14,997	16,338	15,000	14,997	16,338	15,000
FF	8,200	7,368	7,626	8,200	6,290	6,510	7,000
FM	37,200	38,887	39,128	37,200	41,419	41,586	40,000
MM	4,500	5,623	5,446	5,000	5,623	5,446	5,000
FFM	10,800	10,661	10,885	10,800	11,739	12,001	12,000
FMM	10,800	12,605	11,878	10,800	13,859	13,140	12,000
FFF	0	0	0	0	0	0	0
MMM	0	0	0	0	0	0	0
Total	114,000	117,591	118,274	116,000	117,591	118,274	116,000

In this example, the constrained minimum distance methods overestimate the total number of households, but give too little weight to the households without males. In general, too much weight is given to households with males.

It should not be concluded that the principal person method always outperforms the constrained minimum distance methods when there is within-household undercoverage. Under other assumptions about coverage, the principal person method may not do so well. In fact, different versions of the principal person method are used for different surveys, based on various assumptions about coverage. Note also that combinations of the principal person method and raking methods are possible; see Scheuren (1981).

Even in this example, the biased weights assigned by the constrained minimum distance methods could be beneficial for estimating some characteristics. If the households in which males are missed tend to under-report the variable of interest, then giving these households too high a weight may tend to counteract response bias associated with the within-household undercoverage.

The most extreme example of this effect is estimation of the total number of males, in which case the MLE-H and MLE-P weights give estimates which agree with the control totals while the principal person weights do not. However, for household characteristics where there would rarely be reporting errors because of the missed male, such as form of tenure (renter/owner), the biased weights would not be desirable. The performance of the weighting methods in situations like these clearly depends on the nature of the survey undercoverage, and its relationship to the variable being estimated. This is discussed further, with additional examples, in Alexander and Roebuck (1986).

Pending further research on survey coverage and its effect on weighting, what recommendations can be made? Among the constrained minimum distance methods considered in this paper, GLS-H, MDI-H, and MLE-H seem unattractive because of their failure to adjust correctly for uniform undercoverage of households. This is in spite of the fact that, if there were no undercoverage, MLE-H seems to be based on a more sensible model than MLE-P, since households rather than persons are the ultimate sampling unit.

The possibility of negative weights raises questions about the appropriateness of GLS-P, even though in some practical applications (such as Zieschang 1986b) there are very few negative weights, so that they could be replaced by positive weights with little effect on the estimates. That leaves MDI-P and MLE-P. Our results give little basis for choosing between these methods. Computational considerations tend to favor the "raking" method MDI-P. Based on limited experience with the algorithms of Section 3, the MLE methods converge more slowly than the MDI methods. Further, there has been considerable research into ways to improve the efficiency of raking for large-scale applications, such as Ireland and Scheuren (1975). Taking all this into account, the raking method, MDI-P, seems to be the most promising of the constrained minimum distance methods.

The constrained minimum distance methods give household weights which are consistent with control totals for person, unlike the principal person method. However, the superiority of the constrained minimum difference methods over the principal person method as an adjustment for undercoverage is far from obvious. Undercoverage is an essential part of the survey weighting problem. The principal person method is an ad hoc solution to the undercoverage problem, based on some very simplistic assumptions about coverage. However, as seen in Section 4, the constrained minimum difference methods may be viewed as "optimal" (i.e., maximum likelihood or the asymptotic equivalent) estimators under models which assume perfect coverage. The choice is thus between an optimal solution to the wrong problem and an ad hoc solution to what may or may not be the right problem. Clearly more research is needed.

6. SOME AREAS FOR FURTHER RESEARCH

6.1 Household Control Totals

If independent estimates of the number of households of different kinds were available, then ordinary post-stratification could be used for household estimates. Household controls by size of household are being investigated, based on updating 1980 census results (Das Gupta *et al.* 1986). The availability of household controls would fundamentally change our ability to deal with the household weighting problem.

Even with household controls, it might be beneficial to also incorporate person controls. The household controls are not likely to include detailed information on the age, race, and sex of the household members. The use of raking to simultaneously control the estimates to independent controls for persons and households is developed by Scheuren (1981), using an estimate of the total number of households. Zieschang (1986a) describes how similar adjustments may be made using generalized least squares.

Household controls clearly have great potential for adjusting for differential coverage of various types of households. There still may be problems is dealing with within-household undercoverage, since this may lead to errors in determining the true household size, which would cause sample households to be placed in the wrong post-stratification cell.

6.2 Research Concerning Coverage

Coverage of persons is measured fairly well by comparing the initial survey estimates $\hat{N}_j$ to the control totals N_j. It is difficult to determine how much of this undercoverage is due to missing entire households and how much is due to missed persons within households. Additional information could be obtained by comparing initial weighted household estimates

to household controls, once these controls become available. In the meantime, 1980 survey estimates by type of household could be compared to the corresponding 1980 census counts.

Even with this additional information, it is not possible to completely distinguish household undercoverage from within-household undercoverage, without making additional assumptions. Alexander and Roebuck (1986) present some preliminary suggestions about how a range of coverage models might be fit to census and survey data. An alternative approach would be to include coverage parameters in a multinomial sampling model such as those described for the MLE-H or MLE-P weighting methods. Other approaches to modelling coverage are presented in Wolter (1986).

6.3 Estimation of Variances

Methods for estimating variances of the weighted estimators have not been investigated for most of the constrained minimum distance methods. For raking estimators, some methods are available; see Arora and Brackstone (1977), Bankier (1978) and Fan *et al.* (1981).

For any of the methods, replication methods for estimating the variance could be applied. These methods have been shown to give reasonable results under fairly general conditions; see for example Krewski and Rao (1985). It remains to be determined whether these conditions can be applied to the constrained minimum distance methods.

6.4 Computational Issues

Zieschang (1986b) has applied the generalized least squares methods to the U.S. Consumer Expenditure Surveys. Scheuren (1981) describes a large-scale application of the raking method to household weighting. The maximum likelihood constrained minimum distance algorithms (MLE-H and MLE-P) have not been tried on large-scale problems of this kind. If they were to be used in actual survey weighting, research may be needed to improve their computational efficiency.

ACKNOWLEDGMENTS

The author would like to thank Michael J. Roebuck for his assistance with portions of this research, and also the associate editor and referees for their helpful comments. The author is also indebted to Brenda Kelly for her diligence in typing this manuscript.

REFERENCES

ALEXANDER, C.H. (1986). The present Consumer Expenditure Surveys weighting method. In *Population Controls in Weighting Sample Units,* Section 1. Washington, D.C.: U.S. Bureau of Labor Statistics, 1-32.

ALEXANDER, C.H., and ROEBUCK, M.J. (1986). Comparison of alternative methods for household estimation. *Proceedings of the Section on Survey Research, American Statistical Association,* 54-64.

ARORA, H.R., and BRACKSTONE, G.J. (1977). An Investigation of the Properties of Raking Ratio Estimates: II. With cluster sampling. *Survey Methodology,* 4, 232-252.

BANKIER, M.D. (1978). An estimate of the efficiency of raking ratio estimators under simple random sampling. *Survey Methodology,* 4, 115-124.

BRACKSTONE, G.J., and RAO, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyā,* Series C., 41, 97-114.

BISHOP, Y.M.M., FIENBERG, S.W., and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, Massachusetts: MIT Press.

CRESSIE, N., and READ, T.R.C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society,* Series B, 46, 440-464.

DARROCH, J.N., and RATCLIFF, D. (1972). Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics,* 63, 1470-1480.

DAS GUPTA, P., GIBSON, C., HERRIOT, R.A., LAMAS, E., and ZITTER M. (1986). New approaches to estimating households and their characteristics for states and counties. Paper presented at the 1986 annual meeting of the Population Association of America.

DEMING, W.E., and STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected margins are known. *Annals of Mathematical Statistics,* 11, 427-444.

FAGAN, J.T., and GREENBERG, B. (1985). Algorithms for making tables additive: raking, maximum likelihood, and minimum chi-square. U.S. Bureau of the Census, Statistical Research Division Report Series No. Census/SRD/RR-85/12.

FAN, M.C., WOLTMAN, H.F., MISKURA, S.M., and THOMPSON, J.H. (1981). 1980 census variance estimation procedure. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 176-181.

FIENBERG, S.E. (1986). Comments on some estimation problems in the Consumer Expenditure Surveys. In *Population Controls in Weighting Sample Units.* Section 5. Washington, D.C.: U.S. Bureau of Labor Statistics, 1-12.

HABER, M., and BROWN, M.B. (1986). Maximum likelihood methods for log-linear models when expected frequencies are subject to linear constraints. *Journal of the American Statistical Association,* 81, 477-482.

HUANG, E.T., and FULLER, W. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of Social Statistics Section, American Statistical Association,* 300-305.

IRELAND, C.T., and SCHEUREN, F.J. (1975). The rake's progress, *Computer Programs for Contingency Table Analysis.* Washington, D.C.: The George Washington University, 155-216.

KREWSKI, D., and RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics* 9, 1010-1019.

LUERY, D.M. (1986). Weighting sample survey data under linear constraints on the weights. *Proceedings of the Social Statistics Section, American Statistical Association,* 325-330.

OH, H.L., and SCHEUREN, F.J. (1978a). Multivariate raking ratio estimation in the 1973 Exact Match Study. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 716-722.

OH, H.L., and SCHEUREN, F.J. (1978b). Some unresolved application issues in raking ratio estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 723-725.

