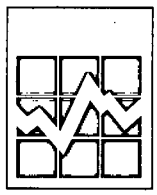


SURVEY METHODOLOGY

Catalogue 12-001

A JOURNAL PUBLISHED BY STATISTICS CANADA

VOLUME 21

NUMBER 2

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A JOURNAL PUBLISHED BY STATISTICS CANADA

DECEMBER 1995 • VOLUME 21 • NUMBER 2

Published by authority of the Minister
responsible for Statistics Canada

December 1995

Price: Canada: $45.00
United States: US$50.00
Other Countries: US$55.00

Catalogue No. 12-001

ISSN 0714-0045

Ottawa

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

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

EDITORIAL POLICY

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

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their manuscripts in either English or French to the Editor, Dr. M.P. Singh, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of Survey Methodology (Catalogue No. 12-001) is $45 per year in Canada, US $50 in the United States, and US $55 per year for other countries. Subscription order should be sent to Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 21, Number 2, December 1995

CONTENTS

Survey Methodology, December 1995
Vol. 21, No. 2, pp. 97–98
Statistics Canada

In This Issue

This issue of *Survey Methodology* contains papers on a variety of topics. The first paper, by Davis and Scott, discusses the impact that interviewer effects may have on comparisons between domain means. Using a components of variance model, it is shown theoretically that the impact depends on the distribution of each interviewer's case load between the domains and on the domain-interviewer interaction. The model is applied to data from a health survey to estimate the magnitude of interviewer effects for comparisons between sexes and between ethnic groups. It was found that in some cases the domain-specific interviewer effects have a large impact on the accuracy of between domain comparisons.

Rivest and Hurtubise examine the usefulness of the Winsorized mean as an estimator of the mean of a population that has a distribution skewed to the right. A Winsorized mean is obtained by replacing all observations greater than a given threshold value R by this same value R, before the mean is calculated. The authors suggest a simple algorithm for calculating R that minimizes the squared error of the estimator. They apply this method to several sample sizes and various sample designs, including stratified sampling and sampling with probabilities proportional to size. They derive direct approximations of the effectiveness of the Winsorized mean. They conclude their article with a Monte Carlo simulation to compare various estimators that reduce the impact of extreme values.

Kish, Frankel and Verma examine the possible incidence and the importance of the design effect (deft) on a set of interrelated statistics. On the basis of 14 surveys conducted in six countries, the authors present an empirical approach relating the design effect of analytical statistics, $\text{deft}(p_i - p_j)$, to the design effects of separate statistics, $\text{deft}(p_i)$ and $\text{deft}(p_j)$, for two of the many categories of the same variable. The proposed approximation must be checked constantly. However, it appears to be widely applicable to the data studied, and it is clearly preferable to the hypotheses put forward thus far on $\text{deft}(p_i - p_j)$.

Dupont discusses the estimation of a total from a two-stage sample where auxiliary information is present. First three regression estimators are presented, each making different use of the auxiliary information. Then four calibration estimators are proposed, each corresponding to a specific strategy for using auxiliary information. Dupont then shows that the calibration strategies can be association with regression modelling. This article also discusses variance estimation for the seven estimators presented, the choice of the estimator where there is nonresponse, and the *a priori* or *a posteriori* use of auxiliary information.

Spisak discusses the use of statistical process control to assure the quality of a frame constructed by a continuous process and used for a survey repeated periodically. The frame sizes constitute a time series for which the appropriate model must be identified in order to estimate the process variance needed for construction of the control charts. The author uses the data from the United States Unemployment Insurance Benefits Quality Control program to illustrate the method.

Binder and Kovacevic show how the estimating equations approach may be used to construct variance estimation procedures that are appropriate when the data come from a survey with a complex design. The approach is most useful when the quantity to be estimated is a complicated non-linear function of the survey population values, as is the case with many common measures of income inequality. Details of the proposed approach are worked out for a number of complex income distribution statistics including the Gini Coefficient, the Lorenz Curve Ordinate, the Quantile Share, and the Low Income Measure. A numerical example is given using data from the Canadian Survey of Consumer Finance.

Ernst and Ikeda present a reduced-size algorithm for maximizing the retention of selected primary sampling units when a new sample (*i.e.*, with a new stratification and allocation) is selected for a repeated survey. First, the transportation procedure developed by Causey, Cox and Ernst (1985) is described. It provides optimal retention of PSU's but the resulting transportation problem may be too large to solve in practice. The authors then expose their algorithm which is an approximation of the previous method but has the advantage of being of smaller size and thus possible to use in many practical situations. Finally, an application of the algorithm to the Survey of Income and Program Participation is presented.

Shrestha and Preston evaluate the consistency of the Census data with the Vital Registration data for the older Americans. First, the data used in the study and their sources of errors are described. Then the authors present the methodology used to evaluate the quality of the old-age statistics and explain how one should interpret the results of the application of that methodology. Finally, results from the application of the methodology to data from 1970 to 1990 are presented.

Dillman, Clark and Sinclair compare different mailout and follow-up strategies with respect to their impact on the response rates for the U.S. Census. The comparison of the strategies includes the use of a factorial design and a sample of 50,000 housing units. The results are analyzed through multiple pairwise comparisons of treatment means and logistic regression.

Forster and Snow evaluate the use of hand-held computers to conduct demographic surveys in developing countries. A data collection test was conducted for comparing the use of paper and computerized questionnaires with the Adult Mortality Survey of people living on the Kenyan coast. The results show that the use of hand-held computers can reduce the data processing time, improve the quality of the data as well as reduce survey costs on the long term.

The Editor

Survey Methodology, December 1995
Vol. 21, No. 2, pp. 99–106
Statistics Canada

The Effect of Interviewer Variance on Domain Comparisons

PETER DAVIS and ALASTAIR SCOTT[1]

ABSTRACT

In this paper we explore the effect of interviewer variability on the precision of estimated contrasts between domain means. In the first part we develop a correlated components of variance model to identify the factors that determine the size of the effect. This has implications for sample design and for interviewer training. In the second part we report on an empirical study using data from a large multi-stage survey on dental health. Gender of respondent and ethnic affiliation are used to establish two sets of domains for the comparisons. Overall interviewer and cluster effects make little difference to the variance of male/female comparisons, but there is noticeable increase in the variance of some contrasts between the two ethnic groupings used in this study. Indeed, the impact of interviewer effects for the ethnic comparision is two or three times higher than it is for gender contrasts. These findings have particular relevance for health surveys where it is common to use a small cadre of highly-trained interviewers.

KEY WORDS: Interviewer variance; Domain comparisons; Design effect.

1. INTRODUCTION

Surveys requiring a high degree of specialist training for interviewers, such as many health studies, are often forced to use a small number of highly-trained interviewers. There has been a substantial amount of work done on estimating the impact of interviewer variability on simple statistics such as means and proportions, and it is well-known that the use of a small number of interviewers, each having a high case load, can lead to a relatively large contribution to the total error. Comprehensive summaries of the literature are given in Groves (1989, chap. 8) and Lessler and Kalsbeek (1992, §11.3). However, most medical and social surveys are primarily interested in more complex questions such as comparisons between sub-groups or estimating the effect of a factor on disease outcome. There is a widespread belief that the effect of interviewer variability is much smaller here, and that the effect of a small number of interviewers is relatively harmless. Following the pioneering work of Kish and Frankel (1974), there has been a great deal of theoretical and empirical work on the effects of clustering on fitting multiple regression models or log-linear models for categorical data. Good accounts of the literature are given in Skinner *et al.* (1989) and Rao and Thomas (1988). There has been some empirical work on the conceptually simpler, yet practically important, problem of comparing sub-group means (see Kish 1987 and Skinner 1989 for example) but relatively little theoretical development.

In this paper we concentrate on comparisons between subgroups (or domains). We first look at theoretical aspects via a straightforward components of variance model. The theory suggests that the impact of interviewer variability depends on two things, the distribution of each interviewer's case load between the domains and the domain-interviewer interaction. Then we apply the theory to data from a reasonably typical health survey, using two sets of domains defined by the sex and ethnic background of the respondent. Unfortunately the study was not designed *a priori* to estimate interviewer effects (most importantly, interviewers were not deployed at random) so the results should be regarded as suggestive rather than definitive. However, they are sufficiently disturbing to indicate that the problem warrants further study. The results from the ethnic comparisons, in particular, suggest that there are cases when we should be concerned about using a small number of interviewers even when comparisons, rather than simple means or proportions, are the main concern of the analysis.

2. THEORY

For simplicity we start with the special case of a two-stage self-weighting design. This is sufficiently complex to illustrate the central ideas, but simple enough to avoid being swamped with extraneous detail. Following Collins and Butcher (1982), we want to address the problems of interviewer variance and clustering together. A simple correlated response model appropriate for observations drawn according to such a design is

$$Y_{ipr} = \mu + a_i + b_p + e_{ipr}, \qquad (1)$$

where i denotes the interviewer, p the primary sampling unit (PSU) and r the individual respondent. Here the

[1] Peter Davis, Department of Community Health, University of Auckland, Private Bag 92019, Auckland, New Zealand; and Professor Alastair Scott, Department of Mathematics and Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand.

mean, μ, is fixed constant and the remaining components, a_i, b_p and e_{ipr}, are assumed to be independent random variables with variances σ_I^2, σ_C^2 and σ^2 respectively. Such models have been used widely in theoretical studies of response variance. See Prasad and Rao (1990) for a recent example. For references to earlier work, see the comprehensive treatment in §11.3 of Lessler and Kalsbeek (1992).

Since the design is self-weighting the sample mean, $\bar{Y}$, is the natural estimator of the population mean. Its variance under the correlated response model (1) is

$$V(\bar{Y}) = (\bar{n}_I\sigma_I^2 + \bar{n}_C\sigma_C^2 + \sigma^2)/n$$

with $\bar{n}_I = \sum_i n_i^2/n$, where n_i is the number of respondents handled by the i-th interviewer and $n = \sum_i n_i$ is the total sample size, and $\bar{n}_C = \sum_p m_p^2/n$ where m_p denotes the number of respondents in the p-th PSU. Note that $\bar{n}_I$ is always larger than the simple arithmetic average of the n_i's and can be considerably larger if the n_i's vary widely.

Now consider what the corresponding expected variance, $V_0(\bar{Y})$ say, would be if the n observations had been generated independently (*e.g.* if we had drawn a simple random sample from a very large population of PSUs using a large pool of interviewers). It follows from (1) that

$$V_0 = \sigma_{t0t}^2/n \tag{2}$$

where

$$\sigma_{t0t}^2 = \sigma_I^2 + \sigma_C^2 + \sigma^2.$$

The inflation in the expected variance due to the combined effects of interviewer variability and intra-cluster correlation is given by the ratio

$$D_0 = V(\bar{Y})/V_0$$
$$= 1 + (\bar{n}_I - 1)\rho_I + (\bar{n}_C - 1)\rho_C \tag{3}$$

where $\rho_I = \sigma_I^2/\sigma_{t0t}^2$ and $\rho_C = \sigma_C^2/\sigma_{t0t}^2$. We shall refer to this ratio as the "design effect" although it differs slightly from the usual definition which is in terms of actual, rather than expected, variances. It is clear from expression (3) that interviewer variability can have a substantial effect on the variance of a sample mean if the average interviewer case-load, $\bar{n}_I$, is large even if the intra-interviewer correlation, ρ_I, is relatively small.

Next suppose that we are interested in the difference between two domain means rather than a single mean. We might, for example, be interested in gender differences or in differences between two ethnic groups. In the simplest extension of the correlated response model (1) we might postulate a model of the form

$$Y_{ipr}^{(d)} = \mu^{(d)} + a_i + b_p + e_{ipr}^{(d)} \tag{4}$$

for observations from the d-th domain. Here the means, $\mu^{(d)}$, may be different for the two domains but the interviewer and cluster effects are assumed to be the same.

Let $p_i^{(d)} = n_i^{(d)}/n^{(d)}$, where $n_i^{(d)}$ is the number of respondents from domain d contacted by the i-th interviewer and $n^{(d)}$ is the total number of respondents from domain d. Similarly, let $q_p^{(d)} = m_p^{(d)}/n^{(d)}$, where $m_p^{(d)}$ is the number of respondents from domain d lying in the p-th PSU. Then, under model (4), the expected variance of $\bar{Y}^{(a)} - \bar{Y}^{(b)}$, the difference between the sample means for the two domains, is

$$V(\bar{Y}^{(a)} - \bar{Y}^{(b)}) =$$
$$(\bar{m}_I\sigma_I^2 + \bar{m}_C\sigma_C^2 + \sigma^2)\left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}}\right), \tag{5}$$

where

$$\bar{m}_I = \sum_i (p_i^{(a)} - p_i^{(b)})^2 \Big/ \left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}}\right) \tag{6}$$

and

$$\bar{m}_C = \sum_p (q_p^{(a)} - q_p^{(b)})^2 \Big/ \left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}}\right). \tag{7}$$

If the observations had been generated independently the corresponding expected variance would be

$$V_1 = \sigma_{t0t}^2\left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}}\right)$$

so that the inflation due to interviewer variability and intra-cluster correlation is now

$$D_1 = \mathrm{Var}(\bar{Y}^{(a)} - \bar{Y}^{(b)})/V_1$$
$$= 1 + (\bar{m}_I - 1)\rho_I + (\bar{m}_C - 1)\rho_C. \tag{8}$$

The size of the effect depends on the way the interviewers' case-loads and the PSUs cut across the domains. At one extreme, when each interviewer contacts the same proportion of people from both domains, (*i.e.* when $p_i^{(a)} = p_i^{(b)}$), $\bar{m}_I$ is zero and the interviewer effect essentially cancels out. At the other extreme, when each interviewer sees only cases from a single domain, $\bar{m}_I$ is similar in size to $\bar{n}_I$ and the interviewer effect for differences is comparable to that for a single mean. Typically interviewers contact people from both domains and $\bar{m}_I$ is rather small, giving some justification to the belief that interviewer variability has a small impact on estimated differences between domains. Similar comments apply to the effect of clustering.

All this depends on the assumption that the interviewer and cluster effects, a_i and b_p, are the same for both domains. It is easy to imagine situations where such an assumption would not be at all reasonable. Some interviewers, for example, might interact very differently with males and females, or with members of different ethnic groups. A model which allows for the possibility of such interactions is

$$Y_{ipr}^{(d)} = \mu^{(d)} + a_i^{(d)} + b_p^{(d)} + e_{ipr}^{(d)}, \qquad (9)$$

where $a_i^{(a)}$ and $a_i^{(b)}$ (respectively $b_p^{(a)}$ and $b_p^{(b)}$) are now assumed to be correlated random variables with correlation $r_I(r_C)$. The naive model (4) corresponds to the special case in which the variances of the effects are equal and r_I and r_C are both equal to one. On the other hand, if there are substantial differences between the interviewer (cluster) effects for the two domains, $r_I(r_C)$ will be small (or even negative in extreme cases). In the rest of this section we suppose for simplicity that the variances of $a_i^{(a)}$ and $a_i^{(b)}$ (respectively $b_p^{(a)}$ and $b_p^{(b)}$) are equal. This may or may not be reasonable in practice but the simplification enables us to concentrate on the essential ideas. The basic form is similar in the more general case but the terms are somewhat messier. Under model (9), the expected variance of $\bar{Y}^{(a)} - \bar{Y}^{(b)}$ is

$$V(\bar{Y}^{(a)} - \bar{Y}^{(b)}) = (\bar{v}_I\sigma_I^2 + \bar{v}_c\sigma_C^2 + \sigma^2)\left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}}\right) \qquad (10)$$

where

$$\bar{v}_I = \sum_i \left(p_i^{(a)2} - 2r_I p_i^{(a)}p_i^{(b)} + p_i^{(b)2}\right) \Big/ \left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}}\right)$$

with a similar definition for $\bar{v}_C$ in terms of $q_p^{(a)}$ and $q_p^{(b)}$.

The variance inflation factor under this model is

$$D_2 = 1 + (\bar{v}_I - 1)\rho_I + (\bar{v}_C - 1)\rho_C. \qquad (11)$$

This is a decreasing function of r_I, the correlation between the interviewer effects for the two domains; the smaller the correlation, the larger the variance inflation. When $r_I = 1$, $\bar{v}_I$ reduces to $\bar{m}_I$ and the interviewer effect is negligible provided all interviewers see a reasonable balance of people from both domains. However, if r_I is small (indicating a strong interaction between the interviewers and domains), $\bar{v}_I$ is the same order of magnitude as $\bar{n}_I$ and the effect of interviewer variability on the variance of domain differences can be substantial.

In practice, the effects will fall between the two extremes and their likely impact is a matter for empirical enquiry. In the next section, therefore, we make a start on building up practical knowledge about the impact using data for a variety of questions drawn from a single health survey that is typical of the genre of research investigation for which domain comparisons are important (although not ideally designed for our purposes!).

3. EXAMPLE

The example is based on data drawn from a survey of the oral health, attitudes and practices of adult New Zealanders. The details of the survey are reported in full elsewhere (Cutress *et al.* 1979). The important features of the study for the purposes of the current investigation are the sample design and the deployment of interviewers.

The sample design was a stratified multi-stage sampling scheme. The country was divided into 256 Territorial Local Authorities (TLAs) and a geographically stratified sample of 68 TLAs was drawn from the 256 with selection probabilities proportional to size (PPS) at the first stage, where size was the estimated number of persons aged 15 and over. Each sampled TLA was split into secondary sampling units (SSUs) comprising existing census mesh-blocks, aggregated where necessary in order to achieve a minimum size of 50. Two SSUs were then selected with PPS from each sampled TLA at the second stage. Finally, a systematic sample of 28 adults was drawn from each sampled SSU. This equalised the final probability of selection for all adults so that the sample design is (approximately) self-weighting.

The key point of the design was the deployment of the interviewers. Thirteen interviewers were employed in the study, with at least three interviewers used within each SSU, and all interviewers carried out at least 10% of their total work-load in one region (Auckland). Ideally the assignment of interviewers would be part of the overall sample design as in Fellegi (1974) or Biemer and Stokes (1985). Unfortunately the study was not designed to estimate interviewer variance, and the assignment of respondents to interviewers was done in a haphazard way, rather than using a formal randomization procedure.

This is fairly typical of large studies. The following quote from Hox (1994) gives a good summary of the situation: "Ideally, in interviewer studies, respondents should be assigned to interviewers at random. In large-scale studies, this is seldom done because it is expensive and complicated to organize. This makes it difficult to use such studies for methodological research because, as a result, interviewer and respondent characteristics might be confounded. Multi-level analysis, as outlined above, offers some remedies for this situation. If the relevant respondent variables are known they can be put in the regression model to equalize interviewers by statistical means... The limitation of this approach is that it relies on statistical control instead of experimental control. It depends on the assumption that all relevant covariates have been included

and have been correctly modeled. Without randomization, it is impossible to conclude that the influence of all confounding variables has been eliminated." In our case, the deployment is such that all the components of variance are formally identifiable, provided that we believe the model and are willing to accept that the assignment of interviewers is independent of the cluster effects. However, because of the lack of formal randomization there always remains the possibility that variations in patterns of response between interviewers could be a function of workload allocation rather than interviewing style. Clearly the empirical results can only be regarded as tentative, pointing out possibilities that will need to be explored further in properly designed studies.

Even if we ignore the lack of randomization in the interviewer deployment, the study design is considerably more complicated than the one assumed for the development of the theory in the previous section, since it involves three stages of sampling and regional stratification of the first stage units. In the full analysis, we fitted a more complex model including fixed effects terms for the stratification, a hierarchical random effects model for the three stages of sampling, and all second-order interaction terms. However it turned out that the TLAs used as the first stage units were so diffuse that the differences between strata and the between-TLA component of variance were negligible for all the variables used in the following analysis. Thus the between-SSU component is dominant and, for all practical purposes, we can treat the design as if it were a two-stage sample with the meshblocks (aggregated where appropriate) as PSUs. We have ignored the other components in the results reported in the next section.

4. RESULTS

We look first at interviewer and cluster effects on a selection of means and proportions. We have used Model (9) for both types of variable. It is now well-known that this leads to an under-estimate of the variance components for binary data (see Anderson and Aitken 1985 and Pannekock 1988 for example), so our estimated design effects for proportions should be regarded as lower bounds. The models are fitted using PROC GLM in SAS. The impact of clustering has been well documented in the literature (Kish 1965; Kish and Frankel 1970; 1974). In general terms, the magnitude of the effects of clustering depends on the type and number of units chosen and is likely to vary with different kinds of social and demographic characteristics. In the current investigation clustering effects were expected to be reasonably high because the census meshblocks used as sampling units are likely to show a fair degree of internal homogeneity. In keeping with this concentration of population characteristics, it was assumed that demographic and related items would show the largest values of ρ_C. Values of ρ_I were expected to be lower because of the intense interviewer training. The literature suggests that these effects are also likely to vary according to the type of questionnaire item, with attitude questions, questions requiring probing, fixed-alternative and forced-choice items, together with poorly-worded and ambiguous questions, being particularly susceptible to interviewer variability (Feather 1973, Groves 1989).

Estimated measures of intra-interviewer and intra-cluster correlation coefficients for a selection of questionnaire items falling under four separate headings (socio-demographic, attitudinal, reports of recent behaviour, and recall of distant behaviour) are outlined in the first two columns of Table 1. These categories were identified as providing natural groupings with the potential to display a wide range of interviewer effects. Within each grouping the items are listed in order of the size of their intra-interviewer correlations. A full description of all questionnaire items (apart from the self-evident socio-demographic category) is provided in the Appendix.

As expected, the socio-demographic variables (except for gender) show the highest values of intra-cluster correlation. The average ρ_C is .07 (.08 if gender is omitted). The average values of ρ_C for the other three categories of item are .02 and less. A few items that might be expected to be closely related to social background - like dental visiting, payment for visits, toothbrushing and certain attitude statements - have higher than average ρ_C values. In general, though, these values fall within the range reported by others. (See, for example, Kish 1965, p. 581 for a series of consumer surveys, Bebbington and Smith 1977 and Verma *et al.* 1984 for the country studies in the World Fertility Survey.)

The corresponding estimated ρ_I values are listed in the first column of Table 1. In general these values are very much smaller than those recorded for cluster effects, being usually less than half, and in some cases a tenth, the size of the ρ_C values for the corresponding items. As expected, some attitude items show higher than average ρ_I values, as do certain reports of behaviour that might be susceptible to a high "social desirability" bias, like toothbrushing and buying sweets and chocolates. Ethnic group and employment status also record relatively high values. The pattern is similar to that found in previous studies, although the values recorded lie at the lower end of the range of typical values reported elsewhere (Feather 1973; Kish 1962; O'Muircheartaigh 1977; O'Muircheartaigh and Wiggins 1981). A comprehensive survey is given in Chapter 8 of Groves (1989). This may partly reflect the intensive training and monitoring of the interviewers that were integral to the field work stage of the study. It may also be influenced by the rigorous post-field work "cleaning" (editing and checking) of the data that was carried out prior to analysis. However it may also simply be due to the attenuation resulting from using Model (1) for proportions that we noted above.

Table 1
Cluster and Interviewer Effects for Means and Proportions

Item Description	$\hat{\rho}_I$	$\hat{\rho}_C$	D_0	%Int
Attitudinal:				
Dentists 1	.014	.014	4.61	91
Visiting	.008	.028	3.42	74
Natural Teeth	.008	.027	3.52	76
Health of Teeth	.007	.015	2.97	84
Dentures	.005	.015	2.67	80
Dentists 2	.004	.033	2.77	57
Health of Gums	.003	.010	1.96	77
Fluoridation	.001	.016	1.66	49
Average	**.006**	**.020**	**2.95**	**73**
Socio-demographic:				
Employment Status	.010	.055	4.20	65
Race	.009	.172	6.87	33
Age	.004	.042	5.98	52
Household Income	.002	.092	3.29	15
Marital Status	.000	.058	2.34	0
Sex of Respondent	.000	.005	1.12	0
Average	**.004**	**.071**	**3.47**	**28**
Recent Behaviour:				
Brushed Teeth	.019	.025	6.16	8
Sweets/Chocolates	.011	.003	3.75	98
Fluoride Toothpaste	.008	.000	3.04	100
Toothpick	.006	.006	2.66	92
Rinse Mouth	.004	.024	2.43	62
Dental Floss	.001	.018	1.60	43
Disclosing Tablet	.000	.027	1.49	0
Mouthwash	.000	.018	1.42	0
Average	**.006**	**.012**	**2.82**	**60**
Distant Behaviour:				
Age First Paid	.004	.029	2.34	57
Visited Dentist	.004	.029	2.51	57
Cost Last Year	.002	.000	1.19	100
Year Last Visit	.000	.014	1.15	0
Average	**.003**	**.018**	**1.80**	**54**

Perhaps more significant than the pattern and values of ρ_I is the impact of interviewer variability on the overall design effect, incorporating both interviewer and clustering effects. This is shown in the third column of Table 1 (D_0), with the final column (% Int) representing the proportionate contribution of interviewer variability to the overall value of D_0. Design effects are substantial, being above two in all but a minority of cases. This is due to the clustering and to the impact of the large interviewer workloads characteristic of the study since, from equation (3), the variance is increased by a factor of $1 + (\bar{n}_I - 1)\rho_I$, where $\bar{n}_I$ is a weighted average of the interviewer workloads. There is a distinct pattern in the contribution to the design effect produced by interviewer variability. For socio-demographic variables it averages just under one half of the contribution from clustering, while for attitudinal items the interviewer contribution to the design effect rises to three times that from clustering. The other two categories of items range in between these two extremes.

What the results outlined in Table 1 confirm is the impact that interviewer workload has on the variance of sample estimates, because of the multiplier effect. In essence, an interviewer component with a very small intra-class correlation can be translated into a major effect if the interviewer workload is high. In the study under review, the logistics of deployment and the requirements of on-going quality control seemed to argue for small interview teams, a practice that appears to be typical of much field work in the health area (for example, Choi and Comstock 1975). This meant that interview workloads averaged over 250. The cost of this strategy is immediately apparent from the results in Table 1; very small differences between interviewers are translated into major reductions in the precision of sample estimates.

Now we turn to the main object of our analysis, viz. the impact of interviewer variability on contrasts between domain means or proportions. In the current analysis, this was assessed for two sets of comparisons, the first set by gender (male/female) and the second set by ethnic group (European/non-European). As we have seen in the discussion following equation (11), the contribution, $1 + (v_I - 1)\rho_I$, to D_0 from interviewer differences depends on the extent to which the interviewer effect is constant across the two domains and on the way the domains cut across individual case-loads. Assuming that the domains cut evenly across interviewer case-loads, then v_I is zero if the interviewer effect is identical in the two domains, in which case the common interviewer effect cancels out completely in the comparison. On the other hand, if the effects in the two domains are weakly correlated then the value of v_I tends to be much higher and in extreme cases may equal the average case-load. In the current study values of v_I fell between 0 and 50 for both gender and ethnic group. Thus the effect of domain-specific interviewer effects on the design effect can be quite substantial. Similar comments apply to the impact of clustering on the comparison; if the effect is the same on both domains then it largely cancels out and the net impact is small, but the impact can be substantial if the clustering effect is domain specific.

Table 2 shows values of ρ_I and ρ_C for comparisons by gender and by ethnic group, together with the overall design effect D_2 and the proportion of this effect due to the impact of interviewer variability. Note that the item on the use of disclosing tablets has been omitted from Table 2. This is because so few respondents either used or knew what this item was that the effective sample size in this case is tiny, thus rendering the results almost meaningless.

The impact of both interviewer and clustering on comparisons by gender is small with design effects little above unity, in spite of the fact that the estimated values of ρ_I and ρ_C are slightly increased when adjusted for this variable.

Table 2

Interviewer and Cluster Effects for Domain Differences

Item Description	By Sex				By Race			
	$\hat{\rho}_I$	$\hat{\rho}_C$	D_2	%Int	$\hat{\rho}_I$	$\hat{\rho}_C$	D_2	%Int
Attitudinal:								
Dentists 2	.004	.043	1.05	0	.010	.027	1.28	46
Visiting	.009	.028	1.08	0	.072	.133	5.19	78
Natural Teeth	.010	.032	1.06	0	.010	.037	1.26	42
Fluoridation	.001	.019	1.12	42	.021	.031	2.13	88
Dentures	.007	.018	1.04	0	.011	.035	1.21	33
Health of Teeth	.012	.022	1.52	85	.010	.045	1.40	20
Dentists 1	.001	.018	1.05	0	.015	.020	1.53	74
Health of Gums	.006	.022	1.07	0	.003	.104	1.46	9
Average	**.006**	**.025**	**1.12**	**16**	**.019**	**.054**	**1.93**	**49**
Socio-demographic:								
Race	.008	.183	1.11	0	–	–	–	
Household Income	.004	.095	1.37	24	.004	.099	1.95	40
Marital Status	.000	.059	1.17	0	.011	.060	1.69	38
Employment Status	.014	.067	1.42	71	.022	.116	2.09	25
Age	.007	.052	1.06	0	.006	.093	1.87	24
Sex of Respondent	–	–	–		.006	.011	1.09	44
Average	**.007**	**.091**	**1.23**	**19**	**.010**	**.076**	**1.74**	**34**
Recent Behaviour:								
Brushed Teeth	.025	.060	1.62	65	.019	.019	1.68	88
Rinse Mouth	.007	.029	1.28	64	.004	.023	1.20	45
Mouthwash	.000	.057	1.20	0	.027	.105	2.69	75
Dental Floss	.003	.021	1.06	0	.015	.036	1.37	32
Toothpick	.006	.010	1.03	0	.006	.046	1.48	63
Sweets/Chocolates	.012	.009	1.02	0	.013	.022	1.31	48
Fluoride Toothpaste	.010	.007	1.11	100	.007	.000	1.02	100
Average	**.009**	**.028**	**1.19**	**33**	**.013**	**.036**	**1.36**	**64**
Distant Behaviour:								
Age First Paid	.003	.033	1.10	0	.029	.141	2.92	71
Visited Dentist	.005	.035	1.04	0	.020	.018	1.26	50
Year Last Visit	.004	.012	1.20	75	.016	.003	1.83	12
Cost Last Year	.007	.021	1.01	0	.076	.117	2.09	42
Average	**.005**	**.025**	**1.09**	**19**	**.035**	**.070**	**2.03**	**44**

A significant gender-specific effect was apparent for only three items, health of teeth and tooth-brushing – for which there may be a unique social acceptability bias – and employment status – which holds quite different connotations for men and women. Note that the interviewer effect is the dominant one in all three of these comparisons.

The impact on comparisions by ethnic group is much higher, with design effects averaging about 1.7. This suggests that there are significant, non-cancelling interviewer and clustering effects associated with the ethnic identity of respondents. There are large ethnic-specific interviewer effects for two hypothetical attitudinal questions (visiting and fluoridation), for one item of recent behaviour, and for age of first payment for dental services. The result is plausible; all the interviewers were European and may have varied systematically in their interactions with respondents of different ethnic backgrounds. Again clustering effects are most marked for the socio-demographic variables. Not only are the design effects on average higher than those recorded for the gender comparisons, but the interviewer component is in general two or three times higher for the ethnic group contrasts.

A referee rightly points out that because of the way the interviewers are deployed (they worked primarily in teams assigned to different parts of New Zealand), there is a real possibility that the interviewer effects might be inflated because of confounding with area effects. The fact that differences between the TLAs were so small gives us some reason to believe that this inflation will be small, but the possibility can never be discounted with this design.

5. DISCUSSION

This paper has applied empirical data from a not untypical health survey to assess the impact of interviewer variability under the assumptions of both simple and extended versions of the correlated response model for the error variance of a multi-stage sample design.

In the first case the simple model analyses the relative impact of cluster and interviewer effects on the estimation of means and proportions. The results of this analysis confirm a number of findings that are well established in the literature: the intra-class correlations for interviewers are generally lower than those for clusters; the intra-class correlations for clusters vary in the expected direction by question type; the overall design effects for these question types vary between 2 and 3.5; a substantial component of this inflation is contributed by interviewer variability and can probably be attributed to the multiplier effect of large interviewer caseloads; finally, the impact of this interviewer component is shown to vary in the expected direction by question type.

In the second case the extended model addresses the analysis of cluster and interviewer effects for the estimation of domain contrasts between means and proportions for two sets of comparisons defined by gender and ethnic group. The effect on contrasts between domain means was smaller but it was still significant for a number of items, particularly for the ethnic comparisons, suggesting that the interviewer effect was different for the two domains. The size of the effect for these items was certainly large enough to suggest that we should be concerned about it in designing similar studies. In general, the impact of interviewer effects was two or three times as great for the ethnic contrasts as it was for the gender comparisons.

The basic deficiencies in the design mean that these results must be regarded as suggestive rather than definitive. They do indicate, however, that there is considerable potential for damage in the use of a small group of interviewers even when interest is centered on domain differences rather than simple means or proportions. This is certainly counter to standard folklore in some fields such as health surveys, and suggests that considerable further empirical work is justified.

On the assumptions of the simple correlated response model a reduction in the impact of interviewer variance

can be achieved by raising the number of interviewers and thus reducing individual interviewer workloads. Of course, this brings with it a potential reduction in the quality of interviewing if training and monitoring procedures have to be tempered. In this instance close attention to question wording and interviewer instruction is clearly crucial. In the case of the extended version of the correlated response model, however, such a strategy is unlikely to be a sufficient one on its own. If comparisions between groups are a major objective of the study, then it is important also to ensure that the interviewers treat the two groups in as similar a way as possible. It is also important to design the study so that each interviewer contacts respondents drawn from both groups. This is likely to be a critical consideration in investigations such as case-control studies in which health outcomes are related to contrasting exposures and in which the control of potential confounder variables may have a significant influence on the magnitude of measures such as the odds ratio.

ACKNOWLEDGEMENTS

This project has received support from the Medical Research Council of New Zealand. The bulk of the detailed computations for this paper were carried out by Joanna Broad.

APPENDIX
Questionnaire Items

Attitudinal

Dentists 1: "Dentists are more interested in their patients than making money."

Dentists 2: "Dentists recommend a lot more things to be done than really need to be done."

Dentures: "Dentures are just as good (or better) than your own teeth."

Fluoridation: "What is your opinion on fluoridating public water supplies?"

Visiting: "Do you think a person should go to the dentist only when they have dental problems or should they go sometimes also when they have no obvious problems?"

Health of Teeth: "If you went to the dentist tomorrow, do you think he would find anything wrong with your teeth?"

Health of Gums: "If you went to the dentist tomorrow do you think he would find anything wrong with your gums?"

Recent Behaviour

"Yesterday did you - use a dislosing tablet/mouthwash/ dental floss/toothpick?
- rinse after eating?
- brush your teeth?"

"Did you buy sweets or chocolates any time last week?"

Distant Behaviour

Age First Paid: "About how old were you when you first went to a dentist for routine treatment for which you or your family had to pay?"

Visited Dentist: "Did you visit a dentist in the last 12 months?"

Year Last Visit: "In what year did you last visit a dentist?"

Cost Last Year: "About how much did you pay for dental treatment in the last 12 months?"

REFERENCES

ANDERSON, D.A. (1985). Variance component models with binary response; interviewer variability. *Journal of the Royal Statistical Society*, Series B, 47, 203-210.

BIEMER, P.B., and STOKES, S.L. (1985). Optimal design of interviewer variance experiments in complex surveys. *Journal of the American Statistical Association*, 80, 158-166.

BEBBINGTON, A.C., and SMITH, T.M.F. (1977). The effect of survey design on multivariate analysis, *The Analysis of Survey Data*, (C.A. O'Muircheartaigh and C. Payne, Eds.), 175-192. New York: John Wiley.

CHOI, I.C., and COMSTOCK, G.W. (1975). Interviewer effects on responses to a questionnaire relating to mood. *American Journal of Epidemiology*, 101, 84-92.

COLLINS, M., and BUTCHER, B. (1992). Interviewer and clustering effects in an attitude survey. *Journal of the Market Research Society*, 25, 39-58.

CUTRESS, T.W., HUNTER, P.B., DAVIS, P.B., BECK, D.J., and CROXSON, L.J. (1979). *Adult Oral Health and Attitudes to Dentistry in New Zealand*, Medical Research Council, Wellington.

DIJKSTRA, W. (Ed.) (1982). *Response Behaviour in the Survey Interview*. New York: Academic Press.

FEATHER, J. (1973). *A Study of Interviewer Variance*, (WHO/ICS-MCU Saskatchewan Study Area Reports Series 2, No. 3). Department of Social and Preventive Medicine. University of Saskatchewan, Saskatoon.

FELLEGI, I.P. (1974). An improved method of estimating correlated response variance. *Journal of the American Statistical Association*, 69, 496-501.

GROVES, R. (1989). *Survey Errors and Survey Costs*. New York: John Wiley.

GROVES, R., and FULTZ, N.H. (1985). Gender effects among telephone interviewers in a survey of economic attitudes. *Sociological Methods and Research*, 14, 31-52.

HOLT, D., SMITH, T.M.F., and WINTER, P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society*, 143, 474-487.

HOX, J.J. (1994). Hierarchical regression models for interviewer and respondent effects. *Sociological Methods and Research*, 22, 300-318.

KISH, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Society*, 57, 92-115.

KISH, L. (1965). *Survey Sampling*. New York: John Wiley.

KISH, L., and FRANKEL, M.R. (1974). Inferences from complex samples. *Journal of the Royal Statistical Society*, Series B, 36, 1-37.

KISH, L. (1987). *Statistical Design for Research*. New York: John Wiley.

LESSLER, J.T., and KALSBEEK, W.D. (1992). *Nonsampling Errors in Surveys*. New York: John Wiley.

O'MUIRCHEARTAIGH, C.A. (1976). Response errors in an attitudinal sample survey. *Quality and Quantity*, 10, 97-115.

O'MUIRCHEARTAIGH, C.A., and PAYNE, C. (Eds.) (1977). *The Analysis of Survey Data*, (Volume 2: Model Fitting). New York: John Wiley.

O'MUIRCHEARTAIGH, C.A., and WIGGINS, R.D. (1981). The impact of interviewer variability in an epidemiological survey. *Psychological Medicine*, 11, 817-824.

PANNEKOEK, J. (1988). Interviewer variance in a telephone survey. *Journal of Official Statistics*, 4, 375-384.

PRASAD, N.G., and RAO, J.N.K. (1990). The estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

RAO, J.N.K., and THOMAS, D.R. (1988). The analysis of cross-classified data from complex sample surveys. *Sociological Methodology*, 18, 213-269.

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

VERMA, V., SCOTT, C., and O'MUIRCHEARTAIGH, C.A. (1980). Sample designs and sampling errors for the World Fertility Survey. *Journal of the Royal Statistical Society*, Series A, 143, 431-473.

Survey Methodology, December 1995
Vol. 21, No. 2, pp. 107–116
Statistics Canada

On Searls' Winsorized Mean for Skewed Populations

LOUIS-PAUL RIVEST and DANIEL HURTUBISE[1]

ABSTRACT

This paper considers the winsorized mean as an estimator of the mean of a positive skewed population. A winsorized mean is obtained by replacing all the observations larger than some cut-off value R by R before averaging. The optimal cut-off value, as defined by Searls (1966), minimizes the mean square error of the winsorized estimator. Techniques are proposed for the evaluation of this optimal cut-off in several sampling designs including simple random sampling, stratified sampling and sampling with probability proportional to size. For most skewed distributions, the optimal winsorization strategy is shown, on average, to modify the value of about one data point in the sample. Closed form approximations to the efficiency of Searls' winsorized mean are derived using the theory of extreme order statistics. Various estimators reducing the impact of large data values are compared in a Monte Carlo experiment.

KEY WORDS: Outliers; Max domain of attraction; Mean square error; Simple random sampling; Stratified sampling.

1. INTRODUCTION

Samples drawn from positively skewed populations often contain outliers with values that are much larger than most sampled values. One usually tries to accomodate these large values when designing the survey (Glasser 1962; Hidiroglou 1987). However, given the multipurpose nature of most surveys, statisticians are often faced with outliers at the estimation stage. These data points make classical survey estimators, such as the sample mean, unstable. It is therefore of interest to study alternative estimators that lower the impact of large data values. Winsorization (Searls 1966) consists in replacing the data values larger than a cut-off value R by R before averaging. Searls suggested to select the value of R which minimizes the mean square error of the winsorized mean. One can also take R equal to the second largest data value in the sample (Rivest 1994). Searls' estimator was best among all the methods to adjust large data values studied by Ernst (1980). Hicks and Fetter (1993) implement Searls' winsorization strategy in an agriculture survey. Other strategies have been proposed for dealing with large observations in survey sampling. Chambers and Kokic (1993) review estimators derived from the theory of "Robust Statistics" (Huber 1981). Fuller (1991, 1993) proposes a preliminary test to detect the presence of extreme values in the sample; the impact of these values is lowered only in samples for which this test is significant. Lee (1994) provides a good review of this expanding literature.

The key to the implementation of Searls' winsorization method is the selection of the cut-off R. A simple algorithm for calculating the optimal cut-off for a known population in simple random sampling and in pps sample is proposed in Section 2. Repeated calculations of the optimal cut-off for several populations and several sample sizes reveal that, in most cases, the optimal scheme winsorizes one data point on average, regardless of the sample size. Section 3 extends the result of Section 2 to stratified sampling. A simple algorithm for the calculation of cut-off values in each stratum is proposed. The rule of winsorizing an average of one data point per sample regardless of sample size is shown to hold also in stratified samples. The efficiencies, with respect to the sample mean, of various winsorized estimators are calculated in Sections 4 and 5. Section 4 derives analytic large sample approximations to the efficiency of Searls' estimator using the theory of extreme order statistics while Section 5 compares, in a Monte Carlo study, estimators for reducing the impact of large data values.

2. SAMPLING PROPERTIES OF THE WINSORIZED MEAN

This section studies winsorized means for data drawn from either a continuous or a discrete distribution. Several families of continuous distributions are available to model positive skewed data. One has the Weibull family, $F_\alpha(x) = 1 - \exp(-(x/\beta)^{1/\alpha})$ for $x > 0$, the log-normal family, $F_\nu(x) = \Phi(\log(x/\beta)/\nu)$ for $x > 0$, and the Pareto family, $F_\gamma(x) = 1 - (1 + x/\beta)^{-\gamma}$ for $x > 0$, where β is a positive scale parameter and α, ν, and γ are positive shape parameters. Discrete skewed distributions arise in survey samping. Let $\{y_1, \ldots, y_N\}$ represent the values of the

[1] Louis-Paul Rivest and Daniel Hurtubise, Département de mathématiques et de statistique, Université Laval, Cité Universitaire, Québec, Canada, G1K 7P4.

variable of interest for the N units of a population to be sampled. If a simple random sample with replacement is drawn, then one can take $F(x) = \sum I(y_i \leq x)/N$ as the underlying distribution where $I(\cdot)$ represents the indicator function. In pps sampling, *i.e.*, sampling with replacement and with probabilities given by $\{p_i, i = 1, \ldots, N\}$, one would take $F(x) = \sum p_i I(y_i/(Np_i) \leq x)$. The standard estimator of $\bar{y}$ under pps sampling,

$$\bar{y}_s = \frac{1}{n} \sum_s \frac{y_i}{Np_i}$$

can then be regarded as the mean of a random sample of size n drawn from distribution F. Fuller (1991) provides examples of survey data having skewed distributions.

Let $X_1, X_2, \ldots, X_n$ denote a sample drawn from $F(x)$. In pps sampling, one would have $X_i = y_i/(Np_i)$ where p_i and y_i are the selection probability and the value of the y-variable for the i-th unit selected in the sample. The population mean μ is to be estimated by a winsorized mean,

$$\bar{X}_R = \frac{1}{n} \sum_1^n \min(X_i, R) = \bar{X} - \frac{1}{n} \sum_{i=1}^n \max(X_i - R, 0), \tag{2.1}$$

where $\bar{X}$ is the mean of the X_i's. The expectation of $\bar{X}_R$ is equal to

$$E(\bar{X}_R) = \mu - \int_R^\infty (x - R) dF(x) = \mu - \int_R^\infty \int_R^X dy dF(x).$$

Changing the order of integration in the above integral proves that $E(\bar{X}_R) = \mu + B(\bar{X}_R)$ where

$$B(\bar{X}_R) = -\int_R^\infty [1 - F(x)]\, dx \tag{2.2}$$

is the bias of the winsorized mean.

By (2.1), an expression for the variance of $\bar{X}_R$ is

$$n\mathrm{Var}(\bar{X}_R) = \sigma^2 - 2\mathrm{cov}[X_1, \max(X_1 - R, 0)] + \mathrm{Var}[\max(X_1 - R, 0)]$$

where X_1 is the first random variable in the sample and σ^2 is the variance of $F(x)$. Manipulations similar to those yielding (2.2) show that

$$E[\max(X_1 - R, 0)^2] = 2\int_R^\infty (x - R)[1 - F(x)]\, dx,$$

and

$$E[\max(X_1 - R, 0) X_1] = 2\int_R^\infty (x - R)[1 - F(x)]\, dx - RB(\bar{X}_R).$$

Thus

$$\mathrm{Var}(\bar{X}_R) = \frac{1}{n}\left\{\sigma^2 - 2\int_R^\infty (x - \mu)[1 - F(x)]\, dx - B^2(\bar{X}_R)\right\},$$

and

$$\mathrm{MSE}(\bar{X}_R) = \frac{\sigma^2}{n} - \frac{2}{n}\int_R^\infty (x - \mu)[1 - F(x)]\, dx + \frac{n-1}{n} B^2(\bar{X}_R). \tag{2.3}$$

Searls (1966) showed that the mean square error of $\bar{X}_R$ has a unique minimum which can be obtained by equating the derivative, with respect to R, of $\mathrm{MSE}(\bar{X}_R)$ to 0. This yields the following equation for the optimal winsorization constant $R(F, n)$,

$$\frac{R - \mu}{n - 1} - \int_R^\infty [1 - F(x)]\, dx = 0. \tag{2.4}$$

This is equivalent to equation (14) in Searls (1966). In the remainder of this work, $\bar{X}_R$ denotes the optimal winsorized mean obtained with the winsorization constant $R(F, n)$ which solves (2.4). Observe that the optimal cut-off point $R(F, n)$ is location and scale equivariant, *i.e.*, if $G(x) = F[(x - b)/a]$, then $R(G, n) = aR(F, n) + b$.

A general algorithm for solving (2.4) is easily constructed. First observe that as a function of R, the left hand side of equation (2.4) is increasing and concave in R since its derivative, $1/(n - 1) + 1 - F(R)$, is positive and decreasing. Therefore, the Newton-Raphson algorithm (Thisted 1988, 164-167) given by

$$R_{j+1} = R_j - \frac{(R_j - \mu) - (n - 1)\int_{Rj}^\infty [1 - F(x)]\, dx}{1 + (n - 1)[1 - F(R_j)]}, \tag{2.5}$$

with $R_0 = 2\mu$ as starting value converges smoothly to the solution of (2.4). For discrete distributions the computations are easily implemented by noting that

$$\int_R^\infty [1 - F(x)]\, dx = E[\max(X - R, 0)].$$

Exact calculations of the optimal cut-off points $R(F,n)$ were carried out for the Weibull, the log-normal, and the Pareto families for samples of size s ranging between 5 and 200. Three distributions, corresponding to coefficients of variation (CV) of 1, 2, and 4, were considered in each family except for the Pareto family where only coefficients of variation of 2 and 4 were considered. The CV measures the skewness of a distribution, with large CVs corresponding to heavy skewness. The corresponding parameter values are given in Table 1.

Table 1

Parameter values of the distributions for which optimal cut-off values $R(F,n)$ were evaluated

CV	Weibull(α)	Log-normal(ν)	Pareto(γ)
1	1	0.83	-
2	1.84	1.27	2.67
4	2.87	1.68	2.13

For each distribution and each sample size, the optimal cut-off point was calculated using algorithm (2.5). Figure 1 presents the expected number of winsorized observations, $m(F,n) = n\{1 - F[R(F,n)]\}$ as a function of n while the corresponding efficiencies are reported in Figure 2. The efficiency of $\bar{X}_R$ is defined as $\mathrm{Var}(\bar{X})/\mathrm{MSE}(\bar{X}_R)$.

In Figure 1 the expected number of winsorized data values under the optimal scheme is, for most skewed distributions, close to 1 even for large sample sizes. Approximating this number by a Poisson distribution with parameter $m(F,n)$ shows there is a non-negligible probability that, under the optimal winsorization scheme, none of the data points is winsorized. This probability increases with the skewness of the distribution since $m(F,n)$ decreases with the CV. Thus, in samples from a highly skewed distribution, it is not always appropriate to winsorize the largest values. Such values should be winsorized only when they are large. As expected, in Figure 2, the largest gains in efficiency are obtained when the skewness is heavy. Therefore monitoring the two or three largest data values in a sample and curtailing their impact when these values are large is the key to a good winsorization strategy.

Figure 1 shows that the expected number of winsorized data values $m(F,n)$ decreases with the skewness of the distribution. This observation can be turned into a rigorous mathematical result. To this end, random variable Y is said to be more skewed than random variable X if Y has the same distribution as $\psi(X)$ where $\psi(x)$ is a convex function

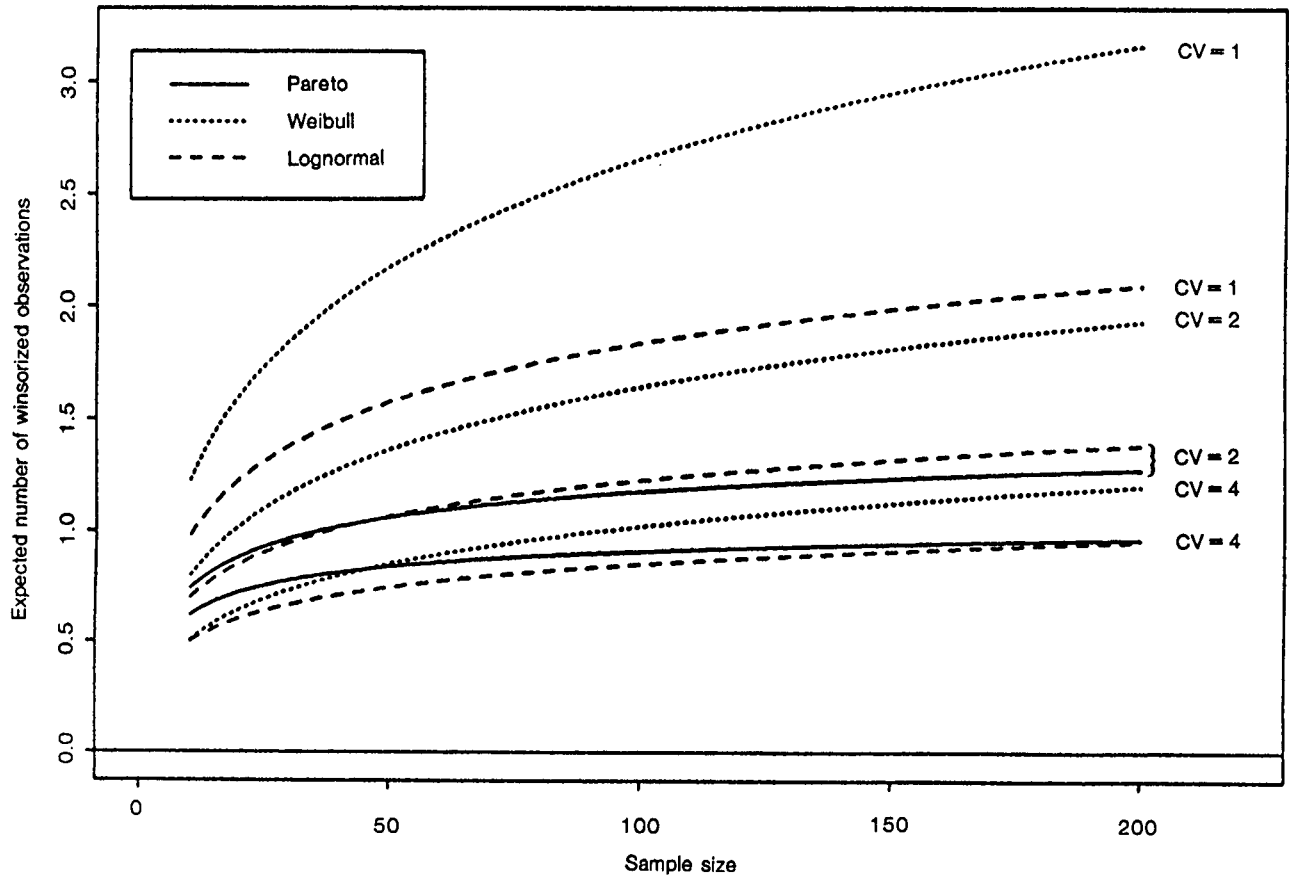


Figure 1. Expected number of winsorized observations for simple and stratified random sampling.

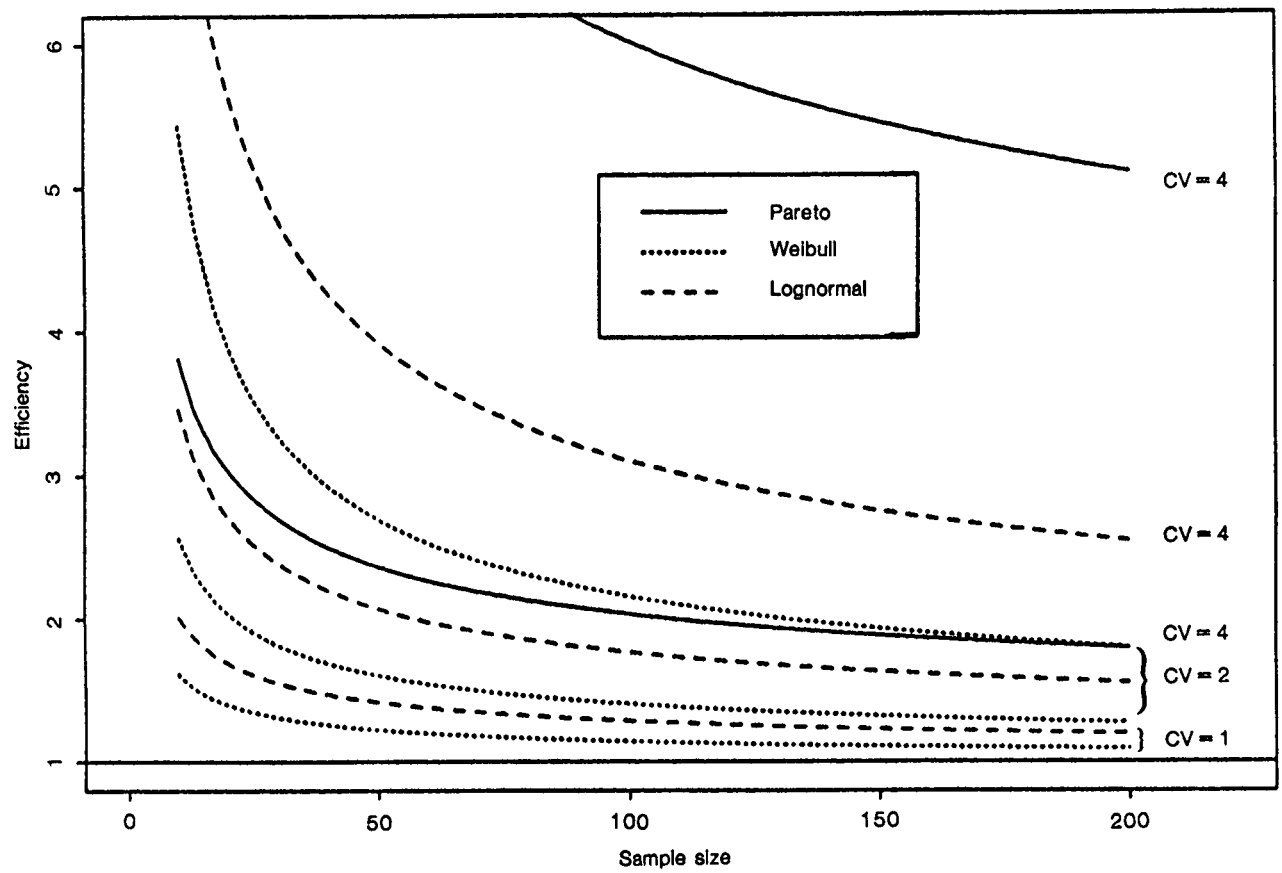


Figure 2. Efficiency of Searls winsorized mean.

of x. Under this definition X^2 is, as should be expected, more skewed than X. This notion of skewness corresponds to the convex partial ordering of van Zwet (Barlow and Proschan 1981). With this definition of skewness, one has the following proposition which is proved in the Appendix together with Propositions 2 and 3.