PUGH, R.E., TYLER, B.S., and GEORGE, S. (1976). Computer-based procedure for N-dimensional adjustment of data - NJUST. U.S. Social Security Administration, Staff Paper No. 24.

SCHEUREN, F.J. (1981). Methods of estimation for the 1973 Exact Match Study. *Studies from Interagency Data Linkages, Report No. 10.,* U.S. Department of Health and Human Services, U.S. Social Security Administration, 9-122.

STEPHAN, F.F. (1942). An iterative method of adjusting sample frequency tables when expected marginal totals are known. *Annals of Mathematical Statistics,* 13, 166-178.

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

ZIESCHANG, K.D. (1986a). Generalized least squares: an alternative to principal person weighting. In *Population Controls in Weighting Sample Units,* Section 2. Washington, D.C.: U.S. Bureau of Labor Statistics, 1-41.

ZIESCHANG, K.D. (1986b). A generalized least squares weighting system for the Consumer Expenditure Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 64-71.

Survey Methodology, December 1987
Vol. 13, No. 2, pp. 199–207
Statistics Canada

An Integrated Method for Weighting Persons and Families

G. LEMAÎTRE and J. DUFOUR[1]

ABSTRACT

Household surveys generally use separate procedures for estimating characteristics of persons and those of families. An integrated procedure is proposed and a least-squares estimator introduced to achieve this end. The estimator is shown to be unbiased under certain general conditions. Using data from the Canadian Labour Force Survey, variances for the estimator are calculated and shown to compare favourably to those from current procedures.

KEY WORDS: Family estimation; Family weighting; Least-squares weighting.

1. INTRODUCTION

It is customary for many household surveys to incorporate in their estimation procedures a post-stratification step in which the design-based estimates of the population, generally by age and sex group, are benchmarked to independent totals obtained from demographic sources. In practice, for ease of tabulation, a weight is normally associated with each responding person, equal to the product of the inverse sampling rate, an adjustment for non-response, and an age/sex ratio adjustment factor. Estimates for a particular characteristic are then obtained by summing up the weights of all responding persons in the sample bearing that characteristic. Because of the age/sex adjustment factors, the weight so assigned will usually differ from person to person within the same household. When estimating characteristics of persons, this may not pose any particular problem; in producing estimates of households or families, however, it is not entirely clear which weight is the appropriate one to use, if any.

To estimate family characteristics, one might well elect to carry out a ratio estimation step using auxiliary information on families as well as persons. However, reliable and timely auxiliary counts of families that could be used in ratio estimation are in general not available. As a result of events such as births, deaths, marriages, divorces and persons leaving or entering a household, characteristics such as family size change from one census to the next, in ways that are less predictable than a characteristic such as age. The administrative records that are the main source of information on post-censal population change (i.e. birth, death and migration records), do not provide information on household-related change. Birth records, for example, do not provide information on the size of a family into which a child is born. Tax records can compensate in part for this deficiency (see Auger 1987); however, such records do not cover the entire population nor are they available in a timely enough fashion to be used in producing current estimates. In the absence of auxiliary counts of families, household surveys generally have adapted the weights obtained from "person-weighting" for use in estimating characteristics of families. For various reasons this is a somewhat less than ideal solution. The present paper proposes a method of estimation that results in a single uniquely defined weight per household which would be appropriate for both individual and family estimation.

[1] G. Lemaître and J. Dufour, Social Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

Techniques to achieve a single household weight have been proposed in the past, with an emphasis on using auxiliary information on persons to improve estimates of families. Oh and Scheuren (1978) proposed a method of "multivariate raking" which consists of successively ratio adjusting population estimates by post-stratum by means of the ratio adjustments calculated for each post-stratum in turn, and then iterating to convergence. The adjustments at each stage are applied to households containing persons in the particular post-stratum being adjusted for. Zieschang (1986) adopted a Generalized Least Squares (GLS) approach in which the sum of weighted squared adjustments to the design weights were minimized, subject to a set of linear constraints. Alexander (1987) examines several constrained minimum distance weighting methods, including the GLS method, and evaluates them in the context of survey undercoverage. Although the above methods were originally proposed as ways of improving estimates of families, the survey weights derived from the various estimators can clearly be used to estimate characteristics of persons as well. This paper argues in favour of adopting such an integrated approach to individual and family estimation. Section 2 discusses the limitations of the current approaches to estimating characteristics of persons and families. Section 3 introduces a model-based estimator adapted from a generalized weighting procedure due to Bethlehem and Keller (1987). Section 4 presents some empirical results taken from the Canadian Labour Force Survey. Section 5 discusses plans for further study.

2. CURRENT ESTIMATION PROCEDURES

The principal mandate of most household surveys traditionally has been to produce estimates for characteristics of persons, particularly of labour force characteristics. Such surveys adopt the household as the ultimate sampled unit essentially for reasons of cost and convenience. Although the household unit is normally respected in preliminary weighting steps (non-response adjustments, rural/urban adjustments, etc.), it is generally ignored in the final weighting step, i.e. no allowance is made for the fact that the members of a household are sampled as a unit. In particular, any coverage biases associated with the sampled unit are not directly taken into account or compensated for in estimation. Undercoverage is thus assumed to be ignorable in the sense of Rubin (1976); every person in an age/sex post-stratum is treated the same in estimation whether he/she is living alone or comes from a multi-person household. One study of non-response in the Labour Force Survey (Paul and Lawes 1983), however, has demonstrated that smaller households, particularly households without children, tend to be underrepresented in the sample. Although no comparable studies exist for missed households in the Labour Force Survey, studies of private household undercoverage in the census have shown that non-enumerated households are indeed smaller on average than enumerated households (Gosselin and Théroux 1980). A missing-at-random type procedure can lead to biases in labour force estimates for persons, particularly if the labour force distribution of persons in smaller households is different from that of persons in larger ones, all things being equal. Intuitively, an estimation procedure which takes into account (even if only indirectly) the fact that smaller households are more subject to non-response and undercoverage than larger ones could correct in part for this deficiency in the sample.

In the absence of auxiliary information on households or families that could be incorporated into an appropriate weighting procedure to produce a well-defined family weight, many current methods adopt as the family weight the weight of a "principal person" in the family. In the Canadian Labour Force Survey, this person is the female spouse if present, otherwise the head. Since such methods do not take household composition into account,

family estimates generated using this weight tend to overestimate larger families and to underestimate unattached persons. In addition many characteristics (e.g., population, income) can be estimated using either the individual weight or the family weight, and the estimates will in general disagree, sometimes substantially. Of course even under ideal sampling and interviewing conditions, with no differential non-response or undercoverage, family and individual-based estimates of the same characteristic will disagree somewhat. With a large enough sample, however, the discrepancies should be small. Under actual, i.e., less than ideal conditions, differences may be too large to explain away by a facile appeal to sampling variability. An estimation procedure that yields a single household weight which, when used as an individual weight, respects the auxiliary population totals will eliminate the awkwardness of having two estimation systems. It is these deficiencies that the estimator described in the following section was designed to deal with.

3. A PROPOSED ESTIMATOR

We begin by introducing a generalized weighting procedure based on linear models due to Bethlehem and Keller (1987) and applying it first to person-based estimation as was done in their paper. A modification of the procedure is introduced which leads to household weights appropriate for estimating characteristics of persons. We will borrow freely from their original presentation in what follows.

Assume a survey target population consisting of N units, an N-vector Y of values of a target variable, and an N by p matrix X of auxiliary variables defined for each unit of the target population. The population totals for each auxiliary variable are assumed to be known and will be denoted collectively by the p-vector x. In our application x will consist of age-sex totals. If the auxiliary variables are correlated with the target variable, then for an appropriate p-vector B, the values of $E = Y - XB$ will vary less than the values of the target variable Y. Ordinary least squares on all units of the target population yields

$$B = (X'X)^{-1}X'Y, \tag{3.1}$$

provided X is of full rank. A sample-based estimate for B is given by

$$\hat{B} = (X'\Pi^{-1}TX)^{-1}X'\Pi^{-1}TY, \tag{3.2}$$

where T is a diagonal matrix whose i-th element is 1 if the i-th unit of the population is in the sample, 0 otherwise, and $E(T) = \pi$.

It can be shown that for large samples $\hat{B}$ will be approximately unbiased. The parameter of interest, however, is not B but the population total y. If we define $\hat{y} = \hat{B}'x$, $\hat{y}$ will be an approximately unbiased estimator of y provided that $B'x = y$, or equivalently, provided the sum of the residuals for the population model $Y = XB + E$ is equal to zero. This will hold if the N-vector whose elements consist of ones is in the space spanned by the columns of X, and in particular, if the auxiliary variables X include an exhaustive and mutually exclusive set of indicator variables (for age/sex groups, for example).

If we write $\hat{y} = \hat{B}'x = Y'\Pi^{-1}TX(X'\Pi^{-1}TX)^{-1}x$, we see that the estimator implicitly defines an N-vector of weights given by

$$W = \Pi^{-1}TX(X'\Pi^{-1}TX)^{-1}x,$$

that do not depend on the particular target variable being estimated. If these weights are used to produce sample estimates for the auxiliary variable characteristics, we have that $X'W = x$, so that the weights do indeed yield the appropriate population totals. Furthermore if X consists exclusively of an exhaustive and mutually exclusive set of indicator variables, then the regression estimator $\hat{y}$ will be equivalent to the ordinary post-stratification estimator. For further details, see Bethlehem and Keller (1987).