Proposition 1 If Y is more skewed than X then $m(F_X,n) > m(F_Y,n)$ where F_X and F_Y are the distributions of X and Y respectively.

The results of this section also apply to simple random sampling without replacement. For this design the mean square error of $\bar{X}_R$ is given by formula (2.3) with n replaced by $n/(1-f)$ where f is the sampling fraction. Algorithm (2.5), with n divided by $(1-f)$, can be used for calculating optimal cut-off values for without replacement simple random sampling.

3. WINSORIZATION IN STRATIFIED SAMPLING

There are many ways to generalize Searls' winsorization strategy to stratified sampling. In this section each stratum has its own cut-off value. Let R_h be the cut-off value in stratum h. The optimal values of $R_1, R_2, \ldots, R_L$, where L is the number of strata, are the ones that minimize the mean square error of $\bar{X}_R = \sum W_h \bar{X}_{Rh}$, where $\bar{X}_{Rh} = \sum \min(X_{hi},R_h)/n_h$, $W_h = N_h/N$ and N_h is the size of stratum h and $N = \sum N_h$. An algorithm for determining these optimal cut-off values is proposed in this section.

Let $F_h(x)$, for $h = 1, \ldots, L$ be the distribution of X in stratum h, and μ_h and σ_h^2 be the mean and the variance of F_h. The derivation of the mean square error of $\bar{X}_R$, under with replacement stratified random sampling, follows that presented in Section 2, it gives

$$\mathrm{MSE}(\bar{X}_R) = \sum_{h=1}^{L} \frac{W_h^2}{n_h} \left(\sigma_h^2 - 2 \int_{R_h}^{\infty} (x - \mu_h)[1 - F_h(x)]\,dx - B^2(\bar{X}_{Rh}) \right) + \left(\sum_{h=1}^{L} W_h B(\bar{X}_{Rh}) \right)^2 \quad (3.1)$$

where $B(\bar{X}_{Rh})$ is the bias of $\bar{X}_{Rh}$ as an estimator of μ_h

$$B(\bar{X}_{Rh}) = -\int_{R_h}^{\infty} [1 - F_h(x)]\,dx.$$

Taking the partial derivatives with respect to R_h, $h = 1, \ldots, L$ yields the following equations for the optimal values:

$$\frac{W_h}{n_h}[R_h - \mu_h - B(\bar{X}_{Rh})] = -\sum_{h=1}^{L} W_h B(\bar{X}_{Rh}), \quad (3.2)$$

for $h = 1, \ldots, L$.

There is no simple way to solve (3.2). An approximate solution can be obtained by noting that $B(\bar{X}_{Rh})/n_h$ is, for all values of h, usually small as compared to the other terms. Dropping these terms leads to

$$\frac{W_h}{n_h}(R_h - \mu_h) = -\sum_{h=1}^{L} W_h B(\bar{X}_{Rh}), \quad (3.3)$$

for $h = 1, \ldots, L$. The solutions to (3.3) overestimate slightly the optimal values satisfying (3.2) since at these solutions the partial derivatives of (3.1) are all positive and since these partial derivatives are increasing functions of R_h, for $h = 1, \ldots, L$. Thus by solving (3.3) to estimate the cut-off values one does not run the risk of winsorizing too many data values. Equations (3.3) imply that $R_h = \mu_h + n_h R/(nW_h)$ where R is some positive constant. A simple equation for R is obtained by changing variable $y = nW_h(x - \mu_h)/n_h$ in the integrals for $B(\bar{X}_{Rh})$, $h = 1, \ldots, L$ where $n = \sum n_h$. This gives

$$-\sum_{h=1}^{L} W_h B(\bar{X}_{Rh}) = \frac{R}{n} = \int_{R}^{\infty} [1 - F(y)]\,dy = -B(\bar{X}_R), \quad (3.4)$$

where $F(y) = \sum n_h F_h[\mu_h + n_h y/(nW_h)]/n$. Equation (3.4) is easily solved using algorithm (2.5) proposed in Section 2 for the single sample case. Therefore simple approximations for Searls' optimal cut-off values in stratified sampling are easily calculated.

Since the distribution F defined above has a zero expectation, the mean square error of the stratified winsorized mean obtained by solving (3.3) is equal to:

$$\mathrm{MSE}(\bar{X}_R) = \frac{1}{n}\left(\sigma_F^2 - 2\int_R^{\infty} y[1 - F(y)]\,dy - B(\bar{X}_R)^2\right) + B(\bar{X}_R)^2 + \left(\frac{1}{n}B(\bar{X}_R)^2 - \sum_{h=1}^{L}\frac{W_h^2 B^2(\bar{X}_{Rh})}{n_h}\right) \quad (3.5)$$

where σ_F^2 is the variance of F. The last term of (3.5) is easily shown to be negative or null; it is null when $B(\bar{X}_R) = nW_h B(\bar{X}_{Rh})/n_h$ for $h = 1, \ldots, L$. The variance of the stratified mean, $\bar{X} = \sum W_h \bar{X}_h$, is equal to σ_F^2/n. Thus a conservative approximation to the efficiency of $\bar{X}_R$ with respect to $\bar{X}$ in stratified sampling is equal to the corresponding efficiency for a random sample of size n drawn from F. Note also that $n[1 - F(R)]$ represents the expectation of the total number of winsorized data points in the L strata.

The optimal winsorization scheme obtained by solving (3.3) has a simple form for many allocation rules. Under proportional allocation, *i.e.*, $n_h = nW_h$ for $h = 1, \ldots, L$, one gets $R_h = \mu_h + R$. Under Neyman optimal allocation, with $n_h = nW_h\sigma_h/(\sum W_h\sigma_h)$ where σ_h is stratum h's standard deviation, one gets $R_h = \mu_h + \sigma_h R/(\sum W_h\sigma_h)$. If in addition, the distributions of X within the strata are equal up to a change in location and scale, *i.e.*, $F_h = F_0[(x - \mu_h)/\sigma_h]$ for some distribution F_0, then $F(x) = F_0[x/(\sum W_h\sigma_h)]$. In this case the characteristics of optimal winsorized means in stratified sampling and in simple random sampling are the same. Thus Figure 1 presents the expected total number of winsorized data points in the L strata as a function of the total sample size n, under Neyman allocation, when F_0 is one of the distributions of Table 1. Figure 2 gives the corresponding efficiencies.

The results of this section are easily generalized to without replacement stratified sampling by replacing n_h by $n_h/(1 - f_h)$ throughout the calculations. The derivation of optimal cut-off values for stratified pps sampling is easily carried out by taking $F_h(x) = \sum p_{hi} I(y_{hi}/(N_h p_{hi}) \leq x)$ where p_{hi} denotes the selection probability for unit the i-th unit of stratum h.

4. LARGE SAMPLE APPROXIMATIONS TO THE EFFICIENCY OF THE WINSORIZED MEAN

For most distributions, equation (2.3) defining the optimal cut-off does not have an explicit solution. This section derives closed form approximations to this solution using the theory of extreme order statistics. This will permit the derivation of explicit approximations to the efficiency of the optimal winsorized mean. Searls' optimal winsorization strategy will then be compared to a simple non parametric winsorization scheme where the largest order statistic is replaced by the second largest (Rivest 1994).

The form of the approximation to $R(F,n)$ depends on the limiting distribution, as the sample size n goes to infinity, of the largest order statistic suitably normalized. For distributions whose support is the positive axis, there are only two possible limiting distributions which are given by Galambos (1987, p. 53-54)

$$H_{1,\alpha}(x) = \exp(-x^{-\alpha}) \quad \text{for} \quad x > 0 \quad \text{and} \quad \alpha > 0$$

and

$$H_{3,0}(x) = \exp[-\exp(-x)] \quad \text{for} \quad x \text{ in } R.$$

For many distributions used for the statistical analysis of positive random variables, for example the Weibull and the log-normal families, the sample maximum suitably normalized converges to $H_{3,0}(x)$. Distributions whose sample maxima converge to $H_{1,\alpha}(x)$ for some $\alpha > 0$ have heavy tails. For such distributions $1 - F(x)$ goes to 0 at a rate of $O(x^{-\alpha})$. The Pareto and the F distributions are in this class.

Distributions whose sample maxima converge to $H_{3,0}(x)$ are considered first.The following characterization is due to von Mises (1964): the sample maximum of a twice differentiable distribution $F(x)$ converges to $H_{3,0}(x)$ if, as x goes to ∞,

$$\lim_x \frac{g'(x)}{g^2(x)} = 0 \tag{4.1}$$

where $f(x)$ is the density of F, $g(x) = f(x)/[1 - F(x)]$ is the failure rate of F, and g' is the derivative of g. An approximation to winsorization constant $R(F,n)$ for this class of distributions is presented next.

Proposition 2 If $F(x)$ is such that (4.1) holds and if, for large values of x, it satisfies:

i) $xg(x)$ increases;
ii) $xg'(x)/g(x)$ is less than some positive constant c;

then the optimal winsorization constant $R(F,n)$ satisfies

$$R(F,n) = F^{-1}\left(1 - \frac{g[F^{-1}(1-1/n)]F^{-1}(1-1/n)[1+o(1)]}{n}\right);$$

and $m(F,n) = g(F^{-1}(1 - 1/n))F^{-1}(1 - 1/n)[1 + o(1)]$. Furthermore, the mean squared error of Searls' winsorized mean is approximately equal to

$$\mathrm{MSE}(\bar{X}_R) \approx \frac{\sigma^2}{n} - \frac{R(F,n)^2}{n^2}.$$

In the Weibull family, $F_\alpha^{-1}(1 - t) = [-\log(t)]^\alpha$, $g(x) = x^{1/\alpha - 1}/\alpha$. The hypotheses of Proposition 2 are met and $m(F_\alpha, n)$, the expected number of winsorized observations in a large Weibull sample, is $\log(n)[1 + o(1)]/\alpha$ which goes to ∞ as n increases. Figure 1 suggests that the convergence is very slow, especially for large coefficients of variation.

Now consider distributions whose sample maxima converge to $H_{1,\alpha}(x)$. This class of distributions has been characterized by Gnedenko (1962): the sample maximum of F converges to $H_{1,\alpha}(x)$ if one can write

$$1 - F(x) = L(x)/x^\alpha \tag{4.2}$$

where as x goes to ∞, $L(x)/L(kx)$ converges to 1, for any constant k. Note that for F to have a finite second moment, one needs $\alpha > 2$ in (4.2). The Pareto distribution satisfies (4.2) with $\alpha = \gamma$.

Proposition 3 If F satisfies (4.2) with parameter α where $\alpha > 2$, then as n goes to infinity, $R(F,n) = F^{-1}[1 - (\alpha - 1)/n][1 + o(1)]$, *i.e.*, $m(F,n) \approx \alpha - 1$. Furthermore,

$$\mathrm{MSE}(\bar{X}_R) \approx \frac{\sigma^2}{n} - \frac{\alpha R(F,n)^2}{n^2(\alpha - 2)}.$$

For distributions satisfying (4.2) a finite number of data points are on average winsorized as the sample size goes to ∞. To some extent, this can be seen in Figure 1 where the curves of $m(F_\gamma, n)$ for the Pareto distribution have $m(F_{2.33}, n) = 1.33$, and $m(F_{2.67}, n) = 1.67$ as asymptotes.

Propositions 2 and 3 shed some light on the estimation of the optimal cut-off value. When F is unknown, a possible estimator for $R(F,n)$ is the value that minimizes an estimator of the mean square error of $\bar{X}_R$. This leads to

$$\frac{R - \bar{X}}{n - 1} = \frac{1}{n}\sum_{i=1}^{n} \max(X_i - R, 0) \tag{4.3}$$

as an estimating function for R. This procedure is questionable when the underlying distribution is highly skewed, *i.e.*, when F satisfies the assumption of Proposition 3. On average, there will only be $\alpha - 1$ non-null terms in the right hand side of equation (3). Thus $\hat{R}$ will, on average be determined by the $\alpha - 1$ largest data values and the sample maximum will have the largest influence on $\hat{R}$. This will make $\hat{R}$ highly unstable and, considering the findings of Figure 1, the second largest sample order statistic should be a better estimator of $R(F,n)$ than the solution of (3.3). This is exemplified in the Monte Carlo simulations of Section 5.

Table 2 compares approximations to the bias and to the mean square error of Searls' winsorized mean $\bar{X}_R$ to those of the once winsorized mean $\bar{X}_1$ obtained by taking the cut-off value R equal to the second largest observation. Rivest (1994) shows this choice of cut-off value yields the optimal non-parametric winsorized mean. He also derives the large sample approximations for the bias and the mean square error of $\bar{X}_1$ appearing in Table 2. The corresponding expressions for $\bar{X}_R$ are taken in Propositions 2 and 3.

Table 2

Approximations to the bias and to the mean square error of the once winsorized mean $\bar{X}_1$ and of Searls' optimal winsorized mean, $\bar{X}_R$, for the Weibull and for the Pareto distribution ($\Gamma(\cdot)$ stands for the gamma function)

		WEIBULL	PARETO
$\bar{X}_R$	MSE	$\frac{\sigma^2}{n} - \frac{(\log n)^{2\alpha}}{n^2}$	$\frac{\sigma^2}{n} - \frac{\gamma}{(\gamma-2)(\gamma-1)^{2/\gamma}n^{2-2/\gamma}}$
	bias	$-\frac{(\log n)^{\alpha}}{n}$	$-\frac{1}{(\gamma-1)^{1/\gamma}n^{1-1/\gamma}}$
$\bar{X}_1$	MSE	$\frac{\sigma^2}{n} - \frac{2\alpha(\alpha-1)(\log n)^{2\alpha-2}}{n^2}$	$\frac{\sigma^2}{n} - \frac{2\Gamma(1-2/\gamma)}{\gamma(\gamma-1)n^{2-2/\gamma}}$
	bias	$-\frac{\alpha(\log n)^{\alpha-1}}{n}$	$-\frac{\Gamma(1-1/\gamma)}{\gamma n^{1-1/\gamma}}$

In Table 2 the mean square error of $\bar{X}_R$ is much smaller than that of $\bar{X}_1$. Indeed, for the Weibull distribution the large sample efficiency of $\bar{X}_R$ with respect to $\bar{X}_1$ is equal to that of $\bar{X}_R$ with respect to $\bar{X}$. Thus non-parametric winsorization reduces the mean square error of estimators of the mean of a skewed population however further reductions in mean square error can be obtained if information concerning the underlying distribution is available. This is illustrated in the Monte Carlo comparisons presented in the next section.

The results of this section apply to stratified sampling. For this design, the large sample solution to equation (3.4) is determined by the stratum with the most skewed distribution. If F_1 is the most skewed distribution then $nW_1R(F_1,n_1)/n_1$ is an approximate solution to (3.4) where an approximation to $R(F_1,n_1)$ is found in Proposition 2 or in Proposition 3 depending on the tail of F_1. In this case only data points in stratum 1 are winsorized in large stratified samples. Searls' winsorized mean is then equal to W_1 times the optimal winsorized mean for stratum one plus a weighted sum of the sample means in the other strata.

5. MONTE CARLO COMPARISONS OF ESTIMATORS OF THE MEAN OF A SKEWED DISTRIBUTION

This section presents Monte Carlo comparisons of the mean square error and of the biases of five estimators of the mean of population CHICKEN of Fuller (1991). This population has 2000 units; its coefficient of variation is 4.46. Further numerical comparisons of the five estimators considered in this section for other distributions, either finite or infinite, are presented in Rivest (1993a and b).

The five estimators under consideration are:

- Searls' winsorized estimator, $\bar{X}_R$, calculated as if the the underlying distribution was known;
- A winsorized estimator where the cut-off value is set equal to the second largest data value of an auxiliary sample of size $2n$; this is an instance where limited auxiliary information concerning the underlying distribution F is available (in the Monte Carlo simulations each simulated sample had its own auxiliary sample);
- The once winsorized mean, $\bar{X}_1$, introduced in Section 4;
- A winsorized estimator where R is estimated from the sample by solving equation (4.3);
- Fuller preliminary test estimator with $j = 3$ (*i.e.*, the numerator of the preliminary test involves the three largest observations), T (the total number of data points involved in the preliminary test) equal to $[4n^{1/2} - 10]$ and K_3, the cut-off value equal to 3.5. A detailed description of this estimator appears in Fuller (1991) and in Rivest (1993a and b). This estimator curtails the largest data values only when a test statistic for detecting extreme data values is significant.

The biases and the efficiencies of $\bar{X}_R$ were calculated exactly. For the other estimators, the biases and the efficiencies presented in Figures 1 and 2 were obtained in Monte Carlo simulations based on 100,000 repetitions.

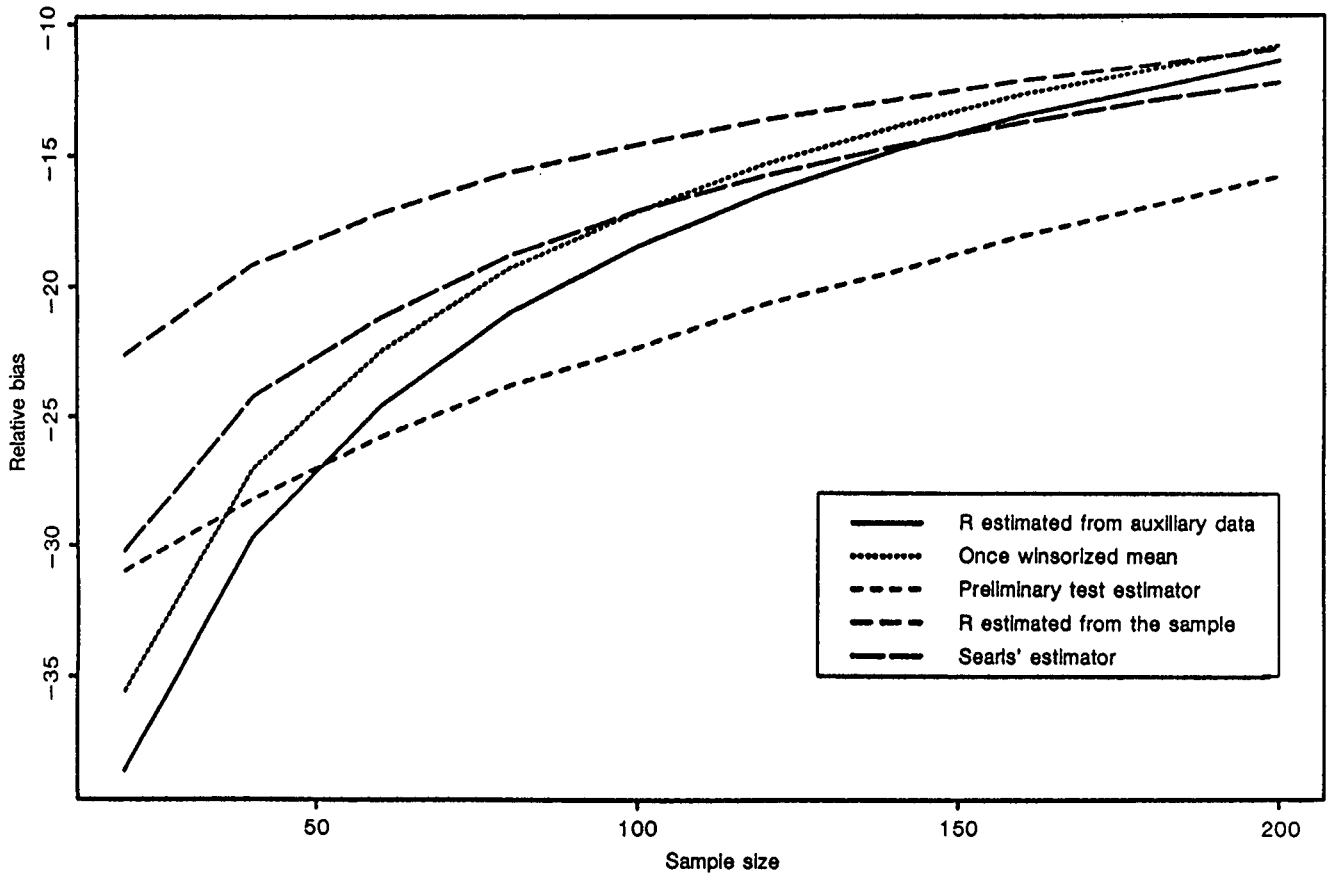


Figure 3. Relative bias of five estimators for the mean of CHICKEN.

Figure 3 indicates that the biases of winsorized estimators are important, even in large samples. Several interesting conclusions can be drawn from Figure 4. First, as expected from Table 2 Searls' estimator is much more efficient than the once winsorized mean. Estimating the optimal cut-off value using limited auxiliary information is highly efficient. This holds true as long as the study variable can be modeled by a superpopulation distribution having a finite variance, see Rivest (1993a) for further discussions. In a sampling context, the auxiliary samples could be data from previous surveys standardized to account for possible changes over time in the distribution of the variable under study.

Among the three estimators of Figure 4 that do not rely on auxiliary information, Fuller estimator is the best. This is in agreement with the simulation results of Fuller (1991). Estimating the cut-off value by minimizing an estimate of the mean square error does poorly especially in small samples. Thus, as shown in Section 4, the resulting estimator is highly sensitive to the wild data values that sometimes appear in small samples. This estimator is not recommended.

6. CONCLUSIONS

Many strategies can be used to accomodate the large values that sometimes arise in surveys. If auxiliary information, such as census data, is available then one can use Searls' estimator in either simple random sampling, stratified sampling, or pps sampling. Since the cut-off values are fixed constant mean square error estimators can be derived from formulae (2.3) and (3.1).

When extra information is not available, the once winsorized mean and Fuller preliminary test estimator can be used. Research is now under way to generalize these estimators to stratified designs. An estimator for the mean square error of the once winsorized mean is proposed in Rivest (1994),

$$v(\bar{X}_1) = \frac{1}{n} S^2 - \frac{1}{n^2} (X_n + X_{n-1} - 2\bar{X}_1) \\ (X_n - 3X_{n-1} + 2X_{n-2})$$

where S^2 denotes the variance of the X-sample and $X_n > X_{n-1} > X_{n-2}$ denote the three largest data values in

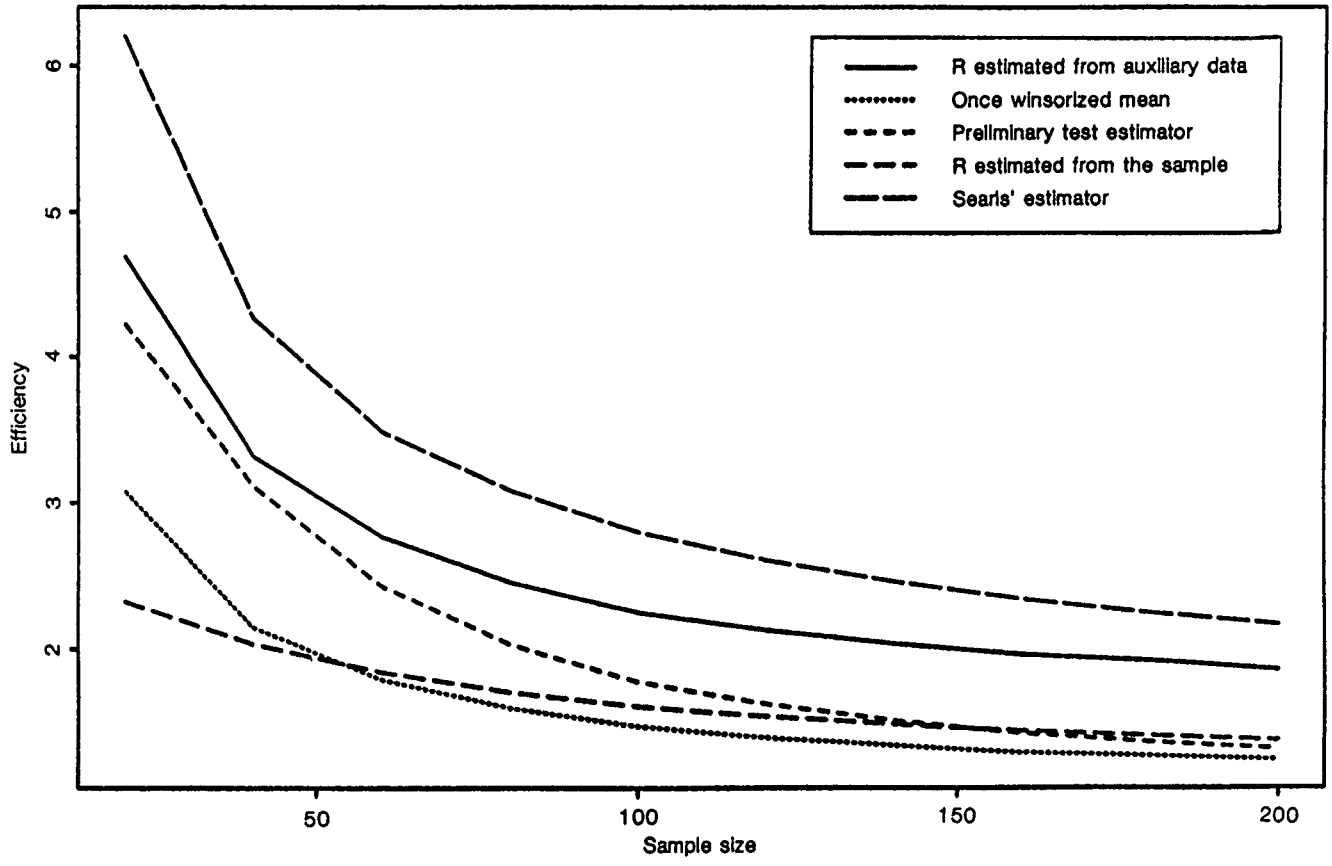


Figure 4. Efficiency of five estimators for the mean of CHICKEN.

that sample. This estimator has a small bias in infinite populations. However the coverage of the standard confidence interval $\bar{X}_1 \pm z_{1-\alpha/2}\sqrt{v(\bar{X}_1)}$ is often well below the nominal $100(1 - \alpha)\%$ level especially when the underlying distribution is skewed. Further research is needed to obtain reliable confidence intervals for estimators of the mean of skewed populations.

ACKNOWLEDGEMENTS

This research was supported by the Natural Science and Engineering Research Council and by the Fond pour la formation des chercheurs et l'aide à la recherche of Québec.

APPENDIX 1

Proof of Proposition 1 The assumption that Y is more skewed than X implies that there exists a convex function ψ such that $\psi(X)$ and Y have the same distribution. Let R denote $R(F_X, n)$. To prove the result, it suffices to show that $\psi(R) < R(F_Y, n)$. This is equivalent to

$$\frac{\psi(R) - E(Y)}{n - 1} < \int_{\psi(R)}^{\infty} [1 - F_Y(x)]\,dx. \quad \text{(A.1)}$$

By Jensen's inequality, $E(Y) = E[\psi(X)] > \psi[E(X)]$. Thus using (2.3), the left hand side of (A.1) is less than or equal to

$$\frac{R - E(X)}{n - 1} \frac{\psi(R) - \psi[E(X)]}{R - E(X)} < \frac{R - E(X)}{n - 1} \psi'(R) = \int_{R}^{\infty} [1 - F_X(y)]\,dy \cdot \psi'(R)$$

where ψ' is the derivative of ψ. Since ψ' is increasing, the left hand side of the above inequality is less than or equal to:

$$\int_{R}^{\infty} \psi'(y)[1 - F_X(y)]\,dy = \int_{\psi(R)}^{\infty} [1 - F_Y(x)]\,dx.$$

This shows that (A.1) holds.

Proof of Proposition 2 The following result obtained by applying Theorems 2.7.5 and 2.7.11 of Galambos (1987) to the distribution $F(z^{p+1})$ is used extensively. If the sample maxima of distribution $F(x)$ converges to $H_{3,0}(x)$, then all the moments of F exist and

$$\int_x^\infty y^p[1 - F(y)]dy \sim \frac{[1 - F(x)]\,x^p}{g(x)} \qquad \text{(A.2)}$$

where $g(x) \sim h(x)$ means that $g(x)/h(x)$ converges to 1 as x goes to infinity. Using (A.2), $R(F,n)$ is obtained by solving

$$\frac{R - \mu}{n - 1} = \frac{1 - F(R)}{g(R)}\,(1 + o(1)).$$

Let $R = F^{-1}(1 - a/n)$, then, up to $(1 + o(1))$, the above equation becomes

$$a = g\left[F^{-1}\left(1 - \frac{a}{n}\right)\right]F^{-1}\left(1 - \frac{a}{n}\right). \qquad \text{(A.3)}$$

Let $a_0 = g[F^{-1}(1 - 1/n)]F^{-1}(1 - 1/n)$ and $a_1 = g[F^{-1}(1 - a_0/n)]F^{-1}(1 - a_0/n)$. Since for large values of $x, xg(x)$ is increasing, $a_0 > a_1$ and the solution to (A.3) belongs to the interval (a_1, a_0). In order to prove the result, one has to show that a_1/a_0 converges to 1 as n goes to ∞.

Since $g(x) = f(x)/[1 - F(x)]$, one can write

$$a_0 = \exp\left[\int_{F^{-1}(1-a_0/n)}^{F^{-1}(1-1/n)} g(t)dt\right] =$$

$$\exp\left[a_0 - a_1 - \int_{F^{-1}(1-a_0/n)}^{F^{-1}(1-1/n)} tg'(t)dt\right],$$

where the second expression is obtained by integrating by parts. Since $tg'(t)/g(t)$ is less then c, one has $a_0 > \exp(a_0 - a_1)a_0^{-c}$. If a_1/a_0 does not converge to 1, say $a_1/a_0 < 1 - \epsilon < 1$ for an infinite sequence of sample sizes, the previous inequality implies that $a_0^{1+c} > \exp(a_0\epsilon)$. This is a contradiction since a_0 tends to ∞ as n becomes large. The approximation for $\mathrm{MSE}(\bar{X}_R)$ is obtained by using (A.2) with $p = 2$.

Proof of Proposition 3 If the sample maxima of distribution $F(x)$ converges to $H_{1,\alpha}(x)$ then F satisfies the following properties (Feller 1971, p. 281):

$$\int_x^\infty y^p[1 - F(y)]dy \sim \frac{[1 - F(x)]x^{p+1}}{\alpha - p - 1} \qquad \text{(A.4)}$$

for any p such that $\alpha - p - 1 \geq 0$. By (A.4), $R(F,n)$ is obtained by solving $F(R) = 1 - [\alpha - 1 + o(1)]/n$. This leads to the approximation for $R(F,n)$. To derive the approximation for $\mathrm{MSE}(\bar{X}_R)$, one applies (A.4) with $p = 1$.

REFERENCES

BARLOW, R.E., and PROSCHAN, F. (1981). *Statistical Theory of Reliability and Life Testing*. Silver Spring MD: To Begin With.

CHAMBERS, R.L., and KOKIC, P.N. (1993). Outlier robust sample survey inference. *Bulletin of the International Statistical Institute. Proceedings of the 49th session*, book 2, 54-72.

ERNST, L.R. (1980). Comparison of estimators of the mean which adjust for large observations. *Sankhyā* C, 42, 1-16.

FELLER, W. (1971). *An Introduction to Probability Theory and its Applications*. Volume II. Second Edition. New York: Wiley.

FULLER, W.A. (1991). Simple estimators for the mean of skewed populations. *Statistica Sinica*, 1, 137-158.

FULLER, W.A. (1993). Estimators for long-tailed distributions. *Bulletin of the International Statistical Institute. Proceedings of the 49th session*, book 2, 39-54.

GALAMBOS, J. (1987). *The Asymptotic Theory of Extreme Order Statistics*. Second edition. Malabar FL: Krieger.

GNEDENKO, B.V. (1962). *The Theory of Probability*. New York: Chelsea.

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

HICKS, S., and FETTER, M. (1993). An evaluation of robust estimation techniques for improving estimates of total hogs. *Proceedings of the International Conference on Establishment Surveys*. Alexandria, Virginia: American Statistical Association, 385-389.

HIDIROGLOU, M.A. (1987). The construction of a self-representing stratum of large units in survey design. *The American Statistician*. 40, 27-31.

HUBER, P.J. (1981). *Robust Statistics*. New York: John Wiley.

LEE, H. (1994). Outliers in Survey Data. Statistics Canada paper.

RIVEST, L.-P. (1994). Some sampling properties of winsorized means for skewed distributions. *Biometrika*, 81, 373-384.

RIVEST, L.-P. (1993a). Winsorization of survey data. *Bulletin of the International Statistical Institute. Proceedings of the 49th session*, book 2, 73-89.

RIVEST, L.-P. (1993b). Winsorization of survey data. *Proceedings of the International Conference on Establishment Surveys*. Alexandria, Virginia: American Statistical Association, 396-401.

SEARLS, D.T. (1966). An estimator which reduces large true observations. *Journal of the American Statistical Association*, 61, 1200-1204.

THISTED, R.A. (1988). *Elements of Statistical Computing*. New York: Chapman and Hall.

VON MISES, R. (1964). *Selected Papers of Richard von Mises*. Volume II. Providence: American Mathematical Society.

Survey Methodology, December 1995
Vol. 21, No. 2, pp. 117–124
Statistics Canada

Design Effects for Correlated $(P_i - P_j)$

LESLIE KISH, MARTIN R. FRANKEL, VIJAY VERMA and NIKO KAĆIROTI[1]

ABSTRACT

We present empirical evidence from 14 surveys in six countries concerning the existence and magnitude of design effects (defts) for five designs of two major types. The first type concerns deft$(p_i - p_j)$, the difference of two proportions from a polytomous variable of three or more categories. The second type uses Chi-square tests for differences from two samples. We find that for all variables in all designs deft$(p_i - p_j) \cong [\text{deft}(p_i) + \text{deft}(p_j)]/2$ are good approximations. These are *empirical* results, and exceptions disprove the existence of mere analytical inequalities. These results hold despite great variations of defts between variables and also between categories of the same variables. They also show the need for sample survey treatment of survey data even for analytical statistics. Furthermore they permit useful approximations of deft$(p_i - p_j)$ from more accessible deft(p_i) values.

KEY WORDS: Design effects; Survey sampling; Sampling errors.

1. DESIGN EFFECTS FOR ANALYTICAL STATISTICS

We explore the existence and the magnitudes of design effects for some special analytical statistics based on data from survey samples. The investigation is both methodological and empirical, with data from several different surveys with different variables and from contrasting populations, hence subject to the risks of inconsistent empirical results. We often hear and read that probability sampling, while necessary for descriptive surveys, is not necessary for analytical surveys. In "Four Obstacles to Representation in Analytic Studies" one of us wrote that "In addition to those four real obstacles, we also encounter another, which is more artificial, in the denials of the need for representation" (Kish 1987, Section 2.7). Sampling investigations show that complex probability selections, especially clustered sampling, have no appreciable influence on descriptive statistics (like means and regression coefficients), but can have drastic effects on inferential statistics, like confidence intervals, tests of significance (Kish and Frankel 1974).

Design effects are defined as deft^2 = actual variance/simple random variance of same n, both estimated. And values of deft > 1 have been shown for sampling errors not only of means, but also for analytical statistics like differences of means (and Chi square tests), regression coefficients *etc.* It is true that considerable reductions and differences of deft values have been found for some analytical statistics. The differing deft values are not mere necessary mathematical consequences of the sample design, which may be deduced once for all. They have empirical content and therefore they need to be replicated with empirical investigations (Kish and Frankel 1974; Kish 1987, 7.1; Kish 1965, 14.1-14.2; Rao and Wu 1985; Scott and Holt 1982; Skinner, Holt, and Smith 1989). In this paper we investigate the possible effects and the magnitudes of design effects for a set of related statistics that have not been investigated before. On the contrary, in several statistical papers the absence of design effects was merely assumed by the authors (all justly famous), and apparently passed on by the journal referees, without warning the readers. We shall see if deft is reduced or eliminated for this set of analytical statistics (Cochran 1950; Mosteller 1952; Scott and Seber 1983; Seber and Wild 1993).

Furthermore, we also propose explicitly, as has been implied before, that the existence of considerable values of deft is strong evidence for the need for probability selections. It would be difficult to assume a model of a population distribution where the selection design was unimportant (or uninformative) but produced considerable design effects. The reverse does not hold: absence of design effects is necessary but not sufficient evidence for license to neglect probability selection. This proposition gives added importance to our study, which relates deft$(p_i - p_j)$ for analytical statistics to deft(p_i) and deft(p_j) for two of several categories of the same variable.

Section 2 describes the five related problems (designs) for which sampling errors are described in Section 3. Section 4 discusses the empirical evidence in the tables. Section 5 places our findings in the context of earlier work on defts for subclasses and their differences.

[1] Leslie Kish, ISR, University of Michigan, Ann Arbor MI 48106, U.S.A.; Martin R. Frankel, NORC and City University of New York; Vijay Verma, University of Essex, Colchester, C04 3SQ, U.K.; Niko Kaćiroti, Institute of Statistics, Tirana, Albania.

2. SIMILAR STATISTICS FOR FIVE DESIGNS

It has been shown that five designs (problems), of two distinct types, can be treated with the same simple statistics (Kish 1965, Section 12.10). For our empirical and simple presentation we use symbols for sample values (like deft, p_i and n_i), even when occasionally capitals for population values would be more appropriate.

The difference of proportions $p_2 - p_0 = n_2/n - n_0/n$ expresses the desired estimate, where $n = n_0 + n_1 + n_2 + \ldots n_k$ is the sample size, with n units selected and weighted equally. Furthermore, under simple random sampling assumptions, the variance of $(p_2 - p_0)$ is $(1 - f)[p_2 + p_0 - (p_2 - p_0)^2]/(n - 1)$.

Type A Comparisons

1. The difference between two categories $(n_2 - n_0)/n = (p_2 - p_0)$ of a polytomy can represent preference between two parties among several (k) in voting surveys, or between two brands of automobiles in market research, or two of several attitudes, opinions, behaviors on one variable, *etc.* The other $(k - 2)$ choices are summed into p_1 and disregarded in the difference. (Also treated by Scott and Seber 1983.)

2. Rank values of -1, 0, $+1$ (or 0, 1, 2 or c, $c + 1$, $c + 2$) can be assigned to an ordered trichotomous variable without a metric, and viewed as a simple form of the difference of two categories. This form is particularly useful for computations of sampling errors, because all the five designs can use -1, 0, $+1$ for instance as a transformed computing variable.

3. The difference of proportions from two different variables (x and y) may be treated as in (1) and (2). Define as positive in x (or success) only those elements that are positive in x but not in y, so that $n_{10} = n(x_1, y_o)$. Similarly define as positive y the $n_{01} = n(x_0, y_1)$. Then $(n_{10} - n_{01})/n = (p_x - p_y)$ is the net difference in the proportion of positives in x and y. Those that are positives or negatives in both x and y do not count in the differences. Thus we have a case of three categories as in (1) and (2). An example is the difference between the proportions who would "stop all nuclear testing," and those who "want complete nuclear disarmament"; or who would "force Iraq to leave Kuwait" and who would "remove Saddam from power," (Wild and Seber 1993). However, the two categories may also come from two *different surveys* of the same n cases, as in a quality check, or from dual frame observations, or from two waves of a sample. These situations resemble those of (4) and (5).

Type B Comparisons

4. Test-retest and before-after are terms for designs in which the same subjects undergo two observations. Then dichotomous answers $n_2 = n_{10}$ denote the number of negative changes; $n_0 = n_{01}$ the number of positive changes; and $n_{11} + n_{00}$ the sum of the unchanged positives and negatives. Positive and negative answers are respectively denoted here as 1 and 0, and the first and second wave by the order of the subscript. The difference $(n_{10} + n_{11}) - (n_{01} + n_{11}) = n_{10} - n_{01} = n_2 - n_0$ measures the change between positives for the two observations; and $p_2 - p_0 = n_2/n - n_0/n$ measures the change in proportions. (McNemar 1949; Cochran 1950; Mosteller 1952).

5. Matched pairs of n pairs of subjects can also be treated as a generalization of the test-retest design (Mosteller 1952). For example n pairs of randomized subjects may represent experimental versus control treatments; or n pairs of boys versus girls matched on control variables. The statistical treatment $(p_{10} - p_{01})$ of the n pairs of matched subjects is the same as for the n pairs of treatments on the same n subjects (4).

The similarity of statistical treatment for these five designs of two distinct types is convenient, and we present empirical results for both types. "It also has heuristic value that has been overlooked in recent publications (Scott and Seber 1983 and Wild and Seber 1993). The Chi-square test for types 4 and 5 was published early (McNemar 1949; Cochran 1950; Mosteller 1952), and the similarity to the categorical cases 1, 2, 3 was shown" (Kish 1965, 12.10). (Kish was wrong in denoting "trichotomies and matched dichotomies," as "Trinomials and Matched Binomials," which terms refer to IID samples only.)

All of these deal with differences of proportions p_i based on count variables n_i. Extensions to correlated differences $(y_i - y_j)$ for other variables are possible, but not within the scope of our study. Practical examples would include the difference in dollar shares (not only numbers n_i) between two automobile makes from a total of $\sum y_i$ sales.

3. SAMPLING ERRORS AND DESIGN EFFECTS

For simple random samples of size n it can be easily shown (Kish 1965, 12.10) that

$$\text{var}(p_2 - p_0) = \left[\frac{(1 - f)n}{(n - 1)}\right][p_2 + p_0 - (p_2 - p_0)^2]/n.$$

Most of the examples found and shown come from large survey samples, where the $(1 - f)$ can be disregarded. It is worth noting that for the element variance

$$p_2 + p_0 - (p_2 - p_0)^2 = p_2q_2 + p_0q_0 + 2p_2p_0,$$

where the last term $\text{cov}(p_2,p_0) = -p_2p_0$ represents the covariance arising because p_2 and p_0 are competitive parts of the same sample, rather than proportions from independent samples. The difference of proportions squared $(p_2 - p_0)^2$ will usually be a small correction term, and without it we have the equivalent of the variance $(p_2 + p_0)/n$ of two independent Poisson samples. Furthermore, note that (Kish 1965, 12.10):

> The Chi-square test has been applied to some of these problems, treated separately (Cochran 1950; Mosteller 1952; McNemar 1962, p. 225). This is essentially $(n_2 - n_0)^2/(n_2 + n_0)$ the square of the difference divided by its variance, under the null hypothesis $n_2 = n_0$. It applies the exact theories available for tests of null hypotheses in small samples, including the "Yates correction," all based on the assumption of simple random sampling. However, there are great advantages in treating these problems in large samples as estimated means with proper standard errors. First, instead of being confined to testing null hypotheses, we can make inferences with the probability intervals $(p_2 - p_0) \pm t_p\, se\,(p_2 - p_0)$. Second, the formulas for standard errors of complex samples can be applied directly to the mean $(p_2 - p_0)$. Third, the logical structure of this statistic $(p_2 - p_0)$ can be seen more clearly in its application to several distinct problems.

Correlated proportions originate usually in data from complex surveys, and the computations of variance should be appropriate to the sample design. The variance formulas for stratified complex samples can be adopted, but the direct formula has eight terms (Kish 1965, 12.10.3). Instead, it is convenient to translate the problem into a trichotomous variable, with values of -1, 0, $+1$ as in design 2 of Section 2; and the computations of Section 4 used that translation.

Then comparisons between variables and between samples can be facilitated by recourse to the design effects:

$$\text{deft}^2(p_2 - p_0) = \frac{\text{computed variance of } (p_2 - p_0)}{[p_2 + p_0 - (p_2 - p_0)^2]/n}.$$

A few words are needed about limitations on the use of deft as a tool for robust approximations. They serve well for clustered and multi-stage samples using ultimate clusters (primary selections) for computing sampling errors. However, we avoided the problem of weighted samples, because their treatment would be too specific and perhaps too complex. Weighting for nonresponse would not be important for the ratio of $\text{deft}(p_i - p_j)$ to $\text{deft}(p_i)$. However weights for gross inequalities of selection probabilities need specific treatments. Nevertheless, inference and experience indicate that deft values are less affected by weights than are the variances and means themselves. Furthermore we conjecture that the relations we found between the values of $\text{deft}(p_i - p_j)$ and $\text{deft}(p_i)$ will hold also for weighted data, if these are not extreme or pathological.

An approximate but dependable relation of $\text{deft}(p_i - p_j)$ to $\text{deft}(p_i)$ and $\text{deft}(p_j)$ would be useful to allow inferences from the latter, which are routinely and easily computed, to the former that are not. Several alternative conjectures may seem reasonable, and none can be mathematically derived, nor excluded.

1. $\text{Deft}(p_i - p_j) = 1$ of no design effect was assumed implicitly in the five publications referenced in Section 1.
2. $\text{Deft}(p_i) > \text{deft}(p_i - p_j) > 1$ denotes persisting but lower effects than for the $\text{deft}(p_i)$ for proportions. This happens for "crossclasses" and their comparisons (Kish 1987, 7.1). This also seemed reasonable to several experienced statisticians we polled.
3. $\text{Deft}(p_i - p_j) = [\text{deft}(p_i) + \text{deft}(p_j)]/2$ is what we actually found to be a good approximation for all of our data, from different populations and designs. This conjecture seems reasonable, because design effects due to clustering for individual p_i can apply similarly to the variable created from the difference $(p_i - p_j)$ of two of them.
4. Inconsistent results would have been possible, but annoying by preventing inference.

4. EMPIRICAL RESULTS FOR Deft$(P_i - P_j)$

Without strong theoretical or mathematical basis for favoring any of the four alternative conjectures, empirical results about $\text{deft}(p_i - p_j)$ become essential, linking these to the computed values for $\text{deft}(p_i)$. These resemble our more familiar conjectures about $\text{deft}(p_i) = \sqrt{1 + roh[\bar{b} - 1]}$; their value depends on several factors that affect *roh*, the coefficient of intraclass correlation, in addition to the average cluster size $\bar{b}$ (Kish 1965, 5.4, 8.2). The values of $\text{deft}(p_i)$ vary greatly between surveys, also between variables for the same survey (Kish, Groves and Krotki 1976; Verma, Scott and O'Muircheartaigh 1980; Verma and Lê 1995). However, survey statisticians gain knowledge from empirical investigations of sampling errors from diverse surveys, which also permit relating the deft values of complex statistics to the simpler $\text{deft}(p_i)$ (Kish L. 1995; Rao and Wu 1985; Rao and Scott 1987). Similarly, to learn about the relation of $\text{deft}(p_i - p_j)$ to $\text{deft}(p_i)$ we have here empirical results from many variables and from many surveys.

In this first essay into this field we present data from fourteen surveys, which represent a great variety of situations. Eleven surveys presented as 5 sets of results (Figures 1 and 2 and Tables 1-3) deal with paired differences of categories from single surveys (Type A). Three sets of results (Tables 1-3) come from social surveys, followed by two sets (Figures 1 and 2) from the Demographic and Health Surveys on population data. Finally three other sets, each dealing with two waves of data, each based on two reinterviews with the same respondents (Tables 4, 5 and 6), represent type B designs of comparisons.

Tables:

1. The National Election Study of 1986 of the Institute for Social Research of the University of Michigan, $n = 2{,}135$.
2. The National Education Longitudinal Study (NELS) of 1988, the National Opinion Research Center of the University of Chicago, $n = 24{,}355$.
3. The National Longitudinal Study of Labor Market Experience of Youth, conducted by the National Opinion Research Center of the University of Chicago, $n = 5{,}857$.
4. National Election Studies Panels 1990 and 1992, Survey Research Center, Institute for Social Research, Ann Arbor, MI 48106.
5. Panel Study of Income Dynamics 1983 and 1987, Survey Research Center.
6. Americans' Changing Lives 1986 and 1989, Survey Research Center.

Figures:

1. Demographic and Health Surveys of Morocco, Niger, and Colombia, MACRO International.
2. Population Census of Indonesia, Rural Java strata (unpublished data).

We note the following important, useful, EMPIRICAL results.

1) First and foremost: The design effects $\text{deft}(p_i - p_j)$ for the differences are usually NO LESS than the $\text{deft}(p_i)$ for the proportions themselves, and $\text{deft}(p_i - p_j) \simeq 0.5\,[\text{deft}(p_i) + \text{deft}(p_j)]$ approximately in all cases. They vary together, along with the considerable variation for deft values between variables, and also with the lesser variation between pairs of categories for the same variables. Researchers who neglect deft commit the usual under-statement of sampling errors for statistics from clustered surveys. This observation is not only interesting but also a useful model for inference, because the other three sources of variation - across variables, categories within variables, and sampling errors of individual statistics - are all greater.

2) We can find these results in all the 14 sets of survey data in the tables and graphs, and we can illustrate them now with Table 1. Note that defts vary from essentially 1.00 for variable D (problems in country) to as high as 2.32 in variable A (religion) which implies $\text{deft}^2 = 2.32^2 = 5.38$. That our empirical rule (1) holds over the range is reassuring. Such variation between variables in the same sample are common and should force us to abandon the practice of using a common average for all defts of a sample (Verma and Lê 1995; Kish 1995).

   Furthermore, we emphasize here the great variation in deft values for the five categories of the same variable from 1.21 to 2.32 (No. 3 for "fundamental" protestants). It follows that $\text{deft}(p_i - p_j)$ is large only when i or j is category 3 for this variable. These variations among the defts for categories of the same variable mean that they should be computed for all categories rather than for only a single "representative" category. These large possible variations between categories of the same variable are an important new finding in our results, that seems to have escaped notice before.

3) There are also sampling errors in the computed values of the defts. Only statisticians who have computed many sampling errors and design effects seem to get the "feel" for how great these can be. They may be mostly responsible for the few cases where $\text{deft}(p_i - p_j)$ fails to fall between $\text{deft}(p_i)$ and $\text{deft}(p_j)$ and either $\text{deft}(p_i) < \text{deft}(p_i - p_j) > \text{deft}(p_j)$ or $\text{deft}(p_i) > \text{deft}(p_i - p_j) < \text{deft}(p_j)$. Incidentally, these cases also show that our results are not mathematical consequences, but empirically based.

The empirical results presented in Figures 1 and 2 further confirm the findings already presented in Tables 1, 2, and 3. Here also we see that: 1) $\text{deft}(p_i - p_j) \cong [\text{deft}(p_i) + \text{deft}(p_j)]/2$ approximately, along the 45° line; that 2) those equalities hold along a wide range of designs effects; and that 3) the variation between variables is large indeed. This large variation is particularly evident for rural Indonesia, with deft values over 4, hence deft^2 values over 16. These large clustering effects are due to the large cluster sizes: with $\bar{b} = 133$ and 137, the values of $roh = 0.12$ are enough for large defts. Note that these empirical results come both from very diverse populations and diverse variables; different from each other and from the data of Tables 1, 2, and 3. Figure 1 has data from 3 countries (Morocco, Niger and Colombia) hence 6 populations, because the urban and rural defts are quite different. Figure 2 shows results for males and females who are quite distinct populations for the occupational variables, though less so for the educational classes.

The empirical data in the tables of studies 4, 5, and 6 were awaited with doubt and anxiety. True that the preceding five sets resulted in similar conclusions, although they dealt with eleven different populations and scores and variables. But studies 1 to 5 all dealt with pairs of categories from polytomies, designs 1 and 2 of Type A. But now we

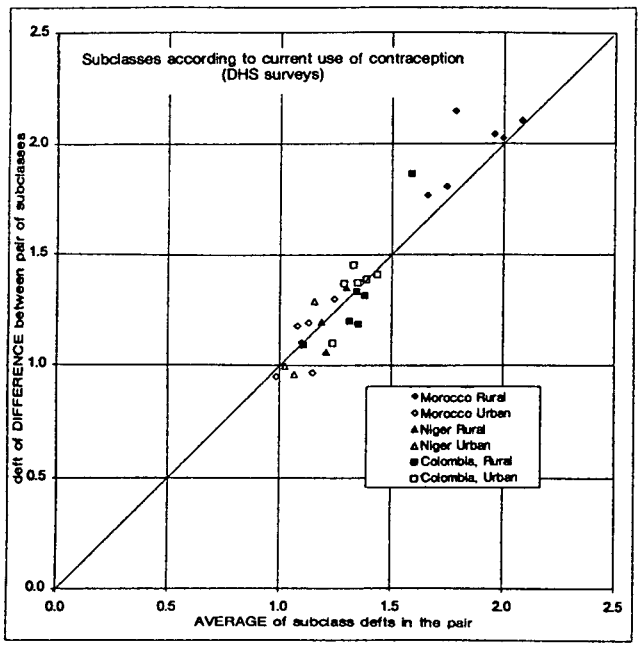


Figure 1. Comparison of deft $(p_i - p_j)$ to the average of deft(p_i),deft(p_j) for categories by current use of contraception*. Illustration of six populations from Demographic and Health Surveys.

* 1 = not using any method of contraception
2 = using only traditional method
3 = using a modern 'reversible' method
4 = sterilised.

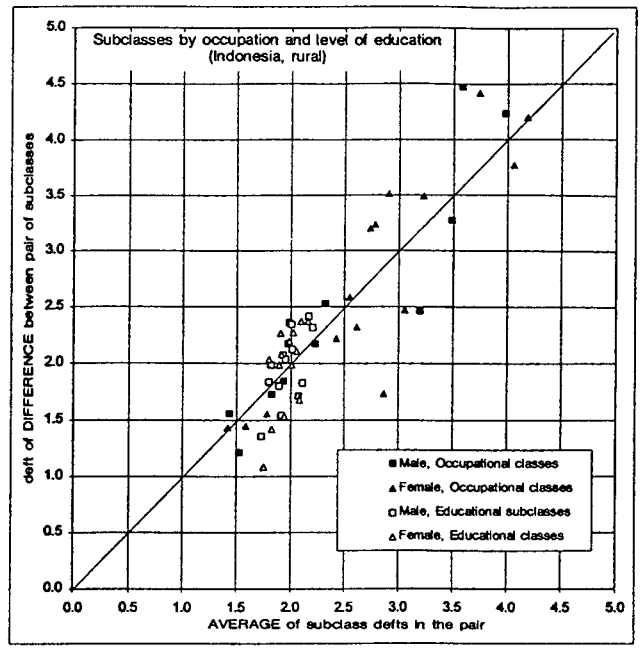


Figure 2. Comparison of deft $(p_i - p_j)$ to the average of deft(p_i),deft(p_j) for categories by occupation and level of education by sex. Illustration from a population census.

Table 1

The National Election Study of 1986 of the I.S.R. of the University of Michigan (n = 2,135)

Categories	Defts for				Categories	Defts for			
$i - j$	P_i	P_j	Average	$(P_i - P_j)$	$i - j$	P_i	P_j	Average	$(P_i - P_j)$
	A. Religion					**B. Abortions Beliefs**			
1-2	1.21	1.42	1.32	1.10	1-2	1.27	.97	1.12	.97
1-3	1.21	2.32	1.77	2.02	1-3	1.27	1.28	1.28	1.32
1-4	1.21	1.50	1.36	1.18	1-4	1.27	1.31	1.29	1.36
1-5	1.21	1.18	1.19	1.17	2-3	.97	1.28	1.12	1.08
2-3	1.42	2.32	1.87	1.93	2-4	.97	1.31	1.14	1.16
2-4	1.42	1.50	1.46	1.57	3-4	1.28	1.31	1.30	1.32
2-5	1.42	1.18	1.30	1.27	*Mean*	*1.17*	*1.24*	*1.21*	*1.20*
3-4	2.32	1.50	1.91	2.03					
3-5	2.32	1.18	1.75	2.04		**D. Problems in Country**			
4-5	1.50	1.18	1.34	1.19	1-2	1.07	.94	1.00	.98
Mean	*1.56*	*1.53*	*1.54*	*1.55*	1-3	1.07	1.04	1.05	1.09
					1-4	1.07	.93	1.00	1.12
	C. Support Reagan				2-3	.94	1.04	.99	1.01
1-2	1.32	1.10	1.21	1.07	2-4	.94	.93	.93	.85
1-3	1.32	.86	1.09	1.26	3-4	1.04	.93	.98	.82
1-4	1.32	1.48	1.40	1.50	*Mean*	*1.02*	*.97*	*.99*	*.98*
2-3	1.10	.86	.98	.96					
2-4	1.10	1.48	1.29	1.38					
3-4	.86	1.48	1.17	1.09					
Mean	*1.17*	*1.21*	*1.19*	*1.21*					
Overall mean	*1.23*	*1.24*	*1.23*	*1.24*					

Table 2

The National Education Longitudinal Study (NELS) of 1988, the National Opinion Research Center of the University of Chicago, (n = 24,355)

Categories	Defts for				Categories	Defts for			
$i - j$	P_i	P_j	Average	$(P_i - P_j)$	$i - j$	P_i	P_j	Average	$(P_i - P_j)$
A. Education Status					**B. Classes are boring**				
1-2	1.38	1.22	1.30	1.11	1-2	.99	1.11	1.05	1.04
1-3	1.38	1.14	1.26	1.16	1-3	.99	1.12	1.06	1.07
1-4	1.38	1.19	1.29	1.30	2-3	1.11	1.12	1.12	1.13
1-5	1.38	1.42	1.40	1.54	*Mean*	*1.03*	*1.12*	*1.08*	*1.08*
2-3	1.22	1.14	1.18	1.11	**C. Freedom to pursue interest**				
2-4	1.22	1.19	1.21	1.24	1-2	1.28	1.10	1.19	1.21
2-5	1.22	1.42	1.32	1.45	1-3	1.28	1.08	1.18	1.28
3-4	1.14	1.19	1.17	1.18	2-3	1.10	1.08	1.09	.97
3-5	1.14	1.42	1.28	1.37	*Mean*	*1.22*	*1.09*	*1.15*	*1.15*
4-5	1.19	1.42	1.31	1.20	**D. School offers good jobs**				
Mean	*1.27*	*1.28*	*1.27*	*1.25*	1-2	1.24	1.07	1.16	1.17
E. Religion					1-3	1.24	1.11	1.18	1.24
1-2	2.48	2.83	2.65	2.74	2-3	1.07	1.11	1.09	1.01
1-3	2.48	2.02	2.25	2.09	*Mean*	*1.18*	*1.10*	*1.14*	*1.14*
2-3	2.83	2.02	2.42	2.59	**F. Dad education**				
Mean	*2.60*	*2.29*	*2.44*	*2.47*	1-2	1.61	1.76	1.69	1.83
I. Feel good about self					1-3	1.61	1.68	1.65	1.65
1-2	1.42	1.28	1.35	1.37	2-3	1.76	1.68	1.72	2.48
					Mean	*1.65*	*1.71*	*1.69*	*1.99*
Overall mean						*1.48*	*1.41*	*1.45*	*1.49*

Table 3

The National Longitudinal of Labor Market Experience of Youth, Conducted by the National Opinion Research Center of the University of Chicago, (n = 5,857)

Categories	Defts for				Categories	Defts for			
$i - j$	P_i	P_j	Average	$(P_i - P_j)$	$i - j$	P_i	P_j	Average	$(P_i - P_j)$
A. Chance is important in my life					**B. Something stops me**				
1-2	1.26	1.20	1.23	1.06	1-2	1.07	1.22	1.14	1.04
1-3	1.26	1.18	1.22	1.30	1-3	1.07	1.12	1.10	1.14
1-4	1.26	1.16	1.21	1.28	1-4	1.07	1.09	1.08	1.09
2-3	1.20	1.18	1.19	1.22	2-3	1.22	1.12	1.17	1.28
2-4	1.20	1.16	1.18	1.25	2-4	1.22	1.09	1.16	1.14
3-4	1.18	1.16	1.17	1.05	3-4	1.12	1.09	1.11	1.07
Mean	*1.23*	*1.17*	*1.20*	*1.19*	*Mean*	*1.13*	*1.12*	*1.13*	*1.13*
C. Have control of my life					**D. I am as worthy as others**				
1-2	1.13	1.06	1.10	1.09	1-2	1.17	1.13	1.15	1.16
1-3	1.13	1.10	1.12	1.13	1-3	1.17	1.07	1.12	1.16
2-3	1.06	1.10	1.08	1.07	2-3	1.13	1.07	1.10	1.08
Mean	*1.11*	*1.09*	*1.10*	*1.10*	*Mean*	*1.16*	*1.09*	*1.12*	*1.13*
E. Plans hardly work out					**F. I am satisfied**				
1-2	1.19	1.07	1.13	1.12	1-2	1.19	1.12	1.16	1.16
1-3	1.19	1.13	1.16	1.20	1-3	1.19	1.13	1.16	1.20
2-3	1.07	1.13	1.10	1.08	2-3	1.12	1.13	1.13	1.09
Mean	*1.15*	*1.11*	*1.13*	*1.13*	*Mean*	*1.17*	*1.13*	*1.15*	*1.15*
I. Mother's work									
1-2	1.49	1.36	1.43	1.41					
1-3	1.49	1.52	1.51	1.53					
2-3	1.36	1.52	1.44	1.44					
Mean	*1.45*	*1.47*	*1.46*	*1.47*					
Overall mean	*1.20*	*1.17*	*1.18*	*1.19*					

Table 4

National Election Studies Panels 1990 and 1992, Survey Research Center, Institute for Social Research, Ann Arbor

Categories before/after (90/92)	Defts for P_i	P_j	Average	$(P_i - P_j)$
Strongly approve Bush	1.14	.93	1.04	1.02
Approve Bush foreign policy	.92	1.05	.99	1.00
Strongly disapprove Bush foreign policy	1.23	1.24	1.24	1.32
Approve Bush economy	.97	.94	.96	.96
Strongly approve Bush economy	1.14	1.04	1.09	1.10
Approve Bush	1.00	1.00	1.00	1.00
Strongly disapprove Bush	1.16	1.10	1.13	1.12
Watch campaign on TV	.89	1.55	1.22	1.40
Mean	*1.06*	*1.11*	*1.08*	*1.11*

Table 5

Panel Study of Income Dynamics, 1983 and 1987, Survey Research Center, Ann Arbor

Categories* before/after (83/87)	Defts for P_i	P_j	Average	$(P_i - P_j)$
Live in South	1.22	1.23	1.23	1.11
Age of head of family	1.28	1.33	1.31	1.37
Family size	1.29	1.43	1.36	1.47
Number of children in family	1.23	1.43	1.33	1.49
Work hours of head	1.12	.84	.98	1.03
Age of youngest child	.93	.91	.92	.87
Mean	*1.18*	*1.20*	*1.19*	*1.22*

* All variables are categorized in two categories.

sought data for Type B comparisons from panel surveys, so that we could investigate the conjectures for the test/retest and before/after experimental designs. Mathematically these can be easily shown to resemble polytomies (*i.e.*, tetratomies), but from that to the empirical values of design effects leads through a "black box." Hence these empirical values are so much more valuable and remarkable. Here we found considerable design effects for Chi square tests for analytical comparisons.

5. PRESENT FINDINGS IN THE CONTEXT OF RELATED RESEARCH

A great deal of empirical information is available from previous work by the authors and by others on design effects for the total sample, for subclasses, and for differences, for diverse variables and designs. It would be useful to put the present findings in the context of that work.

It has been found that nature of the survey variables being estimated is a major (often the main) determinant of the magnitude of the design effects: vastly differing defts can occur for different types of variables even with the same samples or with similar designs. For this reason we have always recommended that defts be computed for many different variables, while it is generally less important to compute them for many different subclasses, especially for different categories of subclasses defined in terms of the same characteristic.

The present findings illustrate that defts can differ greatly also among different categories of the same survey variable, estimated with the total sample as the common base. Therefore each individual category and each difference between pairs of categories, even when defined in

Table 6

Americans' Changing Lives, 1986 and 1989, Survey Research Center, Ann Arbor

Categories before/after	Defts for P_i	P_j	Average	$(P_i - P_j)$
A. Get together with friends				
Once a week	1.30	1.26	1.28	1.28
2-3 a month	.88	1.00	.94	1.02
Mean	*1.09*	*1.13*	*1.11*	*1.15*
C. How Satisfy Are You				
Very satisfy	1.28	1.21	1.25	1.33
Not satisfy	1.04	1.16	1.10	1.00
Mean	*1.16*	*1.19*	*1.18*	*1.17*
E. How often work in garden				
Often	1.40	1.16	1.28	1.19
Rarely	.91	1.11	1.01	1.18
Never	1.66	1.17	1.42	1.26
Mean	*1.32*	*1.15*	*1.24*	*1.21*
Overall mean	*1.25*	*1.26*	*1.26*	*1.18*

Categories before/after	Defts for P_i	P_j	Average	$(P_i - P_j)$
B. How often do you exercise				
Often	1.51	1.67	1.59	1.26
Never	1.62	1.97	1.80	1.41
Mean	*1.56*	*1.82*	*1.70*	*1.34*
D. How do you like your home				
Very much	1.24	.90	1.07	.91
Not much	1.33	.98	1.16	1.12
Mean	*1.29*	*.94*	*1.12*	*1.02*
F. I have a positive attitude				
Agree	1.10	1.33	1.22	1.19
Disagree	1.05	1.28	1.17	1.21
Mean	*1.08*	*1.31*	*1.20*	*1.20*

terms of the same survey variable, needs to be regarded, in a sense, as a separate variable in its own right for the purpose of computing and analyzing design effects.

As to the relationship between defts for subclasses and subclass differences, previous research has mostly dealt with the following situation. With the total sample n partitioned into subclasses i of size $n_i = p_i.n$, deft (r_i) values for statistics r_i (such as a proportion m_i/n_i, mean $\sum y_i/n_i$, ratio $\sum y_i/\sum x_i$), estimated over subclass elements n_i as the base, are related to deft (r) for the same variable estimated with the total sample as the base. Similarly, deft $(r_i - r_j)$ for subclass differences are related to deft (r_i),deft (r_j) based on individual subclasses and to deft (r) based on the total sample. Numerous computations confirm these relationships to be in accord with our conjecture (2) of section 3:

$$\text{deft}(r) > \text{deft}(r_i); \quad \text{and} \quad \text{deft}(r_i) > \text{deft}(r_i - r_j) > 1.$$

These effects of covariances on design effects of clustered samples are essentially empirical (even sociological in a broad sense); and they must be so verified.

Similarly with our newly discovered relationship for $(p_i - p_j)$ for two categories, which are so different from the above. The relations $\text{deft}(p_i - p_j) \cong [\text{deft}(p_i) + \text{deft}(p_j)]/2$ are also empirical and approximate and they must be verified over and over again. But they seem to be widely applicable in our data, and clearly better than the other assumptions, such as $\text{deft}(p_i - p_j) = 1$ that have been often assumed until now.

ACKNOWLEDGEMENTS

The authors wish to thank the editor and referees whose suggestions made this paper both shorter and better.

REFERENCES

COCHRAN, W.G. (1950). The comparison of percentages in matched samples. *Biometrika*, 37, 256-66.

DEMING, W.E. (1953). On the distinction between enumerative and analytic studies. *Journal of the American Statistical Association*, 48, 244-45.

KISH, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.

KISH, L. (1987). *Statistical Research Design*. New York: John Wiley and Sons.

KISH, L. (1995). Methods for design effects. *Journal of Official Statistics*, 11, 55-77.

KISH, L., and FRANKEL, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society*, (B), 36, 1-37.

KISH, L., GROVES, R.M., and KROTKI, K. (1976). *Sampling Errors for Fertility Surveys*. Occasional Paper No. 17, *World Fertility Surveys*. International Statistical Institute: The Hague.

LÊ, T., and VERMA, V. (1995). *Sample Designs and Sampling Errors for the DHS*. Calverton MD: MACRO International.

McNEMAR, Q. (1949). *Psychological Statistics*. New York: John Wiley and Sons.

MOSTELLER, F. (1952). Some statistical problems in measuring the subjective responses to drugs. *Biometrika*, 8, 220-226.

RAO, J.N.K., and SCOTT, A.J. (1987). On simple adjustments to Chi-square tests with sample survey data. *Annals of Statistics*, 15, 385-397.

RAO, J.N.K., and WU, C.F.S. (1985). Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.

SCOTT, A.J., and HOLT, D. (1982) The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77, 848-54.

SCOTT, A.J., and SEBER, G.A.F. (1983). Difference of proportions from the same survey. *The American Statistician*, 37, 319-20.

SKINNER, C.J., HOLT, D., and SMITH, T.M.F. (1989). *Analysis of Complex Surveys*. New York: John Wiley and Sons.

VERMA, V., and LÊ, T. (1995). Sampling errors for the DHS survey. 50th Session of the International Statistical Institute, Beijing.

VERMA, V., SCOTT, C., and O'MUIRCHEARTAIGH, C. (1980). Sample designs and sampling errors for the World Fertility Surveys. *Journal of the Royal Statistical Society (A)*, 143, 431-473.

WILD, C.J., and SEBER, G.A.F. (1993). Comparing two proportions for the same survey. *The American Statistician*, 47, 178-181.

Survey Methodology, December 1995
Vol. 21, No. 2, pp. 125–135
Statistics Canada

Alternative Adjustments Where There Are Several Levels of Auxiliary Information

F. DUPONT[1]

ABSTRACT

Regression estimation and its generalization, calibration estimation, introduced by Deville and Särndal in 1993, serves to reduce *a posteriori* the variance of the estimators through the use of auxiliary information. In sample surveys, there is often useable supplementary information that is distributed according to a complex schema, especially where the sampling is realized in several phases. An adaptation of regression estimation was proposed along with its variants in the framework of two-phase sampling by Särndal and Swensson in 1987. This article seeks to examine alternative estimation strategies according to two alternative configurations for auxiliary information. It will do so by linking the two possible approaches to the problem: use of a regression model and calibration estimation.

KEY WORDS: Auxiliary information; Regression estimator; Calibration estimator; Two-phase sampling.

1. INTRODUCTION

Using the regression estimator studied by Fuller (1975), Cassel, Särndal and Wretman (1976), Särndal (1980), Gourieroux (1981), Isaki and Fuller (1982), and Wright (1983), it is possible to improve *a posteriori* – that is, after the sampling has been completed – the estimate of a total of a variable of interest on the basis of auxiliary variables $x_1, \ldots, x_k$ for which additional information is available. The variance in relation to the Horwitz-Thompson estimator is reduced by using the regression estimator, provided that one knows the true value of the target population totals of the auxiliary variables, which will constitute the additional information referred to as auxiliary information. Deville and Särndal in 1992 proposed a class of estimators derived from a reweighting approach that addresses the same issue of variance reduction: calibration estimators. By calibrating sampling weights it is possible to incorporate *a posteriori* auxiliary information of the type totals $X_1, \ldots, X_k$ of k variables $x_1, \ldots, x_k$ into the estimate made on the basis of the new weightings and thus to improve the estimate. This approach generalizes regression estimation, which is one of the elements of the class.

However, in surveys based on sampling, there is often usable additional information that is distributed according to a more complex schema than what has been described above, especially when the sampling is carried out in several phases. This article looks at different strategies for using this complex auxiliary information in the framework of two-phase sampling, with the possibility of generalizing to more than two phases.

When the sampling plan entails two phases, the auxiliary information consists of information known for the entire population, but also of information known for the sample resulting from the first sampling phase. These two bodies of information may concern different variables.

In their 1987 article, Särndal and Swensson propose an estimator that uses all the auxiliary information available for a two-phase sampling, with different auxiliary information for the total population and the population obtained from the first-phase sampling. This is an estimator adapting the principle of the regression estimator when the information known for the individuals obtained from the first-phase sampling is considered to be substitutable for the aggregated information and to be of better quality than the information available for the target population as a whole, for purposes of estimating the variable of interest. However, in practice it often happens that these two bodies of information are complementary rather than substitutable. We have thus sought in this study to develop the regression estimate in a context in which the bodies of auxiliary information are complementary.

Furthermore, insofar as calibration estimation generalizes regression estimation when the auxiliary information is at only one level, we have sought to adapt calibration estimation to this context. We review the various calibration strategies in order to propose the most suitable ones, seeking to relate them to generalizations of regression estimation that are possible in this context.

We show (Section 2) that the joint use of two different bodies of auxiliary information leads to two regression models and three associated decompositions of the variable of interest. The regression model assisted approach (RMAA) thus enables us to derive 3 alternative estimators.

In turn, the calibration approach (CA) (Section 3) enables us to derive 4 estimators. Each of these estimators may be related to (associated with) the three estimators derived from the regression model approach.

[1] F. Dupont, Unité Méthodes Statistiques, Institut National de la Statistique et des Études Économiques, (INSEE), 18 Blvd. Adolphe Pinard, 75675 Paris Cedex.

Thus (Section 4), the two approaches may be linked together and result in three classes of estimators, each associated with a decomposition of the variable of interest. The estimators of a given class have the same asymptotic variance.

When strategies are evaluated on the basis of the sampling plan alone, our choice is directed toward the third class of estimators, which is superior to the other two from the standpoint of variance.

When strategies are evaluated on the basis of a modelling of the variable of interest, the preferable class of estimators is the one associated with the modelling adopted.

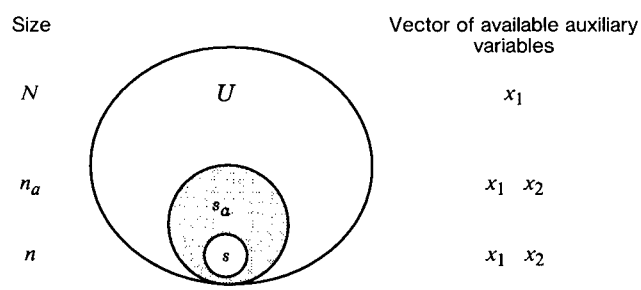
In a situation in which we wish to adjust a survey, and in which we wish simultaneously to correct the biases that would result from the use of gross weightings and to reduce the variance, the findings must be adapted: the changes introduced in the weightings to correct the biases are greater than the corrections for variance reduction. Hence the variables will be incorporated into the calibration once it appears that they are affecting the probability of selection and thus participating in the creation of the bias.

When the auxiliary variables are qualitative, the choice between *a priori* and *a posteriori* use of the auxiliary information - that is, between its use at the sampling stage and at the adjustment stage - still rests on the distinction between the two modellings of the variable of interest.

These findings may be extended to samplings of more than two phases.

2. NOTATIONS

The framework is that of a two-phase sampling. Assume that auxiliary information is available at two different levels: the target population and the population obtained from the first-phase sampling. The situation may be diagrammed as follows:



where U represents the target population for which the values of the vector of variable x_1 are known or, failing that, the total $X_1 = \sum_{i \in U} x_{i1}$. s_a represents an intermediate level of sampling for which the values of the vectors of k_1 variables x_1 and k_2 variables x_2 are known for all individuals. We denote as π_{ia} the probability of selection from the sample associated with the first phase of the sampling. s represents the final sample for which are available the values of the variable y, the total of which we are trying to estimate, as well as the values of the vectors of the auxiliary variables x_1 and x_2. This is denoted as $\pi_i = P(i \mid s_a)$.

We hope to make optimum use of all this auxiliary information in order to improve the estimates that will be made on the basis of the data gathered from the sample that results from the second sampling phase s.

An obvious first idea is to try to generalize the regression estimator in this context.

3. REGRESSION ESTIMATION APPROACH

3.1 The Information Contained in x_1 is Considered to be Substitutable for the Information Contained in x_2 for Estimating y and to be of Lesser Quality

In their work, Särndal, Swensson and Wretman propose the following regression estimator for estimating the total of y:

$$\hat{Y}_1 = \sum_{i \in s} \frac{y_i}{\pi_i \pi_{ai}} + \left(\sum_{i \in s_a} \frac{x'_{i2} \hat{b}_2}{\pi_i} - \sum_{i \in s} \frac{x'_{i2} \hat{b}_2}{\pi_i \pi_{ai}} \right) + \left(\sum_{i \in U} x'_{i1} \hat{b}_1 - \sum_{i \in s_a} \frac{x'_{i1} \hat{b}_1}{\pi_{ai}} \right)$$

where the second term is the correction for poor estimation on s_a and the third is the correction for poor estimation on s.

The estimation can also be written:

$$\hat{Y}_1 = \sum_{i \in U} x'_{i1} \hat{b}_1 + \sum_{i \in s_a} \frac{(x'_{i1} \hat{b}_1 - x'_{i2} \hat{b}_2)}{\pi_{ai}} + \sum_{i \in s} \frac{(y_i - x'_{i2} \hat{b}_2)}{\pi_i \pi_{ai}}$$

where the second term is the correction for poor approximation of y_i by $x'_{i1} \hat{b}_1$ and the third is the correction for poor approximation of y_i by $x'_{i2} \hat{b}_2$;

with $$\hat{b}_1 = \left(\sum_{i \in s_a} \frac{x_{i1} x'_{i1}}{\pi_{ai}} \right)^{-1} \left(\sum_{i \in s} \frac{x_{i1} y_i}{\pi_i \pi_{ai}} \right)$$

and $$\hat{b}_2 = \left(\sum_{i \in s} \frac{x_{i2} x'_{i2}}{\pi_i \pi_{ai}} \right)^{-1} \left(\sum_{i \in s} \frac{x_{i2} y_i}{\pi_i \pi_{ai}} \right).$$

The underlying idea is that we have two concurrent models for y, namely:

(1) $y_i = x_{i1}'b_1 + u_{i1}$ with $E(u_{i1}) = 0$ and $V(u_{i1}) = \sigma_1^2$

and

(2) $y_i = x_{i2}'b_2 + u_{i2}$ with $E(u_{i2}) = 0$ and $V(u_{i2}) = \sigma_2^2$

the second of which we believe is *a priori* better for predicting the value of y_i. Thus in this model-based perspective, x_1 functions as a proxy of x_2. A situation of this type corresponds, for example, to a case in which x_2 represents the update - that is, the update to the date of the survey - of the variable retrieved from the x_1 sampling frame. In other words, if x_2 were available at the level of the entire population, the estimator used would be

$$\sum_{i\in U} x_{i2}'\hat{b}_1 + \sum_{i\in s} \frac{(y_i - x_{i2}'\hat{b}_2)}{\pi_i \pi_{ai}}.$$

Let us now imagine the case of a two-phase sampling survey of households. Assume that the sampling frame is made up of dwellings for which we have information consisting of dwelling size, denoted as x_1, which is therefore known for all individuals in the target population. If all the individuals obtained from the first sampling phase are questioned on the composition of the household, denoted as x_2, in particular on the number of children in the household, the two bodies of information appear to be complementary rather than substitutable for purposes of studying the household budget. This is further reinforced if instead of household composition, the information collected is the age or occupation of the head of household.

In a model-based perspective, the alternative situation, in which the information contained in x_1 is considered complementary to that contained in x_2 for estimating y, thus naturally suggests itself.

3.2 The Information Contained in x_1 is Considered to be Complementary to the Information Contained in x_2 for Estimating y

3.2.1 Decomposition $y_i = x_{i1}'a_1 + x_{i2}'a_2 + u_i$

The underlying model is then:

$y_i = x_{i1}'a_1 + x_{i2}'a_2 + u_i$ with $E(u_i) = 0$ and $V(u_i) = \sigma_1^2$.

The estimator to be used is then:

$$\hat{Y}_2 = \sum_{i\in U} x_{i1}'\hat{a}_1 + \sum_{i\in s_a} \frac{x_{i2}'\hat{a}_2}{\pi_{ai}} + \sum_{i\in s} \frac{(y_i - x_{i1}'\hat{a}_1 - x_{i2}'\hat{a}_2)}{\pi_i \pi_{ai}}$$

with:

$$\begin{pmatrix}\hat{a}_1\\ \hat{a}_2\end{pmatrix} = \begin{pmatrix}\sum_{i\in s} \frac{x_{i1}x_{i1}'}{\pi_{ai}\pi_i} & \sum_{i\in s} \frac{x_{i1}x_{i2}'}{\pi_{ai}\pi_i}\\ \sum_{i\in s} \frac{x_{i2}x_{i1}'}{\pi_{ai}\pi_i} & \sum_{i\in s} \frac{x_{i2}x_{i2}'}{\pi_{ai}\pi_i}\end{pmatrix}^{-1} \begin{pmatrix}\sum_{i\in s} \frac{x_{i1}y_i}{\pi_{ai}\pi_i}\\ \sum_{i\in s} \frac{x_{i2}y_i}{\pi_{ai}\pi_i}\end{pmatrix}.$$

The variable here is broken down into three components $y_i = x_{i1}'\hat{a}_1 + x_{i2}'\hat{a}_2 + \hat{u}_i$. The total of y is thus broken down into three components, each of which is estimated at the highest level, that is, with the greatest precision possible:

- U for $x_{i1}'\hat{a}_1$,
- s_a for $x_{i2}'\hat{a}_2$, and
- s for $\hat{u}_i$.

3.2.2 Decomposition $y_i = x_{i1}'c_1 + M_{x_1}(x_{i2})'c_2 + u_i$

If we wish to make maximum use of the information contained in x_1 available on U, it is natural to introduce another formulation of the same model $y_i = x_{i1}'a_1 + x_{i2}'a_2 + u_i$ which isolates everything which in y can be taken into account through x_1. It is written as follows:

$$y_i = x_{i1}'c_1 + M_{x_1}(x_{i2})'c_2 + u_i \quad \text{with}$$

$$E(u_i) = 0 \quad \text{and} \quad V(u_i) = \sigma_1^2,$$

where M_{x_1} represents the orthogonal projection, in the metric associated with the weights $1/\pi_{ai}$, on the orthogonal of the vector space generated in s_a (similar to $\Re^n$) by the group of variables x_1.

$M_{x_1}(x_{i2})$ is defined by:

$$M_{x_1}(x_{i2}) = x_{i2}' - \left(\sum_{i\in s_a} \frac{x_{i2}x_{i1}'}{\pi_{ai}}\right)\left(\sum_{i\in s_a} \frac{x_{i1}x_{i1}'}{\pi_{ai}}\right)^{-1} x_{i1}.$$

The associated natural estimator is then:

$$\hat{Y}_3 = \sum_{i\in U} x_{i1}'\hat{c}_1 + \sum_{i\in s_a} \frac{(M_{x_1}x_{i2}'\hat{c}_2)}{\pi_{ai}} + \sum_{i\in s} \frac{(y_i - x_{i1}'\hat{c}_1 - M_{x_1}x_{i2}'\hat{c}_2)}{\pi_i \pi_{ai}},$$

where $\hat{c} = \begin{pmatrix}\hat{c}_1\\ \hat{c}_2\end{pmatrix}$ is the regression coefficient $y = x'c_1 + (M_{x_1}x_2)'c_2 + u$ estimated over s with weights $1/\pi_{ai}\pi_i$ (which differs slightly from $\begin{pmatrix}\hat{b}_1\\ \hat{a}_2\end{pmatrix}$).

3.3 The Three Estimators Derived from the Model-based Approach

The modelling approach has enabled us to construct 3 estimators that can be rewritten synthetically by introducing

new notations. Throughout what follows, for a vector of a given variable z, the following notation will be used:

$$\hat{\hat{Z}} = \sum_{i \in s} \frac{1}{\pi_{ai}\pi_i} z$$

$$\hat{Z} = \sum_{i \in s_a} \frac{1}{\pi_{ai}} z .$$

With these notations, the three estimators are rewritten as follows:

$$\hat{Y}_1 = [X_1' \hat{b}_1] + [\hat{X}_2' \hat{b}_2 - \hat{X}_1' \hat{b}_1] + [\hat{\hat{Y}} - \hat{\hat{X}}_2' \hat{b}_2]$$

associated with the models:

$$(1) \quad y_i = x_{i1}' b_1 + u_{i1}$$

and

$$(2) \quad y_i = x_{i2}' b_2 + u_{i2}$$

$$\hat{Y}_2 = [X_1' \hat{a}_1] + [\hat{X}_2' \hat{a}_2] + [\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{a}_1 - \hat{\hat{X}}_2' \hat{a}_2]$$

associated with the model

$$y_i = x_{i1}' a_1 + x_{i2}' a_2 + u_i$$

$$\hat{Y}_3 = [X_1' \hat{c}_1] + [M_{x_1} \hat{X}_2' \hat{c}_2] + [\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{c}_1 - M_{x_1} \hat{\hat{X}}_2' \hat{c}_2]$$

associated with the model

$$y_i = x_{i1}' c_1 + M_{x_1}(x_{i2})' c_2 + u_i .$$

In the same manner as the regression estimator is generalized by calibration estimators, the problem of the use of auxiliary information at several levels may be dealt with through calibration theory, be attempting to construct calibration strategies adapted to the auxiliary information configuration examined in this article.

4. CALIBRATION APPROACH

4.1 Different Strategies Possible

When we try to generalize the calibration estimate proposed in a context in which auxiliary information is present at a single level - that of the entire population - several strategies naturally suggest themselves:

Strategy 1

a) calibrate the structure of the 1st-phase sample s_a on that of the total population U in terms of variable x_1, then,

b) calibrate the structure of the 2nd-phase sample s on that of the 1st phase sample s_a in terms of variable x_2.

Note: For the latter operation, it is better to take account of the preceding calibration in terms of x_1 in order to determine the reference value in the calibration in terms of x_2 on s_a. If the preceding calibration is not taken into account, only the estimates made at the level of s_a will benefit from the improvement made by stage a. A good way to convince oneself of this is to consider the specific extreme case where $x_1 = x_2$.

This strategy corresponds to the following calibration equations:

Stage a:

$$\sum_{i \in s_a} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}} x_{i1} = \sum_{i \in U} x_{i1} = X_1$$

which determines β_1, then

Stage b:

$$\sum_{i \in s} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}\pi_i} F(x_{i2}'\beta_2) x_{i2} = \sum_{i \in s_a} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}} x_{i2} = \hat{X}_2^*$$

which determines β_2,

where F designates, as throughout this article, the function which is used in the calibration and which may be linear, exponential, truncated linear or logit (see Deville, Särndal 1993).

Strategy 2

Calibrate the structure of the 2nd-phase sample s simultaneously in terms of variables x_1 and x_2, that is,

- on the structure of the total population U as regards x_1
- on the structure of s_a for x_2.

This second strategy leads us to the following calibration equations:

$$\sum_{i \in s} \frac{F(x_{i1}\alpha_1 + x_{i2}\alpha_2)}{\pi_{ai}\pi_i} x_{i1} = \sum_{i \in U} x_{i1} = X_1, \quad \text{and}$$

$$\sum_{i \in s} \frac{F(x_{i1}'\alpha_1 + x_{i2}'\alpha_2)}{\pi_{ai}\pi_i} x_{i2} = \sum_{i \in s_a} \frac{x_{i2}}{\pi_{ai}} = \hat{X}_2,$$

which determines α_1 and α_2.

The first strategy offers the advantage of correcting the 1st phase weightings, that is, of incorporating the auxiliary information at the highest level. The second strategy, for its part, makes it possible to correct the weightings that will actually be used in the estimation, and in particular to obtain a perfect estimate of the total of x_1.

A third strategy may be proposed; it combines the advantages of the above two strategies and would therefore seem preferable to them:

Strategy 3

a) calibrate the structure of the 1st phase sample s_a on that of the total population U in terms of variable x_1, then

b) calibrate the structure of the 2nd phase sample s simultaneously in terms of variables x_1 and x_2, that is,
   - on the structure of the total population U as regards x_1
   - on the structure of s_a modified by taking account of the preceding calibration for x_2.

This strategy leads to the following calibration equations:

Stage a:

$$\sum_{i\in s_a} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}} x_{i1} = \sum_{i\in U} x_{i1}$$

which determines β_1, then

Stage b:

$$\sum_{i\in s} \frac{F(x_{i1}'\gamma_1 + x_{i2}'\gamma_2)}{\pi_{ai}\pi_i} x_{i1} = \sum_{i\in U} x_{i1} = X_1, \quad \text{and}$$

$$\sum_{i\in s} \frac{F(x_{i1}'\gamma_1 + x_{i2}'\gamma_2)}{\pi_{ai}\pi_i} x_{i2} = \sum_{i\in s_a} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}} x_{i2} = \hat{X}_2^*,$$

which determines γ_1 and γ_2.

Lastly, a fourth strategy may be proposed; it may be seen as a variant of the preceding strategy:

Strategy 4

a) calibrate the structure of the 1st phase sample s_a on that of the total population U in terms of variable x_1, then

b) calibrate the structure of the 2nd phase sample s simultaneously in terms of variables x_1 and x_2, on the basis of the weights modified by the preceding calibration, that is,
   - on the structure of the total population U as regards x_1
   - on the structure of s_a modified taking account of the preceding calibration for x_2.

This strategy leads to the following calibration equations:

Stage a:

$$\sum_{i\in s_a} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}} x_{i1} = \sum_{i\in U} x_{i1},$$

which determines β_1, then

Stage b:

$$\sum_{i\in s} \frac{F(x_{i1}'\beta_1)F(x_{i1}'\delta_1 + x_{i2}'\delta_2)}{\pi_{ai}\pi_i} x_{i1} = \sum_{i\in U} x_{i1} = X_1, \quad \text{and}$$

$$\sum_{i\in s} \frac{F(x_{i1}'\beta_1)F(x_{i1}'\delta_1 + x_{i2}'\delta_2)}{\pi_{ai}\pi_i} x_{i2} = \sum_{i\in s_a} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}} x_{i2} = \hat{X}_2^*,$$

which determines δ_1 and δ_2.

When the calibration function is exponential, it is clear that strategies 3 and 4 coincide.

In this calibration-based approach, the viewpoint adopted is that of reduction of variance based on the characteristics of the sampling plan, without consideration of the model. Two questions then naturally arise:

- Can each of these four strategies be linked to a model-based approach?
- Can these four strategies be compared in terms of variance?

We will first examine the link between the three strategies defined by a calibration approach and the strategies defined by a model-based or regression approach, after which we will focus on calculating the variances of the estimators associated with each of the strategies.

4.2 Link Between the Different Possible Strategies and the Regression Approach

When F is linear, each of the estimators associated with the four strategies may be rewritten simply.

Notations

Throughout the rest of this article we will use the following notations for a vector of any variable z:

$$\hat{\hat{Z}}^* = \sum_{i\in s} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}\pi_i} z_i \qquad \hat{Z}^* = \sum_{i\in s_a} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}} z_i.$$

We will also omit the i indexes in order to lighten the presentation when there is no ambiguity.

Strategy 1

The weightings are of the form

$$w_i^4 = \frac{F(x_{i1}'\beta_1)}{\pi_{ai}\pi_i} F(x_{i2}'\beta_2),$$

the associated estimator $\hat{Y}_4$ may be rewritten by translating the effect of the second calibration on x_2:

$$\hat{Y}_4 = \hat{\hat{Y}}^* + [\hat{X}_2^* - \hat{\hat{X}}_2^*]' \hat{B}_2 \quad \text{with}$$

$$\hat{B}_2 = \left(\sum_s \frac{F(x_1'\beta_1)}{\pi_a \pi} x_2 x_2' \right)^{-1} \left(\sum_s \frac{F(x_1'\beta_1)}{\pi_a \pi} x_2 y \right),$$

then by translating the effect of the first calibration on x_1:

$$\hat{Y}_4 = \hat{\hat{Y}} + [X_1 - \hat{X}_1]' \hat{B}_1 + [\hat{X}_2^* - \hat{\hat{X}}_2^*]' \hat{B}_2,$$

or:

$$\hat{Y}_4 = [X_1' \hat{B}_1] + [\hat{X}_2^{*\prime} \hat{B}_2 - \hat{X}_1' \hat{B}_1] + [\hat{\hat{Y}} - \hat{\hat{X}}_2^{*\prime} \hat{B}_2],$$

$$\text{with} \quad \hat{B}_1 = \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a} \right)^{-1} \left(\sum_s \frac{x_1 y}{\pi_a \pi} \right).$$

Now, $\hat{Y}_1$ is rewritten:

$$\hat{Y}_1 = [X_1' \hat{b}_1] + [\hat{X}_2^{*\prime} \hat{b}_2 - \hat{X}_1' \hat{b}_1] + [\hat{\hat{Y}} - \hat{\hat{X}}_2^{*\prime} \hat{b}_2].$$

We thus obtain an estimator similar to the estimator $\hat{Y}_1$ that is obtained from the model-based approach in cases where the information contained in x_1 is considered to be substitutable for the information contained in x_2 for estimating y and also to be of lesser quality. The differences between $\hat{Y}_1$ and $\hat{Y}_4$ concern the following points:

1. $\hat{B}_2$ is estimated by incorporating the changes from the calibration on x_1, unlike $\hat{b}_2$.
2. The estimate $\hat{B}_1 = \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a} \right)^{-1} \left(\sum_s \frac{x_1 y}{\pi_a \pi} \right)$ of B_1, is made in part on s_a, unlike $\hat{b}_1$.
3. Lastly, we use the adjusted weights $F(x_1\beta_1)/\pi_a\pi$ in the sums in x_2 on s and on S_a in $\hat{Y}_4$ in unlike what was done for $\hat{Y}_1$: the estimation on x_2 is improved by the knowledge of x_1.

Thus the underlying modelling here is indeed: (1) $y_i = x_{i1}' b_1 + u_{i1}$ and (2) $y_i = x_{i2}' b_2 + u_{i2}$, the second of which we think is *a priori* better for predicting the value of y_i.

Strategy 2

We obtain weights

$$w_i^5 = \frac{F(x_{i1}'\alpha_1 + x_{i2}'\alpha_2)}{\pi_{ai}\pi_i},$$

the associated estimator is rewritten as follows:

$$\hat{Y}_5 = [X_1' \hat{a}_1] + [\hat{X}_2' \hat{a}_2] + [\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{a}_1 - \hat{\hat{X}}_2' \hat{a}_2].$$

We thus obtain exactly the estimator $\hat{Y}_2$ proposed in the regression model approach in the case in which the information contained in x_1 is considered complementary to the information contained in x_2 for estimating y. The underlying model here is indeed $y_i = x_{i1} a_1 + x_{i2} a_2 + u_i$.

Strategy 3

We obtain weights

$$w_i^6 = \frac{F(x_{i1}'\gamma_1 + x_{i2}'\gamma_2)}{\pi_{ai}\pi_i},$$

the associated estimator is rewritten as:

$$\hat{Y}_6 = \hat{\hat{Y}} + [X_1 - \hat{\hat{X}}_1]' \hat{a}_1 + [\hat{X}_2^* - \hat{\hat{X}}_2]' \hat{a}_2$$

thus:

$$\hat{Y}_6 = [X_1' \hat{a}_1] + [\hat{X}_2^{*\prime} \hat{a}_2] + [\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{a}_1 - \hat{\hat{X}}_2' \hat{a}_2].$$

Now,

$$\hat{X}_2^* = \sum_{s_a} \frac{x_2}{\pi_a} + \left(\sum_{s_a} \frac{x_2 x_1'}{\pi_a} \right)^{-1} \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a} \right) \left[X - \sum_{s_a} \frac{x_1}{\pi_a} \right].$$

From this it can be deduced by replacing in $\hat{Y}_6$ that:

$$\hat{Y}_6 = [X_1' \hat{C}_1] + [M_{x_1} \hat{X}_2' \hat{a}_2] + [\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{a}_1 - \hat{\hat{X}}_2' \hat{a}_2],$$

with

$$\hat{C}_1 = \hat{a}_1 + \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a} \right)^{-1} \left(\sum_{s_a} \frac{x_1 x_2'}{\pi_a} \right) \hat{a}_2.$$

We thus obtain an estimator that is close to the estimator $\hat{Y}_3$ proposed in the regression model approach in the case in which the information contained in x_1 is considered complementary to the information contained in x_2 for estimating y. The underlying model here is $y_i = x_{i1}' c_1 + M_{x_1}(x_{i2})' c_2 + u_i$. The differences between $\hat{Y}_3$ and $\hat{Y}_6$ concern the estimated coefficients: $\begin{pmatrix} \hat{C}_1 \\ \hat{a}_2 \end{pmatrix}$ differs slightly from $\begin{pmatrix} \hat{c}_1 \\ \hat{c}_2 \end{pmatrix}$ and $[\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{a}_1 - \hat{\hat{X}}_2' \hat{a}_2]$ differs slightly from $[\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{c}_1 - M_{x_1} \hat{\hat{X}}_2' \hat{c}_2]$. On the other hand, these quantities are asymptotically equivalent.

Strategy 4

We obtain weights

$$w_i^7 = \frac{F(x_{i1}'\beta_1) F(x_{i1}'\delta_1 + x_{i2}'\delta_2)}{\pi_{ai}\pi_i},$$

the associated estimator is rewritten as follows:

$$\hat{Y}_7 = \hat{\hat{Y}}^* + [X_1 - \hat{\hat{X}}_1^*]' \hat{a}_1^* + [\hat{X}_2^* - \hat{\hat{X}}_2^*]' \hat{a}_2^*.$$

By changing the initial weights in $d_i = F(x_{i1}\beta_1)/\pi_{ai}\pi_i$ we obtain in the same manner:

$$\hat{Y}_7 = [X_1' \hat{a}_1^*] + [\hat{X}_2^{*\prime} \hat{a}_2^*] + [\hat{\hat{Y}}^* - \hat{\hat{X}}_1^{*\prime} \hat{a}_1^* - \hat{\hat{X}}_2^{*\prime} \hat{a}_2^*].$$

By replacing $\hat{X}_2^*$ by its expression found above, we obtain:

$$\hat{Y}_7 = [X_1' \hat{C}_1^*] + [M_{x_1} \hat{X}_2' \hat{a}_2^*] + [\hat{\hat{Y}}^* - \hat{\hat{X}}_1'^* \hat{a}_1^* - \hat{\hat{X}}_2^{*\prime} \hat{a}_2^*],$$

with

$$\hat{C}_1 = \hat{a}_1^* + \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a}\right)^{-1} \left(\sum_{s_a} \frac{x_1 x_2'}{\pi_a}\right) \hat{a}_2^*.$$

Finally, $\hat{Y}_7$ and $\hat{Y}_6$ are asymptotically equivalent.

Say that $w = y - x_1' \hat{a}_1^* - x_2' \hat{a}_2^*$. Then $\hat{Y}_7 = \hat{Y}_6 + [\hat{\hat{W}}^* - \hat{\hat{W}}]$. Now, asymptotically $[\hat{\hat{W}}^* - \hat{\hat{W}}]$ is an infinitely small negligible before $\hat{Y}_6$:

$$[\hat{\hat{W}}^* - \hat{\hat{W}}] = \left(\sum_s \frac{w x_1'}{\pi_a \pi}\right) \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a}\right)^{-1} [X_1 - \hat{\hat{X}}_1],$$

and

$$\left(\sum_s \frac{w x_1'}{\pi_a \pi}\right) \text{ tends toward zero and } [X_1 - \hat{\hat{X}}_1] = O\left(\frac{1}{\sqrt{m}}\right).$$

Ultimately we obtain $\hat{Y}_7 \cong \hat{Y}_6$.

In conclusion, when the calibration function is exponential, the estimator $\hat{Y}_7$ coincides exactly with the preceding. When F is linear, $\hat{Y}_7$ is close to the preceding and thus still corresponds to the regression model approach in the case in which the information contained in x_1 is considered complementary to the information contained in x_2 for estimating y and in which the decomposition $y_i = x_{i1}' c_1 + M_{x_1}(x_{i2})' c_2 + u_i$ is used.

Conclusion: The Three Classes of Estimators

We have just seen that the four strategies derived from a calibration approach could be associated with regression modelling. We thus obtain three classes of estimators:

$Y_4 \cong Y_1$ associated with the models

$$(1) \quad y_i = x_{i1}' b_1 + u_{i1},$$

and

$$(2) \quad y_i = x_{i2}' b_2 + u_{i2}$$

$\hat{Y}_5 = \hat{Y}_2$ associated with the model

$$y_i = x_{i1}' a_1 + x_{i2}' a_2 + u_i$$

$\hat{Y}_6 \cong \hat{Y}_3$ and $\hat{Y}_7 \cong \hat{Y}_3$ associated with the model

$$y_i = x_{i1}' c_1 + M_{x_1}(x_{i2})' c_2 + u_1.$$

The approximation $\cong$, which indicates that the estimators are attached to the same regression model, takes on its full meaning when we are interested in calculating the variance of these different estimators, since the estimators that are attached to the same regression model have the same asymptotic variance.

5. ESTIMATION OF VARIANCES

Let us consider the variances of the different estimators $\hat{Y}_1, \ldots, \hat{Y}_7$ defined above. AV designates the asymptotic variance of an estimator that is obtained when N, n and m tend toward infinity in a constant relationship.

5.1 Estimator $\hat{Y}_1$ and $\hat{Y}_4$: model $y_i = x_{i1}' b_1 + u_{i1}$ and (2) $y_i = x_{i2}' b_2 + u_{i2}$.

• Estimator $\hat{Y}_1$

The variance of this estimator and its estimate are given in the work of Särndal, Swensson and Wretman (1991). The variance breaks down into two terms that measure the amounts of variance due respectively to the first and the second phase of the sampling.

$$\mathrm{AV}(\hat{Y}_1) = \left(\sum_{i,j \in U} \Delta_{ij}^1 \frac{u_{1i} u_{1j}}{\pi_i \pi_j}\right) + \left(E_{s_a} \sum_{i,j \in s_a} \Delta_{ij}^2 \frac{u_{2i} u_{2j}}{\pi_i \pi_{ai} \pi_j \pi_{aj}}\right),$$

with: $\Delta_{ij}^2 = \pi_{ij} - \pi_i \pi_j$,

$\Delta_{ij}^1 = \pi_{aij} - \pi_{ai} \pi_{aj}$,

$u_{1i} = y_i - x_{i1}' b_1$,

$u_{2i} = y_i - x_{i2}' b_2$,

$$b_1 = \left(\sum_{i \in U} x_{i1} x_{i1}'\right)^{-1} \left(\sum_{i \in U} x_{i1} y_i\right),$$

$$b_2 = \left(\sum_{i \in U} x_{i2} x_{i2}'\right)^{-1} \left(\sum_{i \in U} x_{i2} y_i\right).$$

Thus the variance estimator also breaks down into two terms that estimate the amounts of variance relating to each of the sampling phases. We find that by construction

of $\hat{Y}_1$, x_1 serves to reduce the variance brought about by the first phase and x_2 serves to reduce the variance brought about by the second phase.

$$\hat{V}(\hat{Y}_1) = \left(\sum_{i\in s}\sum_{j\in s}\frac{\Delta^1_{ij}}{\pi_{ij}\pi_{aij}}\frac{\hat{u}_{1i}\hat{u}_{1j}}{\pi_i\pi_j}\right) + \left(\sum_{i,j\in s}\frac{\Delta^2_{ij}}{\pi_{aij}}\frac{\hat{u}_{2i}\hat{u}_{2j}}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

1st phase 2nd phase

with: $\hat{u}_{1i} = y_i - x'_{i1}\hat{b}_1,$

$\hat{u}_{2i} = y_i - x'_{i2}\hat{b}_2.$

Such a decomposition is based on the expression $V(\hat{Y}_1) = V(E[\hat{Y}_1 \mid s_a]) + E(V[\hat{Y}_1 \mid s_a])$, which will apply for all the other estimators.

• **Estimator $\hat{Y}_4$**

The terms of the development to the first order in $1/\sqrt{m}$ of $\hat{Y}_1$ and $\hat{Y}_4$ coincide exactly. We can therefore give a more precise meaning to the expression $\hat{Y}_4 \cong \hat{Y}_1$. We deduce from this that $\mathrm{AV}(\hat{Y}_1) = \mathrm{AV}(\hat{Y}_4)$. Thus:

$$\hat{V}(\hat{Y}_4) = \left(\sum_{i,j\in s}\sum_{j\in s}\frac{\Delta^1_{ij}}{\pi_{ij}\pi_{aij}}\frac{\tilde{u}_{1i}\tilde{u}_{1j}}{\pi_i\pi_j}\right) + \left(\sum_{i,j\in s}\frac{\Delta^2_{ij}}{\pi_{aij}}\frac{\tilde{u}_{2i}\tilde{u}_{2j}}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

with: $\tilde{u}_{1i} = y_i - x'_{i1}\hat{B}_1,$

$\tilde{u}_{2i} = y_i - x'_{i2}\hat{B}_2.$

5.2 Estimators $\hat{Y}_2 = \hat{Y}_5$: model $y_i = x'_{i1}a_1 + x'_{i2}a_2 + u_i$

It is easy to show (see Dupont 1994) that:

$$\mathrm{AV}(\hat{Y}_2) = \mathrm{AV}(\hat{Y}_5) \cong \left(\sum_{i,j\in U}\Delta^1_{ij}\frac{v_iv_j}{\pi_i\pi_j}\right) + \left(E_{s_a}\sum_{i,j\in s_a}\Delta^2_{ij}\frac{u_iu_j}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

with: $v_i = y_i - x'_{i1}a_1,$

$u_i = y_i - x'_{i1}a_1 - x'_{i2}a_2$

$$a = \begin{pmatrix}a_1\\a_2\end{pmatrix} = \begin{pmatrix}\sum_{i\in U}x_{i1}x'_{i1} & \sum_{i\in U}x_{i1}x'_{i2}\\ \sum_{i\in U}x_{i2}x'_{i1} & \sum_{i\in U}x_{i2}x'_{i2}\end{pmatrix}\begin{pmatrix}\sum_{i\in U}x_{i1}y_i\\ \sum_{i\in U}x_{i2}y_i\end{pmatrix}$$

From this we deduce that:

$$\hat{V}(\hat{Y}_2) = \hat{V}(\hat{Y}_5) = \left(\sum_{i,j\in s}\frac{\Delta^1_{ij}}{\pi_{ij}\pi_{aij}}\frac{\hat{v}_i\hat{v}_j}{\pi_i\pi_j}\right) + \left(\sum_{i,j\in s}\frac{\Delta^2_{ij}}{\pi_{aij}}\frac{\hat{u}_i\hat{u}_j}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

with: $\hat{v}_i = y_i - x'_{i1}\hat{a}_1,$

$\hat{u}_i = y_i - x'_{i1}\hat{a}_1 - x'_{i2}\hat{a}_2.$

In this formulation we find that by construction of $\hat{Y}_2 - \hat{Y}_5$, x_1 reduces the variance brought about by the first phase and x_1 and x_2 are used simultaneously to reduce the variance brought about by the second phase.

5.3 Estimators $\hat{Y}_3$, $\hat{Y}_6$ and $\hat{Y}_7$: model $y_i = x'_{i1}c_1 + M_{x_1}(x_{i2})'c_2 + u_i$

We show that $\mathrm{AV}(\hat{Y}_6) = \mathrm{AV}(\hat{Y}_7) = \mathrm{AV}(\hat{Y}_3)$. Thus,

$$\mathrm{AV}(\hat{Y}_3) = \mathrm{AV}(\hat{Y}_6) = \mathrm{AV}(\hat{Y}_7) \cong \left(\sum_{i,j\in U}\Delta^1_{ij}\frac{u_{1i}u_{1j}}{\pi_i\pi_j}\right) + \left(E_{s_a}\sum_{i,j\in s_a}\Delta^2_{ij}\frac{u_iu_j}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

$u_{1i} = y_i - x'_{i1}c_1 = y_i - x'_{i1}b_1,$

$u_i = y_i - x'_{i1}c_1 - M_{x_1}x'_{i2}c_2 = y_i - x'_{i1}a_1 - x'_{i2}a_2.$

From this we deduce the three variance estimators, which differ owing to different estimated coefficients:

$$\hat{V}(\hat{Y}_3) = \left(\sum_{i,j\in s}\frac{\Delta^1_{ij}}{\pi_{ij}\pi_{aij}}\frac{\hat{u}_{1i}\hat{u}_{1j}}{\pi_i\pi_j}\right) + \left(\sum_{i,j\in s}\frac{\Delta^2_{ij}}{\pi_{aij}}\frac{\hat{u}_i\hat{u}_j}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

$\hat{u}_{1i} = y_i - x'_{i1}\hat{c}_1,$

$\hat{u}_i = y_i - x'_{i1}\hat{c}_1 - M_{x_1}x'_{i2}\hat{c}_2,$

$$\hat{V}(\hat{Y}_6) = \left(\sum_{i,j\in s}\frac{\Delta^1_{ij}}{\pi_{ij}\pi_{aij}}\frac{\tilde{u}_{1i}\tilde{u}_{1j}}{\pi_i\pi_j}\right) + \left(\sum_{i,j\in s}\frac{\Delta^2_{ij}}{\pi_{aij}}\frac{\tilde{u}_i\tilde{u}_j}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

$\tilde{u}_{1i} = y_i - x'_{i1}\hat{C}_1,$

$\tilde{u}_i = y_i - x'_{i1}\hat{a}_1 - x'_{i2}\hat{a}_2,$

$$\hat{V}(\hat{Y}_7) =$$

$$\left(\sum_{i,j\in s} \frac{\Delta_{ij}^1}{\pi_{ij}\pi_{aij}} \frac{\tilde{u}_{1i}\tilde{u}_{1j}}{\pi_i\pi_j}\right) + \left(\sum_{i,j\in s} \frac{\Delta_{ij}^2}{\pi_{aij}} \frac{\tilde{u}_i\tilde{u}_j}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

$$\tilde{u}_{1i} = y_i - x_{i1}'\hat{C}_1^*,$$

$$\tilde{u}_i = y_i - x_{i1}'\hat{a}_1^* - x_{i2}'\hat{a}_2^*,$$

$$\hat{C}_1 = \hat{a}_1^* + \left(\sum_{s_a} \frac{x_1x_1'}{\pi_a}\right)^{-1}\left(\sum_{s_a} \frac{x_1x_2'}{\pi_a}\right)\hat{a}_2^*.$$

We find that by construction of $\hat{Y}_3$, $\hat{Y}_6$ and $\hat{Y}_7$, x_1 is used to achieve *maximum* reduction of the variance brought about by the first phase and x_2 serves to reduce the variance brought about by the second phase.

6. CHOICE OF ESTIMATORS WHERE THERE IS SELECTION BIAS

In practice, when a survey is adjusted, it is not unusual to want not only to improve the estimation, but also and more especially to correct the biases introduced by uncontrolled selections of individuals, such as nonresponse.

We shall examine the case of a two-phase sampling in which the second phase is equivalent to total nonresponse. The weights π_i of the second-phase sampling are thus unknown. The calibration of s will enable us to estimate these probabilities, while reducing the variance (*cf.* Deville and Dupont 1993). However, asymptotically, the corrections of bias to be made to the weights are greater than the changes to be made in order to improve the estimators. It is therefore the implicit response model that will guide the choice between the different estimators:

The implicit response model for the first class of estimators is $p_i = 1/F(x_{i2}'B_2)$.

The implicit response model for the second and third classes of estimators is $p_i = 1/F(x_{i2}'A_2 + x_{i1}'A_1)$.

- Whatever the response model, an evaluation of the three classes of estimators on the basis of the sampling plan alone still indicates that the third is preferable, since it is appropriate for all the response models.
- If the strategies are evaluated on the basis of regression modelling, we will use the first class of estimators only if the response mechanism is well explained by x_2, that is, $p_i = 1/F(x_{i2}'B_2)$. Now, we have seen that the modelling associated with the first class of estimators takes on its meaning when the variables x_1 and x_2 are highly correlated. It is therefore fairly probable that in this context, the variable x_2 will be sufficient to explain the response mechanism. Should this not be the case, it will be necessary to turn to the third class of estimators.

The comparison between the three strategies may thus be adapted in a context in which we wish to correct the biases introduced by uncontrolled selections. The conclusions remain largely the same.

According to the same principle, it is of course possible to make comparisons between alternative adjustment strategies in the context of samplings that entail more than two phases and one or more uncontrolled selections.

7. *A PRIORI* AND *A POSTERIORI* USE OF AUXILIARY INFORMATION

The calibration estimator enables us to improve the estimate *a posteriori*, by reducing the variance and correcting the bias, as noted above. However, we may want to incorporate the auxiliary information *a priori*, at the sampling stage rather than *a posteriori* at the estimation stage. We then encounter, in a more complex context, the classical opposition between stratification and poststratification, well known in the case of single-phase sampling, when all the auxiliary variables are qualitative.

It is possible to transpose the terms of the choice between using the information *a priori* and *a posteriori*, in the sampling and auxiliary information configuration studied, when the auxiliary variables are qualitative. When the auxiliary variables are qualitative, a calibration corresponds exactly to poststratification.

We saw earlier that in order to determine the proper adjustment procedure, it was necessary to distinguish two possible modellings of the variable of interest, depending on whether the information in x_1 and the information in x_2 were considered substitutable or complementary. Each of these two modellings then led to one or more different adjustment procedures. Similarly, these two modellings arise when it is a matter of identifying the best sampling strategy for incorporating the auxiliary information:

- When the information in x_1 and the information in x_2 are substitutable, the modelling of the variable of interest is as follows:

  (1) $y_i = x_{i1}b_1 + u_{i1}$ and

  (2) $y_i = x_{i2}b_2 + u_{i2}$ where the second model is better for predicting the value of y_i.

  We have seen that the use of the auxiliary information *a posteriori* leads to calibration strategy No. 1, that is, to the first class of estimators. If we wish to take account of the auxiliary information at the sampling stage, it is natural to propose a sampling stratified on x_1 for the first phase and a sampling stratified on x_2 for the second phase.

  However, the parallel between the adjustment procedure and the sampling procedure is not complete: in a calibration, only the marginal information in x_1 can be used.