The weight of an arbitrary sample person i under this procedure can be expressed generally as

$$W_i = \sum_j \frac{x_{ij} b_j}{\pi_i}, \tag{3.3}$$

where $(b_1, \ldots, b_p) = (X'\Pi^{-1}TX)^{-1}x$ and π_i is the inclusion probability for person i. This suggests that the estimation method described above can be adapted to yield the desired weights by defining the auxiliary variables in the same way for all household members. An obvious way to do this is to define auxiliary variables at the household level, for example by replacing the corresponding variables defined at the person level by the household mean. More formally let Z be an N by p matrix defined for person i $(i = 1, \ldots, N)$ belonging to household h $(h = 1, \ldots, H)$ by

$$Z_{ij} = \frac{U_{hj}}{n_h},$$

where U_{hj} is the total for characteristic j in household h, i.e. $U_{hj} = \Sigma_k X_{kj}$, with the summation being over all members k of household h, n_h = size of household h, and $\Sigma_h n_h = N$. Let Y again be an N-vector of values for an arbitrary target variable defined on *persons*. As in person-level estimation, we work with the population model $Y = ZC + E$ and apply least squares to the sample data to obtain an estimate

$$\hat{C} = (Z'\Pi^{-1}TZ)^{-1}Z'\Pi^{-1}TY. \tag{3.4}$$

We define $\hat{y} = \hat{C}'x$ where x is again the vector of population totals for the auxiliary variables. $\hat{y}$ will be an approximately unbiased estimator of y provided the N-vector of ones is in the space spanned by the columns of Z. In a manner analogous to (3.3), the weight for an arbitrary sampled person in household h will be given by

$$W_h = \sum_j \frac{U_{hj} c_j}{\pi_h n_h}. \tag{3.5}$$

Since each household member contributes the same row vector to Z and since each has the same first order inclusion probability, each person within a household will have the same weight. Furthermore the use of the household weight as a person weight yields the correct auxiliary population totals. Although it is possible to obtain negative weights under this procedure (if some of the c_j's are less than zero), for well-behaved samples (i.e., not subject to serious non-response or undercoverage) households whose weights are changed substantially by this procedure tend to be households of unusual composition that are uncommon in the sample and in the population at large. Recently in weighting twenty-four months of Labour Force

Survey data under this procedure, only one household had a (small) negative weight attributed to it. Negative weights are problematic because it is difficult to attach the usual meaning one assigns to weights, that is, the number of persons/households in the population at large represented by a particular sampled person/household. However, under the formulation described above, the final weights are defined only implicitly and indeed could be viewed as merely a convenient means of generating estimates. In practice even with some negative weights, it is unlikely that a meaningful estimate of level for a characteristic of interest would turn out negative. The problem of explaining a negative weight to a mystified user is of course a different question.

The variance of the estimator $\hat{y} = \hat{C}'x$ described in this paper can be obtained using methods described in Fuller (1975). In addition the estimator can be shown to be equivalent to the GLS estimators proposed by Zieschang (1986) and Alexander (1987) when the space spanned by the auxiliary variables Z contains a vector of ones. Further properties of this type of estimator can be found in Wright (1983).

4. EMPIRICAL RESULTS

The Canadian Labour Force Survey is a monthly rotating panel survey of approximately 48,000 households across Canada (see Platek and Singh 1976 and Singh, Drew, and Choudhry 1984). Households once selected remain in the sample for six consecutive months before being replaced. The primary geographic strata are the ten provinces. Sample sizes vary from a low of 1500 households in Prince Edward Island, the smallest province, to about 9000 households in Ontario, the most populous one. The survey collects data concerning the labour market situation of respondents during a reference week each month and publishes a wide variety of estimates related to the nation's labour supply.

A preliminary evaluation of the estimator described above was carried out using data from one of the monthly surveys. May 1981 was chosen to permit comparisons to results from the 1981 census held at about that time. Although we have been using the terms "household" and "family" interchangeably up to now, user interest is often focused on estimates of "economic families", which consist of all persons in a household related by blood, marriage, or adoption. For weighting purposes it is conceptually more appealing to deal with the actual sampled unit, i.e. the household. However, the empirical results presented here will be based on estimates for economic families. The evaluation carried out focused on both characteristics of persons (labour force status) and of families (number of economic families and number of unattached persons). The least-squares weighting was carried out for two sets of five-year age/sex groups, with persons seventy and over being grouped according to sex. The first set of (twenty-four) age/sex groups excluded children 0 to 14 years of age from the weighting, to permit a comparison to a standard person-based post-stratification estimator using the same auxiliary information. The second set included children grouped into six age/sex groups and was used only for least-squares weighting, since under standard post-stratification the weighting of children would have no effect on the weighting of persons 15 and over.

Although all estimators considered are approximately unbiased for estimates of characteristics of persons, each makes different assumptions about the nature of under-coverage and non-response. (The Labour Force Survey's non-response adjustment procedure assumes that non-responding households are missing at random within geographic area). The post-stratification estimator implicitly assumes that any differential non-response and under-coverage depends only on age and sex and is therefore adequately compensated for by

person-based estimation using auxiliary information on these characteristics. Under least-squares weighting, the weight of a person will depend on the age/sex composition of the household (without children in one case, with children in the other). Thus, all things being equal, one would expect the design weight of a person belonging to an age/sex group subject to substantial undercoverage to be adjusted less if that person is living with persons belonging to age/sex groups well covered by the sample than if he/she is living alone.

Since the auxiliary population totals by age and sex are available by province, estimation was carried out separately for each province. However, the smaller provinces have been collapsed into two groups in the following tables.

In general the three estimators do not yield substantially different estimates, particularly A and B. The inclusion of children in the weighting does appear to lead to slightly higher estimates of employment and of unattached persons and slightly lower estimates of economic families nationally and in the larger provinces (compare results from Scheuren *et al.* 1981). This is in line with expectations, although there is still some ground to cover vis-a-vis census results, which show (rounded to thousands) 6,369,000 economic families and 2,583,000 unattached persons at the national level. The moral of the tale is that, although the least-squares estimator does take us part of the way home (when the presence of children is taken into account), it will require accurate and timely auxiliary information to eliminate the residual bias.

Table 1

Number of Persons Employed and Unemployed, Number of Economic Families and Unattached Persons, Labour Force Survey, May 1981 (In Thousands)

Estimator[a]		Employed	Unemployed	Economic Families	Unattached Persons
Canada	A	11,094	850	6,424	2,432
	B	11,090	850	6,446	2,442
	C	11,120	851	6,410	2,495
Atlantic Region	A	819	102	563	156
	B	819	102	570	154
	C	821	102	569	156
Quebec	A	2,725	304	1,723	587
	B	2,724	304	1,725	596
	C	2,735	305	1,714	614
Ontario	A	4,198	274	2,325	863
	B	4,200	273	2,325	861
	C	4,211	273	2,310	881
Prairie Region	A	2,074	83	1,078	506
	B	2,072	84	1,089	510
	C	2,074	83	1,085	517
British Columbia	A	1,277	88	735	319
	B	1,276	88	738	321
	C	1,280	88	734	327

[a] A = post-stratification/principal person, B = least squares with children excluded from weighting and C = least squares with children included in weighting.

The expected performance of the least-squares estimator with regard to efficiency is not altogether obvious. Certainly, if one were to base a prediction on the results observed above, then the similarity of the estimates to those produced by the post-stratification estimator would lead one to expect it to perform as well as the latter. On the other hand, one might expect efficiency gains for estimates of economic families, because of the fact that the least-squares estimator makes use of the auxiliary population totals in determining the household weight. However, a single weight per household is not achieved without some redistribution of weights at the micro level.

Table 2

Distribution of Percent Deviations of Final Weights Relative to the Design Weights, Labour Force Survey, May 1981

Percent Deviation	Percentage of Total Sample		
	Post-Stratification	Least-Squares	Least-Squares (With Children)
> −30%	0.0	0.1	0.2
−30 to −20%	0.0	0.5	0.9
−20 to −10%	0.6	3.0	5.3
−10 to 0%	23.9	20.4	27.1
0 to 10%	53.9	44.6	37.3
10 to 20%	20.6	26.3	21.6
20 to 30%	0.6	4.4	6.2
30 to 40%	0.1	0.4	0.9
40 to 50%	0.0	0.0	0.2
< 50%	0.0	0.0	0.2

Note: Sample size is N = 159014.

Table 3

Estimated Efficiencies of Least-Squares Estimators Relative to Post-Stratification Estimator, Labour Force Survey, May 1981

	Estimator[a]	Employed	Unemployed	Economic Families	Unattached Persons
Canada	B	1.044	0.999	1.565	1.038
	C	1.066	0.999	1.616	1.036
Atlantic Region	B	1.110	0.977	1.266	0.998
	C	1.193	0.992	1.567	1.070
Quebec	B	1.059	1.005	1.553	1.020
	C	1.063	0.992	1.582	0.992
Ontario	B	1.028	1.011	1.825	1.064
	C	1.059	1.010	1.828	1.037
Prairie Region	B	1.001	1.009	1.205	1.009
	C	1.072	1.066	1.420	1.134
British Columbia	B	1.038	0.964	1.248	1.048
	C	1.053	0.978	1.203	1.045

[a] B = least squares with children excluded from weighting and C = least squares with children included in weighting.