This results in incomplete poststratification (Särndal and Deville 1992). On the other hand, in the sampling procedure proposed as an *a priori* alternative, we are obliged to use all the cross-tabulations of the x_1 variables. The *a priori* equivalent of a calibration would accordingly be a sampling balanced on the margins of the vector of variables x_1.

- When the information contained in x_1 and the information contained in x_2 are complementary, the modelling of the variable of interest is y is $y_i = x_{i1}b_1 + x_{i2}b_2 + u_i$. We have seen that in this case the use of *a posteriori* auxiliary information led to calibration strategies 2, 3 and 4 in estimator classes 2 and 3. If we wish to take account of the auxiliary information at the sampling stage, it is natural to propose a sampling stratified on $x1$ for the first phase and a sampling stratified on x_1 and x_2 for the second phase.

  As before, there is no exact parallel between the *a priori* and *a posteriori* procedures, since the use of the information *a priori* mobilizes all the cross-tabulations between the variables x_1 and x_2.

Thus it is possible to make a choice between incorporating the information either *a priori* or *a posteriori*, and indeed to optimize the sampling plan, when the auxiliary variables are qualitative. The terms of the choice are the same as in a single-phase sampling with a single level of information. An additional consideration is the multiplicity of strata created by the cross-tabulations of x_1 and x_2 in the case in which the modelling used is $y_i = x_{i1}b_1 + x_{i2}b_2 + u_i$, which reinforces the advantages of using the information *a posteriori*.

When the auxiliary variables are quantitative, the choice depends on their conversion into qualitative variables, it not being possible to generalize correctly except by using the parallel between calibration and balanced sampling (*cf.* Deville 1992).

8. CONCLUSION

In a two-phase sampling, when two different sets of information are available for the total population on the one hand and the sample resulting from the first phase on the other hand, several strategies are possible when one wishes to use the auxiliary information to improve the estimation of totals.

Two different natural approaches have been used to derive estimators: a regression model assisted approach, which seeks to adapt the idea of the regression estimator; and a calibration approach, which attempts to adapt the idea of calibration. The estimators obtained by the two approaches may be linked together. We generated three alternative underlying modellings to which the various estimators obtained may be attached. Thus we obtained three classes of estimators. Several conceivable calibration strategies were eliminated at the outset as irrelevant.

We have shown that the estimators of a given class, that is, the estimators attached to a given model, are asymptotically equivalent; we gave the form of the variances derived in the case of a linear calibration function, but with asymptotic equivalences, these results remain valid for any calibration function.

For purposes of evaluating strategies, the form of the variances indicates, as intuition would suggest, that one of the classes of estimators (estimators 3, 6 and 7 (calibration strategies 3 and 4)) is preferable to the other from the standpoint of variance when the evaluation is based on the sampling plan alone. When it is based on a modelling of the variable of interest, it suggests that the preferable class of estimators is the one associated with the modelling adopted.

In a situation in which the goal is to adjust a survey and to simultaneously correct the biases that would arise from the use of gross weightings and reduce the variance, the conclusions must be adapted. The changes introduced in the weighting to correct the biases are greater than the corrections to reduce variance. Hence the variables will be incorporated into the calibration once it appears that they affect the probability of selection and thus participate in the creation of bias.

When the auxiliary variables are qualitative, the choice between *a priori* and *a posteriori* use of auxiliary information, that is, between using it at the sampling stage or at the adjustment stage, still rests on the distinction between the two modellings of the variable of interest.

These results may easily be generalized to the case of sampling involving more than two phases.

ACKNOWLEDGEMENTS

I am deeply grateful to Jean-Claude Deville, Louis Meuric and Carl-Erik Särndal for their many helpful suggestions regarding this article.

REFERENCES

CASSEL, C.M., SÄRNDAL, C.-E., and WRETMAN, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite population. *Biometrica*, 63, 615-620.

DEVILLE, J.-C. (1992). Constrained samples, conditional inference, weighting: three aspects of the utilisation of auxiliary information. *Proceedings of the Workshop on Uses of Auxiliary Information in Surveys*, October 1992, Örebro.

DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators and generalized raking techniques in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DEVILLE, J.-C., SÄRNDAL, C.-E., and SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

DEVILLE, J.-C., and DUPONT, F. (1993). Calage et redressement de la non-réponse totale. Journées de Méthodologie.

DUPONT, F. (1994). Redressements alternatifs en présence de plusieurs niveaux d'information auxiliaire. Working paper of Direction des Statistiques Démographiques et Sociales, F9409.

FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā* C, 37, 117-132.

GOURIEROUX, C. (1981). *Théorie des sondages*. Edition Economica Paris.

ISAKI, C.T., and FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

SÄRNDAL, C.-E. (1980). On π inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.

SÄRNDAL, C.-E., and SWENSSON, B. (1987). A general view of estimation for two phases of selection with applications to two-phases sampling and nonresponse. *International Statistical Review*, 55, 279-294.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1991). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

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

Survey Methodology, December 1995
Vol. 21, No. 2, pp. 137–145
Statistics Canada

Estimating Some Measures of Income Inequality from Survey Data: An Application of the Estimating Equations Approach

DAVID A. BINDER and MILORAD S. KOVACEVIC[1]

ABSTRACT

We summarize some salient aspects of the theory of estimation functions for finite populations. In particular, we discuss the problem of estimation of means and totals and extend this theory to estimating functions. We then apply this estimating functions framework to the problem of estimating measures of income inequality. The resulting statistics are nonlinear functions of the observations. Some of them depend on the order of observations or quantiles. Consequently, the mean squared errors of these estimates are inexpressible by simple formulae and cannot be estimated by conventional variance estimation methods. We show that within the estimating function framework this problem can be resolved using the Taylor linearization method. Finally, we illustrate the proposed methodology using income data from Canadian Survey of Consumer Finance and comparing it to the 'delete-one-cluster' jackknifing method.

KEY WORDS: Complex survey design; Gini family coefficient; Lorenz curve ordinate; Low income measure; Quantile share.

1. INTRODUCTION

The measurement and analysis of economic inequality are well covered in econometrics literature from both, theoretical and applied aspects, although the theoretical issues prevail. Estimation of inequality measures and the impact of the design of sample surveys have gotten less attention. Variance estimation, unavoidable in statistical inference based on these measures, is seldom an issue in the relevant econometric literature. It is usually addressed under very strong assumptions and under unsustainable simplifications of the design or the formulae for the approximate variance. In this paper we present a method that can handle with ease both the estimation of the measures of income inequality and the variance estimation of the resulting non-linear statistics. This method is applicable under a variety of sampling designs.

In general, a population distribution can be described by its cumulative distribution function, $F(y) = \Pr\{Y \leq y\}$, where Y is the random variable corresponding to selecting one population unit at random. Throughout this paper, we assume that Y is non-negative. If Y represents income then we are interested in the properties of an income distribution, such as income concentration, income shares for different population shares, low income proportions, *etc.* We are also interested in the quantile function $\xi(p) = F^{-1}(p) = \inf\{y \mid F(y) \geq p\}$.

The Lorenz curve, for example, depicts the cumulative income against the population share. The formal definition of the ordinate of the Lorenz curve corresponding to the 100 p-th percentile of the population is

$$L(p) = \frac{\int_0^{\xi_p} y dF(y)}{\mu_Y}, \tag{1.1}$$

where

$$\int_0^{\xi_p} dF(y) = p, \quad \text{and} \quad \int_0^{\infty} y dF(y) = \mu_Y.$$

The finite population form of the expression (1.1), more familiar to survey statisticians, is given by

$$L(p) = \sum_U Y_i I\{Y_i \leq \xi_p\} \Big/ \sum_U Y_i,$$

where U represents a finite population and $I\{.\}$ is an indicator function.

The income (quantile) share is defined as the percentage of total income shared by the population allocated to the certain income quantile interval $[\xi_{p_1}, \xi_{p_2}], p_1 \leq p_2$. It is equal to the difference of Lorenz curve ordinates

$$Q(p_1,p_2) = L(p_2) - L(p_1).$$

In Figure 1 we give a graph of the Lorenz curve for the Weibull distribution with shape parameter $\alpha = 1.6$, along with the 45° axis. For example, one can read from the graph that not more than 25% of the total income is allocated to the poor half of population, or that the richest 10% of the population earn 20% of the total available income.

[1] David A. Binder, Director, Business Survey Methods Division, and Milorad S. Kovacevic, Senior Methodologist, Household Survey Methods Division, Statistics Canada, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

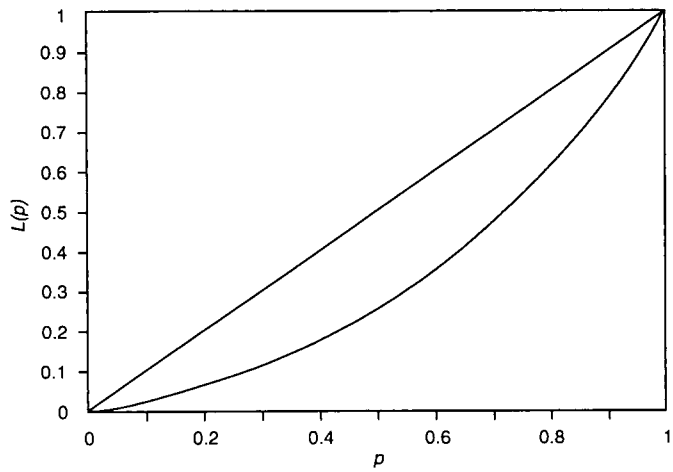


Figure 1. Lorenz Curve for the Weibull Distribution with Shape Parameter $\alpha = 1.6$.

The Gini coefficient measures the degree of the inequality in income distribution. One definition of the Gini coefficient is a linear function of the area between the Lorenz curve and the 45° axis, normalized to lie between 0 and 1. The Gini coefficient in Figure 1 is 0.35. The formal definition of the Gini coefficient (Nygård and Sandström 1981) is

$$G = 1 - 2\int_0^1 L(p)dp = \frac{1}{\mu}\int_0^\infty [2F(y) - 1]ydF(y).$$

A more general family of Gini coefficients, given in Nygård and Sandström (1981) is

$$G_J = \frac{1}{\mu_Y}\int_0^\infty J[F(y)]ydF(y), \tag{1.2}$$

where J is a bounded and continuous function. For the usual Gini coefficient, $J(p) = 2p - 1$.

Another measure of income inequality used by some economists is the Low Income Measure. This is defined as the proportion of the population units whose income is less than half the median income for the population. Formally, this is

$$\Theta = \int_0^{M/2} dF(y), \tag{1.3a}$$

where M is the median defined by

$$\int_0^M dF(y) = \frac{1}{2}. \tag{1.3b}$$

For all these measures, we can express the parameter of interest, Θ, as the solution to the equation

$$\int u(y,\Theta)dF(y) = 0,$$

where $u(y,\Theta)$ is the kernel of the estimating equation. This estimating equation formulation will be discussed in Section 2. In Sections 3, 4, and 5 we give the estimating equations for the above measures along with the approximation of their mean squared error estimates. In Section 6 we present estimators of these measures based on the complex sample design. Section 7 contains an illustration based on the Canadian Survey of Consumer Finance data.

2. USE OF ESTIMATING EQUATIONS FOR FINITE POPULATIONS

The theory for estimating means and totals from finite populations is now well established in the statistical literature. A formulation which encompasses most estimators used in practice is given in Särndal, Swensson, and Wretman (1992). In this section, we briefly review this theory and show how it can be applied to more complex statistics through the use of estimating equations, as described by Binder (1991) and Binder and Patak (1994).

We begin the exposition of the main idea by reviewing the estimation of the population total T_Y and the finite population distribution function $F(y)$. The estimation of the population total is the core of the estimation equations approach of Binder (1991) and Binder and Patak (1994). Let the population total of the variable Y, be defined as

$$T_Y = N\int ydF(y).$$

Note here that $F(y)$ is a step function corresponding to the distribution function for the finite population. We consider estimators of the form:

$$\hat{T}_Y = \sum_{i\in s} w_i(s)y_i = \sum_{i=1}^{N} w_i(s)Y_i, \tag{2.1}$$

where $w_i(s)$ is zero whenever the i-th unit is not in the sample. Expression (2.1) gives, for example, the Horvitz-Thompson (HT) unbiased estimator if

$$w_i(s) = \begin{cases} 1/\pi_i, & i\in s, \\ 0, & i\notin s, \end{cases}$$

or the generalized regression estimator if

$$w_i(s) = \begin{cases} [1 + (T_X - \hat{T}_X)x_i/\hat{T}_{X2}]/\pi_i, & i\in s, \\ 0, & i\notin s, \end{cases}$$

where T_X is the population total of X, and $\hat{T}_X$ and $\hat{T}_{X^2}$ are the HT estimates of the totals of X and X^2 variables, respectively.

Similarly, an estimator for the distribution function is given by

$$N\hat{F}(y) = \sum_{i\epsilon s} w_i(s) I\{y_i \leq y\},$$

where

$$I\{y_i \leq y\} = \begin{cases} 1 & \text{if } y_i \leq y, \\ 0 & \text{if } y_i > y. \end{cases}$$

We note that $\hat{F}(y)$ is uniformly and asymptotically design consistent for $F(y)$, but it is not necessarily a true distribution function, unless

$$\sum_{i\epsilon s} w_i(s) = N.$$

In general, and under certain regularity conditions for complex designs (Francisco and Fuller 1991),

$$\hat{F}(y) - F(y) \to_p 0, \text{ for any } y.$$

That is, the finite population distribution function, $F(y)$, allows a consistent estimator, $\hat{F}(y)$. This property of the $\hat{F}(y)$ will be used later in proving the consistency of the linearized variance estimators for different income statistics.

Now, we review the application of the estimating equations theory to the estimation of any finite population parameter Θ_o that can be expressed as the solution to

$$\int u(y,\Theta_o) dF(y) = 0.$$

We define the estimating equation estimate for Θ_o as that value of $\hat{\Theta}$ for which

$$\int \hat{u}(y,\hat{\Theta}) d\hat{F}(y) = 0, \tag{2.2}$$

where $\hat{u}(y,\Theta)$ is an estimate of $u(y,\Theta)$.

We can rewrite (2.2) as

$$\begin{aligned} 0 &= \int \hat{u}(y,\hat{\Theta}) d\hat{F}(y) \\ &= \int [\hat{u}(y,\hat{\Theta}) - u(y,\Theta_o)] dF(y) + \int u(y,\Theta_o) d\hat{F}(y) + R, \end{aligned} \tag{2.3}$$

where

$$R = \int [\hat{u}(y,\hat{\Theta}) - u(y,\Theta_o)] [d\hat{F}(y) - dF(y)].$$

The decomposition in (2.3) is the basic starting point for all the derivations of variance in the paper. For each parameter considered we will prove that the remainder term, R, is asymptotically negligible.

Binder (1983) considered the case where $\hat{u}(y,\Theta) = u(y,\Theta)$ and where, for large samples,

$$\begin{aligned} &\int [u(y,\hat{\Theta}) - u(y,\Theta_o)] dF(y) \\ &= (\hat{\Theta} - \Theta_o) \left. \frac{\partial E\{u(y,\Theta)\}}{\partial \Theta} \right|_{\Theta=\Theta_o} + o_p(|\hat{\Theta} - \Theta_o|). \end{aligned}$$

Note that the remainder term R from the decomposition (2.3) should be of order $o_p(|\hat{\Theta} - \Theta_o|)$ to be considered as asymptotically negligible.

For most applications $u(y,\Theta)$ does not need to be estimated by $\hat{u}(y,\Theta)$. However, for some applications such as the Gini coefficient, the function $u(y,\Theta)$ is estimated so that formula (2.2) allows for these cases in general.

Using these approximations, we have

$$\begin{aligned} \hat{\Theta} - \Theta_o \approx &- \left[\left. \frac{\partial E\{u(y,\Theta)\}}{\partial \Theta} \right|_{\Theta=\Theta_o} \right]^{-1} \\ &\times \int u(y,\Theta_o) d\hat{F}(y) = \int u^*(y) d\hat{F}(y), \end{aligned} \tag{2.4}$$

where

$$u^*(y) = - \left[\left. \frac{\partial E\{u(y,\Theta)\}}{\partial \Theta} \right|_{\Theta=\Theta_o} \right]^{-1} u(y,\Theta_o).$$

Once we have obtained the expression for $u^*(y)$, the derivation of the variance of $\hat{\Theta}$ becomes straightforward. Since we have approximated $\hat{\Theta} - \Theta_o$ as an estimator of a population total of $u^*(y_i)$'s, we can use the mean squared error calculations for the estimate of total to obtain the variance estimate of $\hat{\Theta}$.

For example, for Θ_o equal to the ratio, T_Y/T_X, we have

$$u = y - \Theta_0 x,$$

$$u^* = \frac{1}{\mu_X} (y - \Theta_0 x).$$

The remainder term in this case is

$$R = \int [y - \hat{\Theta}x - (y - \Theta_0 x)]\,[d\hat{F}(y) - dF(y)].$$

Therefore,

$$-\frac{R}{\hat{\Theta} - \Theta_o} = [\hat{F}(y) - F(y)]x \to_p 0,$$

for any y and any finite x.

Similarly, for population quantiles, we have

$$\begin{aligned} u &= I\{y \le \Theta_0\} - p, \\ u^* &= -\frac{1}{f(\Theta_0)}\,[I\{y \le \Theta_0\} - p], \end{aligned} \tag{2.5}$$

where $f(\Theta_0)$ is the value of the density function at Θ_0. The second expression in (2.5) is an extension of the Bahadur representation for sample quantiles, as described by Francisco and Fuller (1991). Result (2.5) will be used for the ordinates of the Lorenz curve and for the Low Income Measure, which are discussed in Sections 4 and 5.

The remainder term R in this case reduces to $R = \hat{F}(\hat{\Theta}) - \hat{F}(\Theta_o) - F(\hat{\Theta}) + F(\Theta_o)$. In the case of the simple random sample design, Randles (1982) showed that $R = o_p(n^{-1/2})$. For the complex design situation, under some regularity conditions, Shao and Rao (1994) established a similar asymptotic result: first they showed that $\hat{\Theta} - \Theta_o = O_p(n^{-1/2})$, then that $R = o_p(n^{-1/2})$, and therefore $R = o_p(|\hat{\Theta} - \Theta_o|)$.

3. GINI FAMILY COEFFICIENT

For the Gini family coefficient, given by (1.2), we can use

$$u(y, G_J) = J[F(y)]y - G_J y.$$

Binder's (1983) approach cannot handle the variance estimation of the Gini coefficient. For the Gini coefficient, rather than deriving the variances by breaking the problem into two parts - one for the ratio estimator and the other for the variance of the numerator - we use the estimating equations approach to solve the problem in one step.

Ignoring the remainder term in (2.3), we have the following approximation:

$$\begin{aligned} 0 &= \int \{J[\hat{F}(y)]y - \hat{G}_J y\}d\hat{F}(y) \\ &\approx \int \{J[\hat{F}(y)] - J[F(y)]\}ydF(y) \\ &\quad - (\hat{G}_J - G_J)\int ydF(y) + \int \{J[F(y)]y - G_J y\}d\hat{F}(y). \end{aligned}$$

Letting

$$\int \{J[\hat{F}(y)] - J[F(y)]\}ydF(y) \approx \int [\hat{F}(y) - F(y)]J'[F(y)]ydF(y),$$

and

$$\begin{aligned} \int \hat{F}(y)J'[F(y)]ydF(y) &= \int\int_0^y J'[F(y)]yd\hat{F}(x)dF(y) \\ &= \int \left[\int_y^\infty J'[F(x)]xdF(x)\right] d\hat{F}(y), \end{aligned}$$

we have that

$$\hat{G}_J - G_J \approx \int u^*(y)d\hat{F}(y),$$

where

$$u^* = \frac{1}{\mu_Y}\left[\int_{F(y)}^1 J'(p)F^{-1}(p)dp + J[F(y)]y - G_J y - E\{F(y)J'[F(y)]y\}\right]. \tag{3.1}$$

For the case of independent and identically distributed observations, this yields the same variance result as described by Glasser (1962) and Sendler (1979). To estimate the variance, it is necessary to use estimates for μ_Y, $F(y)$, and G_J in the expression for u^*.

We investigate the asymptotic behaviour of the remainder term R for the usual Gini coefficient G. The remainder is

$$R = \int \{2y[\hat{F}(y) - F(y)] - y(\hat{G} - G)\} \times [d\hat{F}(y) - dF(y)].$$

Denoting the difference $\hat{F}(y) - F(y)$ by $\hat{D}(y)$, the remainder can be expressed as a sum of two integrals

$$R = \int 2y\hat{D}(y)d\hat{D}(y) - \int (\hat{G} - G)yd\hat{D}(y).$$

The first integral is reduced to zero by the integration by parts, so that the remainder is approximated by

$$R \approx -(\hat{G} - G)(\hat{\mu}_Y - \mu_Y)$$
$$= -(\hat{G} - G)o_p(n^{-1/2+\delta}), \quad 0 < \delta < 1/2.$$

Therefore, we can say that $R = o_p(|\hat{G} - G|)$.

4. LORENZ CURVE ORDINATE AND QUANTILE SHARE

The ordinate of the Lorenz curve was defined in (1.1). In terms of estimating equations, the following two equations are required:

$$u_1(y,L(p)) = I\{y \le \xi_p\}y - L(p)y,$$
$$u_2(y) = I\{y \le \xi_p\} - p.$$

The second equation defines the $100p$-th percentile of the distribution; whereas the first equation defines the ordinate of the Lorenz curve in terms of the $100p$-th percentile. Ignoring the remainder term in (2.3), we have the following approximation:

$$0 = \int [I\{y \le \hat{\xi}_p\} - \hat{L}(p)]yd\hat{F}(y)$$
$$\approx \int_{\xi_p}^{\hat{\xi}_p} ydF(y) - [\hat{L}(p) - L(p)]\int ydF(y)$$
$$+ \int [I\{y \le \xi_p\} - L(p)]yd\hat{F}(y).$$

The first term of this expression can be further approximated as

$$\int_{\xi_p}^{\hat{\xi}_p} ydF(y) \approx (\hat{\xi}_p - \xi_p)\xi_p f(\xi_p),$$

and from (2.5) we see that

$$\hat{\xi}_p - \xi_p \approx -\int \frac{1}{f(\xi_p)} [I\{y \le \xi_p\} - p]d\hat{F}(y), \quad (4.1)$$

so that

$$(\hat{\xi}_p - \xi_p)\xi_p f(\xi_p) \approx -\int \xi_p[I\{y \le \xi_p\} - p]d\hat{F}(y).$$

Therefore, to estimate the variance of the ordinate of the Lorenz curve, the appropriate linearization is given by using

$$u^*(y) = \frac{1}{\mu_Y}[(y - \xi_p)I\{y \le \xi_p\} + p\xi_p - yL(p)].$$

This yields the same result as described by Beach and Davidson (1983) for variances and covariances of ordinates of the Lorenz curve in the case of independent and identically distributed random variables. To estimate the variance it is necessary to use $\hat{\xi}_p$ and $\hat{L}(p)$ in the expression for $u^*(y)$.

To estimate the quantile share $Q(p_1, p_2)$ we need three equations

$$u_1(y,Q(p_1,p_2)) = I\{\xi_{p_1} < y \le \xi_{p_2}\}y - Q(p_1,p_2)y,$$
$$u_2(y) = I\{y \le \xi_{p_1}\} - p_1,$$
$$u_3(y) = I\{y \le \xi_{p_2}\} - p_2.$$

Using the same arguments as before, we arrive at

$$u^*(y) = \frac{1}{\mu_Y}[(y - \xi_{p_2})I\{y \le \xi_{p_2}\}$$
$$- (y - \xi_{p_1})I\{y \le \xi_{p_1}\}$$
$$+ p_2\xi_{p_2} - p_1\xi_{p_1} - yQ(p_1,p_2)].$$

5. LOW INCOME MEASURE

The Low Income Measure was defined in (1.3). In terms of estimating equations, the following two equations are required:

$$u_1(y,\Theta) = I\left\{y \le \frac{M}{2}\right\} - \Theta,$$
$$u_2(y) = I\{y \le M\} - \frac{1}{2},$$

where M denotes the median of the distribution defined by the second equation, whereas the first equation defines the Low Income Measure in terms of the median. Ignoring the remainder term in (2.3), we have the following approximation:

$$0 = \int \left(I\left\{ y \le \frac{\hat{M}}{2} \right\} - \hat{\Theta} \right) d\hat{F}(y)$$

$$\approx \frac{1}{2}(\hat{M} - M) f\left(\frac{M}{2}\right) - (\hat{\Theta} - \Theta) + \int \left(I\left\{ y \le \frac{M}{2} \right\} - \Theta \right) d\hat{F}(y).$$

Using result (4.1) to substitute for $\hat{M} - M$, and solving for $\hat{\Theta} - \Theta$, we obtain

$$\hat{\Theta} - \Theta \approx \int u^*(y) d\hat{F}(y),$$

where

$$u^* = -\frac{f\left(\frac{M}{2}\right)}{2f(M)} \left(I\{y \le M\} - \frac{1}{2} \right) + I\left\{ y \le \frac{M}{2} \right\} - \Theta. \quad (5.1)$$

The problem with applying this result to estimate the variance of the estimated Low Income Measure is that it is necessary to estimate $f(M)$ and $f(M/2)$. To accomplish this, we could use

$$\hat{f}(\xi) = \frac{\hat{F}\left(\xi + \frac{h}{2}\right) - \hat{F}\left(\xi - \frac{h}{2}\right)}{h},$$

for some suitably small h. Alternatively, we could perform the following calculations, as suggested by Francisco and Fuller (1991) for another problem. For a given value of ξ, we estimate the corresponding percentile, $100p$. We then construct the Woodruff interval for that percentile. This is determined by first solving for h_1 and h_2 in

$$\inf_{h_1} \left[\frac{\int [I\{y \le \xi - h_1\} - p] d\hat{F}(y)}{\left[\text{mse}\left\{ \int [I\{y \le \xi\} - p] d\hat{F}(y) \right\} \right]^{1/2}} \le -z_{1-\alpha/2} \right],$$

$$\inf_{h_2} \left[\frac{\int [I\{y \le \xi + h_2\} - p] d\hat{F}(y)}{\left[\text{mse}\left\{ \int [I\{y \le \xi\} - p] d\hat{F}(y) \right\} \right]^{1/2}} \ge z_{1-\alpha/2} \right],$$

where $z_{1-\alpha/2}$ is the $100(1 - \alpha/2)$-th percentile from the standard normal distribution. Then we compute

$$\hat{f}(\xi) = \frac{2z_{1-\alpha/2} \left[\text{mse}\left\{ \int [I\{y \le \xi\} - p] d\hat{F}(y) \right\} \right]^{1/2}}{h_1 + h_2}. \quad (5.2)$$

This calculation uses the asymptotic equivalence of $\hat{\xi} - \xi$ and the estimated sum of the $u^*(y)$'s given by (2.5).

We see that the estimated variance for the Low Income Measure may be somewhat complex to compute. The estimating functions framework has however provided us with the appropriate formulae.

The discussion about the remainder term in the decomposition (2.3) of the low income measure is analogous to that made for the case of the quantile estimation (2.5).

6. ESTIMATION WITH A COMPLEX SURVEY

Let us assume a stratified multistage design with a large number of strata, H, with a few primary sampling units (clusters), $n_h (\ge 2)$, sampled from each stratum. For example, in the Canadian Survey of Consumer Finance (SCF) which uses the Labour Force Survey (LFS) vehicle, the number of strata is several hundreds and the number of clusters per stratum is on average less than six. Let w_{hci} be the normalized weight attached to the i-th ultimate unit in the c-th cluster of the h-th stratum such that the appropriate estimator of mean and the consistent estimator of its mean squared error are

$$\hat{\mu} = \sum_s w_{hci} y_{hci}$$

$$\text{mse}(\hat{\mu}) = \sum_h \frac{n_h}{n_h - 1} \sum_c (u^*_{hc} - \bar{u}^*_h)^2 \quad (6.1)$$

where $u^*_{hc} = \sum_i w_{hci}(y_{hci} - \hat{\mu})$ and $\bar{u}^*_h = 1/n_h \sum_c u^*_{hc}$. We use $\sum_s = \sum_h \sum_c \sum_i$ to denote summation over all ultimate units in the sample incorporating all stages of sampling. We assume that PSU's are selected with replacement.

This paper is not concerned with the efficiency of the estimators but rather the properties of commonly used estimators. An analysis of more complex estimators found in the econometric literature is beyond the scope of our study.

An estimator of the finite population distribution function is

$$\hat{F}(y) = \sum_s w_{hci} I\{y_{hci} \le y\}.$$

A consistent estimator of the approximation of the mean squared error of the distribution function estimated in y takes the form (6.1) where $u^*_{hc} = \sum_i w_{hci}[I\{y_{hci} \leq y\} - \hat{F}(y)]$.

The usual estimate of the finite population quantile is the sample quantile

$$\hat{\xi}_p = \inf\{y_{hci} \epsilon s : \hat{F}(y_{hci}) \geq p\}$$

which is the solution of the estimating equation

$$\sum_s w_{hci}[I\{y_{hci} \leq \hat{\xi}_p\} - p] = 0.$$

Accordingly, using result (2.5), the estimator of the mean squared error of the p-th quantile has the form (6.1) with

$$u^*_{hc} = \frac{1}{[\hat{f}(\hat{\xi}_p)]^2} \sum_i w_{hci}[I\{y_{hci} \leq \hat{\xi}_p\} - p].$$

If the expression (5.2) is used for the estimation of the density function $f(\xi)$, the MSE estimate of the quantile $\hat{\xi}_p$ becomes

$$\text{mse}_\alpha(\hat{\xi}_p) = \left(\frac{D_\alpha(\hat{\xi}_p)}{z_{1-\alpha/2}}\right)^2 \qquad (6.2)$$

where $D_\alpha(\hat{\xi}_p) = (h_1 + h_2)/2 = (\hat{\xi}_U - \hat{\xi}_L)/2$ is the half length of the $100(1 - \alpha)\%$ confidence interval for $\hat{\xi}_p$. In a complex sample design, h_1 and h_2 are obtained as solutions of

$$\hat{\xi}_L = \hat{\xi}_p - h_1 = \inf\left\{y_{hci} \epsilon s : \hat{F}(y_{hci}) \geq p - z_{1-\alpha/2}\sqrt{\text{mse}[\hat{F}(\hat{\xi}_p)]}\right\}$$

$$\hat{\xi}_U = \hat{\xi}_p + h_2 = \inf\left\{y_{hci} \epsilon s : \hat{F}(y_{hci}) \geq p + z_{1-\alpha/2}\sqrt{\text{mse}[\hat{F}(\hat{\xi}_p)]}\right\}.$$

The estimator (6.2) was also used by Francisco and Fuller (1991). Generally speaking the motivation for (5.2) and consequently for (6.2) comes from Woodruff's (1952) confidence interval for individual quantiles. Francisco and Fuller (1986) and Rao and Wu (1987) used these intervals to derive variance estimators. Although the estimator depends on the confidence coefficient, they showed that it is asymptotically consistent for any significance level α. Rao and Wu (1987) studied the standard errors of quantiles for the cluster samples estimated in this manner. Their Monte Carlo results suggest that 95% confidence interval works well as a basis for extracting the standard error. Binder and Patak (1994) obtained a similar form of the variance estimator by using the estimating equations approach.

The estimate of the usual Gini coefficient is the solution of the following estimating equation

$$\sum_s w_{hci}\{[2\hat{F}(y_{hci}) - 1]y_{hci} - \hat{G}y_{hci}\} = 0$$

and takes the form

$$\hat{G} = \frac{2}{\hat{\mu}} \sum_s w_{hci}\hat{F}(y_{hci})y_{hci} - 1$$

where $\hat{\mu} = \sum_s w_{hci}\, y_{hci}$.

The estimate of the MSE of the Gini coefficient can be computed using expression (6.1) by replacing u^*_{hc}, originally defined by (3.1), with its complex survey form. After some algebraic manipulation we obtain the following expression:

$$u^*_{hc} = \frac{2}{\hat{\mu}} \sum_i w_{hci}\left[A(y_{hci})y_{hci} + B(y_{hci}) - \frac{\hat{\mu}}{2}(\hat{G} + 1)\right]$$

where

$$A(y) = \hat{F}(y) - \frac{\hat{G} + 1}{2}$$

and

$$B(y) = \sum_s w_{hci}\, y_{hci}\, I\{y_{hci} \geq y\}.$$

The Lorenz curve ordinates could be obtained by solving a system of estimating equations

$$\sum_s w_{hci}[I\{y_{hci} \leq \hat{\xi}_p\}y_{hci} - \hat{L}(p)y_{hci}] = 0$$

$$\sum_s w_{hci}[I\{y_{hci} \leq \hat{\xi}_p\} - p] = 0.$$

The resulting estimate is

$$\hat{L}(p) = \frac{1}{\hat{\mu}} \sum_s w_{hci}\, y_{hci}\, I\{y_{hci} \leq \hat{\xi}_p\}.$$

To estimate the mean squared error of the Lorenz curve ordinates we simply use the values of u^*_{hc} defined by (6.3) in (6.1)

$$u^*_{hc} = \frac{1}{\hat{\mu}} \sum_i w_{hci}\left[(y_{hci} - \hat{\xi}_p)I\{y_{hci} \leq \hat{\xi}_p\} + p\hat{\xi}_p - y_{hci}\hat{L}(p)\right]. \qquad (6.3)$$

Similarly, the mse of the quantile share

$$\hat{Q}(p_1,p_2) = \frac{1}{\hat{\mu}} \sum_s w_{hci}\, y_{hci}\, I\{\hat{\xi}_{p_1} < y_{hci} \leq \hat{\xi}_{p_2}\}$$

is approximated by (6.1) using

$$u^*_{hc} = \frac{1}{\hat{\mu}} \sum_i w_{hci}\big[(y_{hci} - \hat{\xi}_{p_2})I\{y_{hci} \leq \hat{\xi}_{p_2}\} - (y_{hci} - \hat{\xi}_{p_1})I\{y_{hci} \leq \hat{\xi}_{p_1}\} + p_2\hat{\xi}_{p_2} - p_1\hat{\xi}_{p_1} - y_{hci}\,\hat{Q}(p_1,p_2)\big].$$

The Low Income Measure defined by (1.3) is estimated as

$$\hat{\Theta} = \hat{F}(\hat{M}/2) = \sum_s w_{hci}\, I\{y_{hci} \leq \hat{M}/2\}.$$

The mean squared error of the low income measure can be estimated approximately by the expression (6.1), where, (from the equation (5.1)):

$$u^*_{hc} = -\frac{\hat{f}(\hat{M}/2)}{2\hat{f}(\hat{M})} \sum_i w_{hci}[I\{y_{hci} \leq \hat{M}\} - 1/2] + \sum_i w_{hci}[I\{y_{hci} \leq \hat{M}/2\} - \hat{\Theta}].$$

7. ILLUSTRATION

The methodology above is illustrated with an application to the family income data collected in the Canadian Survey of Consumer Finance (SCF). We use the file on the Disposable Income of Economic Families obtained for the province of Ontario in 1988. Disposable income is defined as total income after tax reported in the survey. The SCF uses the framework of the Canadian Labour Force Survey which is based on a stratified, multistage design. For more details on the sample design see Singh *et al.* (1990).

We estimated the median M, the Gini coefficient G, the Low Income Measure Θ, Lorenz Curve Ordinates and quintile shares $Q(0,.2), Q(.2,.4), Q(.4,.6), Q(.6,.8)$, $Q(.8,1.0)$. Their standard errors are obtained using the proposed methodology and the jackknife 'delete-one-cluster' method.

We present a brief description of the jackknife 'delete-one-cluster' method used for this illustration. First, we assume that the estimate of the unknown parameter Θ can be expressed as $\hat{\Theta} = \mathcal{L}(\hat{F})$, where $\hat{F}$ is the estimated distribution function. The estimate of the distribution function $\hat{F}_{(hj)}$ obtained from the sample after removing the j-th sampled cluster of the h-th stratum ($j = 1, \ldots, n_h$, $h = 1, \ldots, H$) is

$$\hat{F}_{(gj)}(y) = \sum_s A_{hci}(g,j)\, w_{hci} I\{y_{hci} \leq y\}$$

where

$$A_{hci}(g,j) = \begin{cases} 1, & h \neq g; \\ \dfrac{n_g}{n_g - 1}, & h = g,\ c \neq j; \\ 0, & h = g,\ c = j. \end{cases}$$

Then $\hat{\Theta}_{(gj)} = \mathcal{L}(\hat{F}_{(gj)})$ and the resulting 'delete-one-cluster' jackknife estimator of the variance of $\hat{\Theta} = \mathcal{L}(\hat{F})$ is

$$\mathrm{var}_J(\hat{\Theta}) = \sum_{g=1}^{H} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{\Theta}_{(gj)} - \hat{\Theta})^2.$$

It is known that the jackknife variance estimator performs poorly for quantiles due to its inconsistency (Kovar *et al.* 1988). There are some recent results (Shao and Wu 1989, Rao, Wu and Yue 1992) that suggest that the 'delete d' jackknife and 'delete-one-cluster', under certain conditions, may have desirable asymptotic properties for the variance estimation of non-smooth statistics like quantiles or the low income measure. On the other hand, for statistics like the Gini coefficient the jackknife estimator of the asymptotic variance is consistent (Shao 1993).

Unlike jackknifing, the estimating equations approach is not computationally intensive. It is simple, explicit and incorporates the sample design. It provides formulae for the asymptotic variance that are easy to program despite their complicated form.

Realizing the limitations imposed by using a single sample to make an objective comparison between different methods, the purpose of this example is to point out differences in the standard errors obtained by the estimating equations approach and a computationally intensive method like the jackknifing. Results are summarized in the table below. The direction of the difference in the estimated standard errors confirms the overall conservativeness of the jackknifing method. The difference can be attributed to the upward bias of the jackknifing method in the case of the median, although the 'delete-one-cluster' jackknife is preferable to the 'delete-1' jackknife. For the quantile shares it can be partly explained by the fact that upper quantile shares may not cut over all primary sampling units but rather perform as separated classes which may affect the jackknifing more than the estimating equations method.

Table 1
Measures of Income Inequality and Their Standard Errors

Measure	Estimate	Standard Error	
		Estimating Equations Approach	Jackknifing 'Delete-One-Cluster'
Median	31705	303.3	569.8
Gini	0.3482	0.005	0.005
Low Income Measure	0.1980	0.00586	0.00613
Lorenz Curve Ordinates			
L(0.2)	0.0561	0.00137	0.00175
L(0.4)	0.1745	0.00166	0.00194
L(0.6)	0.3522	0.00246	0.00285
L(0.8)	0.5982	0.00317	0.00393
Quintile Shares			
Q(0, 0.2)	0.0561	0.00137	0.00167
Q(0.2, 0.4)	0.1186	0.00159	0.00221
Q(0.4, 0.6)	0.1775	0.00157	0.00282
Q(0.6, 0.8)	0.2461	0.00158	0.00337
Q(0.8, 1.0)	0.4017	0.00395	0.00451

8. SUMMARY

The problem of estimating the variance of complex statistics, such as measures of income inequality, have eluded statisticians for years. Replication methods such as the jackknife are often suggested for estimation. The advantage of the linearization approach is that it can be used under a wide class of sampling designs and does not suffer from the need for intensive computations which methods such as the bootstrap entail. Through the method of estimating functions and the decomposition given in (2.3), we find that some difficult problems can be solved more easily. A discussion about the order of the remainder term for some of these measures is given as well. A more rigorous proof for a complex sample design can be established along the lines given in Shao and Rao (1994).

REFERENCES

BEACH, C.M., and DAVIDSON, R. (1983). Distribution-free statistical inference with Lorenz curves and income shares. *Review of Economic Studies*, 50, 723-735.

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

BINDER, D.A. (1991). Use of estimating functions for interval estimation from complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 34-42.

BINDER, D.A., and PATAK, Z. (1994). Use of estimating functions for interval estimation from complex surveys. *Journal of the American Statistical Association*, 89, 1035-1044.

FRANCISCO, C.A., and FULLER, W.A. (1986). Estimation of the Distribution function with a complex survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 37-45.

FRANCISCO, C.A., and FULLER, W.A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, 19, 454-469.

GLASSER, G.J. (1962). Variance formulas for the mean difference and coefficient of concentration. *Journal of the American Statistical Association*, 57, 648-654.

KOVAR, J.G., RAO, J.N.K., and WU, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics*, 16, 25-45.

NYGÅRD, F., and SANDSTRÖM, A. (1981). *Measuring Income Inequality*. Stockholm: Almqvist and Wiksell International.

RANDLES, R.H. (1982). On the Asymptotic Normality of Statistics with Estimated Parameters. *Annals of Statistics*, 10, 462-474.

RAO, J.N.K., and WU, C.F.J. (1987). Methods for Standard Errors and Confidence Intervals from Survey Data: Some Recent Work. *Proceedings of the 46th session, International Statistical Institute*, 3, 5-19.

RAO, J.N.K., WU, C.F.J., and YUE, K. (1992). Some Recent Work on Resampling Methods for Complex Surveys. *Survey Methodology*, 18, 209-217.

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

SENDLER, W. (1979). On statistical inference in concentration measurement. *Metrika*, 26, 109-122.

SHAO, J., and RAO, J.N.K. (1994). Standard Errors for Low Income Proportions Estimated from Stratified Multi-Stage Samples. *Sankhyā, B*, (to appear).

SHAO, J. and WU, C.W.J. (1989). A general Theory for Jackknife Variance Estimation. *Annals of Statistics*, 17,1176-1197.

SHAO, J. (1993). Inferences Based on *L*-statistics in Survey Problems: Lorenz Curve, Gini Family and Poverty Proportion. In *Proceedings of the Workshop on Statistical Issues in Public Policy Analysis*, Carleton University and University of Ottawa.

SINGH, M.P., DREW, J.D., GAMBINO, J.G., and MAYDA, F. (1990). *Methodology of the Canadian Labour Force Survey*, Catalogue No. 71-526, Statistics Canada.

WOODRUFF, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47,635-646.

Survey Methodology, December 1995
Vol. 21, No. 2, pp. 147–157
Statistics Canada

A Reduced-Size Transportation Algorithm for Maximizing the Overlap Between Surveys

LAWRENCE R. ERNST and MICHAEL M. IKEDA[1]

ABSTRACT

When redesigning a sample with a stratified multi-stage design, it is sometimes considered desirable to maximize the number of primary sampling units retained in the new sample without altering unconditional selection probabilities. For this problem, an optimal solution which uses transportation theory exists for a very general class of designs. However, this procedure has never been used in the redesign of any survey (that the authors are aware of), in part because even for moderately-sized strata, the resulting transportation problem may be too large to solve in practice. In this paper, a modified reduced-size transportation algorithm is presented for maximizing the overlap, which substantially reduces the size of the problem. This reduced-size overlap procedure was used in the recent redesign of the Survey of Income and Program Participation (SIPP). The performance of the reduced-size algorithm is summarized, both for the actual production SIPP overlap and for earlier, artificial simulations of the SIPP overlap. Although the procedure is not optimal and theoretically can produce only negligible improvements in expected overlap compared to independent selection, in practice it gave substantial improvements in overlap over independent selection for SIPP, and generally provided an overlap that is close to optimal.

KEY WORDS: Linear programming; Sample redesign; Survey of Income and Program Participation.

1. INTRODUCTION

The problem of maximizing the expected number of primary sampling units (PSUs) retained in sample when redesigning a survey with a stratified design for which the PSUs are selected with probability proportional to size was introduced to the literature by Keyfitz (1951). Typically, the motivation for maximizing the overlap of PSUs is to reduce additional costs, such as the training of a new interviewer for a household survey, incurred with each change of sample PSU. Procedures for maximizing overlap do not alter the unconditional probability of selection for a set of PSUs in a new stratum, but conditions its probability of selection in such a manner that the probability of a PSU being selected in the new sample is generally greater than its unconditional probability when the PSU was in the initial sample and less otherwise.

Overlap procedures are applicable when the redesign results in either a restratification of the PSUs or a change in their selection probabilities. Keyfitz (1951) presented an optimal procedure, but only for one-PSU-per-stratum designs in the special case when the initial and new strata are identical, with only the selection probabilities changing. Causey, Cox and Ernst (1985) obtained an optimal solution to the overlap problem under very general conditions by formulating it as a transportation problem, which is a special form of linear programming problem. This procedure imposes no restrictions on changes in strata definitions or number of PSUs per stratum. (A similar result had been independently obtained by Arthanari and Dodge (1981), although they did not discuss the issue of changes in strata definitions. Both sets of authors obtained their results by generalizing work of Raj (1968).) However, there are at least two other difficulties with the procedure of Causey, Cox and Ernst which can make it unusable in practice, one which is the focus of Ernst (1986), and the other the focus of the current paper.

The first difficulty is that, if the initial sample of PSUs was not selected independently from stratum to stratum, the information necessary to compute all the joint probabilities required by this method may not be available in practice. An alternative linear programming procedure, for use in such cases, was developed by Ernst (1986). The Bureau of the Census has used linear programming to overlap its demographic surveys on five occasions. On four of these occasions (the selection of the 1980s and 1990s Current Population Survey (CPS) designs, and the 1980s and 1990s National Crime Victimization Survey (NCVS) designs) the procedure in Ernst (1986) was used because the initial design was not selected independently from stratum to stratum. In particular, as explained in Ernst (1986), if the initial sample was itself selected by overlapping with a still earlier design then this independence assumption generally does not hold, which was the key reason why it did not hold for these four redesigns.

The second difficulty with the optimal procedure is that the transportation problem may be too large to solve in practice. The Bureau of the Census also used linear

[1] Lawrence R. Ernst, Chief, Research Group, Office of Compensation and Working Conditions, Bureau of Labor Statistics, Washington, DC 20212, U.S.A.; Michael M. Ikeda, Mathematical Statistician, Statistical Research Division, Bureau of the Census, Washington, DC 20233, U.S.A.

programming to overlap the 1990s Survey of Income and Program Participation (SIPP) design with the 1980s SIPP design, both two-PSUs-per-stratum designs. The initial sample for SIPP was selected independently from stratum to stratum. However, the transportation problem for the optimal procedure would have been too large to practically solve for many strata. This is because for each new stratum to be overlapped consisting of n PSUs, the number of variables in the transportation problem for the optimal procedure can be as large as $2^n \times \binom{n}{2}$. The largest value of n for which a transportation problem with that many variables can be solved with the computer facilities that we have used is approximately $n = 15$.

This paper presents a reduced-size formulation of the overlap procedure as a transportation problem which decreases the numbers of variables in the SIPP problem to $\left(\binom{n}{2} + n + 1\right) \times \binom{n}{2}$, a striking reduction for moderate to large values of n. The procedure assumes that the initial sample was selected independently from stratum to stratum, and hence could not have been used instead of the procedure of Ernst (1986) to overlap the CPS and NCVS designs. This reduced-size procedure has been successfully run for strata with as many as 68 PSUs. In contrast, for $n = 68$, the $2^{68} \times \binom{68}{2}$ possible number of variables for the unreduced formulation is far beyond the size of problem that can be solved by any current computer. Furthermore, though the reduced-size procedure sacrifices optimality in exchange for its size reduction, it does appear in practice to yield results fairly close to optimal, as we will show. The reduced-size procedure is the procedure that was used to overlap SIPP.

In Section 2 the procedure of Causey, Cox and Ernst (1985) is reviewed, to provide background for the presentation of the reduced-size procedure.

The reduced-size procedure is presented in Section 3. Although the approach has general applicability, for ease of presentation it is only described in detail for the case when both the initial and new designs are two-PSUs-per-stratum without replacement. A small, artificial example of the reduced-size procedure is also presented in Section 3. This example serves to illustrate the procedure and to demonstrate that the ordering of the pairs of PSUs in a new design stratum, a key step in the algorithm, affects the expected overlap. We also outline in this section some analytical results on the comparison between the reduced-size procedure and the optimal procedure. Upper bounds on the loss in expected overlap from using the reduced-size procedure instead of the optimal procedure are stated. It is also explained that in certain situations this loss can approach two PSUs for two-PSUs-per-stratum designs, the worst possible situation. Further details and proofs of the results in this section as well as some results in other sections are presented in Ernst and Ikeda 1994.

In Section 4 the performance of the reduced-size procedure is presented, both for the actual SIPP production overlap and for earlier, artificial simulations of the SIPP overlap. The expected overlap for this procedure is compared to that for independent selection of the new sample PSUs and to an upper bound on the optimal expected overlap. The results show that for this application, in contrast with some of the theoretical results described in Section 3, the expected overlap with the reduced-size procedure is much larger than if independent selection had been used to select the new sample PSUs, and nearly as large as the optimal expected overlap. Also presented are computer running times for the reduced-size procedure as a function of stratum size.

Finally, our conclusions are stated in Section 5.

2. REVIEW OF THE OVERLAP PROCEDURE OF CAUSEY, COX AND ERNST (1985)

The overlap procedure of Causey, Cox and Ernst (1985), like all overlap procedures, conditions the selection of sample PSUs in each new stratum in some way on which PSUs in the stratum were in the initial sample. This particular overlap procedure attains true optimality by making complete use of this information and formulating the procedure as a transportation problem. We proceed to present this procedure.

First, however, we introduce some notation that will be used throughout the paper. Let S denote a stratum in the new design. Each such stratum corresponds to a separate overlap problem. Let n denote the number of PSUs in S and let $A_1, \ldots, A_n$ denote the PSUs in S. Let I denote the random subset of $\{1, \ldots, n\}$ such that $k \in I$ if and only if A_k was in the initial sample, and let N denote the corresponding set with respect to the new sample. For example, if A_2 and A_3 were the PSUs in S that were in the initial sample and A_1 and A_3 are the PSUs in the new sample, then $I = \{2,3\}$ and $N = \{1,3\}$. Let m^*, n^* denote the number of possible values for I and N, respectively. Let J_i, $i = 1, \ldots, m^*$, denote the possible values for I and let S_j, $j = 1, \ldots, n^*$, denote the possible values for N. The goal of all overlap procedures is to maximize the expected number of PSUs in $N \cap I$, while preserving the values of the $P(S_j)$'s.

To illustrate some of these concepts further, consider an example for which $n = 3$. Then $n^* = 3$ if the new design is either 1 or 2 PSUs per stratum with the values for N, that is the S_j's, consisting of $\{1\},\{2\},\{3\}$ in the 1 PSU per stratum case and $\{1,2\},\{1,3\},\{2,3\}$ in the two PSUs per stratum case. Suppose PSUs A_1 and A_2 were in one initial stratum and PSU A_3 was in another initial stratum and there were three PSUs in each of these initial strata. If the initial design was 1 PSU per stratum, then $m^* = 6$, with the values of I, that is the J_i's, consisting of $\varnothing, \{1\},\{2\},\{3\},\{1,3\},\{2,3\}$; if the initial design was 2 PSUs per stratum then $m^* = 6$, with the J_i's consisting of $\{1\},\{2\},\{1,2\},\{1,3\},\{2,3\},\{1,2,3\}$.

We now present the transportation problem for the overlap procedure of Causey, Cox and Ernst (1985). Abbreviate by $P(J_i)$ the probability that $I = J_i$ and by $P(S_j)$ the probability that $N = S_j$. In addition, let x_{ij} be the variable denoting the joint probability of these two events, and let c_{ij} denote the number of elements in $J_i \cap S_j$. The $P(J_i)$'s, $P(S_j)$'s and c_{ij}'s are known values, while the x_{ij}'s are variables for which the optimal values are to be determined. Then the transportation problem to solve is to determine $x_{ij} \geq 0$ which maximize

$$\sum_{i=1}^{m^*} \sum_{j=1}^{n^*} c_{ij} x_{ij} \tag{2.1}$$

subject to

$$\sum_{j=1}^{n^*} x_{ij} = P(J_i), \quad i = 1, \ldots, m^*, \tag{2.2}$$

$$\sum_{i=1}^{m^*} x_{ij} = P(S_j), \quad j = 1, \ldots, n^*. \tag{2.3}$$

Note that in this transportation problem, the objective function (2.1) is the expected number of PSUs in S that are in $N \cap I$. Also note that the constraints (2.2) and (2.3) are required by the definitions of the $P(J_i)$'s, $P(S_j)$'s and the x_{ij}'s.

Once the optimal x_{ij}'s have been obtained, the conditional probability that $N = S_j$ given that $I = J_i$ is then $x_{ij}/P(J_i)$ for all i,j.

We present an example to illustrate the use of the formulation (2.1)-(2.3) in the case where both the initial and new designs are two-PSUs-per-stratum without replacement. In this example, and throughout the paper, p_i, π_i denote the predetermined probability that $i \in I$ and $i \in N$, respectively.

Consider a final stratum S with $n = 3$. All of the PSUs were in different initial strata. Let $p_1 = .6$, $p_2 = .75$, $p_3 = .7$, $\pi_1 = .5$, $\pi_2 = .8$, $\pi_3 = .7$. Since the PSUs were all in different initial strata, there are 8 different possibilities for I, with probabilities given in Table 1.

Table 1
Probabilities for Possible Sets of Initial Sample PSUs

i	1	2	3	4	5	6	7	8
J_i	{1,2,3}	{1,2}	{1,3}	{2,3}	{1}	{2}	{3}	$\varnothing$
$P(J_i)$	.315	.135	.105	.21	.045	.09	.07	.03

Since the new design is two-PSUs-per-stratum without replacement, there are 3 different possibilities for N, namely the pairs $S_1 = \{1,2\}$, $S_2 = \{1,3\}$, $S_3 = \{2,3\}$, and hence $P(S_1) = .30$, $P(S_2) = .20$, $P(S_3) = .50$.

Furthermore, the values of c_{ij} are then as given in Table 2. Upon maximizing (2.1) subject to (2.2) and (2.3) with the given $P(J_i)$'s, $P(S_j)$'s and c_{ij}'s, an optimal set of x_{ij}'s, presented in Table 2, is obtained. Finally, by dividing each of the x_{ij} entries in row i of Table 2 by $P(J_i)$, an optimal set of conditional probabilities $P(S_j \mid J_i)$, is obtained. For example, since $x_{12} = .025$ and $P(J_1) = .315$, it follows that $P(S_2 \mid J_1) = 5/63$.

Table 2
Values of c_{ij} and Values of x_{ij} that Maximize Overlap for Optimal Procedure

	c_{ij}			x_{ij}		
	j			j		
i	1	2	3	1	2	3
1	2	2	2	.000	.025	.290
2	2	1	1	.135	.000	.000
3	1	2	1	.000	.105	.000
4	1	1	2	.000	.000	.210
5	1	1	0	.045	.000	.000
6	1	0	1	.090	.000	.000
7	0	1	1	.000	.070	.000
8	0	0	0	.030	.000	.000

For this example, as can be computed from (2.1) and Table 2, the expected overlap under the optimal procedure is 1.735 PSUs. In comparison, the expected overlap if the initial and final designs are selected independently is $p_1\pi_1 + p_2\pi_2 + p_3\pi_3 = 1.39$ PSUs.

For two-PSU-per-stratum without replacement problems, the possible values for N are always the $\binom{n}{2}$ subsets of $\{1, \ldots, n\}$ of size 2, that is $n^* = \binom{n}{2}$. However m^* can vary widely. $m^* = \binom{n}{2}$ when the PSUs in S comprise a single initial stratum. The upper bound of 2^n on m^* is attained when all the PSUs in S were in different initial strata, as illustrated by the previous example, and in some other situations. A general, exact expression for m^* is presented in Ernst and Ikeda (1994).

For the two-PSUs-per-stratum without replacement overlap problem, the number of variables in the transportation problem for the optimal procedure is m^*n^* which can be as large as $2^n\binom{n}{2}$. For $n = 15$, $2^n\binom{n}{2} = 3{,}440{,}640$, which is about as large a transportation problem as can be solved with the computer facilities that we used. However, $n > 15$ for nearly half the nonselfrepresenting strata (that is strata consisting of noncertainty PSUs) in our SIPP application, and consequently it was necessary to develop a procedure, described in the next section, which reduces the size of the transportation problem, while still producing nearly maximal expected overlap in practice.

3. THE ALGORITHM FOR THE REDUCED-SIZE PROCEDURE

Previous work on reducing the size of the transportation problem (2.1)–(2.3) has focused on accomplishing the size reduction while retaining optimality. For example, the approach of Aragon and Pathak (1990) retains optimality and reduces the size of the problem by 75 percent when $m^* = n^*$. Unfortunately, when m^* is much larger than n^*, which is when size reduction is most needed, their method produces negligible size reduction in relative terms. A generalization of this approach is presented in Pathak and Fahimi (1992), but there is no indication that their procedure always yields a size reduction that is substantial in relative terms.