As Table 2 illustrates, the least-squares weights have a somewhat greater dispersion than those based on standard post- stratification methods. Including children in the weighting results in an even greater dispersion. The movement in the weights essentially reflects the extent to which the age/sex household size composition of the sample fails to mirror that existing in the general population. Since the objective of a single weight per household imposes an additional constraint on the estimation procedure, one might expect variances to suffer somewhat, particularly if no additional auxiliary information is brought to bear in estimation.

Variances for the post-stratification estimator were estimated using the Keyfitz method (1957) with PSU's (primary sampling units) or collapsed PSU's as replicates. The least-squares variances were estimated using the method described in Fuller (1975). To ensure comparability, variances for several characteristics estimated by means of post-stratification were calculated using the Fuller technique and compared to those from the Keyfitz approach. In all cases the two sets of variance estimates were very close (within one or two percent).

Table 3 summarizes the estimated efficiencies of the least-squares estimators relative to post-stratification for the characteristics considered in Table 1. The efficiency gains for estimates of economic families are substantial. Estimates of persons employed and of unattached persons also appear to gain somewhat; however, the variance reductions for these characteristics are small, with the exception of employed in the Atlantic Region, particularly when children are included in the weighting. Interestingly average family sizes in the Atlantic Region are higher than in the rest of the country, although it is not clear how this would affect estimates of employed persons. The variances for the characteristic unemployed are essentially unaffected by the least-squares procedure. One can probably expect these results to hold in general, i.e. for arbitrary characteristics. Although the one-weight-per-household criterion is a restrictive one for estimates of characteristics of persons, the least-squares estimators appear to compensate through the additional "explanatory" variables of the linear model, i.e. the household means of all auxiliary variables. The above preliminary results suggest that individual and family estimation could be integrated at little or no loss in efficiency for estimates of persons.

5. PLANS FOR FURTHER STUDY

The results presented in this paper are preliminary, and a more extensive empirical evaluation of the properties of the least-squares estimator is currently under way, with particular attention being given to the behaviour of estimates over time and to efficiencies for a larger group of characteristics relative to estimates produced with the Labour Force Survey's current raking ratio estimator. The foregoing results have suggested that at least for some characteristics of persons, the "explanatory power" of the age-sex composition of a household is at least as great as that of the age-sex group alone. It will be instructive to see if the relative efficiencies will be as favourable for characteristics more strongly correlated with age-sex. In addition although in practice negative weights have been uncommon, it is likely that some procedure must be developed to deal with them when they occur. Among the possibilities one might consider would be to accord them outlier treatment or perhaps to forestall their occurrence by imposing some bound on changes to the weights (Zieschang 1987). Finally it would be useful to make explicit the undercoverage model underlying the least-squares estimator to permit an evaluation of the model on its own merits.

ACKNOWLEDGEMENTS

The author would like to thank F. Scheuren for his comments and suggestions regarding this paper.

REFERENCES

ALEXANDER, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology,* 13, 183-198.

AUGER, E. (1987). Family data from the Canadian personal income tax file. In *Statistics of Income and Related Administrative Record Research: 1986-1987,* (Eds. W. Alvey and B. Kilss), Washington, D.C.: Internal Revenue Service, 177-184.

BETHLEHEM, J.C., and KELLER, W.A. (1987). Linear weighting of sample survey data. *Journal of Official Statistics,* 3, 141-153.

FULLER, W.A. (1975). Regression analysis for sample surveys. *Sankhyā,* C37, 117-132.

GOSSELIN, J.-F., and THÉROUX, G. (1980). 1976 Census of Canada Quality of Data Series I: Sources of Error - Coverage. Catalogue No. 99-840, Statistics Canada.

KEYFITZ, N. (1957). Estimates of sampling variance where two units are selected from each stratum. *Journal of the American Statistical Association,* 52, 503-510.

OH, H.L., and SCHEUREN, F. (1978). Multivariate raking ratio estimation in the 1973 exact match study. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 716-722.

PAUL, E.C., and LAWES, M. (1982). Characteristics of respondent and non-respondent households in the Canadian Labour Force Survey. *Survey Methodology,* 8, 48-85.

PLATEK, R., and SINGH, M.P. (1976). Methodology of the Canadian Labour Force Survey. Catalogue No. 71-526, Statistics Canada.

RUBIN, D.B. (1976). Inference on missing data. *Biometrika,* 63, 581-592.

SINGH, M.P., DREW, J.D., and CHOUDHRY, G.H. (1984). Post '81 censal redesign of the Canadian Labour Force Survey. *Survey Methodology,* 10, 127-140.

SCHEUREN, F., OH, H.L., VOGEL, L., and YUSKAVAGE, R. (1981). Studies from Interagency Data Linkages, Report No. 10: Methods of Estimation for the 1973 Exact Match Study. U.S. Department of Health and Human Services, Social Security Administration, SSA Publication No. 13-11750.

WRIGHT, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association,* 78, 879-884.

ZIESCHANG, K.D. (1986). A generalized least squares weighting system for the Consumer Expenditure Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

ZIESCHANG, K.D. (1987). Sample weighting methods and estimation of totals in the Consumer Expenditure Survey. Unpublished manuscript, U.S. Bureau of Labor Statistics.

Survey Methodology, December 1987
Vol. 13, No. 2, pp. 209–219
Statistics Canada

Modified Raking Ratio Estimation

H. LOCK OH and FRITZ SCHEUREN[1]

ABSTRACT

A hybrid technique is described that employs both conventional and raking ratio estimation to handle the case when the population frequencies N_{ij} in a two-dimensional table are known, but some of the observed frequencies n_{ij} are small (or zero). Results are provided on the approach taken as it has evolved in the Corporate Statistics of Income Program over the last several years. Changes are still being considered and these will be discussed as well.

KEY WORDS: Raking ratio estimation; Conventional ratio estimation; Conditional bias and variance.

1. INTRODUCTION

Raking ratio estimation, or simply "raking," is a widely used technique in sample surveys. Applications differ depending on the nature of the sample design, the extent of the auxiliary information available and the presence of various nonsampling errors (such as might arise because of nonresponse or undercoverage).

Raking was first proposed by Deming and Stephan (1940) as a way of assuring consistency between complete count and sample data from the 1940 U.S. Census of Population. The originators themselves elaborated their ideas early on (Deming 1943; Stephan 1942). Since then, perhaps because of the basic intuitive appeal of the iterative algorithm employed, there have been several wholly independent rediscoveries of the technique (Fienberg 1970).

Advances and modifications have also been numerous. For example, important theoretical work on convergence of the algorithm was done by Ireland and Kullback (1968). As might be expected, practitioners at Statistics Canada, and also at the U.S. Bureau of the Census, have deeply studied the application of raking in census and survey taking, especially in situations where the raking is not allowed to proceed to complete convergence (e.g., Brackstone and Rao 1979; Fan *et al.* 1981). A reasonably complete bibliography of the statistical research on raking prior to 1978 can be found in Oh and Scheuren (1978b).

In many treatments of raking, it is assumed that two (or more) sets of marginal population totals, say $N_{i.}$ and $N_{.j}$, are known, but that the interior of the table N_{ij} can only be estimated from the sample. When the N_{ij} are also known, the usual ratio estimator with weights N_{ij}/n_{ij} would be the natural choice, unless the corresponding sample sizes n_{ij} are "too small."

The present paper describes a hybrid technique that employs both conventional and raking ratio estimations to handle the case when the population cell frequencies N_{ij} are known, but some of the observed frequencies n_{ij} are small (or zero). In Section 2, we describe our approach. Some empirical results from the application of the method to our Corporate Statistics of Income Program are covered in Section 3. In Section 4, we conclude with a brief summary and some plans for the future.

[1] H. Lock Oh and Fritz Scheuren, Statistics of Income Division, Internal Revenue Service, 1111 Constitution Avenue N.W., Washington, D.C. 20224, U.S.A.

2. RAKING RATIO ESTIMATION

2.1 General Considerations

Raking ratio estimation usually assumes that two (or more) marginal population totals, say, $N_{i.}$ and $N_{.j}$ are known, but that the interior of the table N_{ij} can only be estimated from the sample by, say, $\tilde{N}_{ij}$, where graphically (Deming 1943) we have

$$
\begin{array}{c|cccccc|c}
 & 1 & 2 & & \cdots & S & & \\
\hline
1 & N_{11} & N_{12} & & \cdots & N_{1S} & & N_{1.} \\
2 & N_{21} & N_{22} & & \cdots & N_{2S} & & N_{2.} \\
\vdots & \vdots & \vdots & \vdots & \boxed{N_{ij}} & \vdots & \vdots & \vdots N_{i.} \\
R & N_{R1} & N_{R2} & & \cdots & N_{RS} & & N_{R.} \\
\hline
 & N_{.1} & N_{.2} & \cdots & N_{.j} & \cdots N_{.S} & & N
\end{array}
$$

with $i = 1, \ldots, R$ and $j = 1, \ldots, S$. The corresponding sample count table is

$$
\begin{array}{c|cccccc|c}
 & 1 & 2 & & \cdots & S & & \\
\hline
1 & n_{11} & n_{12} & & \cdots & n_{1S} & & n_{1.} \\
2 & n_{21} & n_{22} & & \cdots & n_{2S} & & n_{2.} \\
\vdots & \vdots & \vdots & \vdots & \boxed{n_{ij}} & \vdots & \vdots & \vdots n_{i.} \\
R & n_{R1} & n_{R2} & & \cdots & n_{RS} & & n_{R.} \\
\hline
 & n_{.1} & n_{.2} & \cdots & n_{.j} & \cdots n_{.S} & & n
\end{array}
$$

In simple random sampling, the raking algorithm begins by setting

$$\tilde{N}_{ij} = \frac{N}{n} n_{ij}, \tag{2.1}$$

and then proceeds by proportionately scaling the $\tilde{N}_{ij}$ such that the relations

$$\sum_{j}^{S} \tilde{N}_{ij} = N_{i.} \tag{2.2}$$

and

$$\sum_{i}^{R} \tilde{N}_{ij} = N_{.j} \tag{2.3}$$

are satisfied in turn. Each step in the algorithm begins with the results of the previous step, with the $\widetilde{N}_{ij}$ continuing to change; the process terminates either after a fixed number of steps or when expressions (2.2) and (2.3) are simultaneously satisfied to the closeness desired. (See Oh and Scheuren (1983) for further details; see Ireland and Scheuren (1975) for generalizations to multi-way tables and the handling of computational efficiency issues.)