In this section a reduced-size procedure is presented which takes a different approach. We sacrifice optimality, at least in theory, in return for an assured size reduction down to a manageable size transportation problem. This size reduction is accomplished, in the case when the initial and new designs are both two PSUs per stratum for example, by ordering all pairs of PSUs in a new stratum and then conditioning the new selection probabilities for any initial set of sample PSUs of size greater than 2 on the first pair of PSUs in the ordering contained in the initial set, rather than conditioning on the entire initial set. That is, each possible initial set of sample PSUs which consists of more than 2 PSUs is combined with a set of size 2. As illustrated in Section 4, this procedure may yield a near optimal overlap in practice; particularly with an appropriate ordering of the pairs of PSUs, as described in Section 3.1.2.

The reduced-size procedure is applicable whenever PSUs in the initial and new designs are selected without replacement. However, the procedure will be described in detail, in Section 3.1, only for the case when both the initial and new designs are two-PSUs-per-stratum. Then, in Section 3.2, the changes necessary to apply this procedure for other initial and new designs will be sketched. Finally, in Section 3.3, some analytical results are outlined on the relationships among the expected overlap for the reduced-size procedure, the optimal procedure and independent selection. It is assumed throughout this section that PSUs in the initial sample were selected independently from stratum to stratum.

3.1 Reduced-Size Procedure When Both Designs Are Two-PSUs-Per-Stratum

The reduced-size procedure to be described includes the following key aspects: the specific ordering of the pairs of PSUs; the reformulation of the transportation problem (2.1)–(2.3) for the reduced size procedure; the computation of the probabilities for the initial outcomes for this formulation; and the computation of the cost coefficients (the c_{ij}'s) in the objective function. In Section 3.1.1 we present a detailed outline of the reduced-size procedure, including the reformulated transportation problem. The ordering of the pairs is described in Section 3.1.2. Finally, the computation of the probabilities for the initial outcomes and the cost coefficients are given in Section 3.1.3.

3.1.1 General Outline of the Procedure

The general outline of the procedure is as follows. First, the $\binom{n}{2}$ subsets of $\{1, \ldots, n\}$ of size 2 are ordered in a manner to be described later. (For now, we simply note that any ordering can be used to reduce the size of the transportation problem. The specific one used is for the purpose of accomplishing the size reduction while also attempting to give up as little as possible of the gains in overlap that the optimal procedure yields.) We let I_i, $i = 1, \ldots, \binom{n}{2}$, denote the i-th element in the ordering; let $I_{\binom{n}{2}+1}, \ldots, I_{\binom{n}{2}+n}$ be the n singleton subsets; and set $I_{\binom{n}{2}+n+1} = \varnothing$. Thus, the I_i's constitute all subsets of $\{1, \ldots, n\}$ of 2 or fewer elements. For each possibility for I, a unique set I^* is associated among these $\binom{n}{2} + n + 1$ subsets and the new selection probabilities conditioned on the associated I^*, rather than on I itself. Therefore, the new selection probabilities are conditioned on $\binom{n}{2} + n + 1$ events instead of a possible 2^n events, which is the reason for the size reduction. The associated I^* is the first I_i for which $I_i \subset I$. That is, if I consists of at least two integers, the associated I^* is the first pair in the ordering contained in I, while if I is a singleton set or empty then $I^* = I$.

The reduced-size transportation problem attempts to retain the PSUs corresponding to elements in the associated set I^* in the new sample, but does not use information on elements in $I \sim I^*$. The form of this reduced-sized transportation problem based on the set of I_i's is as follows. Let p_i^* be the probability that $I^* = I_i$, $i = 1, \ldots, \binom{n}{2} + n + 1$, and abbreviate $\pi_j^* = P(S_j)$, $j = 1, \ldots, \binom{n}{2}$. For each i,j, the variable x_{ij} is the joint probability that $I^* = I_i$ and that $N = S_j$, while c_{ij} is the expected number of elements in $I \cap S_j$ given $I^* = I_i$. The problem to solve is to determine $x_{ij} \geq 0$ that maximize

$$\sum_{i=1}^{\binom{n}{2}+n+1} \sum_{j=1}^{\binom{n}{2}} c_{ij} x_{ij}, \tag{3.1}$$

subject to

$$\sum_{j=1}^{\binom{n}{2}} x_{ij} = p_i^*, \quad i = 1, \ldots, \binom{n}{2} + n + 1, \tag{3.2}$$

$$\sum_{i=1}^{\binom{n}{2}+n+1} x_{ij} = \pi_j^*, \quad j = 1, \ldots, \binom{n}{2}. \tag{3.3}$$

Once the optimal x_{ij}'s have been obtained, then the conditional new selection probabilities for $S_j, j = 1, \ldots, \binom{n}{2}$, given $I^* = I_i$, are x_{ij}/p_i^*. Note that the number of variables, x_{ij}, in the formulation (3.1)-(3.3) is $\left(\binom{n}{2} + n + 1\right) \times \binom{n}{2}$, in comparison with a maximum of $2^n \times \binom{n}{2}$ in the formulation (2.1)-(2.3).

It remains to explain the general method for obtaining the ordering of the $\binom{n}{2}$ pairs and the procedures for computing the p_i^*'s and c_{ij}'s. Before doing this, we present an example of the reduced-size procedure, namely the two-PSUs-per-stratum example used in Section 2 to illustrate the transportation problem formulation for the optimal procedure.

The ordering of the pairs for this example, as will be shown later, is $\{2,3\}$, $\{1,2\}$, $\{1,3\}$. Consequently, the I_i's, are as given in Table 3. Note that if $I = \{1,2,3\}$ or $I = \{2,3\}$, then the associated set is $I_1 = \{2,3\}$. For the other six possibilities for I the associated set is I itself.

Consequently, from Table 1 we obtain that

$$p_1^* = P(I = \{1,2,3\}) + P(I = \{2,3\}) = .525, \quad (3.4)$$

$p_i^* = P(J_i), i = 2,3$, and $p_i^* = P(J_{i+1}), i = 4, \ldots, 7$, yielding the values in Table 3. Since $\pi_j^* = P(S_j)$, we have $\pi_1^* = .30$, $\pi_2^* = .20$, $\pi_3^* = .50$.

Table 3
Probabilities of Associated Sets: Reduced-Size Procedure

	i						
	1	2	3	4	5	6	7
I_i	{2,3}	{1,2}	{1,3}	{1}	{2}	{3}	∅
p_i^*	.525	.135	.105	.045	.09	.07	.03

The c_{ij} values for this example are given in Table 4. In order to obtain these values, we simplified the computation by letting

$$b_{it} = P(t \in I \mid I^* = I_i),$$

$$i = 1, \ldots, \binom{n}{2} + n + 1, \quad t = 1, \ldots, n, \quad (3.5)$$

and noting that if $S_j = \{s,t\}$ then

$$c_{ij} = b_{is} + b_{it}. \quad (3.6)$$

That is, the expected number of elements in $I \cap S_j$ given $I^* = I_i$ is simply the sum of the probabilities that each of the two elements in S_j was in I given $I^* = I_i$. Also observe that while the transportation problem for the optimal procedure knows the exact value for I and hence knows with certainty whether each element in S_j was in I, this is not the case for the reduced-size procedure, since only the associated set I_i is known. To illustrate, consider the first row of Table 4. Since $I_1 = \{2,3\}$, we know that $2 \in I$ and $3 \in I$, and hence $b_{12} = b_{13} = 1$. However, we do not with certainty whether $1 \in I$ since I_1 is the associated set for both $I = \{1,2,3\}$ and $I = \{2,3\}$. In fact, from Table 1,

$$b_{11} = \frac{P(I = \{1,2,3\})}{P(I = \{1,2,3\}) + P(I = \{2,3\})} = .6.$$

Then $c_{11} = b_{11} + b_{12} = 1.6$, with c_{12}, c_{13} computed similarly. For the remaining six rows in Table 4, $I_i = I$ and hence it is known with certainty which integers were in I. Consequently, the c_{ij}'s for these six rows are easily computed.

Finally, we maximize the expected overlap (3.1) subject to (3.2) and (3.3), obtaining the x_{ij} values in Table 4. The conditional probabilities $P(N = S_j \mid I^* = I_i)$ in Table 5 are then obtained by dividing each of the x_{ij} entries in the i-th row of Table 4 by p_i^*.

Table 4
Values of c_{ij} and Values of x_{ij} that Maximize Overlap for the Reduced-Size Procedure

		c_{ij}			x_{ij}		
		j			j		
i	I_i	1	2	3	1	2	3
1	{2,3}	1.6	1.6	2.0	0.000	0.025	0.500
2	{1,2}	2.0	1.0	1.0	0.135	0.000	0.000
3	{1,3}	1.0	2.0	1.0	0.000	0.105	0.000
4	{1}	1.0	1.0	0.0	0.045	0.000	0.000
5	{2}	1.0	0.0	1.0	0.090	0.000	0.000
6	{3}	0.0	1.0	1.0	0.000	0.070	0.000
7	∅	0.0	0.0	0.0	0.030	0.000	0.000

Table 5
Conditional Probabilities for the Reduced-Size Procedure

		j		
i	I_i	1	2	3
1	{2,3}	0	1/21	20/21
2	{1,2}	1	0	0
3	{1,3}	0	1	0
4	{1}	1	0	0
5	{2}	1	0	0
6	{3}	0	1	0
7	∅	1	0	0

The expected overlap for the reduced-size procedure is .01 less than optimal, that is 1.725 PSUs. The deviation from optimality arises solely because the expected overlap is 1.6 for the joint event that $I^* = \{2,3\}$ and $N = \{1,3\}$. Since the probability of this joint event is .025, and the optimal procedure for this example always produces an overlap of 2 when at least 2 of the PSUs were in the initial sample, the deviation from optimality is $.025(2 - 1.6) = .01$.

The reason that the reduced-size procedure is not able to obtain optimality is that the pair $\{2,3\}$ has a smaller probability of selection in the new sample than in the initial sample. As a result, both the optimal procedure and the reduced-size procedure must sometimes select another pair (always $\{1,3\}$ for both procedures in this example) when $\{2,3\}$ was in the initial sample. The distinction between the two procedures is that the optimal procedure only selects $\{1,3\}$ when $1 \in I$. The reduced-size procedure is unable to use the information about whether $1 \in I$. As a result, when $\{2,3\} \subset I$, $1 \in N$ independently of whether $1 \in I$. This results in a deviation from the optimal overlap.

3.1.2 The Ordering of the Pairs

We now proceed to show in general how the ordering of the pairs is obtained. We use the additional notation here that p_{st}, π_{st}, s, $t = 1, \ldots, n$, $s \neq t$, is the joint probability that s, $t \in I$ and s, $t \in N$, respectively.

The motivation for the ordering of the pairs is as follows. If the i-th pair in the ordering is $\{s,t\}$ then it would be possible for the transportation problem to retain this pair in the new sample when $I^* = I_i$ with conditional probability $\min\{1, \pi_{st}/p_i^*\}$. (The conditional retention probability cannot be any higher than this, since a higher value would result in an unconditional selection probability for the pair in the new design exceeding π_{st}.) Therefore, roughly the goal in the ordering is to make these conditional probabilities as large as possible on average over all pairs.

To illustrate how the ordering of the pairs affects the expected overlap we consider the example of Table 3. Our ordering procedure, as will be shown later, produces the indicated ordering and yields an expected overlap of 1.725 PSUs. Next consider the following alternative ordering for this example. Let the first pair in the ordering be $\{1,3\}$, the second pair be $\{1,2\}$ and the last pair be $\{2,3\}$. With this alternative ordering, $I^* = \{1,3\}$ whenever either $I = \{1,2,3\}$ or $I = \{1,3\}$. Therefore, for this ordering p_1^* is the probability that $I^* = \{1,3\}$, which is now .42. Furthermore, for this alternative ordering, $p_3^* = P(I^* = \{2,3\}) = P(I = \{2,3\}) = .21$, while the other 5 columns in Table 3 remain unchanged. The alternative ordering results in a table of conditional probabilities similar to Table 5, except that in row 1 the I_j, $j = 2$ and $j = 3$ columns now become $\{1,3\}$, 10/21 and 11/21, respectively, and in row 3 the corresponding columns are now $\{2,3\}$, 0 and 1, respectively.

It can be calculated, using the same approach used for the original ordering that the expected overlap for the alternative ordering is 0.055 less than optimal, that is 1.68 PSUs. The reason that this alternative ordering results in a lower expected overlap is as follows. In general a later placement of a pair in the ordering, results in a lower value for the corresponding p_i^*, and hence a higher conditional retention probability when $I^* = I_i$. That is, with $\{1,3\}$ first in the ordering, $\pi_{13}/p_1^* = 10/21$, which is the conditional retention probability for this pair when $I^* = \{1,3\}$; while when $\{1,3\}$ is third in the ordering, $\pi_{13}/p_3^* > 1$ and this pair is retained with certainty. Now the conditional retention probability for the pair $\{2,3\}$ when $I^* = \{2,3\}$ also increases to 1 when $\{2,3\}$ is moved from first to third in the ordering, but the increase is only from 20/21, and hence the original ordering in Table 3 produces a higher expected overlap than the alternative ordering.

Thus, as this example illustrates, the goal of the ordering is to place pairs earlier in the ordering that have a relatively high conditional retention probability even with an early placement. To obtain the desired ordering of the pairs of integers, an ordering $f(1), \ldots, f(n)$ of $\{1, \ldots, n\}$ will first be obtained by recursion. Then corresponding to each $k = 1, \ldots, n-1$, an ordering $g_k(1), \ldots, g_k(n-k)$ of $\{1, \ldots, n\} \sim \{f(1), \ldots, f(k)\}$ will be constructed by recursion. A linear ordering of the distinct pairs in $\{1, \ldots, n\}$ would then be determined as follows. Each such pair can be represented uniquely as an ordered pair $(f(k), g_k(\ell))$ for some $k \in \{1, \ldots, n-1\}$, $\ell \in \{1, \ldots, n-k\}$. A second pair representable in the form $(f(k'), g_{k'}(\ell'))$ precedes $(f(k), g_k(\ell))$ if and only if either $k' < k$, or $k' = k$ and $\ell' < \ell$. To illustrate, for the example just considered it will be shown later that $f(1) = 2$, $f(2) = 3$, $f(3) = 1$, $g_1(1) = 3$, $g_1(2) = 1$, $g_2(1) = 1$, and hence the ordering of the pairs is $\{2,3\}$, $\{2,1\}$, $\{3,1\}$. Both the f ordering and the g_k ordering will be constructed to meet the goal stated at the beginning of this paragraph.

To obtain the ordering $f(1), \ldots, f(n)$, recursively define $f_1(k)$, $k = 1, \ldots, n$, by choosing $f(k) \in T_k$ satisfying

$$\pi_{f(k)}/p_{f(k)}^{(k)} = \max\{\pi_i/p_i^{(k)} : i \in T_k\},$$

where

$$T_1 = \{1, \ldots, n\}, \quad T_k = T_{k-1} \sim \{f(k-1)\},$$
$$k = 2, \ldots, n, \quad p_i^{(k)} = P(i \in I \text{ and } I \subset T_k),$$
$$k = 1, \ldots, n, \quad i \in T_k. \qquad (3.7)$$

Since $p_i^{(1)} = p_i$, the ordering just defined corresponds to placing first a PSU with the greatest value of π_i/p_i^*. For all k, $p_{f(k)}^{(k)}$ is the probability that $f(k)$ was in I and none of the $k-1$ elements preceeding $f(k)$ in the f ordering were in I, and hence $p_{f(k)}^{(k)}$ is the probability that

an attempt is made to retain $A_{f(k)}$ in the new sample either as the first member of an ordered pair of initial sample PSUs or as the only initial sample PSU in S. Generally, the larger $\pi_{f(k)}/p_{f(k)}^{(k)}$ is, the greater the probability that this attempt would be successful. Thus, the motivation for the f ordering of the individual PSUs is the analog of the motivation for the ordering of the pairs of PSUs that we previously discussed.

It remains to explain how to compute $p_i^{(k)}$ for $k \geq 2$. To this end, let r denote the number of initial strata with PSUs in common with S and let F_α, $\alpha = 1, \ldots, r$, denote a partition of $\{1, \ldots, n\}$ such that i and j are in the same F_α if and only if A_i and A_j were in the same initial stratum. Then let

$$p'_\alpha(T) = P(I \cap F_\alpha \subset T), \quad \alpha = 1, \ldots, r, \quad T \subset \{1, \ldots, n\}, \tag{3.8}$$

$$p''_{i\alpha}(T) = P(i \in I \text{ and } I \cap F_\alpha \subset T), \quad \alpha = 1, \ldots, r, \quad T \subset \{1, \ldots, n\}, \quad i \in F_\alpha \cap T, \tag{3.9}$$

and observe that

$$p'_\alpha(T) = 1 - \sum_{i \in F_\alpha \sim T} p_i + \sum_{\substack{i,j \in F_\alpha \sim T \\ i<j}} p_{ij}, \tag{3.10}$$

$$p''_{i\alpha}(T) = p_i - \sum_{j \in F_\alpha \sim T} p_{ij}, \tag{3.11}$$

and finally, as established in Ernst and Ikeda (1994),

$$p_i^{(k)} = p''_{i\alpha}(T_k) \prod_{\substack{\ell=1 \\ \ell \neq \alpha}}^{r} p'_\ell(T_k), \quad k = 1, \ldots, n, \quad i \in F_\alpha \cap T_k. \tag{3.12}$$

Next, for each $k = 1, \ldots, n-1$, the ordering $g_k(\ell)$, $\ell = 1, \ldots, n-k$, is recursively defined by choosing $g_k(\ell) \in T_{k\ell}$ satisfying

$$\pi_{f(k),g_k(\ell)}/p_{f(k),g_k(\ell)}^{(\ell)} = \max\{\pi_{f(k),j}/p_{f(k),j}^{(\ell)} : j \in T_{k\ell}\},$$

where

$$T_{k1} = \{1, \ldots, n\} \sim \{f(1), \ldots, f(k)\},$$

$$T_{k\ell} = T_{k(\ell-1)} \sim \{g_k(\ell-1)\}, \quad \ell = 2, \ldots, n-k,$$

$$T^*_{k\ell} = T_{k\ell} \cup \{f(k)\}, \quad \ell = 1, \ldots, n-k,$$

$$p_{f(k),j}^{(\ell)} = P(f(k), \quad j \in I \text{ and } I \subset T^*_{k\ell}), \quad \ell = 1, \ldots, n-k, \quad j \in T_{k\ell}. \tag{3.13}$$

Note that $p_{f(k),j}^{(\ell)}$ is thus the joint probability that $f(k)$ is the first integer in the f ordering in I, that none of the first $\ell - 1$ integers in the g_k ordering are in I, and that $j \in I$. Consequently, $p_{f(k),g_k(\ell)}^{(\ell)}$ is the probability that $I^* = \{f(k), g_k(\ell)\}$. Furthermore, if $I_i = \{f(k), g_k(\ell)\}$ then $p_i^* = p_{f(k),g_k(\ell)}^{(\ell)}$, and hence the choice of $g_k(\ell)$ results in the largest value of $\pi_{f(k),g_k(\ell)}/p_i^*$ among the elements in $T_{k\ell}$ in accordance with the previously stated goal for the ordering of the pairs of PSUs.

To compute $p_{f(k),j}^{(\ell)}$, it is established in Ernst and Ikeda (1994) that if $f(k) \in F_\alpha$, $j \in F_\beta$, then

$$p_{f(k),j}^{(\ell)} = p_{f(k),j} \prod_{\substack{t=1 \\ t \neq \alpha}}^{r} p'_t(T^*_{k\ell}) \text{ if } \alpha = \beta,$$

$$= p''_{f(k),\alpha}(T^*_{k\ell}) p''_{j\beta}(T^*_{k\ell}) \prod_{\substack{t=1 \\ t \neq \alpha,\beta}}^{r} p'_t(T^*_{k\ell}) \text{ if } \alpha \neq \beta. \tag{3.14}$$

We illustrate the computations used in obtaining the ordering for the example that we have been considering. First note that $f(1) = 2$ since the largest value of π_i/p_i occurs for $i = 2$. Next we find $g_1(1)$ which, since $f(1) = 2$, is the $j \in \{1,3\}$ with the maximum value of $\pi_{2j}/p_{2j}^{\{1\}}$. To find this j, first let $F_\alpha = \{\alpha\}$, $\alpha = 1,2,3$, and note that $T^*_{11} = \{1,2,3\}$. From (3.14) with $\alpha = 2$, $\beta = 1$, it then follows that

$$p_{21}^{\{1\}} = p''_{22}\{1,2,3\} p''_{11}\{1,2,3\} p'_3\{1,2,3\} = p_2 p_1 \cdot 1 = .45,$$

and similarly it can be obtained that $p_{23}^{\{1\}} = .525$. Hence $g_1(1) = 3$, since $.5/.525 > .3/.45$. Therefore, the first pair in the ordering is $\{f(1), g_1(1)\} = \{2,3\}$. Then $g_1(2) = 1$, since 1 is the only integer remaining to be used in the g_1 ordering, and consequently the second pair in the ordering is $\{f(1), g_1(2)\} = \{2,1\}$. It is not really necessary to determine $f(2)$, since $\{1,3\}$ is the only remaining pair, and hence the last pair, but to further illustrate the computations, observe that $T_2 = \{1,3\}$, $p_1^{(2)} = p''_{11}\{1,3\} p'_2\{1,3\} p'_3\{1,3\} = p_1(1-p_2) \cdot 1 = .15$ by (3.12), and similarly $p_3^{(2)} = p_3(1-p_2) \cdot 1 = .175$. Hence $f(2) = 3$, since $.7/.175 > .5/.15$. Consequently, $g_2(1) = 1, f(3) = 1$.

3.1.3 Computation of p_i^* and c_{ij}

Next we explain the computation of the p_i^*'s. If I_i consists of the pair of integers $I_i = \{f(k), g_k(\ell)\}$ then, as previously noted, $p_i^* = p_{f(k),g_k(\ell)}^{(\ell)}$. Consequently, p_i^* can be computed from (3.14) with $j = g_k(\ell)$.

If I_i is a singleton set $\{t\}$ for some $t \in F_\alpha$, then, as established in Ernst and Ikeda (1994),

$$p_i^* = p''_{t\alpha}(\{t\}) \prod_{\substack{u=1 \\ u\neq\alpha}}^{r} p'_u(\varnothing). \tag{3.15}$$

Finally, if $I_i = \varnothing$, then

$$p_i^* = \prod_{u=1}^{r} p'_u(\varnothing).$$

It remains only to explain how to compute the c_{ij}'s which, by (3.5) and (3.6), reduces to computing b_{it}, $i = 1, \ldots, \binom{n}{2} + n + 1$, $t = 1, \ldots, n$.

To compute b_{it}, observe that

$$b_{it} = 0 \quad \text{if} \quad I_i = \varnothing,$$

$$= 1 \quad \text{if} \quad I_i = \{v\} \quad \text{and} \quad t = v,$$

$$= 0 \quad \text{if} \quad I_i = \{v\} \quad \text{and} \quad t \neq v,$$

while if $I_i = \{f(k), g_k(\ell)\}$ and $f(k) \in F_\alpha$, $g_k(\ell) \in F_\beta$, $t \in F_\gamma$, then

$$b_{it} = 1 \quad \text{if} \quad t = f(k) \quad \text{or} \quad t = g_k(\ell), \tag{3.16}$$

$$= 0 \quad \text{if} \quad t \notin T^*_{k\ell}, \tag{3.17}$$

$$= 0 \quad \text{if} \quad t \in T_{k\ell} \sim \{g_k(\ell)\} \quad \text{and} \quad \gamma = \alpha = \beta, \tag{3.18}$$

$$= \frac{p_{f(k),t}}{p''_{f(k),\alpha}(T^*_{k\ell})} \quad \text{if} \quad t \in T_{k\ell} \sim \{g_k(\ell)\} \quad \text{and} \quad \gamma = \alpha \neq \beta, \tag{3.19}$$

$$= \frac{p_{g_k(\ell),t}}{p''_{g_k(\ell),\beta}(T^*_{k\ell})} \quad \text{if} \quad t \in T_{k\ell} \sim \{g_k(\ell)\} \quad \text{and} \quad \gamma = \beta \neq \alpha, \tag{3.20}$$

$$= \frac{p''_{t\gamma}(T^*_{k\ell})}{p'_\gamma(T^*_{k\ell})} \quad \text{if} \quad t \in T_{k\ell} \sim \{g_k(\ell)\} \quad \text{and} \quad \gamma \neq \alpha, \gamma \neq \beta. \tag{3.21}$$

In Ernst and Ikeda (1994) it is demonstrated how (3.16)–(3.21) were obtained.

In the actual implementation for the SIPP application, modifications of the reduced-size procedure were needed to overlap the 1990s SIPP design with the 1980s SIPP design. The modifications were necessary because the PSU definitions in the 1980s and 1990s designs were not identical. As a result, some PSUs in the 1990s design could intersect more than one 1980s design PSU. These modifications are detailed in Ernst and Ikeda (1994).

3.2 Modifications of Reduced-Sized Procedure for Other Designs

In general, consider any m'-PSUs-per-stratum without replacement initial design and any m-PSUs-per-stratum without replacement final design, where m', m are any positive integers. Although the reduced-size procedure in Section 3.1 was only presented for the case $m = m' = 2$, it is actually applicable for any m, m'. We will sketch the modifications necessary when $m \neq 2$ or $m' \neq 2$.

A different value of m' only requires modification of some of the computations. For example, if $m = 2$, but $m' \neq 2$, then the computations for $p_i^{(k)}$, $p^{(\ell)}_{f(k),j}$ and c_{ij} would be different but their definitions would not change.

If $m = 3$, then, regardless of the value of m', the set of all distinct triples, instead of pairs, of integers in $\{1, \ldots, n\}$, is ordered. If I consists of at least three integers, then the new selection probabilities are conditioned only on the first listed triple in the ordering contained in I. Otherwise, the new selection probabilities are conditioned on I itself. Thus the new selection probabilities are conditioned on $\binom{n}{3} + \binom{n}{2} + n + 1$ events.

To obtain the desired ordering of the triples of integers, first the orderings $f(1), \ldots, f(n)$ and $g_k(1), \ldots, g_k(n-k)$ are constructed exactly as in the case $m = 2$. Then, corresponding to each $k = 1, \ldots, n-2$, $\ell = 1, \ldots, n-k-1$, an ordering $h_{k\ell}(1), \ldots, h_{k\ell}(n-k-\ell)$ of $\{1, \ldots, n\} \sim \{f(1), \ldots, f(k), g_k(1), \ldots, g_k(\ell)\}$ is constructed in a manner similar to the construction of $g_k(1), \ldots, g_k(n-k)$. For example, in defining $h_{k\ell}(v)$ for $v \geq 2$, $p^{(\ell)}_{f(k),j}$ in the definition of $g_k(\ell)$ is replaced by

$$P\big(f(k), g_k(\ell), j \in I \quad \text{and} \quad I \subset (T^*_{k\ell} \cup g_k(\ell)) \sim \{h_{k\ell}(1), \ldots, h_{k\ell}(v-1)\}\big).$$

A linear ordering of the distinct triples in $\{1, \ldots, n\}$ is then determined by representing each triple uniquely as an ordered triple of the form $(f(k), g_k(\ell), h_{k\ell}(v))$. A second triple $(f(k'), g_{k'}(\ell'), h_{k'\ell'}(v'))$ precedes the first if and only if either $k' < k$, or $k' = k$ and $\ell' < \ell$, or $k' = k$ and $\ell' = \ell$ and $v' < v$.

For $m \geq 4$, ordered m-tuples would be defined in a similar manner and the new selection probabilities conditioned on $\binom{n}{m} + \binom{n}{m-1} \ldots + n + 1$ events.

For $m = 1$, the new selection probabilities are conditioned on the first member of the ordering $f(1), \ldots, f(n)$ in I if $I \neq \varnothing$, or on $\varnothing$ if $I = \varnothing$.

Note that if $m > m'$, it is possible that at least some ordered m-tuples cannot be subsets of I, in which case all such subsets should be excluded from the ordering and the set of events on which the new selection probabilities are conditioned. If no m-tuple can be a subset of I, then the new selection probabilities are conditioned on I itself.

It is not necessary to limit the initial events used in the transportation problem to subsets of I of size m or less. For example, if $m = 2$ and $\binom{n}{3} + \binom{n}{2} + n + 1$ is sufficiently small, then a procedure conditioned on subsets of three or less can be used, resulting in a generally higher expected overlap. Conversely, if $\binom{n}{m} + \binom{n}{m-1} \ldots + n + 1$ is too large, the new selection probabilities can be conditioned on subsets of I of size m'' or less, where $m'' < m$, although with a generally smaller expected overlap.

3.3 Relationship Between Expected Overlap for the Reduced-Size Procedure, the Optimal Procedure and Independent Selection

Let Ω_I, Ω_R, Ω_O denote the expected overlap for the independent selection, the reduced-size procedure, and the optimal procedure, respectively. In Ernst and Ikeda (1994) the relationship between these quantities is explored. We briefly summarize here some of the results.

It is established that $\Omega_I \leq \Omega_R \leq \Omega_O$ for any m, m' where m, m' are as in Section 3.2. In addition, for the case that we have been focusing on, $m = m' = 2$, lower bounds are established on Ω_R and upper bounds are established on Ω_O and $\Omega_O - \Omega_R$.

For example, let μ_2 denote the probability that there are at least two elements in I, μ_1 denote the probability that I is a singleton set, and

$$\lambda = \min\{\min\{\pi_i/p_i : i = 1, \ldots, n\}, \min\{\pi_{ij}/p_{ij} : i,j = 1, \ldots, n, i \neq j\}, 1\}.$$

Then $\Omega_O \leq 2\mu_2 + \mu_1$, $\Omega_R \geq \lambda(2\mu_2 + \mu_1/2)$, and $\Omega_O - \Omega_R \leq 2(1 - \lambda)\mu_2 + (1 - \lambda/2)\mu_1$.

Unfortunately these bounds are not always very tight. However, in certain circumstances they are useful. For example, if $\pi_{ij} \geq p_{ij}$ for all i,j and the probability is 1 that there is at least two elements in I, then it follows from these bounds that $\Omega_R = \Omega_O = 2$.

Finally, an example is presented to illustrate a worst case situation for Ω_R in relation to Ω_O for the case m, $m' = 2$. It shows that Ω_O may equal 2, while Ω_R is arbitrarily close to 0. Thus, at least in theory, the reduced-size procedure can be ineffective. However, in practice, as will be shown in the next section, Ω_R is much closer to Ω_O than to Ω_I, at least for the SIPP application.

4. APPLICATION OF REDUCED-SIZE PROCEDURE TO SIPP

Results from simulations of the SIPP overlap, done prior to production for research and testing purposes, are presented, as well as results from the actual SIPP production overlap. Further details are given in Ernst and Ikeda (1992b, 1994).

In the implementation of the reduced-size overlap procedure, minimum cost flow (MCF) optimization software, written by Darwin Kingman and John Mote at the University of Texas at Austin, was used to solve the required transportation problem. A FORTRAN program was written to produce input to and process output from the MCF software.

To test the software prior to production, the program was used to overlap two stratifications, based on 1970 census data, of the SIPP Midwest region with the actual 1980s design stratification for the SIPP Midwest region. (At the time of this test, 1990 census data was not yet available.) The 1970-based stratifications were produced by stratifying the 1980s SIPP noncertainty PSUs in the Midwest region using 1970 data. Both of the 1970-based stratifications partitioned the noncertainty PSUs into 31 strata, using different sets of stratification variables. The stratifications based on 1980 and 1970 data were treated as "initial" and "final" stratifications for the purposes of the overlap algorithm.

In the actual implementation, as noted in Section 3.1 and detailed in Ernst and Ikeda (1994), a modification of the reduced-size procedure was used to overlap the 1990s SIPP design with the 1980s SIPP design, because the PSU definitions in the 1980s and 1990s designs were not identical. The modified reduced-size procedure was used to overlap 103 final (1990s design) nonselfrepresenting strata in SIPP.

The expected overlap was calculated for the reduced-size maximum overlap algorithm, for independent selection of final PSUs, and for an upper bound to the expected overlap for the optimal procedure. An upper bound was calculated instead of the actual optimal overlap, since the optimal overlap cannot be calculated for the larger strata. For the simulation, the upper bound used is the one stated in Section 3.3, $\mu_2 + 2\mu_1$, while for the production SIPP, a different upper bound, described in Ernst and Ikeda (1994), was required because the PSU definitions in the 1980s and 1990s were not identical.

The results from the two final stratifications in the simulation were generally similar to each other. Combining the results from both stratifications, the mean expected overlap for this set of 62 strata was 1.552, 1.569 and 0.480 PSUs/stratum for the reduced-size procedure, the upper bound to the optimal overlap and independent selection respectively. For the actual SIPP implementation, the corresponding number was 1.523, 1.647 and 0.582, respectively, while the corresponding expected number of PSUs overlapped for the 103 strata was 156.9, 169.6 and 59.9, respectively. Thus, in both the simulations and the production SIPP, the reduced-size procedure yielded results reasonably close to the upper bound for the optimal procedure.

The reduced-size algorithm took a fairly short time to run on most strata. The CPU times in the simulation for

final strata with different numbers of PSUs are given below. The reduced-size program was run on a Solbourne 5/605 computer. The median number of PSUs in a stratum, for the entire group of 62 strata, was 17 PSUs. The 68 PSUs stratum was the largest stratum.

Table 6
CPU Times for Reduced-Size Procedure

Number of PSUs	CPU Time (hrs:min:sec)
18	0:36
37	5:44
49	24:05
68	2:23:43

We also calculated for the actual SIPP implementation, that of the 103 final strata overlapped by the modified reduced-size procedure, 41 would not have run under the optimal procedure. This calculation was based on our estimate that the maximum size transportation problem, in terms of number of variables, that could have run in production was 4×10^6. The number of variables for the optimal procedure was less than 4×10^6 for all 56 strata for which $n \leq 14$, but exceeded this limit for all but 6 of the 47 strata with $n \geq 15$, including two with $n = 15$. The maximal size of the transportation for the optimal procedure among the 103 strata occurred for a stratum with $n = 46$, for which there were 3.61×10^{12} variables. In contrast, there were 1.03×10^6 variables for the modified reduced-size procedure for this stratum.

Another question of interest is the overlap effectiveness of the reduced-size procedure in comparison with the overlap procedure of Ernst (1986). In general it is believed that the reduced-size procedure should produce a higher overlap in situations when both are usable, since the reduced-size procedure makes use of the stratum-to-stratum independence in the initial design. However, although the procedure in Ernst (1986) is applicable to two-PSU-per-stratum designs, no computer program has ever been written at the Census Bureau (or anywhere else that the authors are aware of) to implement this procedure for such designs, since there has not yet been a production application for this program. Consequently, we cannot make a direct comparison of these two methods on the same data. However, a crude comparison can be made from the results of the reduced-size overlap procedure for SIPP data and the results of the overlap using the procedure in Ernst (1986) for the overlap of 1990s CPS and NCVS designs with their respective 1980s designs. (Both the 1980s and 1990s designs for CPS and NCVS are one-PSU-per-stratum designs.)

For CPS, the overlap procedure resulted in an average increase in expected overlap, in comparison with independent selection, of .26 PSUs/stratum, and for NCVS the overlap procedure resulted in an average increase in expected overlap of .30 PSUs/stratum. This compares with an increase of .94 PSUs/stratum for the reduced-size procedure over independent selection for SIPP. If the two overlap procedures are equally effective, then one might expect that the increase in overlap per stratum for SIPP would be roughly twice as large as for CPS and NCVS, since SIPP has a two-PSUs-per-stratum design. By this standard, the reduced-size procedure program performs better than the procedure in Ernst (1986). However, since the stratifications were quite different for these three surveys, the validity of this comparison is open to question.

For the example considered in Sections 2 and 3, a valid comparison of the different overlap procedures can be made, since the expected overlap values for the procedure in Ernst (1986)), 1.625, was easily calculated by hand. For the reduced-size procedure the corresponding overlap value is 1.725, and for the optimal procedure it is 1.735.

CONCLUSIONS

The reduced-size overlap procedure presented in this paper meets its two key objectives in practice. It reduces the size of the transportation problems to a usable size, as evidenced both by the size of the transportation problem in the formulation (3.1)-(3.3), and the fact that it has actually been implemented in the redesign of a major survey. In addition, the procedure accomplishes the size reduction while yielding nearly optimal overlap, at least for the SIPP application. It can only be used when the PSUs in the initial design are selected independently from stratum to stratum, but when this condition is met we believe it is the overlap procedure of choice for large strata.

ACKNOWLEDGEMENTS

The programming assistance of Todd Williams is gratefully acknowledged. The authors would also like to thank the referees and the editors for their constructive comments. The views expressed in this paper are attributable to the authors and do not necessarily reflect those of the Bureau of Labor Statistics and the Census Bureau.

REFERENCES

ARAGON, J., and PATHAK, P.K. (1990). An algorithm for optimal integration of two surveys. *Sankhyā: The Indian Journal of Statistics*, 52, 198-203.

ARTHANARI, T.S., and DODGE, Y. (1981). *Mathematical Programming in Statistics*. New York: John Wiley and Sons.

CAUSEY, B.D., COX, L.H., and ERNST, L.R. (1985). Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80, 903-909.

ERNST, L.R. (1986). Maximizing the overlap between surveys when information is incomplete. *European Journal of Operational Research*, 27, 192-200.

ERNST, L.R. (1989). Further Applications of Linear Programming to Sampling Problems. Bureau of the Census, Statistical Research Division, Research Report Series, No. RR-89/05.

ERNST, L.R., and IKEDA, M. (1992a). Modification of the Reduced-Size Transportation Problem for Maximizing Overlap When Primary Sampling Units Are Redefined in the New Design. Bureau of the Census, Statistical Research Division, Technical Note Series, No. TN-91/01.

ERNST, L.R., and IKEDA, M. (1992b). Summary of the Performance of the Maximum Overlap Algorithms for the 1990's Redesign of the Demographic Surveys. Bureau of the Census, Statistical Research Division, Technical Note Series, No. TN-92/01.

ERNST, L.R., and IKEDA, M. (1994). A Reduced-Size Transportation Algorithm for Maximizing the Overlap Between Surveys. Bureau of the Census, Statistical Research Division, Research Report Series, No. RR-93/02.

GLOVER, F., KARNEY, D., KLINGMAN, D., and NAPIER, A. (1974). A computation study on start procedures, basic change criteria and solution algorithms for transportation problems. *Management Sciences*, 20, 793-813.

KEYFITZ, N. (1951). Sampling with probabilities proportional to size: Adjustment for changes in probabilities. *Journal of the American Statistical Association*, 46, 105-109.

KISH, L., and SCOTT, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.

PATHAK, P.K., and FAHIMI, M. (1992). Optimal integration of surveys. In *Essays in Honor of D. Basu*. Eds. M. Ghosh, and P.K. Pathak. Hayward, California: Institute of Mathematical Statistics, 208-224.

PERKINS, W.M. (1970). 1970 CPS Redesign: Proposed Method for Deriving Sample PSU Selection Probabilities Within 1970 NSR Stata. Memorandum to Joseph Waksberg, Bureau of the Census.

RAJ, D. (1968). *Sampling Theory*. New York: McGraw Hill.

Survey Methodology, December 1995
Vol. 21, No. 2, pp. 159–165
Statistics Canada

How Prenotice Letters, Stamped Return Envelopes and Reminder Postcards Affect Mailback Response Rates for Census Questionnaires

DON A. DILLMAN, JON R. CLARK and MICHAEL D. SINCLAIR[1]

ABSTRACT

In a 1992 National Test Census the mailing sequence of a prenotice letter, census form, reminder postcard, and replacement census form resulted in an overall mailback response of 63.4 percent. The response was substantially higher than the 49.2 percent response rate obtained in the 1986 National Content Test Census, which also utilized a replacement form mailing. Much of this difference appeared to be the result of the prenotice - census form - reminder sequence, but the extent to which each main effect and interactions contributed to overall response was not known. This paper reports results from the 1992 Census Implementation Test, a test of the individual and combined effectiveness of a prenotice letter, a stamped return envelope and a reminder postcard, on response rates. This was a national sample of households (n = 50,000) conducted in the fall of 1992. A factorial design was used to test all eight possible combinations of the main effects and interactions. Logistic regression and multiple comparisons were employed to analyze test results.

KEY WORDS: Mail survey; Response rates; Multiple comparisons; Logistic regression.

1. INTRODUCTION

A decline of 10 percentage points from 75 to 65 in the mailback response rates for the 1990 U.S. Decennial Census has stimulated the conduct of research aimed at finding ways to improve response. Each percentage point gain in response has the potential for saving approximately $16 million in personal visit enumeration costs (Miskura 1992). From an earlier experiment it was learned that respondent-friendly construction and asking somewhat fewer questions than posed in the 1990 Census short questionnaire improved mailback response rates by 8.0 percentage points (Dillman, Clark and Sinclair 1993). An experimental census form with these features was returned by 71.4 percent of households, compared to 63.4 percent of those which had received the 1990 Census short form as a control. Response rates for both of these forms were substantially higher than had previously been obtained in similar non-census year tests. For example, in the 1986 National Content Test which utilized a questionnaire equivalent to the 1990 Census short form, a 49.2 response rate was obtained. It was hypothesized that part of the high response observed in the recent experiment was due to a multiple contact implementation strategy which consisted of a prenotice letter, a reminder postcard and a replacement questionnaire.

The purpose of this paper is to report results of the 1992 Implementation Test (IT), a test designed to determine the relative and combined contribution to mailback response of the prenotice letter and reminder postcard used in the previously reported experiment (Dillman *et al.* 1993). Also included in the test is the effect of including a stamped return envelope (*vs.* business reply) with the mailed census form.

The 1990 U.S. Decennial Census required surveying over 100,000,000 households. Cost considerations alone suggest the importance of learning the extent to which each of these three response-inducing techniques might be employed in improving household response. Although past research has suggested that each of the three elements can be important to improving response, little information is available on potential interactions among them. The study was designed in such a way as to explore the extent to which their combined uses are additive and/or interactive.

1.1 Past Research

Numerous studies have confirmed that the most important determinant of overall response to mail surveys is the number of contacts (*e.g.*, Scott 1961 and Heberlein and Baumgartner 1978). Both prenotices and reminders have been demonstrated as being effective promoters of response (*e.g.*, Kanuk and Berenson 1975, Linsky 1975 and Fox *et al.* 1988). However, past research has provided minimal insight into their relative importance as inducers of response.

Past research is generally consistent in suggesting that inclusion of a stamped return envelope (*vs.* a business reply envelope) improves response (Scott 1961, Kanuk and Berenson 1975, Duncan 1979, Harvey 1987 and Fox *et al.* 1988). A noteworthy exception is a regression analysis of previous studies by Heberlein and Baumgartner, which

[1] Don A. Dillman, Washington State University, Pullman, WA, U.S.A.; Jon R. Clark, U.S. Bureau of the Census, Washington DC 20233, U.S.A.; and Michael D. Sinclair, Response Analysis Corp., Princeton, NJ, U.S.A.

found no significant effect for the inclusion of stamped return envelopes (1979). A review study by Armstrong and Luske reported 20 studies in which alternatives to business reply envelopes had been tested (1987). In each of these comparisons the absolute level of response to the alternative was significantly higher in 15 of the 20 cases, by an average of 9.2 percentage points. Six studies of metered marks *vs.* envelopes with real stamps were reported. On average they showed a 3.4 percentage point advantage for stamps. Finally, four studies in which a constellation of response inducing factors was used to insure high overall response rates showed a 2-4 percentage point advantage for stamped over business reply envelopes (Dillman 1978).

The three response stimuli to be tested here are among the top eight techniques reported consistently in the research literature as factors which improve mailback response rates. Others include financial incentives, special postage, choice of sponsor, personalization and interest (or salience) (Dillman 1991).

Two of these eight factors, financial incentives and special postage (*e.g.*, certified or two day priority mail) were judged impractical for use in a census of more than 100,000,000 households. A third factor, sponsorship by the U.S. Bureau of the Census, was considered desirable from the standpoint of encouraging response. A fourth factor, respondent interest, or question salience could not be manipulated in the sense that the survey questions are specified by federal laws. The fifth factor, personalization of correspondence was limited by the fact that Census forms cannot be addressed to individuals and are necessarily sent to only household addresses. By examining the individual and combined response effects of the prenotice, stamped return envelope and reminder, we hoped to learn whether the use of one or more of these elements would substitute for another, therefore making it possible to improve response at less cost.

1.2 Design and Integration of Treatment Elements

Certain features of the census form mailout packet suggest that it may be overlooked or ignored by those to whom it is sent. By necessity it is sent only to household addresses; names cannot be used to address any of the letters. Accurate processing of returned questionnaires requires identification of the household address on the questionnaire itself. Separately addressing an outside envelope, letter and questionnaire and being sure that the correct components are inserted into the appropriate envelope presents a serious quality control problem in a large census. Therefore it is considered important to print addresses only on one of the pieces that has to be merged together for the mailout package. Consequently, a windowed envelope through which the address on the questionnaire can be seen is used to deliver it.

The combined effect of the inability to use resident names plus size and outward appearance of the windowed envelope suggest that it contains unimportant material or perhaps, "junk mail." Also, research on nonresponse to the 1990 Census revealed that some people did not recall receiving their census questionnaire in the mail, or saw it, but did not open it, both of which might have resulted from a mass mailing appearance (Kulka *et al.* 1991).

In this experimental test the prenotice letter and reminder postcard were designed to bring attention to the envelope containing the census form. This was accomplished in five ways. First, the prenotice was developed as a letter, and the reminder as a postcard. It was reasoned that people were more likely to look at two pieces of mail which appeared different from one another. The letter format was chosen for the prenotice in order to save the more convenient postcard format for the reminder.

Second, the prenotice letter consisted of a letter from the Director of the Census Bureau with the notation "To the residents at" and the address imaged onto stationery in the normal inside address position. Our goal was to communicate that the census questionnaire which would soon arrive was specifically for people at that address. This address also doubled as an outside address, being visible through a windowed envelope, thereby avoiding the quality control concern noted for the census form mailing of merging separately addressed components.

Third, the prenotice was scheduled to be delivered a few days before the envelope containing the census form itself, and the reminder was scheduled to arrive just a few days afterwards. The mailout dates were September 21st, 24th, and 29th, respectively. It was reasoned that to be effective, a reminder (without a replacement questionnaire) should arrive within a few days of the questionnaire, before normal household cleaning would have resulted in unopened mail being thrown out.

Fourth, the wording of the prenotice, "Within the next few days you should receive..." and the reminder, "A few days ago you should have received..." were designed to encourage recipients to look for the census form. Fifth, the use of the Director's letterhead stationary and white postcard stock which showed the seal of the Department of Commerce above the reminder message, were aimed at communicating that the census questionnaire was from the government and not from some other group attempting to emulate a governmental appearance, as is sometimes done by noting, *e.g.*, "this is your official notice."

The stamped return envelope's positive influence, if any, on response may result from encouraging trust that the request is legitimate and important (otherwise why would the sender "waste" a stamp, which could be torn off and used for another purpose and/or a recipient's reluctance to throw away something of value, *i.e.*, an uncanceled stamp). The prenotice, and to some extent the reminder, could enhance the stamp's effect by getting the

envelope containing the census form opened. Also, once opened, the awareness of an uncancelled stamp could discourage throwing away the contents so that the effect of the reminder is enhanced.

In order for the prenotice, stamped return and reminder to mutually support one another, it was deemed important that first class mail be used. Had bulk rate been used, and the mailings been closely spaced, it was likely that in some households a later mailing would have arrived before an earlier one.

In sum, this test involved more than simply juxtaposing three separate test elements from the literature. The elements were operationalized in ways that improved the likelihood that each would augment effects of the others, and be feasible for use in large scale mailings. Practically, we hoped to learn whether one or more of the elements might be eliminated without a significant loss of response, thus showing how to save costs for a census mailing.

2. EXPERIMENTAL DESIGN

A factorial design, consisting of all eight of the possible combinations of the three main effects, was used for the experiment. The treatments were as follows:

1. None (control),
2. Prenotice letter only,
3. Stamped return envelope only,
4. Reminder postcard only,
5. Letter plus stamped return,
6. Stamped return plus reminder,
7. Letter plus reminder, and
8. Letter plus stamped return plus reminder.

2.1 Sample Design

The sampling universe consisted of all housing units in the questionnaire mailback areas identified by Census Bureau address files. The 449 district office (DO) areas for the 1990 Census were selected as the geographic units for defining the strata for the test. Two strata were defined. Due to the high correlation between the minority rate (minority is defined as including all Black and Hispanic classifications) and the 1990 Census mail response rate, the stratification objectives were met by ranking the DOs by their percent minority. DOs with a combination of high minority (Black and/or Hispanic origin) population and low 1990 questionnaire mail response rates were defined as "low response areas" (LRA) and made up the first stratum. The remaining DOs were classified as "high response areas" (HRA) and constituted the second stratum.

The first stratum, consisting of 67 DOs, had a combined minority population of about 64 percent and encompassed about 11 percent of all housing units in the census mailback areas. The second stratum of 382 DOs had a combined minority population of about 15 percent. The HRA stratum had a cumulative mail response rate in the 1990 Census of approximately 10 percentage points higher than the LRA stratum.

A sample of 50,000 housing units was selected with 25,000 units in each stratum. The LRA stratum was oversampled to concurrently study factors related to differential undercount, which falls outside the scope of this paper. Each stratum was divided into eight equally sized panels to test the eight different treatments. A systematic sample of 3,125 housing units was selected from each panel/stratum combination. Once a housing unit was selected, the seven subsequent units were also selected. The resulting households in each of the eight unit clusters were randomly allocated to a panel. Hence, all eight neighbors got different treatments. The sample was clustered to reduce the sampling variance in the panel-to-panel comparison.

The sample size selected for this study was developed by extensive data simulations which indicated that the 50,000 unit sample would be sufficient for detecting a minimum of a 3 percent difference in all pairwise treatment comparisons.

Table 1

Implementation Test Final Rates National and Stratum Level Estimates

Treatment	Response Rate (%) Estimates and Standard Errors (%)					
	National		1990 High Response Areas		1990 Low Response Areas	
	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
1. Control	50.0	0.8	51.9	0.9	36.3	0.9
2. Prenotice Letter Only	56.4	0.8	58.6	0.9	40.5	0.9
3. Stamped Return Envelope Only	52.6	0.8	54.5	0.9	37.9	0.9
4. Reminder Card Only	58.0	0.8	60.2	0.9	42.0	0.9
5. Letter and Stamp	59.8	0.8	62.1	0.9	43.0	0.9
6. Stamp and Reminder	59.5	0.8	61.8	0.9	42.6	0.9
7. Letter and Reminder	62.7	0.8	65.0	0.9	45.4	0.9
8. Letter, Stamp and Reminder	64.3	0.8	66.5	0.9	47.8	0.9

3. FINDINGS

The major results from this study are presented through two analytical methods, first through multiple pairwise comparisons of treatment means and secondly through logistic regression. See Appendix for estimation procedures. Both methods provide consistent results. The overall response rates and standard errors for each of the treatments at the national and stratum levels are presented in Table 1. They range from 50.0 percent for the control group to 64.3 percent when all three main effects are applied together.

3.1 Multiple Comparisons of Mail Response Rates

Twenty eight comparisons are presented in Table 2 corresponding to all possible pairwise comparisons of the 8 treatments. Given the space restrictions in the table, the following abbreviations were used: C = control, L = prenotice letter, S = stamped return envelope, R = reminder postcard.

The first three comparisons in Table 2 illustrate the improvements in response that main effect components added to response individually above and beyond the control treatment. The estimated improvement in response due to the prenotice letter was 4.2 percent in the LRA stratum, 6.7 percent in the HRA stratum and 6.4 percent at the national level. The estimated improvement due to the reminder card was 5.7 percent in the LRA stratum, 8.3 percent in the HRA stratum and 8.0 percent at the national level. All of these improvements are significant. Thus, the principal finding of this study is that both the prenotice letter and the reminder card increased mail response at the national and stratum level. No significant improvements were noted for the stamped return envelope at the national or stratum level.

Table 2

Differences in Response Rates - Each Component in the Presence of Another Component

Experimental Comparisons	Response Rate Differences (%) and 90% Confidence Intervals (C.I.)					
	National		1990 Low Response Areas (LRA)		1990 High Response Areas (HRA)	
	Difference	90% C.I.	Difference	90% C.I.	Difference	90% C.I.
1. L - C	6.4	3.3 to 9.5*	4.2	0.9 to 7.5*	6.7	3.2 to 10.2*
2. S - C	2.5	−0.5 to 5.6	1.7	−1.7 to 5.0	2.7	−0.8 to 6.1
3. R - C	8.0	4.9 to 11.1*	5.7	2.4 to 9.1*	8.3	4.9 to 11.7*
4. LS - C	9.8	6.7 to 12.9*	6.8	3.4 to 10.1*	10.2	6.7 to 13.7*
5. SR - C	9.5	6.4 to 12.5*	6.4	3.0 to 9.7*	9.9	6.5 to 13.3*
6. LR - C	12.7	9.6 to 15.7*	9.2	5.8 to 12.5*	13.2	9.7 to 16.6*
7. LSR - C	14.2	11.2 to 17.2*	11.5	8.2 to 14.8*	14.6	11.3 to 18.0*
8. L - S	3.8	0.8 to 6.9*	2.5	−0.9 to 5.9	4.1	0.6 to 7.5*
9. R - L	1.6	−1.5 to 4.8	1.5	−1.9 to 5.0	1.6	−1.96 to 5.10
10. R - S	5.5	2.4 to 8.5*	4.1	0.7 to 7.5*	5.6	2.2 to 9.0*
11. LS - L	3.4	0.3 to 6.5*	2.6	−0.9 to 6.0	3.5	0.03 to 7.0*
12. SR - L	3.1	0.03 to 6.2*	2.2	−1.3 to 5.6	3.2	−0.3 to 6.6
13. LR - L	6.3	3.2 to 9.3*	5.0	1.5 to 8.4*	6.4	3.0 to 9.9*
14. LS - S	7.3	4.2 to 10.3*	5.1	1.7 to 8.5*	7.6	4.1 to 11.0*
15. SR - S	6.9	3.8 to 10.1*	4.7	1.2 to 8.2*	7.2	3.8 to 10.7*
16. LR - S	10.1	7.1 to 13.2*	7.5	4.1 to 11.0*	10.5	7.0 to 13.9*
17. LS - R	1.8	−1.3 to 4.9	1.1	−2.4 to 4.5	1.9	−1.6 to 5.4
18. SR - R	1.5	−1.6 to 4.5	0.7	−2.8 to 4.1	1.6	−1.8 to 5.0
19. LR - R	4.7	1.6 to 7.7*	3.5	−0.02 to 6.9	4.9	1.5 to 8.3*
20. LSR - L	7.9	4.8 to 10.9*	7.3	3.9 to 10.7*	7.9	4.5 to 11.4*
21. LSR - S	11.7	8.7 to 14.7*	9.8	6.4 to 13.3*	12.0	8.6 to 15.4*
22. LSR - R	6.2	3.2 to 9.3*	5.8	2.3 to 9.3*	6.3	2.9 to 9.7*
23. LSR - LS	4.4	1.4 to 7.5*	4.7	1.2 to 8.2*	4.4	1.0 to 7.8*
24. LSR - SR	4.8	1.7 to 7.8*	5.1	1.7 to 8.6*	4.7	1.3 to 8.2*
25. LSR - LR	1.6	−1.4 to 4.5	2.3	−1.1 to 5.8	1.5	−1.8 to 4.8
26. SR - LS	−0.3	−3.3 to 2.7	−0.4	−3.8 to 3.1	−0.3	−3.7 to 3.1
27. LR - LS	2.9	−0.2 to 6.0	2.4	−1.1 to 5.9	2.9	−0.6 to 6.4
28. LR - SR	3.2	0.2 to 6.2*	2.8	−0.6 to 6.2	3.3	−0.1 to 6.6

A C.I. marked with an * indicates the difference was statistically significant at $\alpha = .10$ (9-in-10 chance that the C.I.s will include the actual differences).

3.2 Logistic Regression Analysis

A model including components for the stratum, prenotice letter, stamp and reminder card including all of the interaction terms was evaluated. Modeling was also performed at the stratum level using only parameters for the component effects and their interactions.

The results of the full model analysis indicate that only the main effects of the letter and the reminder card along with the intercept and stratum term are statistically significant in the model. Given these results, additional modeling at the national level was accomplished with a reduced model including only the stratum main effect, the individual components and the component interactions. The results of this modeling are presented in Table 3 below.