By an application of the theory of minimum discrimination information (Kullback 1968), it can be shown (e.g., Ireland and Kullback 1968) that, under some regularity conditions if only the $N_{i.}$ and $N_{.j}$ are known, the $\widetilde{N}_{ij}$ obtained by raking to convergence are asymptotically unbiased, normally distributed and minimum variance (i.e., best asymptotically normal, or BAN, estimators). Theoretical results of this kind are partly what motivates the raking estimator for a general survey characteristic Y_{ijk} (e.g., income or assets), where we are interested in estimating the population total

$$Y = \sum_{i}^{R} \sum_{j}^{S} \sum_{k}^{N_{ij}} Y_{ijk} \tag{2.4}$$

with, say, the statistic

$$\widetilde{Y} = \sum_{i}^{R} \sum_{j}^{S} \frac{\widetilde{N}_{ij}}{n_{ij}} \left(\sum_{k}^{n_{ij}} Y_{ijk} \right). \tag{2.5}$$

Typically, of course, in survey processing a raking weight

$$\widetilde{W}_{ij} = \frac{\widetilde{N}_{ij}}{n_{ij}} \tag{2.6}$$

is placed on each individual record on the file for ease of handling. It is important to note that a feature of the raking algorithm is that if $n_{ij} = 0$ then necessarily $\widetilde{N}_{ij} = 0$. For convenience, let $\widetilde{W}_{ij} = 0$ in such cases as well.

Our interest below will be mainly on the conditional properties of the various estimators being examined. Such an approach has considerable appeal, as advocated by Holt and Smith (1979) and Rao (1985). (As an aside, it may be worth noting that Brackstone and Rao (1979), among others, have looked at the conditional behavior of the raking estimator. They conditioned, however, on the sample marginals $n_{i.}$ and $n_{.j}$.)

2.2 Conditional Bias

Following Oh and Scheuren (1983) we focus primarily in this paper on the conditional properties of $\widetilde{Y}$, given $\underline{n} = (n_{11}, n_{12} \ldots, n_{RS})$. In particular, let $\bar{Y}_{ij}$ be the population mean for the ij-th subgroup. Then the conditional expected value of $\widetilde{Y}$ is

$$\mathrm{E}(\widetilde{Y} \mid \underline{n}) = \sum_{i}^{R} \sum_{j}^{S} \widetilde{N}_{ij} \bar{Y}_{ij} = Y + \sum_{i}^{R} \sum_{j}^{S} (\bar{Y}_{ij} - \bar{Y}) (\widetilde{N}_{ij} - N_{ij}). \tag{2.7}$$

Thus $\widetilde{Y}$ is conditionally biased with the importance of the bias depending on the structure of the population and whether or not the raking is to convergence. (Of course, when raking to convergence, unconditionally $E(\widetilde{N}_{ij}) = N_{ij}$ asymptotically.)

Employing the usual analysis of variance conventions (e.g., Scheffé 1959)

$$(\bar{Y}_{ij} - \bar{Y}) = (\bar{Y}_{i.} - \bar{Y}) + (\bar{Y}_{.j} - \bar{Y}) + (\bar{Y}_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}); \tag{2.8}$$

hence the conditional bias, given $\underline{n}$, is expressible as

$$\text{Bias}\,(\tilde{Y} \mid \underline{n}) = \sum_{i}^{R} (\bar{Y}_{i.} - \bar{Y})\,(\tilde{N}_{i.} - N_{i.}) + \sum_{j}^{S} (\bar{Y}_{.j} - \bar{Y})\,(\tilde{N}_{.j} - N_{.j})$$

$$+ \sum_{i}^{R} \sum_{j}^{S} (\bar{Y}_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y})\,(\tilde{N}_{ij} - N_{ij}). \tag{2.9}$$

If the raking is to convergence, then the first two terms of the conditional bias become zero. For the third term of the conditional bias to be zero for either form of raking, it is sufficient that the Y_{ij} be such that there is no interaction. In large-scale surveying with many variables, this is unrealistic to assume; nonetheless, in practice the interaction is often a minor part of the decomposition of Y_{ij}; consequently, the raking ratio estimator may, in many cases, have small biases even in moderate sample sizes.

2.3 Conditional Variance

Conditional and unconditional approaches to the variance of the raking ratio estimator have been extensively examined (e.g., Binder 1983; Causey 1972; Bankier 1986; Fan *et al.* 1981; Brackstone and Rao 1979). In our own early work (described in Section 3.2), we have employed replication techniques (e.g., Leszcz, Oh and Scheuren 1983). The replication methods used (which were equivalent to conditioning on the sample marginals) proved expensive, unwieldy, and somewhat unstable, leading us to a simpler attack on the conditional variance estimation problem (albeit the level of conditioning was deeper).

To motivate the approach we are currently taking, consider the conditional variance of $\tilde{Y}$, given $\underline{n}$. Now it can be shown by a slight extension of Oh and Scheuren (1983) that

$$\text{Var}\,(\tilde{Y} \mid \underline{n}) = \sum_{i}^{R} \sum_{j}^{S} n_{ij}\,(\tilde{W}_{ij})^2 \left(1 - \frac{n_{ij}}{N_{ij}}\right) V_{ij} \tag{2.10}$$

where the V_{ij} are the population variances of the ij-th subgroup and if $N_{ij} = 0$ or 1 we define $V_{ij} = 0$. (We are also employing the convention in expression (2.10) that $0/0 = 0$.)

Expression (2.10) holds whether or not the raking goes to convergence. Despite this it has been little studied because it cannot be readily adapted to estimate the conditional variance. The principal difficulty, of course, lies in our inability to calculate stable estimators of the V_{ij} when the n_{ij} are small. To overcome this problem we began looking at collapsing techniques based on the size of the raking weight. First, we let $\tilde{W}_{ij}$ approximate N_{ij}/n_{ij} which gives us

$$\text{Var}\,(\tilde{Y} \mid \underline{n}) \cong \sum_{i}^{R} \sum_{j}^{S} n_{ij}\,\tilde{W}_{ij}\,(\tilde{W}_{ij} - 1)\,V_{ij}. \tag{2.11}$$

Now if the $\tilde{W}_{ij}$ are ordered from smallest to largest and if they vary over a narrow range, then averaging them into (ordered) groups of, say, about $n_g \geq 25$ observations each will

alter the value of expression (2.11) very little. It will, however, allow us to calculate collapsed post-stratum variance estimates for the V_{ij}. This is the approach we have taken in Section 3.

One final point should be noted. The alternative proposed here is stable and fairly easy to calculate. Our limited empirical work, however, is inconclusive on the method's utility and, while we feel the method is worthy of discussion, we are in no sense advocating its general use at this time.

2.4 Modified Raking Estimation

As we have noted, under fairly general conditions the $\widetilde{N}_{ij}$ are BAN estimators. This does not mean, however, that $\widetilde{Y}$ will share all these properties. Indeed, if the variables used in the raking are not highly correlated with the characteristic Y, the estimator $\widetilde{Y}$ may suffer some degradation in variance relative, say, to a simple ratio estimator

$$\widetilde{\widetilde{Y}} = \sum_{j}^{S} \left(\frac{N_{.j}}{n_{.j}}\right)\left(\sum_{i}^{R} \sum_{k}^{n_{ij}} Y_{ijk}\right). \tag{2.12}$$

Typically, of course, experience has shown that both positive and negative impacts may occur in the same sample. The practitioner's problem is somehow to keep the positive effects while minimizing the negative ones.

There seems to be no general solution to this dilemma but we have had some limited successes, in our application settings, with two techniques that may be of wider interest (see Subsections 3.2 and 3.3 for results).

In most treatments of raking, it is assumed that the marginal population totals $N_{i.}$ and $N_{.j}$ are known; and that the interior of the table N_{ij} can only be estimated from the sample. In our setting we actually have the population values N_{ij} and are employing raking as a way of systematically handling cells in the table where the n_{ij} are small. Conventional collapsing alternatives exist here, of course (e.g., Cochran (1977) Fuller (1966)); but seemed unsuitable for reasons that will be explained later.

It may be possible to agree that raking is a satisfactory way of handling the small cells in this setting; but what about the larger ones? Surely it would be better to use the conventional simple ratio estimator in the large cells. Indeed, if this were done, the conditional bias for these "large" cells would be zero; but what would be the effect on the rest of the cells? This line of reasoning suggested that we employ a hybrid estimation method where, for cells where the n_{ij} was large, the conventional simple ratio estimator is used. These cells are then removed from the population and sample tables, and the remaining sample cells are raked to the adjusted population marginals.