Table 3

Analysis of Weighted Least Squares Logistic Regression Modeling Reduced Model, no Stratum by Component Interactions

Model Parameters	Estimated Parameters and 90% Bonferroni Confidence Intervals (C.I.)	
	Estimate	90% C.I.
Intercept, β_0	−.61	−.686 to −.545*
Stratum, β_1	.738	.689 to .789*
Letter, β_2	.227	.130 to .324*
Stamp, β_3	.090	−.006 to .186
Reminder, β_4	.291	.194 to .387*
Letter/Stamp, β_5	.036	−.101 to .173
Letter/Reminder, β_6	−.054	−.192 to .083
Reminder/Stamp, β_7	−.043	−.179 to .093
Let/Reminder/Stmp, β_8	−.003	−.197 to .191

A C.I. marked by an * indicates the difference was statistically significant at $\alpha = .10$.

The results of both modelings show that significant improvements were realized from the prenotice letter and reminder post card, but not from the stamped return envelope for the national and within stratum models. These results correspond to those presented by the multiple comparisons above. None of the interaction terms were statistically significant, indicating the effect of the components are basically additive in nature.

4. DISCUSSION AND CONCLUSIONS

The prenotice letter, stamp and reminder postcard individually improved response rates by 6.4, 2.5 and 8.0 percentage points, respectively. The increase of 2.5 was not statistically significant. The effects of the elements were also found to be mostly additive, and did not interact with one another. In comparison to the control group, the combination of letter-stamp improved response 9.8 percentage points, the stamp and reminder, 9.5 percent, and the letter and reminder, 12.7 percent. All three elements together improved response by 14.3 percent. Each use of the letter and reminder added significantly to response, but the stamp only added significantly when used with a prenotice and no reminder. The most important conclusion from this experiment was that both the prenotice letter and reminder postcard are important to achieving a high response and that neither eliminates the effect of the other.

Although the individual effect (2.5 percent overall) of the stamped return envelope is slightly smaller than needed for significance, it is of similar magnitude to what has been found significant in past research (Armstrong and Luske 1987; Dillman 1978, 1991). In light of the preponderance of past research showing its effectiveness, this technique should probably not be completely dismissed as being ineffective. It also appears that the stamped return envelope relates differently to the prenotice and reminder. When used alone with the prenotice, the effect of the stamped return is significant (3.4 percentage points), but it is clearly insignificant (1.6 percentage points) when a reminder is included in the mailout procedures. The reminder compensates for the lack of a stamped return envelope, whereas the prenotice appears to amplify its effect. It may be that a prenotice alerts people to notice and open the census form mailout package, and once opened, people are then encouraged to respond by the presence of the stamped return envelope. This differential connection to the mailings that precede and those that follow, appears not to have been examined in past research. A practical implication for the Census is that if a prenotice letter and no reminder is used, a stamped return envelope might add significantly to response, but be of less importance if a reminder postcard is used, as was done in the last census.

There are at least two significant barriers to the direct application of this research to conduct of the 2000 Census. First, it is important to recognize that these tests are being done in non-census years. In the past the Census Bureau has obtained much lower response rates in non-census years than in census years. For example, the 1986 National Content Test, obtained only a 49.2 percent response employing a replacement questionnaire, while the 1990 Census without employing a replacement questionnaire, achieved a 65 percent response rate. The usual explanation for this difference is "census climate," a succinct explanation of the combination of media attention, advertising, and cultural sense of participation that seems to build each decade during the census year.

The response rates obtained in our tests with the use of the five elements found to increase response are much higher than normally obtained in non-census years, but are close to the same, or perhaps a little lower, than those obtained during the last decennial census when none of these elements were used. We do not know whether the

existence of a "census climate" will substitute for the effects of these elements or add to the response likely to be obtained in a census year. Certainly a 30 percentage point increase will not be realized in the 2000 Census since that would suggest a response of nearly 100 percent. Therefore, considerable uncertainty remains with respect to the exact implications of the present findings for the 2000 Census.

APPENDIX
Estimation Procedures

Analytical results are derived from two separate methods, multiple comparisons among the mail response rates by treatment group, and logistic regression analysis. Each method has advantages over the other in terms of ease of interpretation and ease of statistical inference; hence a combined approach was utilized to bring forth the best of both methods for presentation.

The national mail response rate estimates for a given panel as presented in this study is computed by dividing the weighted total of the number of questionnaires returned by the weighted total number of forms mailed out less weighted postmaster returns (mostly vacant units).

Multiple comparisons of the 8 treatment mail response rates were reviewed to determine the level of increase in the mail response to each of the treatments. These comparisons involved a pairwise assessment of each of the treatments with the control panel and with each other.

The logistic regression procedures provide a quick and effective means for evaluating whether or not observed increases from each of the components, especially interactions, are the result of sampling variation or imply a true increase, and if these increases are influenced by the presence of other components. However, parameter estimates cannot be easily equated to the mail response rates. A detailed overview of the logistic regression methodology is provided in Thompson 1993.

Response rates were calculated for each of the treatment groups within stratum and at the national level (stratum 1 and stratum 2 combined). Standard errors for the national estimates were computed using the stratified jackknife variance procedure (Wolter 1985). The estimates were produced by the VPLX statistical software package. Standard errors for the within stratum estimates were computed using the formula for the simple random sampling jackknife variance procedure.

The primary analysis involved pairwise comparisons of the differences between response rates for eight treatments, both overall at the national level and for the two strata, LRA and HRA.

Because of the various hypotheses being tested, all possible pairwise comparisons (28 total) between the eight treatments are analyzed in the experiment. In the logistic regression framework 8 or more model parameters are tested for significance. The more comparisons that are made, the greater the potential that some of these comparisons will be incorrectly declared significant. In this case, additional statistical measures are employed to control the overall error of the decision process.

The analysis has been carried out so that statements about the entire "family" of 28 pairwise comparisons or the logistic regression parameters are made while maintaining the 90 percent (a Census Bureau standard) confidence level simultaneously for all comparisons. All 90 percent confidence intervals for the pairwise comparisons were adjusted using Dunnett's C-procedure for comparing pairwise contrasts of the test panel estimates (Hochberg and Tamhane 1987). Bonferroni simultaneous inference procedures were used to evaluate the statistical significance of the logistic regression parameters.

REFERENCES

ARMSTRONG, J.S., and LUSKE, E.J. (1987). Return postage in mail surveys: A meta analysis. *Public Opinion Quarterly*, 51 (1) 233-248.

DILLMAN, D.A., CLARK, J., and SINCLAIR, M. (1993). Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. *Public Opinion Quarterly*.

DILLMAN, D.A., SINCLAIR, M., and CLARK, J. (1992). Mail-back response rates for simplified decennial census questionnaires. *Proceeding of the Section on Survey Research Methods, American Statistical Association*, 776-783.

DILLMAN, D.A. (1991). The design and administration of mail surveys. *Annual Review of Sociology*, 17, 225-249.

DILLMAN, D.A. (1978). *Mail and Telephone Surveys: The Total Design Method*. New York: Wiley-Interscience.

DUNCAN, W.J. (1979). Mail questionnaires in survey research: A review of response inducement techniques. *Journal of Management*, 5, 39-55.

FOX, R.J., CRASK, M.R., and KIM, J. (1988). Mail survey response rate: A meta-analysis of selected techniques for inducing response. *Public Opinion Quarterly*, 52, 467-491.

HARVEY, L. (1987). Factors affecting response rates to mailed questionnaires: A comprehensive literature review. *Journal of the Market Research Society*, 29, 3, 342-353.

HEBERLEIN, T., and BAUMGARTNER, R. (1978). Factors affecting response rates to mailed questionnaires: A quantitative analysis of the published literature. *American Sociological Review*, 43, 447-462.

HOCHBERG, Y., and TAMHANE, A.C. (1987). *Multiple Comparison Procedures*. New York: John Wiley and Sons.

KANUK, L., and BERENSON, C. (1975). Mail surveys and response rates: A literature review. *Journal of Marketing Research*, 12, 440-453.

KULKA, R.A., HOLT, N.A., CARTER, W., and DOWD,K.L. (1991). Self reports of time pressures, concerns for privacy and participation in the 1990 Mail Census. *Proceedings of the 1991 Annual Research Conference*. U.S. Bureau of the Census, 33-54.

LINSKY, A.S. (1975). Stimulating responses to mailed questionnaires: A review. *Public Opinion Quarterly*, 39, 82-101.

MISKURA, S.M. (1992). Estimating the Full Cycle Costs for the Simplified Questionnaire Test (SQT), 2KS Memorandum Series, Design 2000, Book I, Chapter 30, #6.

SCOTT, C. (1961). Research in mail surveys. *Journal of Royal Statistical Society*, 143-205.

THOMPSON, J.H. (1993). Final Results of the Mail Response Evaluation for the Implementation Test (IT), DSSD 2000 Census Memorandum Series, #E-32.

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

Survey Methodology, December 1995
Vol. 21, No. 2, pp. 167–177
Statistics Canada

Consistency of Census and Vital Registration Data on Older Americans: 1970-1990

LAURA B. SHRESTHA and SAMUEL H. PRESTON[1]

ABSTRACT

Major uncertainties about the quality of elderly population and death enumerations in the United States result from coverage and content errors in the censuses and the death registration system. This study evaluates the consistency of reported data between the two sources for the white and the African-American populations. The focus is on the older population (aged 60 and above), where mortality trends have the greatest impact on social programs and where data are most problematic. Using intercensal cohort analysis, age-specific inconsistencies between the sources are identified for two periods: 1970-1980 and 1980-1990. The U.S. data inconsistencies are examined in light of evidence in the literature regarding the nature of coverage and content errors in the data sources. Data for African-Americans are highly inconsistent in the 1970-1990 period, likely the result of age overstatement in censuses relative to death registration. Inconsistencies also exist for whites in the 1970-1980 intercensal period. We argue that the primary source of this error is an undercount in the 1970 census relative to both the 1980 census and the death registration. In contrast, the 1980-1990 data for whites, and particularly for white females, are highly consistent, far better than in most European countries.

KEY WORDS: Age misreporting; Coverage; Mortality; Census evaluation; Death registration; Data quality; Mortality crossover, United States.

1. INTRODUCTION

Conventional methods of estimating levels of mortality in more developed countries use data from two different sources. The numerators of death rates are normally counts of deaths derived from vital statistics. The denominators are usually derived from census counts of persons alive. The accuracy of calculated rates depends on the quality of data from both sources.

This paper reports the results from a test of data quality applied to United States data for two intercensal periods: 1970-1980 and 1980-1990. In particular, we examine the consistency of reported changes in the size of a cohort between two censuses and the recorded number of intercensal deaths for that cohort, with allowance for intercensal cohort migration. All data refer to the population in single years of age and separate tests are conducted for the black and white populations.

Our focus is on the older population (aged 60 and above), where mortality trends have the greatest impact on social programs (Preston 1993) and where data quality is most problematic. The white population of the United States appears to have lower death rates above age 80 than any other industrialized country (Vaupel 1993). If valid, this comparison would have important implications for evaluating the relative quality of medical systems. But the African-American population of the United States has even lower rates than the white population above age 80, reflecting the well-known crossing over of the age patterns of mortality between the races somewhere between ages 75 and 85. Whether either set of mortality rates can be accepted at face value depends, of course, on the quality of the data. Data on blacks has elicited considerable skepticism (*e.g.*, Zelnik 1969; Coale and Kisker 1990), although most observers appear to accept the validity of the crossover (Manton *et al.* 1986; McCord and Freeman 1990).

In the process of constructing new model mortality patterns for low mortality countries, Condran, Himes, and Preston (1991) report similar data quality tests for 68 intercensal periods in 18 industrialized countries. In general, consistency was very good for cohorts aged 65 at the second census (66 of 68 data sets passed the consistency check). Consistency deteriorated with age; only about half of the data sets showed consistency at age 85 and fewer than 15% did so at age 95 (Condran *et al.* 1991: Table 7). The United States was not among the countries included in these tests because it lacked published data on deaths by single year of age. We are now able to fill in this important gap because we have processed data tapes on each individual death registered in the United States from 1970 through 1988. (The single year death distribution for 1989 (full year) and 1990, January to March only, is estimated using published group data from the National Center for Health Statistics and the 1988 single-year death distribution. Details are provided in Appendix A.) These tapes are produced by the National Center for Health Statistics

[1] Laura B. Shrestha, The World Bank, Human Development Department, 1818 H Street, N.W., Washington, DC 20433, U.S.A.; Samuel H. Preston, Population Studies Center, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104, U.S.A.

(NCHS) and are distributed by the Inter-University Consortium for Political and Social Research. For the years we have included, they contain approximately 50 million deaths.

2. STUDY POPULATION AND DATA

2.1 Background

Three major sources of data are utilized: (1) national-level census enumerations from the U.S. Bureau of the Census for the years 1970, 1980 and 1990; (2) annual death registration data produced by NCHS; and (3) unpublished estimates of net immigration obtained from the U.S. Bureau of the Census. While the data sources are described in more detail in Appendix A, a brief description of the data and significant adjustments is warranted.

2.2 The Census Enumerations

We utilize census tabulations which are classified by race (black/white), sex, and single years of age (open-ended at age 100). The tabulations refer to the resident population of the 50 states and the District of Columbia. Included in the enumerations are: the institutionalized population, Americans travelling abroad temporarily, and foreign citizens having their usual residence (legally or illegally) in the United States (except foreign military and diplomatic personnel). Specifically excluded are: Americans overseas for an extended period and foreign citizens temporarily visiting the U.S. The official statistics do not adjust for census undercount, *e.g.*, the failure to find and enumerate legal residents and undocumented resident immigrants.

The term "resident population" implies that both the legal population and undocumented immigrants are included in the census tabulations. While undocumented persons were residing in the U.S. at the time of the 1970 census, it appears that only a negligible number were counted. Hence, the legal resident population approximated the total resident population in the 1970 census. In the 1980 count, however, the U.S. Bureau of the Census estimates that, for the first time ever, a significant number of undocumented persons were enumerated. Estimates indicate that the count equalled 2.06 million undocumented persons. Of this number, in the age group 60 and above, 10 thousand white males were enumerated; 19 thousand white females, 3 thousand black males, and 6 thousand black females (U.S. Bureau of the Census 1988).

The official 1970 census tabulations are known to contain errors, the most conspicuous of which is the gross overstatement of the number of persons aged 100 years or more. Although the census enumerated 106,000 persons in this age group, indirect demographic estimates indicated that the correct centenarian count should have been in the range of 3,000 to 8,000 with a preferred estimate of 4,800 (Siegel 1974; Siegel and Passel 1976; U.S. Bureau of the Census 1974). We utilize unpublished U.S. Census Bureau tabulations of the 1970 census, which correct for the centenarian overcount. Use of the corrected estimates is justified by two conditions: first, without adjustment, the excess is large enough to bias results at the oldest ages. Second, it appears that the overcount was not due to systematic misreporting of age into the centenarian population. Rather, it was the result of misunderstanding of the census form wherein individuals confused the columns intended for month of birth and year of birth (Siegel and Passel 1976).

In both the 1980 and 1990 censuses, a large number of individuals enumerated chose to write in a response to the race question as opposed to selecting one of the specified all-inclusive race categories. For the total population, 6.8 million individuals, largely of Spanish-origin, were affected in 1980, whereas the number increased to 9.3 million in the 1990 census. The official census tabulations are not directly comparable with other data sources since only the census enumerations contain a residual race category. To allow comparison with other data systems, the Census Bureau modified the 1980 and the 1990 enumerations to conform to historical categories of the racial groupings. The 1990 modification at the Census Bureau also involved "correction" for an age-related problem (for details, see Word and Spencer 1991). The decision was made to use the race-modified statistics for 1980 and 1990 from the Census Bureau for this research. The choice is justified by the sheer magnitude of individuals that would be excluded by use of the unmodified data, particularly for the white population.

2.3 Death Registration Data

The U.S. death registration data represent every death registered in the 50 states and the District of Columbia, classified by race, sex and age (single years of age to 125+). To insure comparability with the census data, deaths of nonresidents of the United States (nonresident foreign nationals and U.S. nationals residing abroad) have been excluded.

Adjustment is made for neither under-registration of deaths nor for misreporting of characteristics on the death certificates. Two problems were identified that affected the utilization of our intended intercensal methodology. The intercensal period covers the interval from April 1 to March 31, whereas the death registration data refer to calendar years. And both the death registration and the U.S. censuses' data are reported by age at last birthday rather than by year of birth. We manage both problems by assigning deaths to triangles of time-age that correspond to "census years" beginning on April 1. For example, deaths reported in the one year interval between census date April 1, 1970 and April 1, 1971 to those aged 60 (last

birthday) at the time of the census can be classified into four categories: (1) deaths to those aged 60 in calendar year 1970; (2) deaths to those aged 60 in calendar year 1971; (3) deaths to those aged 61 in calendar year 1970; and (4) deaths to those aged 61 in calendar year 1971. Using data on the date of death from the NCHS tapes, we assigned deaths to triangles of time-age that corresponded to the census year beginning on April 1, 1970. In doing so, we assume that deaths within each triangle are evenly distributed. This assumption is necessitated by the lack of reliable birth data for most of the cohorts considered in the paper, data that could be used to apportion deaths more accurately among adjacent birth cohorts. For a more detailed description of the methodology, see Shrestha 1993.

2.4 Net Immigration Statistics

We utilize unpublished net immigration statistics obtained from the U.S. Bureau of the Census. While the quality of net immigration statistics in the U.S. is widely acknowledged to be suspect (Hill 1985), the estimation of population size at the older ages is quite robust to variations in estimates of intercensal migration. This robustness results both from the smaller flow of net migrants at the older ages relative to younger ages and from the greater magnitude of deaths as a source of decrement in the older age groups relative to changes as a result of net migration. For instance, the net immigration data list an inflow of 64 black males for the cohort aged 75 and above (in 1970) during the 1970 to 1980 decade. For comparison, over 141 thousand deaths were recorded for the same cohort.

Estimates of the flow of undocumented residents are not included in the constructed net immigration series, but will be considered in the interpretation of results. Their exclusion was precipitated by a number of factors. Estimates of the size and age-sex distributions of the illegal alien population vary widely due to insufficient data collection instruments in the U.S. But even the most exaggerated estimates of the number of undocumented migrants are minuscule relative to deaths at the older ages.

We have described a number of adjustments that we have made to the basic data: use of unpublished 1970 census tabulations because of a gross overcount of the centenarian population in the official statistics, use of race-modified tabulations of the 1980 and 1990 census, and exclusion of estimates of the undocumented alien population. In order to judge the effect of these adjustments on our results, we carried out numerous sensitivity analyses using uncorrected data. While only modest differences were observed between the results using official statistics and those with corrected tabulations (except at ages 100 and above), the intercensal cohort analyses using uncorrected data generally produced greater deviations in our final results, implying the overall appropriateness of these corrections.

3. SOURCES OF ERROR IN CENSUSES AND DEATH DATA

Errors in demographic data have been classified into coverage errors and content errors. Coverage refers to the completeness with which persons or events that fall within the defined universe of a particular data system are recorded. Content refers to the quality of information about the persons or events that are in fact recorded. Either type of error in any data source can create inconsistencies between intercensal change in cohort size and intercensal deaths. However, if both censuses and death registration suffer from the same net omission rate, then the sources will be consistent with one another; but under these circumstances, recorded death rates will also be accurate.

Identical patterns of age misreporting in censuses and death registration will not, in general, produce consistency between changes in cohort size between the censuses and recorded numbers of intercensal deaths. The reason is that, because death rates rise with age, the age distribution of deaths at older ages is older than the age distribution of population. For example, if 10% of both persons and deaths at true ages 75-79 are misreported into the age interval 80-84, then the proportionate impact on population counts will be greater than the proportionate impact on death counts. Such a pattern of age misreporting would distort death rates, and would also be visible in the consistency tests that we apply.

The Census Bureau has used demographic and statistical procedures to estimate the completeness of census coverage. Demographic procedures compare estimates of the true numbers of births minus estimated cohort deaths and migrations to census counts (see the summary in Robinson *et al.* 1993 and Himes and Clogg 1992). Statistical procedures match a group of individuals identified in an alternative data source (such as the Current Population Survey) to individual-level records from the Census. A third approach is to compare the Census count of older persons to the count of individuals in Medicare files.

A number of general conclusions for the old-age population were reached in the evaluation studies of the 1970 census undertaken by the U.S. Bureau of the Census (1973, 1974, 1975). First, the magnitude of net error (combination of coverage and content errors) in the old-age statistics is greater than for the younger population. Second, females exhibited higher net error rates than males, largely the result of higher levels of age misreporting. But, gross omission rates (which are only one component of net error) were higher for males. Third, levels of net error, of gross omission, and of misreporting of demographic characteristics are considerably higher for the U.S. black population than for the white. Fourth, the evidence suggests that considerable age misreporting exists in the official statistics. For example, it is interesting to note that,

for all four race-sex groups at ages 65-69 years in 1970, the estimates derived by demographic analysis suggest net census overcounts, whereas the Medicare linkage study found gross census omissions in the magnitude of 2.1% (for white females) to 12.6% (for males of the black and other races category). This comparison implies that, while these groups have gross omissions in the number of persons enumerated at ages 65-69, other larger errors (presumably, especially age overstatement among persons below age 65) are operating in the other direction to inflate the net overcount estimates at these ages. One implication is that the characteristics of a substantial part of the population reported as 65 and over in the Census relate to persons who are in fact under age 65 (U.S. Bureau of the Census 1976).

Relative to the 1970 census, the net error rates in 1980 in most of the age-race-sex groups were significantly lower. As noted by the Bureau of the Census (1988), however, results from the Post Enumeration Program (PEP) and from the 1980 Housing Unit Enumeration Duplication study affirm that a considerable proportion of the total census count, likely in excess of 1.1%, represented duplicate enumerations of individuals already in the census. Evidence implies much lower levels of duplication in earlier censuses. Thus, "regrettably, duplication receives dubious credit for part of the improvement in 1980 in net census coverage" (U.S. Bureau of the Census 1988:10).

The Census Bureau plans exhaustive evaluations of the quality of the 1990 Census, but the release of such analyses has been fragmentary to date. It does appear that the gross undercount was lower in 1980 than in 1990 (Robinson *et al.* 1993), but this may be the result of a higher degree of duplications in the 1980 census. A number of generalizations can be made regarding the pattern of net undercount in the 1990 census for the aged population. First, following its historical trend, the net error estimates for African-Americans surpass those of whites by a wide margin. The largest differential is noted for males aged 60-64. The net undercount rate for black males equals 10.3 percent, surpassing the white male estimate of 2.6 percent by 7.7 percentage points. Second, whereas undercounts are observed for all of the male aged categories, overcounts are noted in many of the female groups. Finally, as noted by Robinson *et al.* (*ibid*), the net coverage patterns are generally consistent across the last three censuses for each race-sex group.

Official death statistics produced by the National Center for Health Statistics are the basic source of annual mortality data in the United States. The figures are generally utilized without adjustment for underregistration or for misreporting of characteristics on the death certificate. It is generally assumed, however, that the death registration system is practically complete (Wilkin 1981; U.S. Bureau of the Census 1984a; National Center for Health Statistics 1968) although no national test of its comprehensiveness has been conducted since the completion of the Death Registration Area in 1933. This assumption is based on the strict legal requirements for registration as well as on the needs of survivors for proof of death in connection with burial, settling estates and collecting insurance benefits (U.S. Bureau of the Census 1984a; Wilkin 1981). Calculations by Coale and Kisker (1990), however, suggest that underregistration of deaths exists, particularly at the older ages. For the nonwhite population, for instance, registered deaths were 7% fewer than Medicare deaths for the male population aged over 80 in 1980, whereas registered female deaths were 10% fewer. These numbers, however, may be reflective of differential age reporting between the two sources, rather than of underenumeration.

The best evidence regarding the consistency of age reporting between censuses and death registration - undoubtedly the most important source of content error affecting our consistency test - matched a sample of death certificates from May to August 1960 with the 1960 census records (NCHS 1968; Hambright 1969). Although the data were collected before the time frame considered for this project, the study's findings provide insight into what may be a continuing pattern of biases present in the census and death statistics. The authors found: (1) for whites, there was fairly high agreement between the sources even with increasing age - for nonwhites, however, there was less agreement; (2) in the event of disagreement, age discrepancies for the white population between the sources were generally within one year - for nonwhites, however, the typical difference was more than one year, particularly at ages 45 and above; and (3) for whites of all ages and nonwhites aged less than 45 years, the age reported on the death certificate was typically older than that reported on the census - for nonwhites aged 45 and above, however, age reported on the death certificate was, on average, younger than on the census.

This study was unable to ascertain which data source, if either, provides the "true" age. To this end, Rosenwaike and Logue (1983) attempted to verify age reporting on the death certificate for the population aged 85 and over in the 1968 to 1972 period. The authors selected a sample of death records from those filed for decedents of extreme age in Pennsylvania and New Jersey. They then linked the individual who died to the 1900 manuscript census of population. A total of 1429 decedents were linked of whom 960 were white and 496 were non-white.

They found that age agreement of matched census records with death certificates decreased as age increased for both racial groups. Striking differences were noted between racial groups. Agreement levels for whites were high, except at ages 100 and over. For nonwhites, however, significantly lower agreement was found. The authors further note that, within race, there was little difference by sex in agreement on age.

4. AN INTERCENSAL METHODOLOGY TO EVALUATE THE QUALITY OF OLD-AGE STATISTICS

This analysis examines the extent of inconsistency in old-age U.S. data sources using an intercensal cohort methodology. The expected size of an open-ended age cohort in the second census can be estimated from its size at the first census and the intercensal deaths occurring to that cohort, after adjustment for migration (Condran, Himes and Preston 1991). Use of an open-ended category allows observation of the ratio trend while dampening error-induced extreme values at particular ages. It is insensitive to any errors of age reporting in deaths or population that occur within the population above the age that begins the open-ended age interval.

Using census enumerations and death and migration statistics for an intercensal period, intercensal cohort analysis allows us to estimate the expected size of each open-ended age cohort in the subsequent census. The previously mentioned statistics, classified by single years of age, by sex, and by two races (white, black), were utilized to calculate the following equation for the expected population at the time of the second census:

$$\hat{N}_x(2) = N_{x-10}(1) - D_{x-10}(1) + M_{x-10}(1) \qquad (1)$$

where

$\hat{N}_x(2)$ = the predicted population aged x and above at the second census, taken 10 years after the first.

$N_{x-10}(1)$ = the enumerated population aged $x - 10$ and above at time 1, the first census.

$D_{x-10}(1)$ = the intercensal deaths which had occurred to the cohort aged $x - 10$ and above (at the first census).

$M_{x-10}(1)$ = intercensal net legal immigration into the cohort aged $x - 10$ and above (at the first census).

Similarly, the expected population *at* a given age (as opposed to at age x and above) can be calculated in an analogous manner. In either circumstance, the ratio of the observed population, enumerated in the subsequent census, to the expected population, can then be calculated (after simplifying the notation and assuming net migration to be zero) as:

$$R_x = \frac{N_x(2)}{N_{x-10}(1) - D}. \qquad (2)$$

The change in the size of the cohort as measured at two successive censuses can be produced only by death or migration. A ratio of 1.00 would indicate complete consistency among the data sources. (Note that a ratio of 1.00, while highlighting consistency, does not assure accuracy. On an individual level, for instance, if a person's age was consistently overstated by n years, the method would fail to capture the misreporting.) In fact, however, the reported count will also be affected by: (1) coverage errors in either or both censuses; (2) under- (or over-) enumeration in the death registration data and/or the immigration statistics; and (3) misreporting of characteristics (age, race, *etc.*) in any or all of the data sources (Ewbank 1981; Shryock and Siegel 1976; Condran *et al.* 1991). The ratio of observed to expected population is a useful diagnostic tool if patterns of deviation from 1.00 can be interpreted in terms of these underlying data errors. It is not a highly precise tool because different forms of error can produce the same pattern of ratios. Nevertheless, it can help discriminate among competing alternatives.

5. HOW PATTERNS OF ERROR WILL AFFECT OBSERVED/EXPECTED RATIOS

Effects of certain types of error are visible directly in the formula for the ratio itself (and have been confirmed by simulations that we have performed). To simplify the exposition, define R_x in equation (2) as the ratio of observed to expected population for age $x+$ at the second census. The following major possibilities for coverage error, and their implications for the age-pattern of ratios, can be distinguished:

1) If $N_{x-10}(1)$ and D are equally complete and $N_x(2)$ has a relative completeness level of $C(2)$, then the age pattern of ratios will be constant with age and its level will be $C(2)$.

2) If $N_x(2)$ and D are equally complete and $N_{x-10}(1)$ has a relative completeness level of $C(1)$, then the age pattern of ratios will be:

   a) Above 1.00 and rising with age if $C(1) < 1.00$

   b) Below 1.00 and falling with age if $C(1) > 1.00$.

The reason why an age trend in R_x results from this pattern of error is that a particular proportionate error in $N_{x-10}(1)$ creates increasingly larger proportionate errors in the denominator as the two offsetting terms (one positive and one negative) in the denominator grow more equal in absolute value. This equalization occurs because a higher fraction of each cohort dies during the intercensal period as age advances.

3) If $N_{x-10}(1)$ and $N_x(2)$ are equally complete and D has a relative completeness level of $C(D)$, then the age pattern of ratios will be:

   a) Above 1.00 and rising with age if $C(D) > 1.00$ (*i.e.*, if completeness of death registration exceeds the completeness of enumeration in both censuses).

   b) Below 1.00 and falling with age if $C(D) < 1.00$.

Once again, an age trend is introduced because an equal proportionate error in D will create larger proportionate errors in the denominator as its two components become more equal in absolute value.

Some of the effects of age misreporting patterns can also be understood by examining the components of this formula. Shrestha (1993) and Condran *et al.* (1991) introduce various errors into simulated errorless data sets typical of the current demographic conditions of the United States and the Netherlands respectively. They show that a pattern of net overstatement of age that is confined to the two censuses will produce a pattern of ratios that hovers around 1.00 until advanced ages, whereupon it falls to very low values. The reason why the ratio declines below 1.00 is, once again, that an error in one component of the denominator (in this case, inflation of $N_{x-10}(1)$ by age overstatement) introduces disproportionate effects in the denominator. Even though the rapid tapering off in the age distribution can result in $N_x(2)$ being more inflated than $N_{x-10}(1)$, eventually the inflation of the denominator exceeds that of the numerator and the ratios fall. (For an illustration, see Figure 1 of Condran *et al.* 1991).

Age overstatement that is confined to deaths will create a pattern of ratios that is above 1.00 and rises with age; the denominator is too low (its negative component is too large) and the proportionate deficit grows with age.

Introducing the *same* pattern of age overstatement into deaths and population figures also creates ratios that eventually rise with age. This important result is robust to the extent of error introduced (Condran *et al.* 1991). It reflects the fact that age distributions taper off more and more rapidly as age advances, so that the *same* percentage of persons who overstate their true age will introduce larger *percentage* errors in the reported age distributions at the very advanced ages. That is, $N_x(2)$ has a larger inflation factor than $N_{x-10}(1)$. In this case, some inflation in $N_{x-10}(1)$ is offset in its effects on the denominator by an inflation in D.

6. RESULTS

Intercensal cohort analysis was carried out for the four sex-race groups in the United States in the 1970-1980 and 1980-1990 periods. Figures 1 and 2 present the calculated ratios of the observed to expected population at selected ages by race, sex, and intercensal period.

In all race-period combinations, the age pattern of ratios is virtually the same for females and males. In all cases, the degree of inconsistency increases with age, although any systematic and significant departure from 1.00 is postponed until age 95 and beyond for whites in 1980-1990. There is clearly a discontinuity in many of these series at age 100, reflecting the idiosyncrasies of age reporting and Census Bureau adjustment procedures among centenarians.

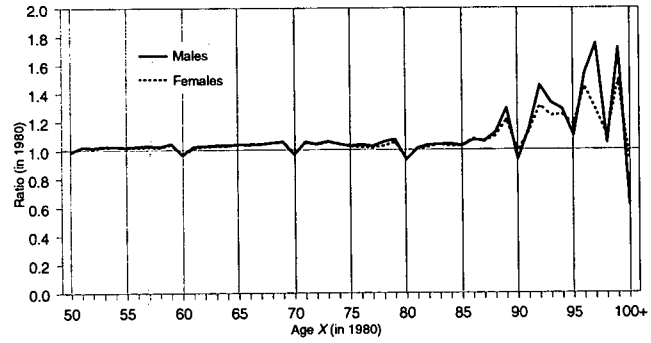


Figure 1A. Intercensal Ratios of Observed to Expected Population: Whites, 1970-1980.

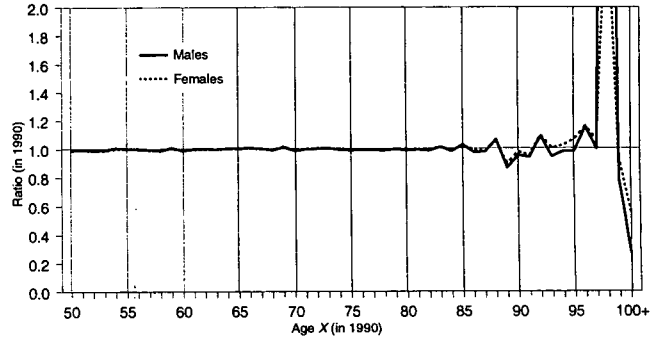


Figure 1B. Intercensal Ratios of Observed to Expected Population: Whites, 1980-1990.

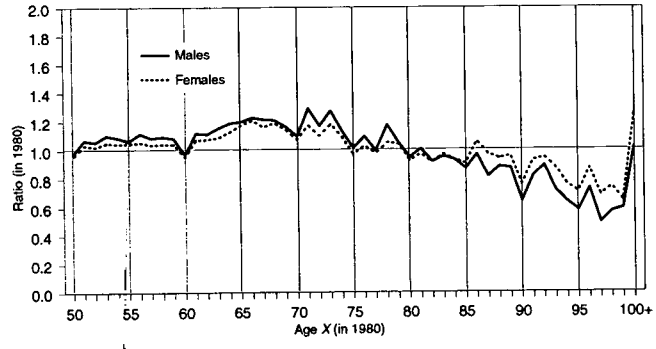


Figure 2A. Intercensal Ratios of Observed to Expected Population: Blacks, 1970-1980.

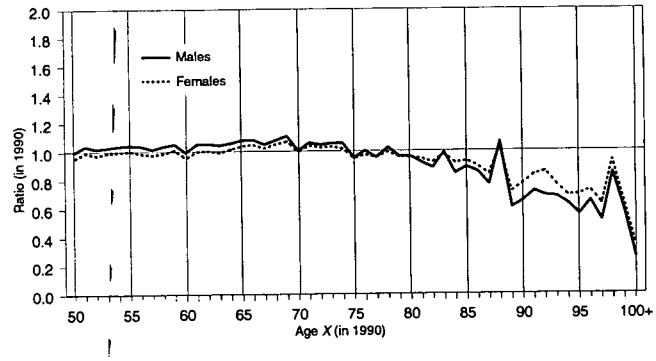


Figure 2B. Intercensal Ratios of Observed to Expected Population: Blacks, 1980-1990.

6.1 Results for Whites

6.1.1 Intercensal Period: 1970-1980

As shown in Figure 1A, the white pattern in 1970-1980 is generally above unity and rising with age (up to age 100). This pattern is consistent with several forms of data error, the two most plausible patterns of which are:

1) Undercount in the 1970 census, relative to both the 1980 census and the death registration.

2) Roughly equal probabilities of age overstatement in deaths and in both censuses.

We believe that the former explanation is more likely to be correct. If the pattern of ratios resulted from similar tendencies for age misstatement in deaths and censuses, one would expect that pattern to continue into the 1980-1990 decade, particularly since the 1980 census is involved in both comparisons. And one would not expect cultural predispositions to misstate age to disappear suddenly. But the 1980-1990 pattern of ratios for whites (Figure 1B) shows remarkable consistency, far better than that in most European countries and equivalent to the pattern of ratios found in Sweden and the Netherlands, countries with highly efficient population registers (Condran *et al.* 1991). The consistency during 1980-1990 is also much greater than that in other English-speaking countries: England and Wales, Canada, Australia and New Zealand.

A second reason for accepting the first explanation is that the Census Bureau has concluded that the 1980 census is more complete than the 1970 census (U.S. Bureau of the Census 1988; Robinson *et al.* 1993). This conclusion is partially based on demographic analysis and hence is not entirely independent of the kind of evidence that we are reviewing. However, their demographic analysis is weighted heavily towards ages that are younger than those considered here. Furthermore, the conclusion that census coverage improved is also supported by their post-enumeration program in which individuals in the census are matched against other data systems.

6.1.2 Intercensal Period: 1980-1990

As noted earlier, the 1980-1990 pattern of ratios for whites (and particularly for white females) is highly consistent, far better than in most European countries. Our investigation seemingly lends support to Vaupel's (1993) contention that the white population of the United States may have lower death rates above age 80 than any other industrialized country. But caution is in order. While our methods clearly highlight the consistency between the censuses and the death registration in 1980-1990, consistency is not equivalent to accuracy. Condran *et al.* (1991) demonstrate one situation in which a pattern of age misreporting can result in a ratio series at exactly 1.00 at all ages. Furthermore, the intercensal methodology fails to capture deliberate misreporting of age by individuals that is consistent over time. As noted by Horiuchi (1993), an initial overstatement of age – *e.g.*, to allow entrance into school or the labor force at a younger age, to avoid being drafted near the upper limit of drafting age, or to receive Social Security, Medicare, or pension payments earlier – may be followed by consistent, intentional overstatement of age. The possibility of such overstatement of age cannot be discounted although we are unable to measure it directly.

6.2 Results for Blacks

In contrast, the pattern of ratios for African-Americans is far more regular over time (see Figure 2A and 2B). The ratios begin falling around age 70 for both sexes in both periods and continue falling through higher ages (until age 100 in 1970-1980). Before age 70, ratios are typically well above unity in 1970-1980, and slightly above 1.00 for African-American males during 1980-1990.

The fact that ratios are generally higher for African-Americans at a particular age in 1970-1980 than in 1980-1990 is consistent with a relative undercount in the 1970 census. As we noted earlier, such an undercount is also likely to have occurred among whites. The undercount, however, is insufficient to explain the persistent pattern of falling ratios above age 70 in both periods. The declining ratio series for African-Americans is consistent with two principal explanations:

1) Deaths are underregistered for the African-American population relative to completeness of census coverage.

2) Age overstatement is greater in censuses than in death registration.

Coale and Kisker (1990) lean toward the former explanation. They note that populations reconstructed from deaths using variable-*r* procedures (Preston and Coale 1982) are too small relative to census counts in 1980 above age 65, suggesting relative underregistration of deaths. They also note that fewer African-American deaths are recorded at advanced ages in vital registration than in Medicare records.

However, both observations are also consistent with ages being overstated in censuses (and Medicare) relative to death registration. That such a pattern exists is strongly supported by a direct match of death certificates in 1960 to records for the same individuals in the 1960 census of population (NCHS 1968; Hambright 1969). For either males or females, the total number of deaths above age 50 when deaths are classified according to census age are within 1% of the total number of deaths when classified according to death certificate age. However, at ages 65+, "census age" deaths are 15.4% greater than "death certificate age" deaths for females and 7.1% greater for males. At age 75+, the disparities are 23.3% and 17.8%, respectively, and at age 85+, 39.2% and 17.6%.

These large discrepancies in age reporting between censuses and deaths are capable of accounting for the declining pattern of ratios above age 70 that is demonstrated in Figure 2. Elo and Preston (1994) calculate the R_x values for African-Americans between 1950-1960 and 1960-1970, periods that bound the 1960 census-death certificate match. They show that, if ages at death are "corrected" to make them consistent with the age reporting in the censuses, the pattern of declining ratios is eliminated.

Reasons why African-American ages are overstated in censuses relative to deaths are not obvious. The pattern does not appear until the 1940 census, the first census after Social Security legislation was passed. At that census, a large surplus of African-American persons aged 65-69 and 70-74 appears, and a deficit of persons aged 50-64 (Elo and Preston 1994). As noted by Wolfenden (1954:56), "the disturbances were so marked in the data for Negroes that special preliminary redistributions of those populations (and deaths) between 55 and 69 were made in the preparation of the [U.S.] life tables." This surplus also appears, although in increasingly attenuated form, in more recent censuses (as shown in Figure 2). Whatever its source, we believe that the principal explanation of the large inconsistencies between censuses and death registration for the African-American population is a pattern of age overstatement in censuses relative to death registration. Such a pattern implies that recorded death rates above age 65 for African-Americans are likely to be seriously underestimated. A cross-over between black and white death rates may indeed occur at advanced ages, but basing such a conclusion on U.S. census and vital registration data is treacherous. These data are simply too inconsistent with one another to allow death rates at advanced ages to be estimated with any confidence.

7. CONCLUSION

Major uncertainties about the quality of elderly population and death enumerations in the United States result from coverage and content errors in the censuses and the death registration system. This study evaluates the consistency of reported data between the two sources for the white and the African-American populations. The focus is on the older population (aged 60 and above), where mortality trends have the greatest impact on social programs and where data are most problematic. Using intercensal cohort analysis, age-specific inconsistencies between the sources are identified for two periods, 1970-1980 and 1980-1990.

In order to evaluate what combinations of coverage completeness and age misreporting patterns would produce the empirical results, a series of simulations were carried out. The U.S. data inconsistencies are examined in light of both the simulation results and evidence in the literature regarding the nature of coverage and content errors in the data sources.

Data for whites in the 1980-1990 intercensal period were found to be remarkably consistent. Data quality up to age 95 approaches that of Sweden and the Netherlands, countries which maintain highly efficient population registers. Less consistency was observed for whites during the 1970-1980 decade. The most likely explanation for this pattern of inconsistencies is the relative net undercount in the 1970 census combined with more complete death statistics. Consequently, mortality estimates at older ages that combine numerators from the death registration with denominators from the 1970 census are likely to overstate mortality.

A different pattern is observed in the African-American data. Above age 70, the enumerated population falls increasingly below the expected population in both 1980 and 1990. It appears that the major reason for this pattern is that ages are overstated in censuses relative to death registration. Such a pattern implies that recorded death rates at older ages for African-Americans are likely to be seriously underestimated. A mortality crossover between black and white death rates may occur at advanced ages, but basing such a conclusion on census and vital registration data is hazardous.

ACKNOWLEDGEMENTS

This research was carried out at the University of Pennsylvania, Population Studies Center, and was supported by a grant from the National Institute of Aging, AG10168, and from the Boettner Institute of Financial Gerontology. For helpful comments on the project and/or paper, we are indebted to Irma Elo, Douglas Ewbank, Shiro Horiuchi, and J. Gregory Robinson. We are especially grateful to J. Gregory Robinson and the U.S. Bureau of the Census for supplying unpublished census and international migration tabulations.

APPENDIX A

Source: Shrestha (1993)

Three major sources of data were utilized in this research: (1) census enumerations for 1970, 1980, and 1990; (2) official death registration data; and (3) net immigration statistics. Sources of the data and adjustments made will be described.

1.A The 1970 Census

Official tabulations of the 1970 population by basic demographic characteristics are presented in *Series B – U.S. Summary of the 1970 Census* (U.S. Bureau of the Census 1972). The official enumerations are known to

contain a number of major inaccuracies which could bias our investigation of the enumerated old-age population in the United States. The first is a conspicuous overcount of the centenarian population. Whereas 106,000 persons were enumerated in the open-ended category, indirect demographic analysis estimates the correct count to be in the range of 3,000 to 8,000 (Siegel 1974; Siegel and Passel 1976). The overcount appears to have been the result of misunderstanding of the census form rather than systematic age misreporting into the centenarian population. The second problem is the result of misclassification of the population by race in the complete-count tabulations, affecting 21,000 individuals aged 65 and above. And finally, the official count omitted over 23,000 individuals (of all ages) whose records were discovered after the initial tabulations were published.

Because of the inherent errors in the official tabulations, we utilize unpublished adjusted tabulations obtained from the U.S. Bureau of the Census. The modified statistics include corrections for the three previously mentioned problems. The data are presented by race (white, black), sex, and age (single years of age 0-94 and grouped data 95-99, 100+). To distribute the grouped data from age group 95-99 to single years of age, we used the sex-and race-specific average age distribution from the 1960 and 1980 censuses for whites, and from the 1950 and 1980 censuses for blacks (data by single years of age is not available in this age range for blacks in the 1960 census).

1.B The 1980 Census

Originally published census tabulations for 1980 were presented in *Series B - U.S. Summary of the 1980 Census* (U.S. Bureau of the Census 1983). In the 1980 census, however, a large number (about 6.8 million) of persons enumerated chose to write-in a response to the race question as opposed to selecting one of the specified all-inclusive race categories. Since only the 1980 census contained a residual race category, the official enumeration was not directly comparable with other data sources (vital registration, earlier censuses, *etc.*). The Census Bureau produced a modified file which conforms to the historical categories of the racial groupings (U.S. Bureau of the Census 1984b). The modification procedure involved macro-level reassignment of race based on detailed cross-tabulation of race and Hispanic-origin from the sample and complete-count census data. The specifics of the Census Bureau modification follow.

For the 219.8 million individuals who chose one of the 14 specified categories, no adjustment was made. Two categories of individuals, totalling 6.7 million, with write-in responses were identified: persons of Hispanic-origin (5.8 million) and persons not of Hispanic-origin (0.9 million). Separate adjustment procedures for the two groups were developed.

Those of Hispanic-origin were distributed only to the white or black categories (and not to American Indian or Asian/Pacific Islander categories). All persons of Mexican origin were reassigned as white. Persons of Puerto Rican, Cuban, and other Spanish origin were assigned to both white and black modified race groups on the basis of the distribution of the same Hispanic-origin individuals who originally specified either a white or black race on the census returns. The calculations were carried out within age-sex-county cells.

Those not of Hispanic-origin were reassigned to all three modified race groups (white, black, other) on the basis of state-specific proportions which are applied to all age-sex-county cells within the state. The proportions are based on sample data from the 1980 census. For a more detailed discussion of the modifications, see U.S. Bureau of the Census 1984b.

The modified tabulations are presented by race, sex, and single years of age (0-99; 100+). We utilize the race-modified statistics in this research, justified by the sheer magnitude of persons transferred from the residual race category to the white or black categories.

1.C The 1990 Census

Published tabulations of the 1990 Census continue to be released by the U.S. Bureau of the Census. The published statistics, however, contain a number of problems that make comparability with earlier censuses and other sources of data difficult. Three problems are apparent: racial classification of 9.3 million individuals in a residual non-specified racial category, inconsistencies in the reporting of age, and a change in allocation procedures for the 1990 census in assigning age to persons with missing data on the characteristic.

A modified 1990 census file, referred to as the MARS (Modified Age and Race Statistics) was produced at the Census Bureau to adjust for the first two problems (Word and Spencer 1991). Modification of the 1990 census was conducted at the micro-level. Hot-deck imputation procedures were utilized to assign a specific race to persons who reported themselves in the "other, not specified" racial category. The method is executed on the individual records of the 100% edited detail file from the 1990 census (Robinson, Word and Spencer 1991).

We again utilize the modified statistics, which are tabulated by race, sex, and single years of age, in this research. The decision to use the modified statistics in both 1980 and 1990 was not clear-cut. See Shrestha 1993 for a more detailed discussion.

2. The Death Registration System

National-level annual death statistics from the National Center for Health Statistics (NCHS) are utilized in this research. The data for 1970 through 1988 are extracted

from NCHS data tapes obtained from ICPSR (NCHS 1970-1988). The data are provided by race (black, white), sex, and single years of age (0-124; 125+). Since the data tapes for calendar year 1989 and for the first three months of 1990 had yet to be released, we developed a procedure to estimate the distribution. Final mortality statistics for 1989 by race and sex were released in published form by NCHS (1992). The grouped age data was distributed to single years of age based on the 1988 death distribution within the grouped age category. Distribution to month of death was based on monthly vital statistics reports (NCHS 1989). Estimates of the death distribution in 1990 are based on monthly advance reports of mortality from NCHS (1990). The preliminary numbers were distributed to single years of age again using the 1988 distribution within the grouped age category.

As noted in the text, we adjusted the available data to correct for two problems. First, the intercensal period covers the interval from April 1 to March 31, whereas the death registration data refer to calendar years. Second, both sets of data are reported by age at last birthday rather than by year of birth. Because the census is on April 1, the latter is preferred because it identifies the birth cohort for use in cohort analysis. To adjust for these two problems, we assume that the three dimensional surface of the number of deaths in age and time is level over the interval. We do not adjust for underregistration nor for misreporting of characteristics in the death statistics.

3. Net Immigration Statistics

We utilize unpublished net immigration statistics obtained from the U.S. Bureau of the Census. The tabulations are categorized in the form of "components of change" for each of the two decades.

Age-, race-, and sex-specific net immigration was calculated on a cohort basis by use of the following equation:

Net immigration = Legal Alien Immigration + Refugees and Parolees + Net Civilian Citizens Immigration + Net Puerto Rican Immigration + Net Foreign Students Immigration + Net Movement of U.S. Armed Forces Overseas − Legal Emigration.

Given the lack of sufficient detail in the raw data provided by the U.S. Census Bureau, a number of adjustments were required. First, the data had been provided with an early terminal age group (age 75 and above at the beginning of the decade). To distribute to five-year age groups (75-79, ..., 95-99, 100+), we assumed that the age-, race-, and sex-specific net immigration rate for ages 75+ remained constant in the open-ended interval beginning at age 75. This admittedly crude estimate is adequate because of small numbers of net immigrants in this age group. Second, to convert the five-year data into single years of age, we used Sprague multipliers or osculatory interpolation (Sprague 1880-81).

REFERENCES

COALE, A.J., and KISKER, E.E. (1990). Defects in data on old-age mortality in the United States: new procedures for calculating mortality schedules and life tables at the highest ages. *Asian and Pacific Population Forum*, 4(1), 1-31.

CONDRAN, G.A., HIMES, C.L., and PRESTON, S.H. (1991). Old-age mortality patterns in low-mortality countries: an evaluation of population and death data at advanced ages, 1950 to present. *Population Bulletin of the United Nations*, 30, 23-60.

ELO, I.T., and PRESTON, S.H. (1994). New estimates of old-age mortality among African-Americans, 1930-1990. *Demography*, 31(3), 427-58.

EWBANK, D.C. (1981). *Age Misreporting and Age-Selective Underenumeration: Sources, Patterns, and Consequences for Demographic Analysis*. National Academy of Sciences, Committee on Population and Demography. Report No. 4. Washington, DC: National Academy Press.

HAMBRIGHT, T.Z. (1969). Comparison of information on death certificates and matching 1960 census records: age, marital status, nativity, and country of origin. *Demography*, 6(4), 413-24.

HILL, K. (1985). Illegal aliens: an assessment. In: Panel on Immigration Statistics. *Immigration Statistics, A Story of Neglect*. Washington, DC: National Academy Press.

HIMES, C.L., and CLOGG, C.C. (1992). An overview of demographic analysis as a method for evaluating census coverage in the United States. *Population Index*, 58(4), 587-607.

HORIUCHI, S. (1993). Personal communication dated April 20, 1993.

MANTON, K.G., STALLARD, E., and VAUPEL, J.W. (1986). Alternative models for the heterogeneity of mortality risks among the aged. *Journal of the American Statistical Association*. 81, 635-644.

MC CORD, C., and FREEMAN, H.P. (1990). Excess mortality in Harlem. *New England Journal of Medicine*. 322, 172-177.

NATIONAL CENTER FOR HEALTH STATISTICS (1968). *Comparability of age on the death certificate and matching census record: United States - May - August 1960*; Vital and Health Statistics: Data Evaluation and Methods Research. By Thea Zelman Hambright. Series 2, No. 29.

NATIONAL CENTER FOR HEALTH STATISTICS (1970-1988). *Mortality Detail Files* (data and codebooks). Data were made available by the Inter-University Consortium for Political and Social Research, University of Michigan.

NATIONAL CENTER FOR HEALTH STATISTICS (1989). Births, marriages, divorces, and deaths for January-December 1989. *Monthly* vital statistics report. Vol. 38, Nos. 1-12. Hyattsville, Maryland: Public Health Service.

NATIONAL CENTER FOR HEALTH STATISTICS (1990). Births, marriages, divorces, and deaths for January-March 1990. *Monthly* vital statistics report. Vol. 39, Nos. 1-3. Hyattsville, Maryland: Public Health Service.

NATIONAL CENTER FOR HEALTH STATISTICS (1992). Advance report of final mortality statistics, 1989. *Monthly* vital statistics report. Vol. 40, No. 8, supp. 2. Hyattsville, Maryland: Public Health Service.

PRESTON, S.H. (1993). Demographic change in the United States, 1970-2050. *Demography and Retirement: The 21st Century*, (Sylvester Scheiber, Ed.). New York: Praeger Press.

PRESTON, S.H., and COALE, A.J. (1982). Age structure, growth, attrition, and accession: a new synthesis. *Population Index*. 48(2), 217-59.

ROBINSON, J.G., AHMED, B., DAS GUPTA, P., and WOODROW, K.A. (1993). Estimation of population coverage in the 1990 United States census based on demographic analysis. *Journal of the American Statistical Association*, 88, 1061-1079.

ROBINSON, J.G., WORD, D.L., and SPENCER, G. (1991). Uncertainty for models to translate 1990 census concepts into historical racial classifications. 1990 Decennial Census, Preliminary Research and Evaluation Memorandum (PREM) No. 81, Demographic Analysis Evaluation Project D8.

ROSENWAIKE, I., and LOGUE, B. 1983. Accuracy of death certificate ages for the extreme aged. *Demography*, 20(4), 569-585.

SHRESTHA, L.B. (1993). Age Misreporting and its Effects on Old-Age Population and Death Registration Estimates: United States, 1970-1990. Unpublished doctoral dissertation, University of Pennsylvania, Population Studies Center.

SHRYOCK, H.S., and SIEGEL, J.S. (1976). *The Methods and Material of Demography*. Orlando, Florida: Academic Press, Inc. (Harcourt Brace Jovanovich, Publishers): Studies in Population Series.

SIEGEL, J.S. (1974). Estimates of coverage of the population by sex, race, and age in the 1970 census. *Demography*, 11(1), 1-23.

SIEGEL, J.S., and PASSEL, J.S. (1976). New estimates of the number of centenarians in the United States. *Journal of the American Statistical Association*. 71, 559-566.

SPRAGUE, T.B. (1880-81). Explanation of a new formula for interpolation. *Journal of the Institute of Actuaries*. 22, 270.

UNITED STATES BUREAU OF THE CENSUS (1972). *General Population Characteristics*. 1970 Census of Population and Housing. Final Report PC(1)-B1. U.S. Summary. Washington, DC: U.S. Government Printing Office.

UNITED STATES BUREAU OF THE CENSUS (1973). *The Medicare Record Check: an Evaluation of the Coverage of Persons 65 Years of Age and Over in the 1970 Census*. 1970 Census of Population: Evaluation and Research Program, PHC(E)-7. Washington, DC: U.S. Government Printing Office.

UNITED STATES BUREAU OF THE CENSUS (1974). *Estimates of Coverage of Population by Sex, Race, and Age: Demographic Analysis*. 1970 Census of Population: Evaluation and Research Program, PHC(E)-4. By Jacob S. Siegel. Washington, DC: U.S. Government Printing Office.

UNITED STATES BUREAU OF THE CENSUS (1975). *Accuracy of Data for Selected Population Characteristics as Measured by the 1970 CPS-Census Match*. 1970 Census of Population: Evaluation and Research Program, PHC(E)-11. Washington, DC: U.S. Government Printing Office.

UNITED STATES BUREAU OF THE CENSUS (1976). *Demographic Aspects of Aging and the Older Population in the United States*. Current Population Reports. Series P-23, No. 59. By J.S. Siegel. Washington, DC: U.S. Government Printing Office.

UNITED STATES BUREAU OF THE CENSUS (1983). *General Population Characteristics*. 1980 Census of Population and Housing. Final Report PC80-1-B1. United States Summary. Washington, DC: U.S. Government Printing Office.

UNITED STATES BUREAU OF THE CENSUS (1984a). *Demographic and Socioeconomic Aspects of Aging in the United States*. Current Population Reports. Series P-23, No. 138. By J.S. Siegel and M. Davidson. Washington, DC: US Government Printing Office.

UNITED STATES BUREAU OF THE CENSUS (1984b). Census of Population: 1980. Race detail file. 100% count. Table IV: modified counts (OMB-consistent) by age, race, and sex. Unpublished tabulations.