For the remaining smaller cells, a second procedure was introduced to reduce the possible negative impacts of the raking on certain variables. We bounded the raking so that the weights $\widetilde{W}_{ij}$ did not vary "too much" from the initial weight. (This kind of constraint is often employed, by the way, in simple ratio estimation, e.g., Hanson 1978.)

The approach to bounded raking ratio estimation is similar to that when "large" sample counts are available in a single cell. That is, it is similar in that, for the cell that is to be constrained, we bound the $\widetilde{W}_{ij}$; then take the estimated population total $\widetilde{N}_{ij} = \widetilde{W}_{ij}\, n_{ij}$ for that cell and the sample n_{ij} for that cell out of the population and out of the sample tables (respectively); and then adjust the remaining observations.

Three problems exist with these partial "solutions." First there is the (uncomfortable) arbitrariness of the definitions of a "large" cell, and of a weighting factor that varies "too much" from its initial value. A related concern was why, if we were willing to use simple ratio estimation for "large" cells, conventional collapsed stratum techniques could not be

used for the remaining cells. The third problem has to do with the properties of the raking algorithm's convergence when we employ this hybrid. It is quite clear, for example, from the research that has been done on raking that tables with too many zeros in them will be very unstable and the raking may not converge (e.g., Oh and Scheuren 1978a and 1978b; Ireland and Scheuren 1975). This is of particular concern since the effect of both our modifications is to introduce zeros into the table. If these zeros are strategically placed, or better, *misplaced*, then this could have a very serious detrimental impact on the rate of convergence and, even, on the quality of the estimators. Our recommendation before starting was, therefore, that the number of times that these procedures were employed would have to be fairly small. It is beyond the scope of the present paper to resolve these concerns in general (if indeed that is possible). In Section 3, however, we will consider them further for the applied setting in which we did this work, and also will return to them in Section 4, when discussing areas for future study.

3. RAKING IN THE CORPORATE STATISTICS OF INCOME PROGRAM

3.1. Background

The U.S. Internal Revenue Service has produced statistics from corporate tax returns annually for over 70 years. Corporate data are, in fact, a mainstay of the so-called Statistics of Income Program, which is the name collectively given to all of the non-administrative statistical series produced by the Internal Revenue Service for public consumption.

Until 1951, corporate statistics were based on a complete census of the returns filed. Since then, a stratified probability sample has been employed, currently running in size at about 90,000 returns annually (from about 3,000,000 returns filed). Assets and income are the principal stratifying variables (Jones and McMahon 1984). Stratification by industry has long been considered, as well, but the quality of the industry coding as self-reported by taxpayers seemed insufficient to justify this step on a wholesale basis. Typically, for example, at the minor industry level perhaps 20 percent or more of the self-reported codes are changed during statistical processing. Nonetheless, because of the importance of industry statistics, efforts to use administrative data by industry to post-stratify the sample still seemed warranted and have been pursued over many years (e.g., Westat, Inc. 1974; Leszcz, Oh, and Scheuren 1983).

In a pilot post-stratification study done by Westat during the early 1970's, substantial improvements in standard errors were achieved for a number of variables, notably Total Receipts (where a reduction of about 12 percent occurred). Some increases in standard errors took place, however, for variables not closely related to industry (e.g., distribution to shareholders), but these were minor. To handle small cells, Westat used conventional collapsed stratum techniques to combine industry post-strata within the then-existing sample strata. Concerns continued to exist about the quality of the administrative industry data, especially for small cells; in any case, due to other operational priorities, the Westat approach was never implemented.

A major series of budget cuts occurred during the 1980-1982 period, and these forced a number of changes in the sample designs and estimation procedures across nearly all the studies that make up the Statistics of Income Program (e.g., Hinkins and Scheuren 1986; Scheuren, Schwartz, and Kilss 1984); in particular, the corporate study experienced sample size cuts during this period which, although later partially rescinded, reopened the issue of post-stratification by industry.

A raking ratio estimation approach to post-stratification seemed to have appeal over what Westat had done. One of the reasons for this was that concerns about the quality of the marginal administrative totals, by industry, were not as great as for the individual cells. The work of implementing a collapsing scheme could be completely avoided, as well.

3.2 Early Modified Raking Results

When we implemented a pure raking scheme for the Tax Year 1979 sample, our principal customers expressed concerns about what we had done. They were particularly worried about the potential for large adjustment factors having an adverse effect on certain statistics. We, in turn, having seen the results ourselves, were concerned that we had not done an adequate job for those industry-sample stratum combinations where the number of sample observations were large. As a consequence, these results were never used and the 1979 Tax Year statistics were published employing normal stratified sampling estimation (NORM).

Research continued, however, and in 1983, a paper was given comparing the root mean square errors of six different variations of raking both with each other and with what we had been doing previously (Leszcz, Oh, and Scheuren 1983). Three "pure" raking alternatives were looked at:

PRRE: "Classical" raking ratio estimation to convergence (Deming and Stephan 1940);

PRRE (200): Simple ratio adjustment of cells with samples of 200 returns or more and "classical" raking of the remaining cells to convergence; and

PRRE (400): Simple ratio adjustment of cells with samples of 400 returns or more and "classical" raking of the remaining cells to convergence.

In addition, three versions of bounded raking ratio estimation were examined, all with the bounds set at $(\sqrt{2/3}, \sqrt{3/2})$. These were:

BRRE: Bounded raking ratio estimation (2 cycles);

BRRE (200): Simple ratio adjustment of cells with samples of 200 and bounded raking (2 cycles) of the remaining cells; and

BRRE (400): Simple ratio adjustment of cells with samples of 400 and bounded raking (2 cycles) of the remaining cells.

For the bounded raking we were initially not sure that complete convergence was possible; hence, we made an operational simplification and only cycled through the constraint equations, e.g., (2.2) and (2.3), twice.

To make the root mean square error (RMSE) comparison, pseudo-replicate half-samples were drawn, each designed in the same way as the overall sample. The procedure involved: (1) construction of the half-samples; (2) two-way classification - by original sample stratum and major industry (post-stratum) - of sample counts for each half-sample; (3) derivation of a set of weights for each half-sample for each estimator; (4) calculation of estimates of selected items by applying the weight to sample values for each half-sample; and (5) calculation of the RMSE, based on the variations in the estimates that each half-sample produced. For cost reasons only 14 sets of half samples were used.

The resultant summary tabulation presented as Table 1 reveals what one would have expected of the number of returns. Near 100 percent reductions occurred for the PRRE, PRRE(200), and PRRE(400) estimates. Application of the bounding limits $\sqrt{2/3}$ and $\sqrt{3/2}$, and not cycling to convergence, decreased the magnitude of these reductions; however, they were still substantial. As Table 1 also indicates, for Total Receipts, a key variable, there were also improvements, although much less sizable.

Table 1

Reduction in Root Mean Square Error (RMSE)
as a Percent of Corresponding Normal Stratified Sampling RMSE

Estimator	Number of Returns	Total Receipts	Jobs Credit
"Pure" raking ratio estimators:			
PRRE	98.6	8.3	−3.09
PRRE (400)	98.6	9.2	−3.09
PRRE (200)	98.6	11.9	−3.09
Bounded raking ratio estimators:			
BRRE	74.0	13.8	+1.09
BRRE (400)	73.4	15.6	+1.09
BRRE (200)	72.3	17.4	+1.09

Note: The percentages shown are simple averages of the percent reductions in each of the 56 major industry groups used in the post-stratification. Notice that the percentage improvements for the "number of returns" column are nearly but *not* 100 percent for the PRRE estimators. This occurs because the raking took place for all corporations, with both the N_{ij} and n_{ij} defined on this basis; however, only active corporations (about 90 percent) were tabulated. The BRRE estimators in the "number of returns" column differ from each other and from the PRRE estimator because the cycling was not to convergence. This has subsequently been changed, beginning with Tax Year 1985.

Jobs Credit results in Table 1 are included to illustrate the expected tradeoff that can exist for items not closely related to industry. In particular, we see that in some cases there are (modest) increases in the root mean square errors for this item, due presumably to the fact that this field is less dependent upon the industry groupings utilized in this research.

It should be noted that, for Total Receipts, the decreases shown in the root mean square error, from the initial (NORM) estimate to that utilizing raking ratio estimation, all compare favorably with the Westat pilot study results. While we are encouraged by this comparison, a great deal has changed over the decade between the earlier Westat results and those in Leszcz, Oh and Scheuren (1983). What would really be telling, and what has not been done, is to compare conventional collapsing schemes with our modified approach to raking *on the same data set.*

One final point about Table 1; it reflects improvements in RMSE when tabulating by the administrative industry information which was used in the post-stratification. Because of differences between the administratively and statistically assigned classifications by industry, the figures shown in this table are therefore likely to overstate the improvements being achieved in our published statistics, since so many entities (over 20 percent) are recoded during the indepth processing done of our corporate sample.

3.3 Current Modified Raking Results

Beginning with Tax Year 1980, we began to regularly produce and publish our corporate statistics using the bounded raking ratio estimator BRRE(200) (U.S. Department of Treasury 1984). For Tax Years 1983 and later, we made the modifications described in Section 2.3 so that approximate conditional variances could be calculated. These were first published for Tax Year 1984 (U.S. Department of Treasury 1987). Also, in an effort to confirm the earlier results, we undertook for Tax Year 1984 to compare the conditional variance of the modified raking method being employed with the variance that would have been estimated had we used normal stratified sampling estimation. Before discussing the limited comparisons made, it might be worthwhile giving some of the application details on the corporate setting for 1984.

In our earlier work (Leszcz, Oh and Scheuren, 1983), and for 1984, the entire corporate return population of IRS Forms 1120 and 1120S was tallied into 58 major industry groups. For 1984, industry was cross-classified by 14 sample strata in each of the two processing years during which the sample had to be selected. Some of the major industries were so sparse that we immediately collapsed the industry detail to 56 groups. This still left a very large table (of 1568 cells).

It may be of interest to note that there were 414 "natural" zero cells in the population and an additional 125 zero cells arising in the sample. Before raking we removed 96 cells that had 200 or more sample observations; these cells were then each ratio adjusted separately. (In all, 57 percent of the Forms 1120 and 1120S corporate sample were so adjusted.) Finally, there were 73 cells that had to be bounded during the raking itself. This meant that altogether in the raking step there were 708 or 45 percent of the cells being treated as zeroes.

The raking was initiated by introducing the normal stratified estimator into each cell of the table. The marginal constraints imposed were (1) by industry and sampling period, and (2) by sample strata and sampling period. In the published statistics for 1984, and in the comparisons made here, the raking did not go to convergence; it was just carried out for two cycles. (Incidentally, concerns about the conditional bias of this approach have led us to rake our 1985 sample data to convergence.)

The results of the efforts for 1984 were to reduce the overall and industry-by-industry standard errors for frequencies by substantial amounts - only about half as much, however, as is shown in Table 1. Similar dampened improvements occurred for Total Receipts (8.7 percent) with many variables like Jobs Credit and Net Income experiencing little or no change in their standard errors overall (see U.S. Department of Treasury 1987, for details). As already noted, conditioning may be part of the reason for this difference (Holt and Smith 1979). The original results were conditional on the sample marginals $n_{i.}$ and $n_{.j}$; the later figures employed a deeper level of conditioning.

We are still examining other possibilities as to why the improvements are more modest than we found in the earlier work. Some obvious possibilities are the way we grouped the data from the smaller cells, including the consequent averaging of the weighting factors $\tilde{W}_{ij}$, and the collapsed variance estimation of the V_{ij}. Tabulating the data using our statistical industry coding, rather than the administrative coding, as in Table 1, may have been a major factor.

4. CONCLUSIONS AND AREAS FOR FURTHER STUDY

4.1 General

The modified raking approach for our corporate sample certainly seems to be an improvement over the normal stratified sampling approach taken formerly. There are, however, a number of unsettling *ad hoc* aspects of the method that trouble us. For instance, the connection between conventional collapsed stratum techniques and our modified raking procedure needs more study. Exploring changes in estimation techniques is not enough, however. More work on the basic sample design appears needed too. Finally, the variance approximation being used needs further looking at. We may well have paid a high price for stability and ease of calculation. As noted earlier, the statistical literature is full of good alternatives, and these deserve to be examined in a full-scale comparison with what we are currently doing.

4.2 Estimation Issues

There is considerable intuitive appeal in developing a post-stratification method that *smoothly* increases the degree of conditioning from just using marginal totals to using some

or all of the interior population counts as well. Our current approach has an embarrassing *ad hoc* flavor. Frankly, we see it just as a stop gap until we can increase the quality of the underlying administrative data by industry. Our main concern is to reduce response variation arising from taxpayer or processing errors. Even if we are unsuccessful in improving the administrative data directly, it may be possible to dampen the response error effects by looking at the tables by industry and sample stratum over several years. This is planned and may allow us to integrate, in a more complete way, raking on the one hand and collapsed post-stratum estimation on the other.

4.3 Design Issues

Improved administrative data by industry has obvious uses at the design stage. At the present time, coefficients of variation differ quite widely by industry, with the smaller industries being very poorly represented. No amount of after-the-fact post-stratification can correct for this completely. Improving the balance by industry, and over time, appear to be top priorities (e.g., Hinkins, Jones and Scheuren 1987).

ACKNOWLEDGMENTS

As is true of nearly all applied work, the authors have many people to thank for the results in this paper. Homer Jones, who has responsibility for the corporate sample, made important contributions, ably assisted by Richard Collins. Nat Shaifer provided illuminating statistics on the degree of comparability of the administrative and statistical coding of industry. Mike Leszcz played the leading role in presenting the earlier results on this topic at the 1983 American Statistical Association meetings in Toronto. The paper profited from conversations with Rod Little and many helpful suggestions were made by the referees. Bettye Jamerson assisted in the manuscript preparation.

REFERENCES

BANKIER, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association,* 81, 1074-1079.

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

BRACKSTONE, G.J., and RAO, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyā,* Ser. C, 41, 97-114.

CAUSEY, B.D. (1972). Sensitivity of raked contingency table totals to changes in problem conditions. *Annals of Mathematical Statistics,* 43, 656-658.

COCHRAN, W.G. (1977). *Sampling Techniques, (3rd ed.).* New York: John Wiley.

DEMING, W.E. (1943). *Statistical Adjustment of Data.* New York: Dover.

DEMING, W.E., and STEPHAN, F.F. (1940). On a least square adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics,* 11, 427-444.

FAN, M.C., WOLTMAN, H.F., MISKURA, S.M., and THOMPSON, J.H. (1981). 1980 Census variance estimation procedure. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 176-181.

FIENBERG, S.E. (1970). An iterative procedure for estimation in contingency tables. *Annals of Mathematical Statistics,* 907-917.

FULLER, W.A. (1966). Estimation employing post strata. *Journal of the American Statistical Association,* 61, 1172-1183.

HANSON, R. (1978). The Current Population Survey: design and methodology. Technical Paper No. 40, U. S. Bureau of the Census.

HINKINS, S., JONES. H., and SCHEUREN, F. (1987). Updating tax return selection probabilities in the corporate Statistics of Income program. A paper presented at the International Symposium on Statistical Uses of Administrative Data, Ottawa, Canada, November 23-25, 1987.

HINKINS, S., and SCHEUREN, F. (1986). Hot deck imputation procedure applied to a double sampling design. *Survey Methodology,* 12, 181-195.

HOLT, D., and SMITH, T. M. F. (1979). Post-stratification. *Journal of the Royal Statistical Society,* Series A, 142, 33-46.

IRELAND, C.T., and KULLBACK, S. (1968). Contingency tables with given marginals. *Biometrika,* 55, 179-188.

IRELAND, C.T., and SCHEUREN, F.J. (1975). The rake's progress. *Computer Programs for Contingency Table Analysis,* Washington, DC: The George Washington University, 155-216.

JONES, H., and MCMAHON, P.B. (1984). Sampling corporation income tax returns for Statistics of Income, 1951 to present. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 269-274.

KULLBACK, S. (1968). *Information Theory and Statistics.* New York: Dover.

LESZCZ, M., OH, H.L., and SCHEUREN, F. (1983). Modified raking estimation in the corporate SOI program. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 107-111.

OH, H.L., and SCHEUREN, F.J. (1978a). Multivariate raking ratio estimation in the 1973 Exact Match Study. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 716-722.

OH, H.L., and SCHEUREN, F.J. (1978b). Some unresolved application issues in raking ratio estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 723-728.

OH, H.L., and SCHEUREN, F.J. (1983). Weighting adjustments for unit nonresponse. In *Incomplete Data in Sample Surveys, Volume 2, Theory and Bibliographies,* (Eds, W.G. Madow, I. Olkin and D.B. Rubin), New York: Academic Press, 143-184.

RAO, J.N.K. (1985). Conditional inference in survey sampling. *Survey Methodology,* 11, 15-31.

SCHEFFÉ, H. (1959). *The Analysis of Variance.* New York: John Wiley, 106-119.

SCHEUREN, F., SCHWARTZ, O., and KILSS, B. (1984). Statistics from individual income tax returns: quality issues and budget cut impact. *Review of Public Data Use,* 12, 55-67.

STEPHAN, F.F. (1942). Iterative method of adjustment sample frequency tables when expected margins are known. *Annals of Mathematical Statistics,* 13, 166-178.

U. S. DEPARTMENT OF TREASURY (1984). *Statistics of Income - 1980 Corporation Income Tax Returns,* Publication 16, Internal Revenue Service, U.S. Department of Treasury.

U. S. DEPARTMENT OF TREASURY (1987). *Statistics of Income - 1984 Corporation Income Tax Returns,* Publication 16, Internal Revenue Service, U.S. Department of Treasury.

WESTAT, INC. (1974). Results of a study to improve sampling efficiency of statistics of corporation income. Bethesda, Maryland (unpublished).

Survey Methodology, December 1987
Vol. 13, No. 2, pp. 221–224
Statistics Canada

Comparison of the Horvitz-Thompson Strategy with the Hansen-Hurwitz Strategy

S.G. PRABHU-AJGAONKAR[1]

ABSTRACT

The Hansen-Hurwitz (1943) strategy is known to be inferior to the Horvitz-Thompson (1952) strategy associated with a number of IPPS (inclusion probability proportional to size) sampling procedures. The present paper presents a simpler proof of these results and therefore has some pedagogic interest.

KEY WORDS: Sampling strategies; Inclusion probability proportional to size; Positive definite quadratic form.