UNITED STATES BUREAU OF THE CENSUS (1988). *The Coverage of Population in the 1980 Census*. 1980 Census of Population and Housing: Evaluation and Research Reports, PHC80-E4. By R.E. Fay, J.S. Passel and J.G. Robinson. Washington, DC: U.S. Government Printing Office.

VAUPEL, J.W. (1993). Verbal presentation at Research Workshop on Oldest Old Mortality, Duke University, Durham, North Carolina. March, 1993.

WILKIN, J.C. (1981). Recent trends in the mortality of the aged. *Transactions of the Society of Actuaries*. Vol. XXXIII, 11-62.

WOLFENDEN, H.H. (1954). *Population Statistics and their Compilation*. Revised Edition. The University of Chicago Press: published for the Society of Actuaries.

WORD, D.L., and SPENCER, G. (1991). Age, sex, race, and Hispanic origin information from the 1990 census: a comparison of census results with results where age and race have been modified. 1990 CHS-L-74. Draft dated August 1991.

ZELNIK, M. (1969). Age patterns of mortality of American Negroes: 1900-02 to 1959-61. *Journal of the American Statistical Association*. 64, 433-451.

Survey Methodology, December 1995
Vol. 21, No. 2, pp. 179–184
Statistics Canada

An Assessment of the Use of Hand-Held Computers During Demographic Surveys in Developing Countries

D. FORSTER and R.W. SNOW[1]

ABSTRACT

Although large scale surveys conducted in developing countries can provide an invaluable snapshot of the health situation in a community, results produced rarely reflect the current reality as they are often released several months or years after data collection. The time lag can be partially attributed to delays in entering, coding and cleaning data after it is collected in the field. Recent advances in computer technology have provided a means of directly recording data onto a hand-held computer. Errors are reduced because in-built checks triggered as the questionnaire is administered reject illogical or inconsistent entries. This paper reports the use of one such computer-assisted interviewing tool in the collection of demographic data in Kenya. Although initial costs of establishing computer-assisted interviewing are high, the benefits are clear: errors that can creep into data collected by experienced field staff can be reduced to negligible levels. In situations where speed is essential, a large number of staff are involved, or a pre-coded questionnaire is used to collect data routinely over a long period, computer-assisted interviewing could prove a means of saving costs in the long term, as well as producing a dramatic improvement in data quality in the immediate term.

KEY WORDS: Hand-held computers; Demographic surveys; Psion.

1. INTRODUCTION

Large scale surveys involving tens of thousands of respondents, such as national censuses, demographic or health surveys, are routinely conducted in developing countries. Their intention is to provide rapid, up-to-date information on population and health issues for evaluation and planning purposes. Their wide scope necessitates numerous personnel comprising trainers, interviewers, supervisors, data entry staff and data managers. Examples of such questionnaire-based surveys include the World Fertility Survey (WFS 1986) and national Demographic and Health Surveys (DHS Kenya 1989). Published dates for the commencement of the WFS surveys in 12 African countries and the dates the first country reports were produced (Table 1) illustrate the time required before data was available for planners to act upon (WFS 1986). On average it took 45.6 months before the final report was released. Survey logistics in developing countries undoubtedly contribute to delays in provision of completed data; so do the mechanics of data processing. The recent Demographic and Health Survey conducted in Kenya required five data entry clerks, two data entry supervisors and a control clerk to process 8,343 household interviews; data collection began in February 1989 and the first draft of the final report was ready for circulation seven months later (DHS Kenya 1989).

Table 1

Summary of Chronology of 12 African WFS Surveys (Source: WFS 1986)

Country	Number of Interviews	Date Survey Started	Date of First Report	Number of Months from Survey Start Till Report Date
Benin	4,018	12/1981	06/1984	30
Cameroon	8,219	01/1978	04/1983	63
Ghana	6,125	02/1979	06/1983	52
Ivory Coast	6,270	08/1980	12/1984	52
Kenya	8,100	08/1977	06/1980	34
Lesotho	3,603	08/1977	12/1981	52
Mauritania	3,500	01/1981	06/1984	41
Morocco	5,800	04/1980	05/1984	49
Nigeria	9,727	10/1981	09/1984	35
Senegal	3,985	05/1978	07/1981	38
Sudan (North)	3,115	12/1978	04/1982	40
Tunisia	4,123	05/1978	06/1983	61

[1] D. Forster, Department of Tropical Medicine, University of Oxford, John Radcliffe Hospital, Headington OX3 9DU, England; R.W. Snow, CRC - Research Unit, Kenyan Medical Research Institute, P.O. Box 230, Kilifi, Kenya.

Surveys of this size involve multiple levels of checking and coding of data collected in the field providing another source of delay. As speed underpins rapid health evaluation (Anker 1991; Vlassoff and Tanner 1992), reducing the time at this check and code stage is a major advantage to the survey process. Advances in computer hardware have led to the development of microcomputers suitable for use in field situations. Together with improved software designed for questionnaire specification and administration, computer-assisted interviewing is now a viable option. National statistics offices in industrialised countries have evaluated the use of this technique, and some now use them on a regular basis (Nicholls and Groves 1986; Lyberg 1985; Denteneer *et al.* 1987; Bench *et al.* 1994). The advantages of these systems are that it reduces recording errors by simplifying skip modules and refusing inappropriate, illogical or inconsistent entries. Furthermore, large numbers of interviews can be stored and simply downloaded to a central computer at the end of every interview session, circumventing the need for data entry clerks.

There is surprising reluctance to adopt this technology in developing countries despite its apparent advantages. There are several possible reasons for this. Firstly, the initial costs may seem daunting and the application deemed inappropriate in countries with scarce resources. Secondly, there have been few attempts to validate their use under field conditions providing little quantifiable evidence of their limitations or advantages over traditional data collection techniques (Reitmaier 1985; Ferry and Cantrelle 1988; Forster *et al.* 1991). This paper presents the results of a comparative study of two methods of field data collection and processing conducted during a demographic survey on the Kenyan Coast.

2. THE ADULT MORTALITY SURVEY

The study was carried out as part of ongoing demographic and epidemiological studies of 60,000 people living on the Kenyan coast. The study population and survey methods employed to monitor demographic events has been described elsewhere (Snow *et al.* 1994). In brief, following an initial census of the population all vital events are monitored by means of 6-weekly house-to-house visits and bi-annual re-censuses of the entire population. During a re-enumeration of the population in November 1993, a survey was undertaken to estimate adult mortality using indirect demographic methods (Timaeus 1991). All women aged between 25 and 44 years were interviewed using the structured questionnaire as shown in Figure 1. The format used precoded closed questions, with logical skips and a consistency check.

Twenty-four field staff, all secondary school leavers, were involved in the survey. All were familiar with survey and census procedures, having had previous formal training in field survey techniques and between 1 and 5 years of field experience. Two days was spent on additional training on the administration of the adult mortality questionnaire. During the survey field staff were divided into two teams, each supervised by a senior fieldworker. Questionnaires completed at the end of each day were checked by field supervisors then passed to the computer staff for data entry. This was done using a screen design reflecting the structure of the paper questionnaire in FoxPro (version 2.0). The same data was independently entered by two data entry clerks and the two completed files compared to identify entry errors, which were subsequently corrected. The completed file was then subjected to logical, range and consistency checks; these included for example, the identification of missing data, incorrect coding (*i.e.*, not using "Y" or "N"), dates inconsistent with the ages of the women and the date of the survey (questions 5 and 6 in Figure 1) and checks that the sums of questions 7, 9, 11 and 13 are consistent with question 15 as shown in Figure 1.

3. COMPUTER DATA COLLECTION TEST

3.1 Computer Hardware and Software

An earlier version of questionnaire-based software was developed for the Psion Organiser II (Forster *et al.* 1991). This model had a limited screen size, 16 characters by 2 lines, but had a fully operational keyboard. The Psion Series 3, used during the present study, offers new possibilities: the screen is much larger, with 40 characters by 8 lines, and integrated graphical capabilities. The machine remains small (165mm by 85mm by 22mm), and weighs 265g including 2 AA sized batteries. The storage devices can store up to 1 megabyte. The keyboard is a 58 key, QWERTY layout. Communications between the Psion Series 3 and a PC entails a simple copy operation between the two storage media.

The software was developed using Psion's in-built programming language, OPL. The paper questionnaire is represented in a structured format in a text file, according to a prescribed format. The questionnaire definition includes a mixture of questions and commands such as skips and range checks. The internal range checks included those developed for the inconsistency checks for the data entered using FoxPro described above. Data entered on the Psion is stored in a separate file, one line for each interview.

To specify a question correctly it must include a question number, the question text and the answer type, which can be a list option, a character input or a number. The definition should also indicate what position in the line the corresponding data entry should be stored and how long the entry is. Numeric answers can also include a prespecified number of decimal points. A range of acceptable inputs

Figure 1. The Adult Mortality Questionnaire

Questionnaire on the survival of relatives
(For all women aged 25-44 years)

Names
Date ID |__|__|-|__|__|__|-|__|__|

I would like to ask you some questions about your natural parents and about your brothers and sisters who have the same mother as you.

1. Is your mother alive? (1 = yes, 2 = no) |__|
2. Is your father alive? (1 = yes, 2 = no) |__|
   INTERVIEWER: If both parents alive (Q1 and Q2 = 1), go to Q7.
3. Have you ever given birth? (1 = yes, 2 = no) |__|
   INTERVIEWER: If she has never given birth (Q3 = 2), go to Q6.
4. Was (*MENTION ALL PARENTS NOT ALIVE NOW*) alive at the time that you gave birth to your first child?

	Yes	No	D/K	
Woman's mother	1	2	9	
Woman's father	1	2	9	

5. In what year was your first child born? |__|__|
6. In what year (*MENTION ALL PARENTS NOT ALIVE NOW*) die?
   Woman's mother |__|__|
   Woman's father |__|__|
7. How many living sisters, born to your mother, do you have? (ALIVE NOW) |__|__|
   INTERVIEWER: If no living sisters (Q7 = 0), go to Q9.
8. How many of these living sisters are less than 15 years old? |__|__|
9. How many of your sisters, born to your mother, have died? |__|__|
   INTERVIEWER: If no dead sisters (Q9 = 0), go to Q11.
10. How many of these dead sisters died before age 15 years? |__|__|
11. How many living brothers, born to your mother, do you have? (ALIVE NOW) |__|__|
    INTERVIEWER: If no living brothers (Q11 = 0), go to Q13.
12. How many of these living brothers are less than 15 years old? |__|__|
13. How many of your brothers born to your mother, have died? |__|__|
    INTERVIEWER: If no dead brothers (Q13 = 0), go to Q15.
14. How many of these dead brothers died before age 15 years? |__|__|
    INTERVIEWER: Sum Q7, 9, 11 and 13:
    Q7 = |__|__|
    Q9 = |__|__|
    Q11 = |__|__|
    Q13 = |__|__|
15. I want to make sure that I have this right. Apart from you, your mother had . . . . . |__|__|
    children altogether? Is that correct?
    INTERVIEWER: In the case of any inconsistency, probe and correct Q7 to Q14 if necessary.

    INTERVIEWER: Please thank the woman for her co-operation.

Fieldworker code |__|__|

is an optional specification for numeric or character answers and will include a minimum, a maximum or both. List options can be used to specify codes and their values.

Command actions can be embedded in question texts, so that they are evaluated at the time of questionnaire administration. For example, the final cross-check question in Figure 1 requires an addition. The syntax allows this instruction to be included within the main body of the question text. Other commands can contain instructions on skipping a question, conducting a cross-check between answers or for moving to a different question.

Thus the software uses a flexible way of defining a questionnaire, which is generally applicable. It incorporates manipulation of entered information, integrating arithmetic functions into questions or command lines. The next step forward would be to design an interface for questionnaire specification which removes the burden of constructing a syntactically correct questionnaire definition. The software is available from the authors.

3.2 Test Design

An additional day's training on the use of the Psion was provided for the two team leaders. This involved an explanation of the hardware and software, as well as practice sessions in the field. Both supervisors had had no previous computing experience. Both conducted interviews using either the Psion or paper questionnaires on alternate days, and these formed the basis of the comparison between the methods.

Errors made by all the 22 fieldworkers using the paper questionnaire were counted and tabulated to estimate the background error rate using this method of data collection. Times taken to check the forms once they had been brought back from the field, and for the data to be entered, verified and corrected following range and consistency checks were recorded throughout the survey. Similar timing assessments were made for the Psion data collection procedures.

4. TEST RESULTS

4.1 Time

The average length of interviews conducted on paper was 5.1 minutes, and for those on computer 5.0 minutes, demonstrating no difference between the two methods (Table 2; note only 215/234 interviews were timed). The length of interviews varied considerably from 1 to 18 minutes and increased time related not simply to the number of skips made on each interview, but whether the respondent gave clear, non-contradictory answers.

Team supervisors required between 2-3 hours per day to check each teams questionnaires. The average time taken to enter 500 records (approximate number of interviews completed per week) was 3 hours 40 minutes per data entry clerk; double entry required 7 hours 20 minutes (Table 3). Verification required an additional 2 hours 23 minutes for the same number of questionnaires. The completed files were edited twice to reflect corrected errors and then verified; this took on average two hours 30 minutes for 500 records.

Table 2

Comparison Between Paper Questionnaire and Psion Series 3 Data Collection Methods

	Length of Interview in Minutes				
	Minimum Overall	Maximum Overall	Average Overall	Average Leader A	Average Leader B
Paper	1	16	5.1 (215)*	5.2 (128)	4.9 (87)
Computer	1	18	5.0 (363)	5.5 (190)	4.5 (173)

* Number of timed interviews stated in brackets.

Table 3

Times for Data Processing 500 Questionnaires

Activity	Average time
Data checking	4 hours 8 minutes
First data entry	3 hours 40 minutes
Second data entry	3 hours 40 minutes
Verification	2 hours 23 minutes
Editing	2 hours 33 minutes
Total time	18 hours 24 minutes

4.2 Errors

The errors made were divided into two periods, to assess the effect of familiarity over time. Excluding the two team leaders, the remaining 22 fieldworkers made 1,704 errors on 1,427 questionnaires in the first period, and 1,049 errors on 1,158 questionnaires in the second period. Thus the average error rate per questionnaire in the first two weeks was 1.19 and in the third and fourth weeks was 0.90. In addition, over the entire period 37 questionnaires (1.2% of all interviews) had to be sent back to the field to be redone, as the errors found were not reconcilable in the office. These questionnaires had between 1 and 6 errors to be corrected, with a total of 61 errors. The highest number of errors were made on question 5 (17 errors) and question 6b (15 errors). Error rates per question are shown in Table 4. Fourteen out of the 22 fieldworkers redid at least one questionnaire. One fieldworker was required to redo 8 questionnaires.

Table 4

Type of Errors Made by 22 Fieldworkers Using Paper Questionnaires (for question specification see Figure 1)

	Period 1 (first fortnight)	Period 2 (second fortnight)
Identification	163	48
Question 1	6	1
Question 2	8	2
Question 3	125	92
Question 4a	201	138
Question 4b	151	93
Question 5	105	61
Question 6a	94	57
Question 6b	65	41
Question 7	14	0
Question 8	109	63
Question 9	51	10
Question 10	178	134
Question 11	13	1
Question 12	108	71
Question 13	53	3
Question 14	204	149
Question 15	19	76
Fieldworker code	37	9
Total errors	**1,704**	**1,049**
Total questionnaires	**1,427**	**1,158**

Errors were detected either manually by final checking by one of the investigators (Forster) or through the range and consistency checks performed in FoxPro on the entered data. Field team leader A made 8 errors on 144 questionnaires (0.06 errors per questionnaire), and team leader B made 18 errors on 90 questionnaires (0.20 errors per questionnaire). Most of these errors occurred in question 10 (4 errors) and question 15 (5 errors). The only errors found from the computer data were errors of respondent identification. There were 7 of these, 2 by leader A and 5 by leader B, giving errors of 0.01 and 0.03 per questionnaire respectively. Such errors could have been circumvented by pre-loading the Psion with a call list of respondents to interview.

4.3 Cost

The differential costs of a survey of this size using Psion-based and paper-based methods are given in Table 5. The Psion prices quoted are the recommended retail prices. Intense competition between retailers means that purchase prices could be up to 20% lower than those quoted here. Prices of hardware products are also decreasing. Current prices indicate that the one off cost of a Psion-based system can be recouped after 12-15 similar paper-based surveys of approximately 7,000 respondents.

Table 5

A Comparative Study of Computer-based and Paper-based Survey Methods (UK £ Sterling)

	Equipment required	Cost
Computed-based survey	20 Psion Series 3	2,539.00
	20 1 MB storage devices	2,039.00
	1 serial communications link	59.45
	80 rechargeable batteries	146.20
	1 battery recharger	15.95
	Total cost	**4,799.59**
Paper-based survey	14 reams of paper for 7,000 interviews	42.00
	Duplicating costs for 7,000 questionnaires (double-sided)	70.00
	20 pens, erasers and correcting fluid	27.40
	20 clipboards	100.00
	2 data entry clerks (two weeks)	70.00
	2 supervisors* (one month plus overtime)	85.00
	Total cost	**394.40**

* Necessary for the manual checking of forms as they come in from the field each day.

5. DISCUSSION

The lowest error rates using a traditional paper questionnaire by senior field workers with five years of data collection experience was on average 0.11 errors per questionnaire with 17 fields. This was reduced to negligible levels using the questionnaire software developed for a Psion Series 3 hand-held computer. This technique eliminated most of the errors made by fieldworkers in the routing of the questionnaire (Table 4) by using pre-defined skip modules, thus reducing the error rate by at least 90%. With the additional implementation of a call list in the software, the rate of respondent identification errors would be even lower.

The field supervisors were keen to use the computer, mastering the unfamiliar QWERTY keyboard, and learnt operating procedures quickly enough to take to the field without supervision after two days. Although no formal investigations were under taken to gauge and quantify interviewees' reactions to the Psion there were surprisingly few comments about the computer and no interview refusals.

Data processing involved two data entry clerks using two IBM machines full time for 92 hours to complete the data entry process for the entire survey. A data manager was on hand to offer assistance where necessary and design

the data entry format. The setting of the present study was such that both data entry clerks were familiar with data entry procedures and the available hardware and software. In situations where this is not the case, closer supervision and involvement by a data manager would be necessary, thus incurring an additional cost. A Psion data collection system would require much less of a data manager's time to down-load each day's data, thereby reducing this component of staff costs. Never-the-less, the initial cost of the Psion Series 3 may be prohibitively expensive when compared to the costs of paper and duplication of questionnaires if it was not envisaged that they form part of future data collection activities.

QUESTOR (Ferry and Cantrelle 1988) offers a suitable software environment for computer-assisted interviewing. However, the hardware required is a portable PC, several times the costs of a hand-held Psion. Our experience demonstrates that it will be worth pursuing the development of an appropriate package using this compact PC compatible technology as a more practical alternative in the field, being easier to handle, more robust and with reduced power consumption.

There is a trade off between error rates, time and cost of a survey. The use of computer-assisted interviewing software can reduce both the error rates and the length of time for data preparation considerably. Such a collection system should reduce the unacceptable delays in first presentation of data experienced during surveys such as the World Fertility Survey (Table 1). The context of the present comparative study differs from many large scale demographic surveys where recruited fieldstaff are unfamiliar with questionnaire procedures. We feel that the results presented here therefore represent a minimum improvement that could be expected in data quality. The initial cost of setting up such a survey mechanism may be daunting, but will be proportionally less for repeated surveys, or in institutions conducting a variety of different surveys over time.

ACKNOWLEDGEMENTS

The authors wish to thank all the field staff involved during the survey especially the two supervisors, Rodgers Chisengwa and Lewis Mitsanze, the data entry clerks Robert Mutai and Monica Omondi. We also acknowledge the support of Dr. Ian Timaeus, who designed the adult mortality questionnaire and Dr. Chris Nevill, who conducted the re-enumeration. We also wish to thank Dr. Kevin Marsh for his support for the computer studies at Kilifi, the Wellcome Trust, UK for financial support; the donation of a Psion Series 3 by Psion PLC, UK; and the Director of KEMRI for permission to publish this work. Dr. Bob Snow is supported as part of The Wellcome Trust Senior Fellowship programme in Basic Biomedical Science.

REFERENCES

ANKER, M. (1991). Epidemiological and statistical methods for rapid health assessment. *World Health Statistics Quarterly*, 44, 94-97.

BENCH, J., CLARK, C., DUFOUR, J., and KAUSHAL, R. (1994). Computer-assisted interviewing for the labour force survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

DHS, KENYA (1989). Report on Kenyan Demographic and Health Survey. Institute for Resource Development/Macro Systems Inc., Columbia, Maryland, USA.

DENTENEER, D., BETHLEHEM, J.G., HUNDEPOOL, A.J., and KELLER, W.J. (1987). The BLAISE System for Computer-Assisted Processing, Automation in Survey Processing. Netherlands Bureau of Statistics, CBS Select No. 4, 67-76.

FERRY, B., and CANTRELLE, P. (1988). The use of microcomputers for collection of demographic data in the field. In *African Population Conference*, Dakar, Senegal. Liege, Belgium: International Union for the Scientific Study of Population (IUSSP), 15-30.

FORSTER, D., BEHRENS, R.H., CAMPBELL, H., and BYASS, P. (1991). Evaluation of a computerized field data collection system for health surveys. *Bulletin of the World Health Organisation*, 69, 107-111.

FORSTER, D., and SNOW, R.W. (1992). Using microcomputers for rapid data collection in developing countries. *Health Policy and Planning*, 7, 667-71.

LYBERG, L. (1985). Plans for computer-assisted data collection at Statistics Sweden. *Bulletin of the International Statistical Institute*, 45th session, Invited Papers, Volume LI, Book 3, Section 18.2.

NICHOLLS, W.L., and GROVES, R.M. (1986). The status of computer-assisted telephone interviewing: Part I - Introduction and impact on cost and timeliness of survey data. *Journal of Official Statistics*, 2, 93-115.

REITMAIER, P., DUPRET, A., and CUTTING, W.A.M. (1987). Better health data with a portable microcomputer at the periphery: An anthropometric survey in Cape Verde. *Bulletin of the World Health Organisation*, 65, 651-657.

SNOW, R.W., MUNG'ALA, V.O., FORSTER, D., and MARSH, K. (1994). The role of the district hospital in child survival on the Kenyan Coast. *African Journal of Health Sciences*, 1, 71-75.

TIMAEUS, I.M. (1991). Measurement of adult mortality in less developed countries: a comparative review. *Population Index*, 57, 552-568.

VLASSOFF, C., and TANNER, M. (1992). The relevance of rapid assessment to health research and interventions. *Health Policy and Planning*, 7, 1-9.

WORLD FERTILITY SURVEY (1986). Final report. International Statistical Institute, Netherlands.

Survey Methodology, December 1995
Vol. 21, No. 2, pp. 185–190
Statistics Canada

Statistical Process Control of Sampling Frames

A.W. SPISAK[1]

ABSTRACT

Statistical process control can be used as a quality tool to assure the accuracy of sampling frames that are constructed periodically. Sampling frame sizes are plotted in a control chart to detect special causes of variation. Procedures to identify the appropriate time series (ARIMA) model for serially correlated observations are described. Applications of time series analysis to the construction of control charts are discussed. Data from the United States Department of Labor's Unemployment Insurance Benefits Quality Control Program is used to illustrate the technique.

KEY WORDS: Autocorrelation; ARIMA models; Control charts; Quality assurance.

1. INTRODUCTION

The integrity of the sampling frame is of paramount importance in survey research. Frame imperfections include missing elements (incomplete frame), element clusters (more than one element in a single listing), blank or foreign elements, and duplicate listings. These imperfections can cause several difficulties by contributing to nonsampling error, reducing the number of sample cases from subclasses of the population, and requiring the use of complex weights to estimate population characteristics. Techniques to minimize frame problems or reduce their impact on the survey are discussed in detail in most textbooks on statistical surveys.

This article focuses on the statistical process control of sampling frames which are constructed periodically (daily, weekly, or monthly, for example) and which consist of elements that are generated by a continuous process. Because of the variation inherent to any dynamic process, the sizes of the sampling frames will vary. How do we know that the changes in the sizes of the sampling frames reflect the random variation of the process and not errors in the construction of the frames? Statistical process control allows survey managers to distinguish between the variation inherent in the process (common causes) and variation which signals a possible problem with frame construction (special causes).

2. PROCESS VARIATION AND STATISTICAL PROCESS CONTROL

Over the last several years managers in the manufacturing, service, and public sectors of the economy increasingly have adopted the quality philosophies developed by W. Edwards Deming, J.M. Juran, Philip B. Crosby, Kaoru Ishikawa, and others. Quality management comprises an array of tools and techniques, including the use of control charts to determine if a process is in statistical control. According to Deming (1982), statistical control is achieved by eliminating special causes of variation, leaving only the random variation of a stable process. The behavior of a process that is in statistical control is predictable.

The distinction between common and special causes of variation is a key principle of statistical process control. Deming (1982) credits Dr. Walter A. Shewhart, who developed many of the principles of statistical process control in the 1920s and 1930s, with originating the concept of special or assignable causes. Special causes are usually attributable to one part of the process, such as a worker, machine, or office. They will reoccur unless they are identified and eliminated. Special causes are signaled by data points that fall outside of the control limits, by consecutive points that fall above or below the process average, or by runs of increasing or decreasing points.

Common causes of variation are inherent to the process; they are present at all times and effect the entire process. Common causes are reduced or eliminated through management actions that change the process.

3. STATISTICAL PROCESS CONTROL APPLICATION TO THE CONSTRUCTION OF SAMPLING FRAMES FOR PERIODIC SURVEYS

3.1 United States Unemployment Insurance Benefits Quality Control

The use of statistical process control as a quality management tool for sampling frames is illustrated by an example from the United States Department of Labor's Unemployment Insurance Benefits Quality Control program. Since 1987, the 50 states, the District of Columbia, and

[1] A.W. Spisak, Mathematical Statistician, Unemployment Insurance Service, U.S. Department of Labor, Washington, DC 20210, U.S.A.

Puerto Rico have conducted the Benefits Quality Control program in cooperation with the United States Department of Labor. The goal of the program is to reduce the overpayment and underpayment of Unemployment Insurance benefits by identifying the causes of payment errors and initiating measures to improve the benefit payment process.

When an individual files a claim for Unemployment Insurance benefits, Unemployment Insurance staff determine whether the claimant has met all of the eligibility requirements – for example, the claimant earned sufficient wages in his or her previous employment to qualify for benefits; the claimant is involuntarily unemployed; and the claimant is able and available to work and is actively seeking employment. If all of the eligibility requirements are satisfied, the state Unemployment Insurance agency issues a benefits check for the week of unemployment claimed.

3.2 Benefits Quality Control Sampling Procedures and Sources of Error

Each state selects weekly random samples of Unemployment Insurance payments that are examined to determine if the correct amount was paid to the claimant. If the amount paid was incorrect, the investigator identifies the types and causes of the errors so that program managers can initiate corrective measures. The sampling frames are constructed each week from the universe of Unemployment Insurance payments that were issued between 12:00 am Sunday and 11:59 pm the following Saturday. A computer program edits the state's database to insure that only payments that meet the program's operational definition of the target population are included in the frame. For example, payments for some temporary or small Unemployment Insurance programs are excluded from the frame.

The volume of Unemployment Insurance checks issued each week (and therefore the size of the sampling frames) varies in response to the number of individuals who claim and receive benefits during that week. However, there are several sources of potential errors which can affect the integrity of the frame. Some of the most serious of these errors are:

- The payments made from some of the local Unemployment Insurance offices might not be picked up for inclusion in the state's central database, due to telecommunication or ADP problems.
- If the state builds a separate file for each day's transactions, the transactions for one or more days might be erroneously omitted from the final cumulative file.
- Incorrect coding of transactions could result in either foreign elements being included in the frame or the editing out of transactions that should be included.

4. DATA ANALYSIS AND MODEL DEVELOPMENT

Figure 1 is a time series plot of sampling frame sizes for a 52 week period. Each week's sampling frame consists of the previous week's Unemployment Insurance benefit recipients who continue to receive benefits, minus the previous week's Unemployment Insurance recipients who have returned to work, exhausted their benefits, or failed to file a claim, plus newly eligible claimants and eligible claimants who did not file a claim or were not compensated for a claim the previous week.

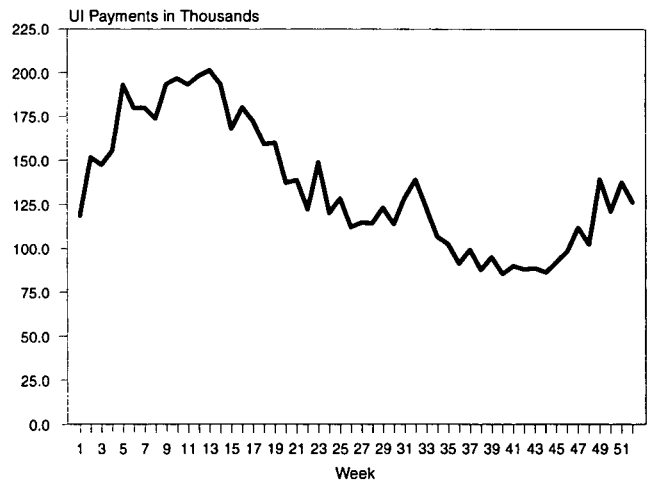


Figure 1. Number of UI payments per week.

Control charts for individual observations assume that the data are independent and identically distributed (i.i.d.). However, if the data are serially correlated, the estimates of the process variance (and therefore the control limits) could be seriously in error. So, before control charts for the Unemployment Insurance sampling frame data can be constructed, we have to determine if the observations are serially correlated.

The plot of the time series in Figure 1 provides *visual* evidence that the observations are not independent. The sampling frame data display distinct trends of increasing values during the first 13-week quarter, decreasing values over the next two quarters, and increasing values during the final 13-week quarter. The serial correlation suggested by the plot of the data in Figure 1 can be tested using methods developed to analyze time series. Although a detailed discussion of the analysis of time series data is beyond the scope of this article, the concepts of stationarity and autocorrelation will be examined, in order to explain the procedures used to identify the appropriate model. Readers who are unfamiliar with the basic principles of time series analysis should consult one of the many texts on the subject, in particular Box and Jenkins (1976).

4.1 Stationarity

We can think of the individual observations that constitute a time series as a collection of jointly distributed random variables - $p(z_1, \ldots, z_n)$ - where p is a probability density function and $z_1, \ldots, z_n$ are random variables. If the joint distribution of the random variables does not vary with respect to time, that is, $p(z_t, \ldots, z_{t+n}) = p(z_{t+m}, \ldots, z_{t+n+m})$, the process is said to be *strictly* stationary. In practice strict stationarity is difficult to establish. In this application, the time series is assumed to be *weakly* stationary. This is also referred to as second-order stationarity, because the first and second moments of the process are invariant with respect to time - $E(z_t) = E(z_{t+m})$, $\mathrm{VAR}(z_t) = \mathrm{VAR}(z_{t+m})$, and $\mathrm{COV}(z_t, z_{t+k}) = \mathrm{COV}(z_{t+m}, z_{t+k+m})$.

Throughout the rest of this article, the terms *stationary* or *stationarity* refer to a process that satisfies the conditions of weak stationarity.

4.2 Autocorrelation

In a stationary time series the covariance between any two observations depends only on the number of time periods (lags) that separate them - $\mathrm{COV}(z_t, z_{t+k}) = \mathrm{COV}(z_{t+m}, z_{t+k+m})$. The correlation of z_t and z_{t+k} equals $\mathrm{COV}(z_t, z_{t+k})/\mathrm{VAR}(z_t)$ and is denoted ρ_k, where k is the number of periods between observations. For example, ρ_1 is the correlation of observations in the time series separated by one period and equals $\mathrm{COV}(z_t, z_{t+1})/\mathrm{VAR}(z_t)$. A correlation for period k is referred to as an autocorrelation, because it is the correlation for observations which constitute a time series. The autocorrelations for the various lags can de displayed in a graph called a correlogram, which is useful in identifying the appropriate model for a time series.

4.3 Time Series Model Identification

Figure 2 is the correlogram for the 52 week time series of the number of Unemployment Insurance payments in the sample frames. The autocorrelations decrease or "die out" very slowly, which is characteristic of a nonstationary process. (Again, the reader is referred to Box (1976) and other texts on time series for a complete discussion of model identification.)

One method to transform a nonstationary series to a stationary series is *differencing*. The symbol B is the backshift operator, which when applied to z_t shifts the subscript back one period. Thus, the first difference of z_t is $(1 - B)z_t = z_t - z_{t-1}$.

Figure 3 is the time series of the differences $z_t - z_{t-1}$ of the Unemployment Insurance sampling frame data. This series appears stationary around a mean of zero. (The estimated sample mean of the differences is 150.8, with a standard error of 2064.0. The test statistic $t = (150.8 - 0)/2064$ equals .07, and the hypothesis that $\mu = 0$ cannot be rejected). First differences might not be sufficient to achieve stationarity for other time series, and transformations such as second differences - $(1 - B)^2 z_t = (z_t - z_{t-1}) - (z_{t-1} - z_{t-2})$, seasonal differences, or logarithmic or other variance stabilizing procedures may be required.

Lag	Auto-Corr.	Stand. Err.	-1 -.75 -.5 -.25 0 .25 .5 .75 1	Box-Ljung	Prob.
1	.915	.135		46.086	.000
2	.900	.133		91.592	.000
3	.830	.132		131.116	.000
4	.778	.131		166.570	.000
5	.699	.129		195.749	.000
6	.631	.128		220.036	.000
7	.557	.127		239.375	.000
8	.468	.125		253.366	.000
9	.381	.124		262.822	.000
10	.298	.122		268.774	.000
11	.216	.121		271.979	.000
12	.157	.119		273.714	.000
13	.084	.118		274.220	.000
14	.056	.116		274.452	.000
15	-.012	.115		274.464	.000
16	-.040	.113		274.586	.000
17	-.084	.112		275.156	.000
18	-.099	.110		275.967	.000
19	-.144	.108		277.737	.000
20	-.178	.107		280.516	.000
21	-.221	.105		284.959	.000
22	-.263	.103		291.419	.000
23	-.297	.102		299.936	.000
24	-.328	.100		310.739	.000
25	-.366	.098		324.658	.000
26	-.371	.096		339.530	.000

Plot Symbols: Autocorrelations * Two Standard Error Limits .

Total cases: 52 Computable first lags: 51

Figure 2. Autocorrelations for UI weeks paid time series.

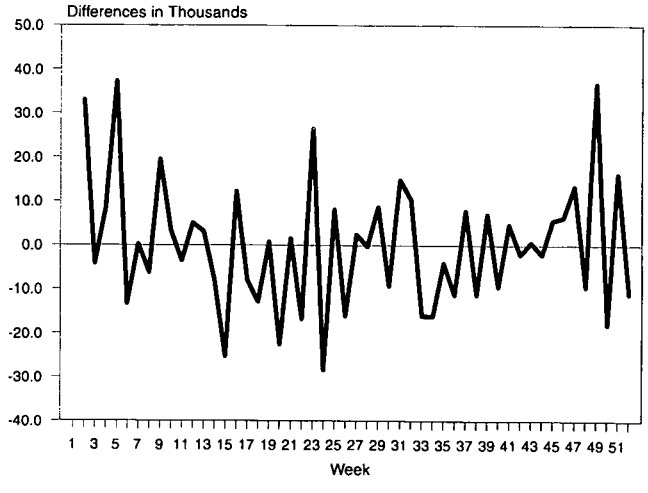


Figure 3. First differences of UI payments.

The autocorrelations of the first differences of the time series, which are displayed in Figure 4, are consistent with a stationary process. The autocorrelations decrease rapidly, while the partial autocorrelations (not displayed) die off after lag 1. This suggests that the data can be modelled with a first-order integrated autoregressive process, ARI (1,1). The AR term indicates that a single autoregressive parameter will be estimated, and the integration term (I) shows that the original time series has been transformed using first differences.

Lag	Auto-Corr.	Stand. Err.	-1 -.75 -.5 -.25 0 .25 .5 .75 1	Box-Ljung	Prob.
1	-.367	.136	**.|****.	7.293	.007
2	.314	.135	.|****.*	12.731	.002
3	-.080	.133	. **|.	13.089	.004
4	.196	.132	.|****.	15.294	.004
5	-.082	.130	. **|.	15.685	.008
6	.027	.129	.|*.	15.728	.015
7	.074	.128	.|*.	16.065	.025
8	.022	.126	.|*.	16.095	.041
9	-.025	.125	. *|.	16.135	.064
10	-.005	.123	.|*.	16.137	.096
11	-.129	.122	. ***|.	17.254	.101
12	.089	.120	.|**.	17.807	.122
13	-.280	.119	**|****.	23.401	.037
14	.204	.117	.|****.	26.443	.023
15	-.222	.115	**|****.	30.133	.011
16	.089	.114	.|**.	30.743	.015
17	-.191	.112	**|****.	33.658	.009
18	.189	.110	.|****.	36.597	.006
19	-.090	.109	. **|.	37.287	.007
20	.075	.107	.|*.	37.776	.009
21	-.048	.105	. *|.	37.983	.013
22	-.016	.104	.|*.	38.007	.018
23	-.049	.102	. *|.	38.241	.024
24	.042	.100	.|*.	38.415	.031
25	-.221	.098	**|****.	43.481	.012
26	.184	.096	.|****.	47.147	.007

Plot Symbols: Autocorrelations * Two Standard Error Limits .

Total cases: 52 Computable first lags after differencing: 50

Figure 4. Autocorrelations for first differences of UI weeks paid.

Lag	Auto-Corr.	Stand. Err.	-1 -.75 -.5 -.25 0 .25 .5 .75 1	Box-Ljung	Prob.
1	.118	.136	.|**.	.757	.384
2	.218	.135	.|****.	3.382	.184
3	.130	.133	.|***.	4.333	.228
4	.205	.132	.|****.	6.753	.150
5	-.007	.130	.|.	6.755	.239
6	.041	.129	.|*.	6.858	.334
7	.128	.128	.|***.	7.858	.345
8	.056	.126	.|*.	8.054	.428
9	-.025	.125	. *|.	8.095	.525
10	-.077	.123	. **|.	8.491	.581
11	-.132	.122	. ***|.	9.676	.560
12	-.069	.120	. *|.	10.009	.615
13	-.242	.119	*****|.	14.175	.362
14	.042	.117	.|*.	14.303	.427
15	-.164	.115	. ***|.	16.331	.360
16	-.071	.114	. *|.	16.721	.404
17	-.133	.112	. ***|.	18.129	.381
18	.119	.110	.|**.	19.283	.375
19	-.009	.109	.|.	19.290	.438
20	.030	.107	.|*.	19.370	.498
21	-.029	.105	. *|.	19.447	.556
22	-.061	.104	. *|.	19.799	.596
23	-.057	.102	. *|.	20.113	.635
24	-.066	.100	. *|.	20.549	.665
25	-.191	.098	****|.	24.352	.499
26	.096	.096	.|**.	25.356	.499

Plot Symbols: Autocorrelations * Two Standard Error Limits .

Total cases: 52 Computable first lags: 50

Figure 5. Autocorrelations for time series model residuals.

4.4 Model Estimation

The model was estimated using the ARIMA procedure of the SPSS Trends software (release 4.0), which is based on the work of Box and Jenkins.

The tentative model is:

$$z_t = (1 + \phi_1)z_{t-1} - \phi_1 z_{t-2} + e_t, \quad \text{or}$$

$$z_t - z_{t-1} = \phi_1(z_{t-1} - z_{t-2}) + e_t,$$

where ϕ_1 is the first-order autoregressive parameter, and e is the error term, which is assumed to be normally distributed with a mean of 0 and variance σ_e^2. The estimated autoregressive parameter, ϕ_1' is $-.4045$, and the estimated residual variance, $\sigma_e^{2\prime}$, is 184,275,853 (with 50 degrees of freedom). The negative sign on the AR parameter is consistent with the alternating signs of the autocorrelations in Figure 4. The model does not include a constant term, because the estimated process mean was not significantly different than zero.

4.5 Model Diagnostics

The adequacy of the estimated model for the observed data can be assessed by examining the model residuals. If the model adequately fits the data, the residuals (e_t) should be "white noise", that is, uncorrelated. Figure 5 displays the autocorrelations of the model residuals. Although the autocorrelation at lag 13 in Figure 5 is significant, the Box-Ljung Q statistic through lag 13 is not significant. (The Q statistic tests the significance of autocorrelations for lags 1 through k. For a detailed discussion, see Box and Pierce (1970)). In addition, none of the partial autocorrelations (not displayed) are significant. These results indicate that the residuals are not serially correlated.

To test the assumption that the model residuals are normally distributed, $N(0,\sigma_e^2)$, a Kolmogorov-Smirnov (K-S) goodness of fit test was conducted. For the estimated variance of 184,275,853, the K-S test statistic equals .591 ($p = .876$), and the hypothesis that the differences are normally distributed cannot be rejected.

For a stationary AR (1) process, the absolute value of the autoregressive parameter must be less than one. To test the hypothesis that $|\phi_1| \geq 1$ for the model, we compute: $t = (|\phi_1'| - 1)/SE(\phi_1')$, where $|\phi_1'|$ is the absolute value of the estimated autoregressive parameter, and $SE(\phi_1')$ is the standard error of ϕ_1'. The model statistics result in $t = (.4045 - 1)/.1295$ or $t = -4.6$. The chance of observing an absolute value of ϕ_1' as small as .4045 if the true absolute value of $\phi_1 \geq 1$ is very small ($< .00001$). The hypothesis that $|\phi_1| \geq 1$ is rejected, and we can conclude that the series of first differences is stationary.

5. USE OF THE ARIMA MODEL IN A CONTROL CHART

5.1 Control Charts for Individual Observations

The control limits for a chart of individual observations are set at $\bar{x} \pm 3\sigma'$, where $\bar{x}$ is the average of observation values and σ' is the estimated standard deviation of the process. Ryan (1989) discusses alternative procedures to estimate the process standard deviation either by computing the average of the moving ranges (the mean of the absolute differences of successive observations) or using the standard deviation (s) of the sample observations, $\sigma' = s/c$, where c is an adjustment constant which depends on the sample size.

When data are serially correlated, the use of either the sample standard deviation or the average moving range can result in poor estimates of σ. The control limits constructed from these estimates can produce seriously

misleading results by either generating false signals that the process is out of control or failing to detect special causes of process variation. The moving range can underestimate σ, because the differences of successive values will tend to be small if the successive observations are highly correlated. The underestimation of σ will result in control limits that are too narrow and an increase in the number of signals of special causes. Ryan notes that using the sample standard deviation to estimate the process standard deviation will result in a better estimate of σ than the average moving range when the data are correlated, provided the sample consists of at least 50 observations. However, the sample standard deviation is an unbiased estimator of σ only when the observations are independent.

Vasilopoulos and Stamboulis (1978) analyzed the effect of serially correlated data on the control limits of $\bar{x}$ and s (standard deviation) charts and developed equations for factors that can be used to adjust the control limits for data generated by an autoregressive process. Alternatively, a time series model can be identified for the correlated data, and a control chart can be constructed using the model residuals to monitor the process. This approach is described by Berthouex, Hunter, and Pallesen (1978) for subgroups of measurements of environmental data collected at water treatment plants. Alwan and Roberts (1988) use the residuals of exponentially weighted moving average (EWMA) models for both stationary and nonstationary time series. Montgomery and Mastrangelo (1991) use the residuals of an autoregressive model in an EWMA chart and contend that EWMA charts can be used to approximate many autocorrelated models, particularly if the observations are positively correlated and the mean does not drift too quickly. The reader is also referred to Maragah and Woodall (1992) and Woodall and Faltin (1993) for additional discussion of the effects of autocorrelation on statistical process control procedures.

5.2 Control Charts for the Unemployment Insurance Data

Figure 6 is a control chart of the residuals ($e_t = z_t - z_t'$) of the ARI (1,1) model identified for the Unemployment Insurance sampling frame data. Since the model diagnostics support the conclusion that the residuals are independent and identically distributed (i.i.d.) $N(0,\sigma_e^2)$, the residuals are standardized, so that the chart's center line is 0 and the control limits are set at ± 3. The chart includes model residuals for the sampling frame sizes in the 52 week baseline period and subsequent calendar quarter. The difference between the size of the sampling frame for week 56 and the value predicted by the model falls outside the upper control limit, signaling a special cause.

As an alternative to charting the model residuals, control charts for the Unemployment Insurance sampling frame

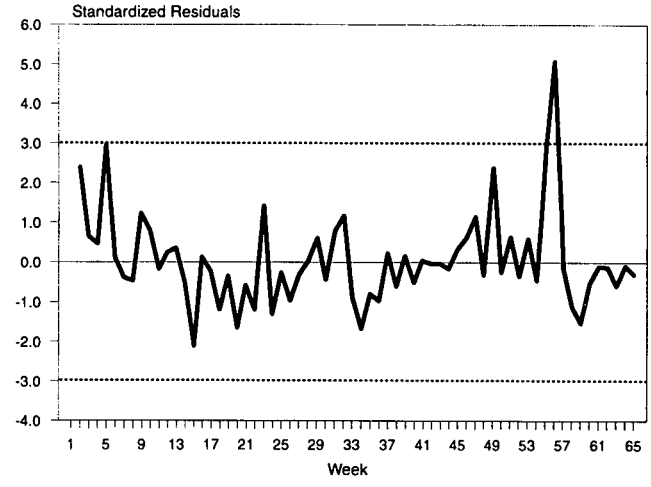


Figure 6. Control chart for model residuals (baseline data + next quarter).

sizes can be constructed. The original observations must be transformed to achieve stationarity, if necessary. The estimated parameters of the time series model are used to construct the mean and control limits of the chart. The variance of an AR(1) process is $\sigma^2 = \sigma_e^2/(1 - \phi_1^2)$. For the time series model of first differences, ϕ_1' is $-.4045$, and the estimated residual variance, $\sigma_e'^2$, is 184,275,853. The estimated process variance is 184,275,853/(1 − .1636) or 220,325,579.4, and the process standard deviation is 14,843.4. The upper and lower control limits are set at $\pm 3\sigma'$ from the estimated mean difference of zero: $\pm$44,530.2. The control chart is shown in Figure 7 and signals a special cause for observation 56, like the control chart for the residuals in Figure 6.

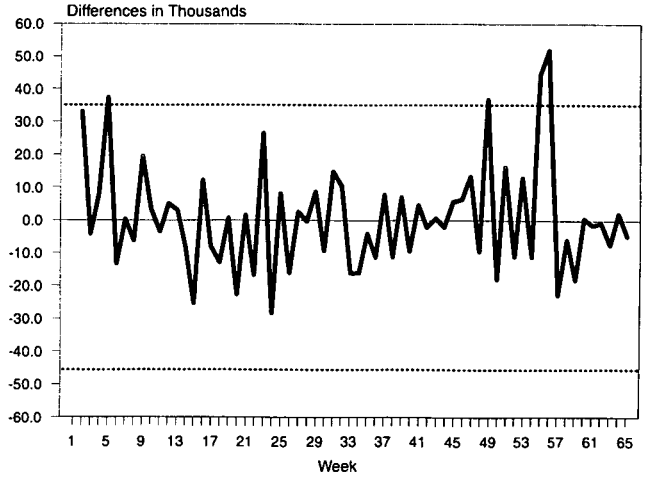


Figure 7. Control chart for UI payments (first differences − baseline + next quarter).

6. CONCLUSIONS

Statistical process control is a useful quality assurance tool for surveys in which samples are selected from frames that are constructed for specified periods from a continuous process. Because the frame sizes constitute a time series, the data may be serially correlated and may have to be transformed in order to achieve stationarity. If the observations are correlated, the appropriate time series (ARIMA) model must be identified in order to estimate the process variance used in setting the control limits. The time series in the preceding example was fitted by a first-order autoregressive integrated (differenced) model - ARI (1,1). More generally, time series may be described by other ARIMA (p,d,q) models, where p is the number of autoregressive terms in the model, d is the degree of differencing to achieve stationarity, and q is the number of moving average terms in the model. Seasonal time series models include additional AR, MA, and differencing parameters for the appropriate lag(s).

Once the model has been identified from baseline data, observations from subsequent periods can be plotted in the control chart. In the control charts in Figures 6 and 7, one calendar quarter (13 weeks) of observations are plotted following the observations from the 52 week baseline. The time series model should be checked periodically, depending on the data collection interval, to determine if the model parameters have changed.

If the statistical process control procedures signal a special cause of variation, survey managers must use other quality management tools to determine the root causes of the frame problems and then implement corrective actions to improve survey procedures. Survey managers can move from troubleshooting and error correction to continuous improvement of the survey process by systematically removing the assignable causes of variation identified through statistical process control.

In the case of the Unemployment Insurance sampling frame data, the special cause was not preventable: the volume of Unemployment Insurance payments spiked during a week which followed a short work week due to a holiday and which coincided with a layoff at a large establishment. The large sampling frame was not the result of a technical problem with the construction of the frame. In other states, at different time periods, statistical process control has detected errors as diverse as data entry mistakes (a frame of 558,432 reported instead of 5,558,432), omission of the Unemployment Insurance transactions for one of five work days, resulting in an approximate 20 percent decrease in the frame size, and the failure to update edits in the sample selection software, which caused foreign elements to enter the frame.

The procedure described in this article is applicable to other areas of survey and information management in addition to the integrity of sampling frames. The procedure can be used to reduce nonsampling error attributable to data recording or data entry for surveys conducted daily, monthly, *etc.* More generally, statistical process control can be used to assure the integrity of databases or management information systems whenever information is collected or reported in subgroups, such as data collected at multiple sites or by several researchers or auditors.

ACKNOWLEDGEMENT

The author wishes to thank the reviewers for their helpful comments and suggestions.

REFERENCES

ALWAN, L.C., and ROBERTS, H.V. (1988). Time series modeling for statistical process control. *Journal of Business and Economic Statistics*, 6, 87-95.

BERTHOUEX, P.M., HUNTER, W.G., and PALLESEN, L. (1978). Monitoring sewage treatment plants: some quality control aspects. *Journal of Quality Technology*, 10, 139-149.

BOX, G.E.P., and PIERCE, D.A. (1970). Distribution of residual autocorrelations in autoregressive moving average time series models. *Journal of the American Statistical Association*, 65, 1509-1526.

BOX, G.E.P., and JENKINS, G.M. (1976). *Time Series Analysis: Forecasting and Control.* San Francisco: Holden-Day.

DEMING, W.E. (1982). *Quality, Productivity, and Competitive Position.* Cambridge: Massachusetts Institute of Technology Center for Advanced Engineering Study.

MARAGAH, H.D., and WOODALL, W.H. (1992). The effect of autocorrelation on the retrospective X-chart. *Journal of Statistical Computation and Simulation*, 40, 29-42.

MONTGOMERY, D.C., and MASTRANGELO C.M. (1991). Some statistical process control methods for autocorrelated data. *Journal of Quality Technology*, 23, 179-204.

RYAN, T.P. (1989). *Statistical Methods for Quality Improvement*, New York: John Wiley and Sons.

VASILOPOULOS, A.V., and STAMBOULIS, A.P. (1978). Modification of control chart limits in the presence of data correlation. *Journal of Quality Technology*, 10, 20-30.

WOODALL, W.H., and FALTIN, F.W. (1993). Autocorrelated data and SPC. *American Society for Quality Control Statistics Division Newsletter*, 13, 18-21.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following persons who have served as referees during 1995. An asterisk indicates that the person served more than once.

C. Alexander, *U.S. Bureau of the Census*
R. Bell, *The Rand Corporation*
*D.R. Bellhouse, *University of Western Ontario*
N. Bennett, *Yale University*
*D.A. Binder, *Statistics Canada*
J.-R. Boudreau, *Statistics Canada*
J. Brehm, *Duke University*
F.J. Breidt, *Iowa State University*
S.J. Butani, *U.S. Bureau of Labor Statistics*
B.D. Causey, *U.S. Bureau of the Census*
R.L. Chambers, *Australian National University*
G. Chen, *University of Regina*
J. Chen, *University of Waterloo*
G.H. Choudhry, *Statistics Canada*
M.L. Cohen, *National Academy of Sciences*
M.J. Colledge, *Australian Bureau of Statistics*
J.L. Czajka, *Mathematica Policy Research*
T. DeMaio, *U.S. Bureau of the Census*
J. Denis, *Statistics Canada*
J.-C. Deville, *INSEE*
J.D. Drew, *Statistics Canada*
J.-J. Droesbeke, *Université Libre de Bruxelles*
D. Findley, *U.S. Bureau of the Census*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
M. Gonzalez, *U.S. Office of Management and Budget*
R.M. Groves, *University of Maryland*
K.P. Hapuarachchi, *Statistics Canada*
M.A. Hidiroglou, *Statistics Canada*
D. Hill, *University of Michigan*
*D. Holt, *Central Statistical Office, U.K.*
C.T. Ireland, *U.S. National Security Agency*
S. Itzhaki, *Hebrew University*
G. Kalton, *Westat, Inc.*
P.N. Kokic, *Australian Bureau of Agricultural and Resource Economics*
P.S. Kott, *National Agricultural Statistical Service*
M. Kovacevic, *Statistics Canada*
R.A. Kulka, *Research Triangle Institute*
S. Kumar, *Statistics Canada*
*N. Laniel, *Statistics Canada*
M. Latouche, *Statistics Canada*
*P. Lavallée, *Statistics Canada*
L. Lazzeroni, *Stanford University*
*H. Lee, *Statistics Canada*
N. Luther, *East-West Center*
*L.Mach, *Statistics Canada*
T.K. Mak, *Concordia University*
*H. Mantel, *Statistics Canada*
A. Mason, *East-West Center*
P. Merkouris, *Statistics Canada*
*D. Pfeffermann, *Hebrew University*
H. Pold, *Statistics Canada*
R.F. Potthoff, *Duke University*
B. Quenneville, *Statistics Canada*
T.E. Raghunathan, *University of Michigan*
É. Rancourt, *Statistics Canada*
*J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
K. Rust, *Westat, Inc.*
I. Sande, *Bell Communications Research, U.S.A.*
C.-E. Särndal, *Université de Montréal*
J. Schafer, *Pennsylvania State University*
*W.L. Schaible, *U.S. Bureau of Labor Statistics*
*F.J. Scheuren, *George Washington University*
*J. Sedransk, *State University of New York - Albany*
R. Sigman, *U.S. Bureau of the Census*
M. Simard, *Statistics Canada*
*A. Singh, *Statistics Canada*
*M.P. Singh, *Statistics Canada*
R.P. Singh, *U.S. Bureau of the Census*
*C.J. Skinner, *University of Southampton*
*D. Stukel, *Statistics Canada*
*A. Théberge, *Statistics Canada*
Y. Tillé, *Université Libre de Bruxelles*
M. Traugott, *University of Michigan*
J. Trépanier, *Statistics Canada*
R. Valliant, *U.S. Bureau of Labor Statistics*
J. Waite, *U.S. Bureau of the Census*
*J. Waksberg, *Westat, Inc.*
S. Wing, *Synetics*
W.E. Winkler, *U.S. Bureau of the Census*
*K.M. Wolter, *National Opinion Research Center*
*A. Zaslavsky, *Harvard University*

Acknowledgements are also due to those who assisted during the production of the 1995 issues: S. Beauchamp (Photocomposition), L. Rousseau and S. Cadieux (Official Languages and Translation). Finally we wish to acknowledge S. DiLoreto, S.F. Bertrand, C. Larabie and D. Lemire of Household Survey Methods Division, for their support with coordination, typing and copy editing.

The Canadian Journal of Statistics **La Revue Canadienne de Statistique**

CONTENTS TABLE DES MATIÈRES

Volume 23, No. 3, September/Septembre 1995

Research papers/Articles

Case study in data analysis
Étude de cas en analyse des données

CONTENTS TABLE DES MATIÈRES

Volume 23, No. 4, December/Décembre 1995

JOURNAL OF OFFICIAL STATISTICS

An International Quarterly Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey Methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 11, Number 1, 1995

Contents Volume 11, Number 2, 1995

Contents Volume 11, Number 3, 1995

Lars Lyberg, U/LEDN, Statistics Sweden, S-115 81 Stockholm, Sweden

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 19, No. 1 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.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, n° 1) et de noter les points suivants:

1. Présentation

1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
1.4 Les remerciements doivent paraître à la fin du texte.
1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) etc.
3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O, 0; 1, l).
3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux

4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois.)

5. Bibliographie

5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
Exemple: Cochran (1977, p. 164).
5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

où T_X représente la population totale de X, et où $\hat{T}_X$ et $\hat{T}_{X^2}$ correspondent à l'estimation HT du total des variables X et X^2, respectivement.