1. INTRODUCTION

Let U be a finite population consisting of N identifiable units $[U_1, U_2, \ldots, U_N]$. With the i-th unit of the population U_i are associated two numbers X_i and Y_i, where X_i's are known and Y_i's are fixed but unknown. Generally, X_i represents a measure of size of U_i which is highly correlated with Y_i.

For estimating the population total $T_y = Y_1 + Y_2 + \ldots + Y_N$, the Hansen and Hurwitz (1943) strategy consists of selecting with replacement n population units with probability proportional to X_i, and using the unbiased estimator

$$t_{HH} = \frac{1}{n} \sum_{r=1}^{n} \frac{y_r}{p_r}$$

where $p_r = X_r / T_x$, $T_x = X_1 + X_2 + \ldots + X_N$, and y_r $(r=1, 2, \ldots, n)$ represents the outcome at the r-th draw. It is easy to show, noting that $\Sigma Z_i = 0$,

$$\mathrm{Var}(t_{HH}) = \sum_{i=1}^{N} \frac{Z_i^2}{np_i} \tag{1}$$

where $Z_i = Y_i - p_i T_y$, $i=1, 2, \ldots, N$.

When population units are selected without replacement, Horvitz and Thompson (1952) proposed the unbiased estimator

$$t_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$$

[1] S.G. Prabhu-Ajgaonkar, Department of Mathematics and Statistics, Marathwada University, Aurangabad 431004, India.

where π_i $(i=1, 2, \ldots, N)$ denotes the probability of including the i-th population unit U_i in the sample. Further, when π_i is proportional to X_i, the sampling procedure is termed an IPPS scheme. For such a sampling procedure,

$$\mathrm{Var}(t_{HT}) = \sum_{i=1}^{N} \frac{Z_i^2}{np_i} + \sum_{i \neq j=1}^{N} Z_i Z_j \frac{\pi_{ij}}{n^2 p_i p_j} \tag{2}$$

where Z_i is given in (1), and π_{ij} $(i \neq j=1, 2, \ldots, N)$ represents the joint probability of including the i-th and j-th population units in the sample. When an IPPS procedure is specified, π_{ij} can be further simplified.

From (1) and (2),

$$\phi = \mathrm{Var}(t_{HT}) - \mathrm{Var}(t_{HH}) = \sum_{i \neq j=1}^{N} Z_i Z_j \frac{\pi_{ij}}{n^2 p_i p_j}. \tag{3}$$

2. COMPARISON OF STRATEGIES

Midzuno (1952), Sen (1952) and Sankaranarayanan (1969) proposed IPPS sampling schemes for estimating T_y, using the Horvitz-Thompson estimator t_{HT}. The Midzuno-Sen scheme is feasible if

$$p_i = \frac{X_i}{T_x} > \frac{n-1}{n(N-1)}, \; i=1, \ldots, N, \tag{4}$$

Sankaranarayanan's scheme requires the weaker condition

$$\sum_{j \in s} p_j > (n-1)/(N-1) \text{ for all } s \in S.$$

For both the schemes, the joint inclusion probabilities are given by

$$\pi_{ij} = \frac{n(n-1)}{N-2}\left(p_i + p_j - \frac{1}{N-1}\right).$$

Hence, from (3),

$$\phi = \frac{n(n-1)}{n^2(N-2)}\left[\sum_{i=1}^{N} \frac{Z_i^2}{p_i}\left(2 - \frac{1}{(N-1)p_i}\right) + \frac{1}{(N-1)}\left(\sum_{i=1}^{N} \frac{Z_i}{p_i}\right)^2\right]. \tag{5}$$

The above expression is nonnegative if

$$P_i > \frac{1}{2(N-1)}, \; i=1, 2, \ldots, N,$$

in which case the Horvitz-Thompson strategy is superior to the Hansen-Hurwitz strategy. The above restriction on X_i^2 was first derived by Rao (1963) when $n=2$ and Midzuno-Sen scheme is employed, but it is interesting to note from (5) that the restriction remains the same even when n is greater than 2.

Chaudhuri (1975) and Mukhopadhyay (1975) independently derived the above for the Midzuno-Sen scheme.

Brewer (1963), Rao (1965) and Durbin (1967) proposed different IPPS schemes, for the case $n=2$, with the same inclusion probabilities,

$$\pi_{ij} = \frac{2p_i p_j}{1+k}\left(\frac{1}{1-2p_i} + \frac{1}{1-2p_j}\right) \text{ where } k = \sum_{i=1}^{N} \frac{p_i}{1-2p_i}.$$

These schemes are free from the restrictions on the p_i's of the previous schemes. From (3),

$$\phi = \frac{1}{1+k}\sum_{i=1}^{N} \frac{Z_i^2}{1-2p_i} \geq 0,$$

so that the Hansen-Hurwitz strategy is again inferior to the Horvitz-Thompson strategy.

ACKNOWLEDGEMENTS

The author is indebted to the Editor, M.P. Singh, and a referee for their many helpful comments.

REFERENCES

BREWER, K.R.W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, 5, 5-13.

CHAUDHURI, A. (1975). On some properties of the sampling scheme due to Midzuno. *Bulletin of Calcutta Statistical Association*, 23, 1-19.

DURBIN, J. (1967). Design of multi-stage surveys for the estimation of sampling errors. *Applied Statistics*, 16, 152-164.

HANSEN, M.H., and HURWITZ, W.N. (1943). On the theory of sampling from finite population. *Annals of Mathematical Statistics*, 14, 333-362.

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.

MIDZUNO, H. (1952). On the sampling system with probability proportionate to sum of sizes. *Annals of Institute of Statistical Mathematics*, 3, 99-107.

MUKHOPADHYAY, P. (1975). PPS sampling schemes to base HTE. *Bulletin of Calcutta Statistical Association*, 23, 21-44.

RAO, J.N.K. (1963). On two systems of unequal probability sampling without replacement. *Annals of Institute of Statistical Mathematics*, 15, 67-72.

RAO, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3, 173-180.

SANKARANARAYANAN, K. (1969). An IPPS sampling scheme using Lahiri's method of selection. *Journal of the Indian Society of Agricultural Statistics*, 21, 58-66.

SEN, A.R. (1952). Further developments of the theory and application of primary sampling units with special reference to the North Carolina agricultural population. Ph.D. Thesis, North Carolina State College, Raleigh.

ACKNOWLEDGEMENTS

The Survey Methodology Journal wishes to thank the following persons who have served as referees between January 1, 1987 and January 31, 1988. An asterisk indicates that the person served more than once.

D.W. Anderson, *National Institute of Health*
J. Armstrong, *Statistics Canada*
H.R. Arora, *Transport Canada*
M. Bankier, *Statistics Canada*
*K.G. Basavarajappa, *Statistics Canada*
G. Brackstone, *Statistics Canada*
D. Bellhouse, *University of Western Ontario*
L. Biggeri, *University of Florence*
*D.A. Binder, *Statistics Canada*
R.D. Burgess, *Statistics Canada*
C.-M. Cassel, *Statistics Sweden*
N. Chinnappa, *Statistics Canada*
*G.H. Choudhry, *Statistics Canada*
D. Dodds, *Statistics Canada*
D. Drew, *Statistics Canada*
M. Eagen, *Goss, Gilroy and Associates*
J.L. Eltinge, *Iowa State University*
I.P. Fellegi, *Statistics Canada*
W.A. Fuller, *Iowa State University*
J.F. Gentleman, *Statistics Canada*
E. Gbur, *University of Arkansas*
*G.B. Gray, *Statistics Canada*
M.A. Hidiroglou, *Statistics Canada*
D. Holt, *University of Southampton*
S. Ingram, *Statistics Canada*
W.D. Kalsbeek, *University of North Carolina*
*G. Kalton, *University of Michigan*
N.J. Kirkendall, *Alexandria, Virginia*
G. Kriger, *Statistics Canada*
H. Lee, *Statistics Canada*
J.M. Lepkowski, *University of Michigan*
*S. Kumar, *Statistics Canada*
R. Lachapelle, *Statistics Canada*
E. Langlet, *Statistics Canada*
S. Michaud, *Statistics Canada*
D.G. Paton, *Statistics Canada*
M. Podehl, *Statistics Canada*
*J.N.K. Rao, *Carleton University*
P.S.R.S. Rao, *University of Rochester*
G. Sande, *Statistics Canada*
*I. Sande, *Statistics Canada*
C.E. Särndal, *Université de Montréal*
*F. Scheuren, *U.S. Internal Revenue Service*
E.A. Schillmoeller, *Nielsen Media Research*
M. Sheridan, *Statistics Canada*
A.C. Singh, *Statistics Canada*
K.P. Srinath, *Statistics Canada*
V. Tremblay, *Statplus*
A. van Baaren, *Statistics Canada*
*K.M. Wolter, *U.S. Bureau of the Census*

Acknowledgements are also due to those who assisted during the production of the 1987 issues: B. Babcock (Text Editing), C. VanBastelaar (Photocomposition), G. Gaulin (Author Services) and M. Haight (Translation Services).

We would like to thank the staff of Social Survey Methods and Business Survey Methods Divisions who assisted in proofreading and verification. Finally we wish to acknowledge J. Clarke, E. Corriveau, J. Dufresne, M. Kent, C. Larabie, D. Lemire and N. Smalldridge for their support with coordination, typing and copy editing.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of Survey Methodology as a guide and note particularly the following points:

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.

1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.

1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.

1.4 Acknowledgements should appear at the end of the text.

1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.

3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.

3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.

3.4 Write fractions in the text using a solidus.

3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).

3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.

4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, 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.