Pareillement, l'équation qui suit permet d'estimer la fonction de distribution

$$N\hat{F}(y) = \sum_{i\in s} w_i(s) I\{y_i \leq y\},$$

où

$$I\{y_i \leq y\} = \begin{cases} 1 & \text{si } y_i \leq y, \\ 0 & \text{si } y_i > y. \end{cases}$$

Soulignons que $\hat{F}(y)$ converge de façon uniforme et asymptotique à $F(y)$, mais n'est pas nécessairement une fonction de distribution au véritable sens du terme, à moins que

$$\sum_{i\in s} w_i(s) = N.$$

En règle générale et sous réserve de certaines conditions de régularité pour les plans d'échantillonnage complexes (Francisco et Fuller 1991),

$$\hat{F}(y) - F(y) \to_p 0 \text{ pour toutes les valeurs de } y.$$

En d'autres termes, la fonction de distribution de la population finie $F(y)$ accepte un estimateur convergent $\hat{F}(y)$. Nous nous servirons plus tard de cette propriété de $\hat{F}(y)$ pour prouver la convergence des estimateurs de variance linéaires à l'égard de diverses statistiques sur le revenu.

Examinons maintenant comment la théorie des équations d'estimation s'applique à l'estimation du paramètre Θ_o d'une population finie, qui est la solution de

$$\int u(y,\Theta_o) dF(y) = 0.$$

La valeur de l'équation d'estimation pour Θ_o correspond à la valeur de $\hat{\Theta}$, pour laquelle

$$\int \hat{u}(y,\hat{\Theta}) d\hat{F}(y) = 0, \tag{2.2}$$

où $\hat{u}(y,\Theta)$ est une estimation de $u(y,\Theta)$.

On peut écrire (2.2) de la façon suivante:

$$\begin{aligned} 0 &= \int \hat{u}(y,\hat{\Theta}) d\hat{F}(y) \\ &= \int [\hat{u}(y,\hat{\Theta}) - u(y,\Theta_o)] dF(y) + \int u(y,\Theta_o) d\hat{F}(y) + R, \end{aligned} \tag{2.3}$$

où

$$R = \int [\hat{u}(y,\hat{\Theta}) - u(y,\Theta_o)][d\hat{F}(y) - dF(y)].$$

La décomposition de (2.3) sert de point de départ à tous les calculs de la variance dans le présent article. Nous prouverons que le reste R est asymptotiquement négligeable pour tous les paramètres examinés.

Binder (1983) a étudié le cas où $\hat{u}(y,\Theta) = u(y,\Theta)$ et où, pour les échantillons importants,

$$\begin{aligned} &\int [u(y,\hat{\Theta}) - u(y,\Theta_o)] dF(y) \\ &= (\hat{\Theta} - \Theta_o) \left. \frac{\partial \mathrm{E}\{u(y,\Theta)\}}{\partial \Theta} \right|_{\Theta=\Theta_o} + o_p(|\hat{\Theta} - \Theta_o|). \end{aligned}$$

Notons que le reste R issu de la décomposition (2.3) devrait avoir pour ordre de grandeur $o_p(|\hat{\Theta} - \Theta_o|)$ afin d'être asymptotiquement négligeable.

La plupart des applications n'exigent pas l'estimation de $u(y,\Theta)$ au moyen de $\hat{u}(y,\Theta)$. Toutefois, dans certains cas (le coefficient de Gini par exemple), on estime la fonction $u(y,\Theta)$ pour que la formule (2.2) accepte les cas en question, en général.

Grâce aux approximations qui précèdent, on obtient

$$\begin{aligned} \hat{\Theta} - \Theta_o \approx &- \left[\left. \frac{\partial \mathrm{E}\{u(y,\Theta)\}}{\partial \Theta} \right|_{\Theta=\Theta_o} \right]^{-1} \\ &\times \int u(y,\Theta_o) d\hat{F}(y) = \int u^*(y) d\hat{F}(y), \end{aligned} \tag{2.4}$$

où

$$u^*(y) = - \left[\left. \frac{\partial \mathrm{E}\{u(y,\Theta)\}}{\partial \Theta} \right|_{\Theta=\Theta_o} \right]^{-1} u(y,\Theta_o).$$

Dès qu'on connaît $u^*(y)$, il est assez facile de trouver la variance de $\hat{\Theta}$. Puisque $\hat{\Theta} - \Theta_o$ donne une estimation de la population totale de $u^*(y_i)$, on peut en appliquer l'erreur quadratique moyenne à l'estimation pour avoir une idée de la variance de $\hat{\Theta}$.

Par exemple, quand Θ_o est égal au rapport T_Y/T_X, on obtient:

$$u = y - \Theta_0 x,$$

$$u^* = \frac{1}{\mu_X} (y - \Theta_0 x).$$

Dans ce cas, le reste correspond à

$$R = \int [y - \hat{\Theta}x - (y - \Theta_0 x)]\,[d\hat{F}(y) - dF(y)].$$

Par conséquent,

$$-\frac{R}{\hat{\Theta} - \Theta_o} = [\hat{F}(y) - F(y)]x \to_p 0,$$

pour toute valeur de y et valeur finie de x.

Parallèlement, pour les quantiles de population,

$$\begin{aligned} u &= I\{y \le \Theta_0\} - p, \\ u^* &= -\frac{1}{f(\Theta_0)}\,[I\{y \le \Theta_0\} - p], \end{aligned} \tag{2.5}$$

où $f(\Theta_0)$ correspond à la densité de probabilité au point Θ_0. Le deuxième élément de (2.5) est le prolongement de la représentation des quantiles de l'échantillon par Bahadur, tel qu'il est décrit par Francisco et Fuller (1991). On se servira des résultats de (2.5) comme ordonnées dans la courbe de Lorenz et comme mesure du faible revenu aux parties 4 et 5.

Après réduction, le reste R devient $R = \hat{F}(\hat{\Theta}) - \hat{F}(\Theta_o) - F(\hat{\Theta}) + F(\Theta_o)$. Avec un plan d'échantillonnage aléatoire simple, Randles (1982) a montré que $R = o_p(n^{-1/2})$. Lorsque le plan est plus complexe, Shao et Rao (1994) obtiennent un résultat asymptotique analogue, sous réserve de certaines contraintes de régularité: ils établissent d'abord que $\hat{\Theta} - \Theta_o = O_p(n^{-1/2})$, puis que $R = o_p(n^{-1/2})$ et enfin que $R = o_p(|\hat{\Theta} - \Theta_o|)$.

3. COEFFICIENT DE LA FAMILLE DE GINI

Pour le coefficient de la famille de Gini donné en (1.2), on peut se servir de

$$u(y,G_J) = J[F(y)]y - G_J y.$$

L'approche de Binder (1983) n'accepte pas l'estimation de la variance du coefficient de Gini. Au lieu de calculer la variance en scindant le problème en deux - un élément pour l'estimateur du ratio et l'autre pour la variance du numérateur - on peut résoudre celui-ci en une fois au moyen des équations d'estimation.

Laissant de côté le reste de (2.3), on obtient l'approximation suivante:

$$\begin{aligned} 0 &= \int \{J[\hat{F}(y)]y - \hat{G}_J y\}d\hat{F}(y) \\ &\approx \int \{J[\hat{F}(y)] - J[F(y)]\}ydF(y) \\ &\quad - (\hat{G}_J - G_J)\int ydF(y) + \int \{J[F(y)]y - G_J y\}d\hat{F}(y). \end{aligned}$$

Soient

$$\begin{aligned} &\int \{J[\hat{F}(y)] - J[F(y)]\}ydF(y) \\ &\qquad \approx \int [\hat{F}(y) - F(y)]J'[F(y)]ydF(y), \end{aligned}$$

et

$$\begin{aligned} &\int \hat{F}(y)J'[F(y)]ydF(y) \\ &\qquad = \int\int_0^y J'[F(y)]yd\hat{F}(x)dF(y) \\ &\qquad = \int \left[\int_y^\infty J'[F(x)]xdF(x)\right] d\hat{F}(y), \end{aligned}$$

il s'ensuit que

$$\hat{G}_J - G_J \approx \int u^*(y)d\hat{F}(y),$$

où

$$\begin{aligned} u^* = \frac{1}{\mu_Y}\Bigg[&\int_{F(y)}^1 J'(p)F^{-1}(p)dp \\ &+ J[F(y)]y - G_J y - E\{F(y)J'[F(y)]y\}\Bigg]. \end{aligned} \tag{3.1}$$

Lorsque les observations sont indépendantes et sont distribuées de façon identique, la variance correspond à celle décrite par Glasser (1962) et Sendler (1979). Pour estimer la variance, il est nécessaire d'estimer μ_Y, $F(y)$, et G_J dans l'expression u^*.

Examinons le comportement asymptotique du reste R pour le coefficient de Gini habituel G. Le reste est égal à

$$\begin{aligned} R = \int &\{2y[\hat{F}(y) - F(y)] - y(\hat{G} - G)\} \\ &\times [d\hat{F}(y) - dF(y)]. \end{aligned}$$

Si on représente la différence $\hat{F}(y) - F(y)$ par $\hat{D}(y)$, le reste peut être exprimé comme la somme de deux intégrales

$$R = \int 2y\hat{D}(y)d\hat{D}(y) - \int (\hat{G} - G)yd\hat{D}(y).$$

La première intégrale se ramène à zéro par une intégration par parties, de telle sorte qu'on peut calculer approximativement le reste par l'équation

$$R \approx - (\hat{G} - G)(\hat{\mu}_Y - \mu_Y)$$
$$= - (\hat{G} - G)o_p(n^{-1/2+\delta}), \quad 0 < \delta < 1/2.$$

Par conséquent, on peut dire que $R = o_p(|\hat{G} - G|)$.

4. ORDONNÉE DE LA COURBE DE LORENZ ET PART DU QUANTILE

L'ordonnée de la courbe de Lorenz a été définie en (1.1). L'estimation exige les deux équations suivantes:

$$u_1(y,L(p)) = I\{y \leq \xi_p\}y - L(p)y,$$

$$u_2(y) = I\{y \leq \xi_p\} - p.$$

La deuxième équation définit le $100p$-ième percentile de la distribution, tandis que la première exprime l'ordonnée de la courbe de Lorenz au point correspondant. Si le reste de (2.3) est négligeable, on obtient l'approximation suivante:

$$0 = \int [I\{y \leq \hat{\xi}_p\} - \hat{L}(p)]ydF(y)$$

$$\approx \int_{\xi_p}^{\hat{\xi}_p} ydF(y) - [\hat{L}(p) - L(p)] \int ydF(y)$$

$$+ \int [I\{y \leq \xi_p\} - L(p)]yd\hat{F}(y).$$

Le premier terme de l'expression peut faire l'objet d'une approximation plus poussée, comme suit:

$$\int_{\xi_p}^{\hat{\xi}_p} ydF(y) \approx (\hat{\xi}_p - \xi_p)\xi_p f(\xi_p).$$

Reprenant (2.5), on constate que

$$\hat{\xi}_p - \xi_p \approx - \int \frac{1}{f(\xi_p)} [I\{y \leq \xi_p\} - p]d\hat{F}(y), \quad (4.1)$$

de telle sorte que

$$(\hat{\xi}_p - \xi_p)\xi_p f(\xi_p) \approx - \int \xi_p [I\{y \leq \xi_p\} - p]d\hat{F}(y).$$

Par conséquent, la linéarisation nécessaire pour estimer la variance de l'ordonnée de la courbe de Lorenz s'exprime de la manière suivante:

$$u^*(y) = \frac{1}{\mu_Y} [(y - \xi_p)I\{y \leq \xi_p\} + p\xi_p - yL(p)].$$

Le résultat est identique à celui obtenu par Beach et Davidson (1983) pour la variance et la covariance des ordonnées de la courbe de Lorenz quand les variables aléatoires sont indépendantes mais distribuées de façon identique. Pour estimer la variance, on doit introduire $\hat{\xi}_p$ et $\hat{L}(p)$ dans $u^*(y)$.

Trois équations sont nécessaires pour estimer la part du quantile $Q(p_1, p_2)$:

$$u_1(y,Q(p_1,p_2)) = I\{\xi_{p_1} < y \leq \xi_{p_2}\}y - Q(p_1,p_2)y,$$

$$u_2(y) = I\{y \leq \xi_{p_1}\} - p_1,$$

$$u_3(y) = I\{y \leq \xi_{p_2}\} - p_2.$$

En reprenant les mêmes arguments qu'auparavant, on parvient à l'équation suivante:

$$u^*(y) = \frac{1}{\mu_Y} [(y - \xi_{p_2})I\{y \leq \xi_{p_2}\}$$
$$- (y - \xi_{p_1})I\{y \leq \xi_{p_1}\}$$
$$+ p_2\xi_{p_2} - p_1\xi_{p_1} - yQ(p_1,p_2)].$$

5. MESURE DU FAIBLE REVENU

Cette mesure a été définie en (1.3). Pour procéder à l'estimation, deux équations sont nécessaires:

$$u_1(y,\Theta) = I\left\{y \leq \frac{M}{2}\right\} - \Theta,$$

$$u_2(y) = I\{y \leq M\} - \frac{1}{2},$$

où M est la médiane de la distribution définie par la deuxième équation, alors que la première définit la mesure du faible revenu par rapport à la médiane. En laissant de côté le reste indiqué en (2.3), on obtient l'approximation suivante:

$$0 = \int \left(I\left\{ y \le \frac{\hat{M}}{2} \right\} - \hat{\Theta} \right) d\hat{F}(y)$$

$$\approx \frac{1}{2}(\hat{M} - M) f\left(\frac{M}{2}\right) - (\hat{\Theta} - \Theta)$$

$$+ \int \left(I\left\{ y \le \frac{M}{2} \right\} - \Theta \right) d\hat{F}(y).$$

Si on remplace $\hat{M} - M$ par le résultat obtenu en (4.1) et résoud $\hat{\Theta} - \Theta$, on obtient

$$\hat{\Theta} - \Theta \approx \int u^*(y) d\hat{F}(y),$$

où

$$u^* = -\frac{f\left(\frac{M}{2}\right)}{2f(M)} \left(I\{y \le M\} - \frac{1}{2} \right) + I\left\{ y \le \frac{M}{2} \right\} - \Theta. \quad (5.1)$$

Pour estimer la variance de la mesure estimative du faible revenu à partir de ce résultat, il faut estimer $f(M)$ et $f(M/2)$. À cette fin, on pourrait se servir de

$$\hat{f}(\xi) = \frac{\hat{F}\left(\xi + \frac{h}{2}\right) - \hat{F}\left(\xi - \frac{h}{2}\right)}{h},$$

pourvu que h soit assez faible. Parallèlement, on pourrait procéder au calcul que voici, comme le proposent Francisco et Fuller (1991) pour un problème analogue. Pour une valeur quelconque de ξ, on estime le percentile correspondant $100p$, puis on établit l'intervalle de Woodruff pour le même percentile en résolvant d'abord h_1 et h_2 dans

$$\inf_{h_1} \left[\frac{\int [I\{y \le \xi - h_1\} - p] d\hat{F}(y)}{\left[\text{mse} \left\{ \int [I\{y \le \xi\} - p] d\hat{F}(y) \right\} \right]^{1/2}} \le -z_{1-\alpha/2} \right],$$

$$\inf_{h_2} \left[\frac{\int [I\{y \le \xi + h_2\} - p] d\hat{F}(y)}{\left[\text{mse} \left\{ \int [I\{y \le \xi\} - p] d\hat{F}(y) \right\} \right]^{1/2}} \ge z_{1-\alpha/2} \right],$$

où $z_{1-\alpha/2}$ est le $100(1 - \alpha/2)$-ième percentile de la distribution normale type. Il suffit ensuite de calculer

$$\hat{f}(\xi) = \frac{2z_{1-\alpha/2} \left[\text{mse} \left\{ \int [I\{y \le \xi\} - p] d\hat{F}(y) \right\} \right]^{1/2}}{h_1 + h_2}. \quad (5.2)$$

Ce calcul utilise l'équivalent asymptotique de $\hat{\xi} - \xi$ et la somme estimée de $u^*(y)$ donnée par (2.5).

Comme on peut le constater, estimer la variance de la mesure du faible revenu s'avère relativement compliqué. La méthode des fonctions d'estimation nous procure toutefois les formules appropriées.

La discussion relative au reste R de la décomposition (2.3) pour la mesure du faible revenu est analogue à celle de l'estimation du quantile (2.5).

6. ESTIMATION DANS LE CADRE D'UNE ENQUÊTE COMPLEXE

Supposons un échantillonnage stratifié à plusieurs degrés comprenant un grand nombre de strates H et quelques unités d'échantillonnage primaires (grappes), $n_h (\ge 2)$, venant de chaque strate. Dans l'Enquête sur les finances des consommateurs (EFC) du Canada, qui utilise elle-même l'Enquête sur la population active (EPA), par exemple, on compte plusieurs centaines de strates et en moyenne moins de six grappes par strate. Supposons que w_{hci} est le poids normalisé de la i-ème unité ultime de la c-ième grappe de la h-ième strate. L'estimateur approprié du total et l'estimateur convergent de l'erreur quadratique moyenne seront

$$\hat{\mu} = \sum_s w_{hci} y_{hci}$$

$$\text{mse}(\hat{\mu}) = \sum_h \frac{n_h}{n_h - 1} \sum_c (u^*_{hc} - \bar{u}^*_h)^2 \quad (6.1)$$

où $u^*_{hc} = \sum_i w_{hci}(y_{hci} - \hat{\mu})$ et $\bar{u}^*_h = 1/n_h \sum_c u^*_{hc}$. On se sert de $\sum_s = \sum_h \sum_c \sum_i$ pour désigner la somme sur toutes les unités ultimes de l'échantillon intégrant tous les degrés d'échantillonnage. On suppose que les unités primaires d'échantillonnage sont sélectionnées avec une unité de remplacement.

Ce n'est pas l'efficacité des estimateurs qui nous intéresse mais les propriétés des estimateurs couramment utilisés. Une analyse des estimateurs plus complexes qu'on retrouve dans les ouvrages d'économétrie déborde du cadre de notre travail.

Voici un estimateur de la fonction de distribution de la population finie:

$$\hat{F}(y) = \sum_s w_{hci} I\{y_{hci} \le y\}.$$

Un estimateur convergent de l'erreur quadratique moyenne approximative pour la fonction de distribution estimée en y prendrait la forme (6.1), où $u^*_{hc} = \sum_i w_{hci}[I\{y_{hci} \leq y\} - \hat{F}(y)]$.

L'estimation habituelle du quantile de la population finie donne le quantile de l'échantillon

$$\hat{\xi}_p = \inf\{y_{hci} \in s : \hat{F}(y_{hci}) \geq p\},$$

c'est-à-dire la solution de l'équation d'estimation

$$\sum_s w_{hci}[I\{y_{hci} \leq \hat{\xi}_p\} - p] = 0.$$

Si on reprend les résultats de (2.5), l'estimateur de l'erreur quadratique moyenne du p-ième quantile correspond à l'équation (6.1) et

$$u^*_{hc} = \frac{1}{[\hat{f}(\hat{\xi}_p)]^2} \sum_i w_{hci}[I\{y_{hci} \leq \hat{\xi}_p\} - p].$$

Quand on utilise l'expression (5.2) pour estimer la fonction de densité $f(\xi)$, l'estimation de l'erreur quadratique moyenne du quantile $\hat{\xi}_p$ devient

$$\text{mse}_\alpha(\hat{\xi}_p) = \left(\frac{D_\alpha(\hat{\xi}_p)}{z_{1-\alpha/2}}\right)^2 \qquad (6.2)$$

où $D_\alpha(\hat{\xi}_p) = (h_1 + h_2)/2 = (\hat{\xi}_U - \hat{\xi}_L)/2$ représente la demi-longueur de l'intervalle de confiance à $100(1 - \alpha)\%$ pour $\hat{\xi}_p$. Dans un plan d'échantillonnage complexe, h_1 et h_2 correspondent respectivement à la solution de

$$\hat{\xi}_L = \hat{\xi}_p - h_1 =$$
$$\inf\left\{y_{hci} \in s : \hat{F}(y_{hci}) \geq p - z_{1-\alpha/2}\sqrt{\text{mse}[\hat{F}(\hat{\xi}_p)]}\right\}$$

$$\hat{\xi}_U = \hat{\xi}_p + h_2 =$$
$$\inf\left\{y_{hci} \in s : \hat{F}(y_{hci}) \geq p + z_{1-\alpha/2}\sqrt{\text{mse}[\hat{F}(\hat{\xi}_p)]}\right\}.$$

Francisco et Fuller (1991) ont utilisé eux aussi l'estimateur (6.2). En règle générale, l'intervalle de confiance de Woodruff (1952) pour les quantiles individuels explique pourquoi on recourt à (5.2), et par conséquent à (6.2). Francisco et Fuller (1986), puis Rao et Wu (1987) se sont servis de cet intervalle pour trouver des estimateurs de la variance. Bien que l'estimateur dépende du coefficient de confiance, ces auteurs ont montré qu'il est convergent par rapport à l'asymptote à un seuil de signification α. Rao et Wu (1987) ont examiné l'erreur-type des quantiles pour les échantillons en grappes estimés de cette manière. Les résultats qu'ils ont obtenus avec la méthode de Monte Carlo suggèrent qu'un intervalle de confiance de 95% donne de bons résultats quand vient le moment d'établir l'erreur-type. Binder et Patak (1994) ont obtenu un estimateur de la variance analogue par l'approche des équations d'estimation.

L'estimation du coefficient de Gini habituel correspond à la solution de l'équation suivante

$$\sum_s w_{hci}\{[2\hat{F}(y_{hci}) - 1]y_{hci} - \hat{G}y_{hci}\} = 0$$

qui se transforme pour donner

$$\hat{G} = \frac{2}{\hat{\mu}} \sum_s w_{hci}\hat{F}(y_{hci})y_{hci} - 1$$

où $\hat{\mu} = \sum_s w_{hci}\, y_{hci}$.

On peut estimer l'erreur quadratique moyenne pour le coefficient de Gini grâce à l'expression (6.1) en remplaçant u^*_{hc}, défini en (3.1) par son équivalent pour une enquête complexe. Après manipulation algébrique, on parvient à l'expression suivante:

$$u^*_{hc} = \frac{2}{\hat{\mu}} \sum_i w_{hci}\left[A(y_{hci})y_{hci} + B(y_{hci}) - \frac{\hat{\mu}}{2}(\hat{G} + 1)\right]$$

où

$$A(y) = \hat{F}(y) - \frac{\hat{G} + 1}{2}$$

et

$$B(y) = \sum_s w_{hci}\, y_{hci}\, I\{y_{hci} \geq y\}.$$

Pour obtenir les ordonnées de la courbe de Lorenz il suffit de résoudre le système d'équations d'estimation suivant:

$$\sum_s w_{hci}[I\{y_{hci} \leq \hat{\xi}_p\}y_{hci} - \hat{L}(p)y_{hci}] = 0$$

$$\sum_s w_{hci}[I\{y_{hci} \leq \hat{\xi}_p\} - p] = 0.$$

L'estimation qui en résulte est donnée par:

$$\hat{L}(p) = \frac{1}{\hat{\mu}} \sum_s w_{hci}\, y_{hci}\, I\{y_{hci} \leq \hat{\xi}_p\}.$$

Pour calculer l'estimateur de l'erreur quadratique moyenne des ordonnées de la courbe de Lorenz, il suffit d'utiliser les valeurs u^*_{hc} définies par (6.3) dans (6.1):

$$u^*_{hc} = \frac{1}{\hat{\mu}} \sum_i w_{hci}\big[(y_{hci} - \hat{\xi}_p)I\{y_{hci} \leq \hat{\xi}_p\}$$
$$+ p\hat{\xi}_p - y_{hci}\hat{L}(p)\big]. \qquad (6.3)$$

De même, l'erreur quadratique moyenne de la part du quantile correspond à

$$\hat{Q}(p_1,p_2) = \frac{1}{\hat{\mu}} \sum_s w_{hci}\, y_{hci}\, I\{\hat{\xi}_{p_1} < y_{hci} \leq \hat{\xi}_{p_2}\}$$

et on en obtient la valeur approximative avec (6.1) en appliquant

$$u^*_{hc} = \frac{1}{\hat{\mu}} \sum_i w_{hci}\big[(y_{hci} - \hat{\xi}_{p_2})I\{y_{hci} \leq \hat{\xi}_{p_2}\} - (y_{hci} - \hat{\xi}_{p_1})I\{y_{hci} \leq \hat{\xi}_{p_1}\} + p_2\hat{\xi}_{p_2} - p_1\hat{\xi}_{p_1} - y_{hci}\,\hat{Q}(p_1,p_2)\big].$$

La mesure du faible revenu définie par (1.3) est estimée grâce à

$$\hat{\Theta} = \hat{F}(\hat{M}/2) = \sum_s w_{hci}\, I\{y_{hci} \leq \hat{M}/2\}.$$

L'erreur quadratique moyenne de la mesure du faible revenu peut être estimée approximativement grâce à l'expression (6.1), où (en partant de l'équation (5.1)):

$$u^*_{hc} = -\frac{\hat{f}(\hat{M}/2)}{2\hat{f}(\hat{M})} \sum_i w_{hci}[I\{y_{hci} \leq \hat{M}\} - 1/2] + \sum_i w_{hci}[I\{y_{hci} \leq \hat{M}/2\} - \hat{\Theta}].$$

7. ILLUSTRATION

Nous illustrerons la méthode décrite auparavant au moyen des données sur le revenu familial recueillies dans le cadre de l'Enquête sur les finances des consommateurs (EFC) du Canada. Nous nous servirons du fichier sur le revenu disponible des familles économiques pour l'Ontario en 1988. Par "revenu disponible", on entend le revenu total après impôt mentionné lors du sondage. L'EFC a le même cadre que l'Enquête sur la population active, qui repose elle-même sur un échantillonnage stratifié à plusieurs degrés. Pour en apprendre davantage au sujet du plan d'échantillonnage, voir Singh et coll. (1990).

Nous avons estimé la médiane M, le coefficient de Gini G, la mesure du faible revenu Θ, les ordonnées de la courbe de Lorenz et la part des quantiles $Q(0,.2), Q(.2,.4), Q(.4,.6), Q(.6,.8), Q(.8,.1.0)$. Les erreurs-types correspondantes ont été obtenues avec la méthode proposée et la méthode du jackknife avec suppression d'une grappe.

Voici une brève description de la dernière méthode, citée aux fins d'illustration. Tout d'abord, on suppose que l'estimation du paramètre inconnu Θ est établie par l'expression $\hat{\Theta} = \mathcal{L}(\hat{F})$, où $\hat{F}$ représente la fonction de distribution estimée. L'estimation de la fonction de distribution $\hat{F}_{(hj)}$ obtenue de l'échantillon après suppression de la j-ième grappe prélevée de la h-ième strate ($j = 1, \ldots, n_h$, $h = 1, \ldots, H$) est

$$\hat{F}_{(gj)}(y) = \sum_s A_{hci}(g,j)\, w_{hci} I\{y_{hci} \leq y\}$$

où
$$A_{hci}(g,j) = \begin{cases} 1, & h \neq g, \\ \dfrac{n_g}{n_g - 1}, & h = g,\ c \neq j; \\ 0, & h = g,\ c = j. \end{cases}$$

Par conséquent, $\hat{\Theta}_{(gj)} = \mathcal{L}(\hat{F}_{(gj)})$ et l'estimateur jackknife, après suppression d'une grappe, de la variance $\hat{\Theta} = \mathcal{L}(\hat{F})$ est

$$\text{var}_J(\hat{\Theta}) = \sum_{g=1}^{H} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{\Theta}_{(gj)} - \hat{\Theta})^2.$$

On sait que l'estimateur jackknife de la variance laisse à désirer avec les quantiles en raison de son manque de convergence (Kovar et coll. 1988). Des résultats plus récents (Shao et Wu 1989; Rao, Wu et Yue 1992) suggèrent que dans certaines conditions, la méthode du jackknife avec suppression du paramètre d ou suppression d'une grappe pourrait avoir les propriétés asymptotiques désirées pour l'estimation de la variance des statistiques non lisses comme les quantiles ou la mesure du faible revenu. Parallèlement, pour d'autres statistiques comme le coefficient de Gini, l'estimateur jackknife de la variance asymptotique est convergent (Shao 1993).

À l'inverse de la méthode du jackknife, la méthode des équations d'estimation n'exige pas d'importants calculs. Elle est simple et explicite, et elle intègre le plan d'échantillonnage. Elle engendre des formules faciles à programmer pour la variance asymptotique, en dépit de leur apparente complexité.

Puisque l'utilisation d'un seul échantillon restreint la comparaison objective de deux méthodes, notre exemple n'a d'autre but que faire ressortir la variation de l'erreur-type telle qu'elle est obtenue avec des équations d'estimation et une méthode de calcul intensif comme celle du jackknife. Les résultats apparaissent sous forme abrégée dans le tableau qui suit. La tendance suivie par la variation de l'erreur-type estimée confirme le caractère généralement classique de la méthode du jackknife. L'écart est attribuable au biais à la hausse introduit par cette méthode dans le cas de la médiane, bien que la méthode du jackknife avec suppression d'une grappe semble préférable à la même méthode avec suppression de la valeur 1. En ce qui concerne la part des quantiles, l'erreur peut s'expliquer en partie par le fait que la part des quantiles supérieurs ne recoupe pas toutes les unités primaires d'échantillonnage mais fonctionne à la manière de classes distinctes dont l'effet sur la méthode du jackknife pourrait être plus marqué que sur la méthode des équations d'estimation.

Tableau 1

Mesure de l'inégalité du revenu et de l'erreur-type

Mesure	Estimation	Erreur-type	
		Méthode des équations d'estimation	Méthode du jackknife avec suppression d'une grappe
Médiane	31705	303.3	569.8
Gini	0.3482	0.005	0.005
Faible revenu	0.1980	0.00586	0.00613
Ordonnées de la courbe de Lorenz			
L(0.2)	0.0561	0.00137	0.00175
L(0.4)	0.1745	0.00166	0.00194
L(0.6)	0.3522	0.00246	0.00285
L(0.8)	0.5982	0.00317	0.00393
Part du quintile			
Q(0, 0.2)	0.0561	0.00137	0.00167
Q(0.2, 0.4)	0.1186	0.00159	0.00221
Q(0.4, 0.6)	0.1775	0.00157	0.00282
Q(0.6, 0.8)	0.2461	0.00158	0.00337
Q(0.8, 1.0)	0.4017	0.00395	0.00451

8. SOMMAIRE

La solution au problème consistant à estimer la variance de statistiques complexes comme la mesure de l'inégalité du revenu a longtemps échappé aux statisticiens. On propose souvent une méthode de répétition comme celle du jackknife pour procéder à l'estimation. L'avantage de la linéarisation est qu'elle se prête à de très nombreux plans d'échantillonnage sans nécessiter de calculs laborieux comme d'autres méthodes, par exemple, la méthode bootstrap. Grâce aux fonctions d'estimation et à la décomposition décrite en (2.3), on se rend compte que certains problèmes délicats peuvent être résolus plus facilement. Le présent article aborde aussi la question de l'ordre de grandeur du reste de certaines mesures. Enfin, on peut effectuer une démonstration plus rigoureuse de la méthode pour un plan d'échantillonnage complexe en suivant les recommandations de Shao et Rao (1994).

BIBLIOGRAPHIE

BEACH, C.M., et DAVIDSON, R. (1983). Distribution-free statistical inference with Lorenz curves and income shares. *Review of Economic Studies*, 50, 723-735.

BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

BINDER, D.A. (1991). Use of estimating functions for interval estimation from complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 34-42.

BINDER, D.A., et PATAK, Z. (1994). Use of estimating functions for interval estimation from complex surveys. *Journal of the American Statistical Association*, 89, 1035-1044.

FRANCISCO, C.A., et FULLER, W.A. (1986). Estimation of the Distribution function with a complex survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 37-45.

FRANCISCO, C.A., et FULLER, W.A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, 19, 454-469.

GLASSER, G.J. (1962). Variance formulas for the mean difference and coefficient of concentration. *Journal of the American Statistical Association*, 57, 648-654.

KOVAR, J.G., RAO, J.N.K., et WU, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *La Revue Canadienne de Statistique*, 16, 25-45.

NYGÅRD, F., et SANDSTRÖM, A. (1981). *Measuring Income Inequality*. Stockholm: Almqvist and Wiksell International.

RANDLES, R.H. (1982). On the Asymptotic Normality of Statistics with Estimated Parameters. *Annals of Statistics*, 10, 462-474.

RAO, J.N.K., et WU, C.F.J. (1987). Methods for Standard Errors and Confidence Intervals from Survey Data: Some Recent Work. *Actes de la 46ième session, Institut Internationale de Statistique*, 3, 5-19.

RAO, J.N.K., WU, C.F.J., et YUE, K. (1992). Quelques travaux récents sur les méthodes de rééchantillonnage applicables aux enquêtes complexes. *Techniques d'enquêtes*, 18, 225-234.

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

SENDLER, W. (1979). On statistical inference in concentration measurement. *Metrika*, 26, 109-122.

SHAO, J., et RAO, J.N.K. (1994). Standard Errors for Low Income Proportions Estimated from Stratified Multi-Stage Samples. *Sankhyā, B*, (à paraître).

SHAO, J. et WU, C.W.J. (1989). A general Theory for Jackknife Variance Estimation. *Annals of Statistics*, 17,1176-1197.

SHAO, J. (1993). Inferences Based on *L*-statistics in Survey Problems: Lorenz Curve, Gini Family and Poverty Proportion. In *Proceedings of the Workshop on Statistical Issues in Public Policy Analysis*, Carleton University et l'Université d'Ottawa.

SINGH, M.P., DREW, J.D., GAMBINO, J.G., et MAYDA, F. (1990). *Méthodologie de l'enquête sur la population active du Canada*, n° 71-526 au catalogue, Statistique Canada.

WOODRUFF, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47,635-646.

Techniques d'enquête, décembre 1995
Vol. 21, n° 2, pp. 161–171
Statistique Canada

Un algorithme de transport à taille réduite pour maximiser le chevauchement des enquêtes

LAWRENCE R. ERNST et MICHAEL M. IKEDA[1]

RÉSUMÉ

Lorsqu'on remanie un échantillon selon un plan stratifié à plusieurs degrés, il est parfois indiqué de maximiser le nombre d'unités primaires d'échantillonnage retenues dans le nouvel échantillon sans modifier les probabilités de sélection inconditionnelle. Il existe à cette fin une solution optimale qui s'appuie sur la théorie du transport pour une classe très générale de plans d'échantillonnage. Toutefois, à la connaissance des auteurs, cette méthode n'a jamais été utilisée pour la refonte d'une enquête. Cela s'explique en partie du fait que même pour une strate de taille modérée, le problème de transport résultant pourrait être trop grand pour se prêter à une solution pratique. Dans le présent article, nous proposons un algorithme de transport modifié à taille réduite permettant de maximiser le chevauchement, ce qui réduit sensiblement les dimensions du problème. Cette méthode a été utilisée lors du remaniement récent de l'enquête sur le revenu et la participation aux programmes (Survey of Income and Program Participation, ou SIPP). Nous décrivons brièvement le rendement de l'algorithme à taille réduite, d'une part pour le chevauchement du SIPP en conditions réelles, et d'autre part pour des simulations artificielles antérieures du chevauchement du SIPP. Même si cette méthode n'est pas optimale et ne risque de produire, en théorie, que des améliorations négligeables du chevauchement attendu comparativement à la sélection indépendante, elle ouvre en pratique la voie à des améliorations importantes du chevauchement par rapport à la sélection indépendante dans le cas du SIPP et produit généralement un chevauchement presque optimal.

MOTS CLÉS: Programmation linéaire; remaniement de l'échantillon; Survey of Income and Program Participation.

1. INTRODUCTION

Le problème de la maximisation du nombre prévu d'unités primaires d'échantillonnage (UPÉ) retenues dans un échantillon lors du remaniement d'une enquête avec un plan stratifié pour lequel les UPÉ sont choisies avec une probabilité proportionnelle à la taille a été abordé pour la première fois dans la documentation scientifique par Keyfitz (1951). Typiquement, la maximisation du chevauchement des UPÉ est motivée par la réduction des coûts supplémentaires tels que ceux requis pour la formation des intervieweurs aux fins d'une enquête auprès des ménages, lesquels s'additionnent à chaque changement d'UPÉ. Les méthodes de maximisation du chevauchement n'influent pas sur la probabilité inconditionnelle de sélection d'un ensemble d'UPÉ dans une nouvelle strate, mais conditionnent sa probabilité de sélection d'une manière telle que la probabilité de sélection d'une UPÉ dans le nouvel échantillon est en règle générale plus grande que la probabilité inconditionnelle lorsque cette UPÉ se trouvait dans l'échantillon initial, et moins grande autrement.

Les méthodes de chevauchement sont applicables lorsque le remaniement entraîne soit une nouvelle stratification des UPÉ, soit un changement de leurs probabilités de sélection. Keyfitz (1951) a proposé une méthode optimale, mais uniquement pour les plans à une UPÉ par strate dans le cas spécial où la strate initiale et la nouvelle strate sont identiques et où seules les probabilités de sélection changent. Causey, Cox et Ernst (1985) ont obtenu une solution optimale du problème du chevauchement sous des conditions très générales en assimilant ce problème à un problème de transport: une forme spéciale de problème de programmation linéaire. Cette méthode n'impose aucune restriction aux changements de définitions des strates ni au nombre d'UPÉ par strate. (Un résultat similaire a été obtenu de façon indépendante par Arthanari et Dodge (1981), même si ces derniers n'ont pas abordé la question des changements dans les définitions des strates. Les deux groupes de chercheurs ont obtenus leurs résultats en généralisant le travail de Raj (1968).) Toutefois, il existe au moins deux autres difficultés avec la méthode de Causey, Cox et Ernst qui peuvent la rendre inutilisable en pratique. La première est examinée par Ernst (1986), et la seconde fait l'objet du présent article.

La première difficulté découle du fait que si l'échantillon initial d'UPÉ n'a pas été sélectionné indépendamment d'une strate à l'autre, l'information nécessaire au calcul de l'ensemble des probabilités conjointes requises par cette méthode risque de ne pas être disponible en pratique. Ernst (1986) a mis au point une méthode de programmation linéaire de rechange à utiliser dans de tels cas. Le Bureau of the Census a eu recours à la programmation linéaire pour réaliser le chevauchement de ses enquêtes démographiques à cinq reprises. À quatre de ces occasions (sélection des plans pour les Current Population Surveys (CPS) et pour les National Crime Victimization Surveys (NCVS) des années 1980 et 1990), la méthode établie par Ernst (1986) a été retenue puisque le plan initial

[1] Lawrence R. Ernst, Chief, Research Group, Office of Compensation and Working Conditions, Bureau of Labor Statistics, Washington, DC 20212, U.S.A.; Michael M. Ikeda, Mathematical Statistician, Statistical Research Division, Bureau of the Census, Washington, DC 20233, U.S.A.

n'était pas sélectionné indépendamment d'une strate à l'autre. En particulier, comme l'avait expliqué Ernst (1986), si l'échantillon initial est lui-même sélectionné par chevauchement avec un plan antérieur, l'hypothèse d'indépendance n'est en général pas valide, ce qui expliquait principalement pourquoi elle n'était pas valide dans ces quatre cas particuliers.

Le second problème posé par la méthode optimale est que le problème du transport risque d'être trop grand pour pouvoir être résolu en pratique. Le Bureau of the Census a également utilisé la programmation linéaire pour le chevauchement du plan du SIPP des années 1990 avec celui des années 1980, deux plans à deux UPÉ par strate. L'échantillon initial du SIPP a été sélectionné indépendamment d'une strate à l'autre. Toutefois, le problème du transport pour la méthode optimale aurait été trop grand pour autoriser une solution pratique pour plusieurs des strates. Cette difficulté se pose du fait que pour chaque nouvelle strate constituée de n UPÉ, le nombre de variables du problème de transport pour la méthode optimale peut atteindre $2^n \times \binom{n}{2}$. La valeur de n la plus grande pour laquelle un problème de transport comportant un tel nombre de variables peut être résolu avec les moyens informatiques que nous avions à notre disposition était approximativement $n = 15$.

Nous présentons ci-après une formule à taille réduite de la méthode de chevauchement, assimilée à un problème de transport, qui réduit le nombre de variables du SIPP à $\left(\binom{n}{2} + n + 1\right) \times \binom{n}{2}$, une réduction surprenante pour les valeurs de n modérées à grandes. Cette méthode part de l'hypothèse que l'échantillon initial a été sélectionné indépendamment d'une strate à l'autre; elle n'aurait de ce fait pas pu remplacer la méthode de Ernst (1986) pour le chevauchement des plans du CPS et du NCVS. Cette méthode à taille réduite a été utilisée avec succès pour des strates comptant jusqu'à 68 UPÉ. À titre de comparaison, pour un $n = 68$, le nombre possible de $2^{68} \times \binom{68}{2}$ variables de la formule non réduite dépasse de loin la taille des problèmes qui peuvent être résolus avec les ordinateurs courants. En outre, même si la réduction de la taille se fait au détriment de l'optimalité, il semble qu'elle donne en pratique des résultats passablement proches des valeurs optimales, comme nous le démontrerons ci-après. La méthode à taille réduite est celle que nous avons utilisée pour le chevauchement du SIPP.

L'examen de la méthode de Causey, Cox et Ernst (1985), présenté dans la section 2, nous sert de toile de fond pour la présentation de la méthode à taille réduite.

La méthode à taille réduite est présentée à la section 3. Même si elle peut se prêter à une application générale, nous nous contenterons, pour simplifier la présentation, de ne décrire en détails que le cas où le plan initial et le nouveau plan comportent deux UPÉ par strate et se font sans remise. Nous présentons également, dans la même section, un petit exemple artificiel de la méthode à taille réduite qui servira à illustrer la méthode et à démontrer que l'ordonnancement des paires d'UPÉ dans la strate d'un nouveau plan, une étape clé de l'algorithme, influe sur le chevauchement attendu. Nous soulignons également dans cette section certains résultats analytiques de la comparaison de la méthode à taille réduite et de la méthode optimale. Nous précisons les limites supérieures de la perte du chevauchement attendu découlant de l'utilisation de la méthode à taille réduite au lieu de la méthode optimale. Nous expliquons également que dans certaines situations, cette perte peut s'approcher de deux UPÉ pour les plans à deux UPÉ par strate - le pire scénario envisageable. On trouvera dans l'article de Ernst et Ikeda (1994) de plus amples détails ainsi que les preuves des résultats présentés dans cette section, en plus de certains résultats d'autres sections du présent article.

Dans la section 4, nous abordons la question du rendement de la méthode à taille réduite, tant pour le chevauchement du SIPP actuel que pour des simulations artificielles du chevauchement des SIPP antérieurs. Le chevauchement attendu avec cette méthode est comparé à celui de la sélection indépendante des UPÉ du nouvel échantillon et à la limite supérieure du chevauchement optimal attendu. Les résultats montrent que pour cette application, et contrairement à certains des résultats théoriques décrits dans la section 3, le chevauchement attendu avec la méthode à taille réduite est beaucoup plus grand que si on avait recouru à une sélection indépendante pour choisir les UPÉ du nouvel échantillon. Ce chevauchement est en fait presque aussi grand que le chevauchement optimal prévu. Nous précisons finalement le temps machine nécessaire pour la méthode à taille réduite en fonction de la taille des strates.

Finalement, les conclusions de notre étude sont présentées à la section 5.

2. EXAMEN DE LA MÉTHODE DE CHEVAUCHEMENT DE CAUSEY, COX ET ERNST (1985)

La méthode de chevauchement de Causey, Cox et Ernst (1985), comme toutes les méthodes de chevauchement, conditionne d'une certaine manière la sélection des UPÉ des échantillons de chaque nouvelle strate au choix des UPÉ de la strate qui faisaient partie de l'échantillon initial. Cette méthode de chevauchement particulière réalise une optimalité réelle en faisant plein usage de cette information et en assimilant la procédure à un problème de transport. Nous en présentons ci-après la description.

Précisons tout d'abord certaines des notations qui nous serviront tout au long de notre exposé. Désignons par S une strate du nouveau plan d'échantillonnage. Chacune de ces strates correspond à un problème de chevauchement séparé. Désignons par n le nombre d'UPÉ comprises dans S, et par $A_1, \ldots, A_n$ les UPÉ en question comprises dans S. Désignons par I le sous-ensemble aléatoire de $\{1, \ldots, n\}$ tel que $k \in I$ si et seulement si A_k était présent dans l'échantillon initial, et désignons par N l'ensemble correspondant du nouvel échantillon. Par exemple, si A_2 et A_3 étaient les UPÉ de S qui faisaient partie de l'échantillon initial et A_1 et A_3 les UPÉ comprises dans le nouvel échantillon, alors, $I = \{2,3\}$ et $N = \{1,3\}$. Désignons par m^*, n^* le nombre de valeurs possibles de I et N respectivement.

Désignons par ailleurs par J_i, $i = 1, \ldots, m^*$ les valeurs possibles de I et par S_j, $j = 1, \ldots, n^*$ les valeurs possibles de N. L'objet de toutes les méthodes de chevauchement est de maximiser le nombre prévu d'UPÉ dans $N \cap I$, tout en préservant les valeurs des $P(S_j)$.

Pour illustrer plus avant certaines de ces notions, supposons un exemple où $n = 3$. Alors, $n^* = 3$ si le nouveau plan compte une ou deux UPÉ par strate, les valeurs de N, c'est-à-dire les S_j, étant $\{1\},\{2\},\{3\}$ dans le cas à une UPÉ par strate et $\{1,2\},\{1,3\},\{2,3\}$ dans le cas à deux UPÉ par strate. Supposons maintenant que les UPÉ A_1 et A_2 étaient dans une strate initiale et que l'UPÉ A_3 était dans une autre, et qu'il y avait trois UPÉ dans chacune de ces strates initiales. Si le plan initial comptait une UPÉ par strate, alors $m^* = 6$, et les valeurs de I, c'est-à-dire les J_i, étaient $\varnothing$, $\{1\},\{2\},\{3\},\{1,3\},\{2,3\}$. Si le plan initial comptait deux UPÉ par strate, alors $m^* = 6$, et les valeurs des J_i étaient $\{1\},\{2\},\{1,2\},\{1,3\},\{2,3\}$, $\{1,2,3\}$.

Examinons maintenant le problème de transport pour la méthode de chevauchement de Causey, Cox et Ernst (1985). Soit $P(J_i)$ la probabilité que $I = J_i$ et $P(S_j)$ la probabilité que $N = S_j$. En outre, désignons par x_{ij} la variable dénotant la probabilité conjointe de ces deux événements, et désignons par c_{ij} le nombre d'éléments dans $J_i \cap S_j$. Les valeurs des $P(J_i)$, des $P(S_j)$ et des c_{ij} sont connues alors que celles des x_{ij} sont des variables dont il s'agit de déterminer les valeurs optimales. Ainsi, le problème de transport à résoudre consiste à déterminer la valeur de $x_{ij} \geq 0$ qui maximise

$$\sum_{i=1}^{m^*} \sum_{j=1}^{n^*} c_{ij} x_{ij} \tag{2.1}$$

sous réserve que

$$\sum_{j=1}^{n^*} x_{ij} = P(J_i), \quad i = 1, \ldots, m^*, \tag{2.2}$$

$$\sum_{i=1}^{m^*} x_{ij} = P(S_j), \quad j = 1, \ldots, n^*. \tag{2.3}$$

Noter que dans ce problème de transport, la fonction objective (2.1) est le nombre prévu d'UPÉ de S qui sont dans $N \cap I$. Noter également que les contraintes (2.2) et (2.3) sont requises par les définitions des $P(J_i)$, des $P(S_j)$ et des x_{ij}.

Une fois obtenues les valeurs optimales x_{ij}, la probabilité conditionnelle que $N = S_j$ étant donné $I = J_i$ est donnée par $x_{ij}/P(J_i)$ pour tous les i,j.

Nous présentons ci-après un exemple de l'utilisation des formules (2.1) à (2.3) dans un cas où le plan initial et le plan nouveau sont tous les deux à deux UPÉ par strate et sans remise. Dans cet exemple comme partout ailleurs dans le présent article, nous désignerons par p_i, π_i la probabilité prédéterminée que $i \in I$ et $i \in N$, respectivement.

Supposons une strate finale S avec $n = 3$. Toutes les UPÉ étaient dans des strates différentes. Soit $p_1 = .6$, $p_2 = .75, p_3 = .7$, $\pi_1 = .5$, $\pi_2 = .8$, $\pi_3 = .7$. Puisque les UPÉ étaient toutes dans des strates initiales différentes, il existe 8 possibilités différentes pour I, dont les probabilités sont présentées dans le tableau 1.

Tableau 1

Probabilités des groupes possibles d'UPÉ de l'échantillon initial

i	1	2	3	4	5	6	7	8
J_i	{1,2,3}	{1,2}	{1,3}	{2,3}	{1}	{2}	{3}	$\varnothing$
$P(J_i)$	.315	.135	.105	.21	.045	.09	.07	.03

Puisque le nouveau plan d'échantillonnage est à deux UPÉ par strate et sans remise, il existe trois possibilités pour N: les paires $S_1 = \{1,2\}$, $S_2 = \{1,3\}$, $S_3 = \{2,3\}$. Ainsi, $P(S_1) = .30$, $P(S_2) = .20$ et $P(S_3) = .50$.

Les valeurs c_{ij} sont présentées dans le tableau 2. En maximisant ensuite (2.1) sous réserve de (2.2) et (2.3) avec les valeurs données des $P(J_i)$, $P(S_j)$ et c_{ij}, on obtient le groupe de valeurs optimales de x_{ij} présentées dans le tableau 2. Finalement, en divisant chacune des entrées x_{ij} d'un rang i du tableau 2 par $P(J_i)$, on obtient un groupe optimal de probabilités conditionnelles $P(S_j \mid J_i)$. Par exemple, puisque $x_{12} = .025$ et $P(J_1) = .315$, il s'ensuit que $P(S_2 \mid J_1) = 5/63$.

Tableau 2

Valeurs de c_{ij} et de x_{ij} qui maximisent le chevauchement pour la méthode optimale

	c_{ij}			x_{ij}		
	j			j		
i	1	2	3	1	2	3
1	2	2	2	.000	.025	.290
2	2	1	1	.135	.000	.000
3	1	2	1	.000	.105	.000
4	1	1	2	.000	.000	.210
5	1	1	0	.045	.000	.000
6	1	0	1	.090	.000	.000
7	0	1	1	.000	.070	.000
8	0	0	0	.030	.000	.000

Pour cet exemple, comme on peut le calculer à partir de (2.1) et du tableau 2, le chevauchement attendu en vertu de la procédure optimale est de 1.735 UPÉ. À titre comparatif, le chevauchement attendu si les plans initial et final sont sélectionnés indépendamment est de $p_1\pi_1 + p_2\pi_2 + p_3\pi_3 = 1.39$ UPÉ.

Pour les plans d'échantillonnage sans remise à deux UPÉ par strate, les valeurs possibles de N sont toujours données par les sous-ensembles $\binom{n}{2}$ de $\{1, \ldots, n\}$ de

taille 2, c'est-à-dire $n^* = \binom{n}{2}$. Toutefois, m^* peut varier largement. $m^* = \binom{n}{2}$ lorsque les UPÉ de S comprennent une seule strate initiale. La limite supérieure de 2^n sur m^* est atteinte lorsque toutes les UPÉ de S étaient dans des strates initiales différentes, comme le montre l'exemple précédent, et comme dans certaines autres situations. Ernst et Ikeda (1994) présentent une expression générale exacte de m^*.

En ce qui concerne le chevauchement des plans d'échantillonnage sans remise à deux UPÉ par strate, le nombre de variables du problème de transport pour la méthode optimale est m^*n^*, et peut atteindre jusqu'à $2^n\binom{n}{2}$. Si $n = 15$, $2^n\binom{n}{2} = 3{,}440{,}640$, ce qui correspond à peu près à la capacité de résolution des ordinateurs que nous avions à notre disposition. Toutefois, $n > 15$ pour près de la moitié des strates qui ne sont pas auto-représentatives (c'est-à-dire celles constituées d'UPÉ sélectionnées avec probabilité) dans l'enquête SIPP, et il s'est donc avéré nécessaire d'élaborer une méthode, décrite dans la prochaine section, qui réduit le problème du transport tout en produisant en pratique un chevauchement attendu presque maximal.

3. L'ALGORITHME DE LA MÉTHODE À TAILLE RÉDUITE

Dans les travaux antérieurs portant sur la réduction de la taille du problème de transport défini par les formules (2.1) à (2.3), on s'est surtout attaché à réaliser la réduction de la taille tout en maintenant l'optimalité. Par exemple, Aragon et Pathak (1990) conservent l'optimalité et réduisent la taille du problème de 75% lorsque $m^* = n^*$. Malheureusement, lorsque m^* est beaucoup plus grand que n^*, c'est-à-dire lorsque la réduction de la taille serait la plus utile, leur méthode produit une réduction de taille négligeable en termes relatifs. Pathak et Fahimi (1992) proposent une généralisation de cette méthode, mais rien n'indique qu'elle puisse aboutir à coup sûr à une réduction de taille sensible en termes relatifs.

Dans la présente section, nous abordons le problème de la réduction de la taille sous un angle différent. Nous sacrifions l'optimalité, à tout le moins en théorie, pour obtenir en retour une réduction du problème de transport qui en autorise une résolution pratique. Nous y arrivons, dans les cas où les plans initial et nouveau comportent tous les deux deux UPÉ par strate par exemple, en ordonnant toutes les paires d'UPÉ dans une nouvelle strate puis en conditionnant les nouvelles probabilités de sélection pour tout ensemble initial d'UPÉ de l'échantillon de taille supérieure à 2 sur la première paire d'UPÉ de l'ensemble initial plutôt que sur la totalité de cet ensemble initial. Autrement dit, chaque ensemble initial possible d'UPÉ de l'échantillon comportant plus de deux UPÉ est combiné à un ensemble de taille 2. Comme nous l'expliquons à la section 4, cette méthode peut en pratique donner un chevauchement presque optimal, en particulier avec un ordonnancement approprié des paires d'UPÉ tel qu'il est décris à la section 3.1.2.

La méthode de réduction de la taille s'applique dans tous les cas où les UPÉ des plans initial et nouveau sont tirées sans remise. Toutefois, nous limiterons notre description détaillée, dans la section 3.1, aux cas où le plan initial et le nouveau plan comportent tous les deux deux UPÉ par strate. Nous expliquerons ensuite brièvement, dans la section 3.2, les changements nécessaires pour appliquer cette méthode à d'autres types de plans initiaux et nouveaux. Finalement, dans la section 3.3, nous présenterons certains résultats analytiques portant sur les rapports entre le chevauchement attendu de la méthode à taille réduite, de la méthode optimale et de la sélection indépendante. Tout au long de la présente section, nous présumons au départ que les UPÉ de l'échantillon initial ont toutes été sélectionnées indépendamment d'une strate à l'autre.

3.1 Méthode de réduction de la taille pour deux plans à deux UPÉ par strate

La méthode que nous décrivons ci-après présente les aspects clés suivants: l'ordonnancement particulier des paires d'UPÉ; la refonte du problème de transport (2.1) à (2.3) aux fins de la méthode à taille réduite; le calcul des probabilités correspondant aux résultats initiaux de la nouvelle formulation; le calcul des coefficients de coût (c_{ij}) de la fonction objective. Dans la section 3.1.1, nous présentons une description détaillée de la méthode à taille réduite, et une nouvelle définition du problème de transport. L'ordonnancement des paires est décrit dans la section 3.1.2. Finalement, nous présentons dans la section 3.1.3 le calcul des probabilités des résultats initiaux et des coefficients de coût.

3.1.1 Description générale de la méthode

Dans une première étape, les sous-ensembles $\binom{n}{2}$ de $\{1, \ldots, n\}$ de taille 2 sont ordonnés selon une méthode que nous décrirons ultérieurement. (Qu'il nous suffise pour l'instant de mentionner que n'importe quel ordonnancement peut servir à réduire la taille du problème de transport. La méthode précise que nous retenons permet de parvenir à ce résultat en conservant la plus grande part possible des gains de la méthode optimale en matière de chevauchement.) Désignons par I_i, $i = 1, \ldots, \binom{n}{2}$, le i-ième élément de l'ordonnancement et par $I_{\binom{n}{2}+1}, \ldots, I_{\binom{n}{2}+n}$ les n sous-ensembles à un seul élément. Posons finalement que $I_{\binom{n}{2}+n+1} = \varnothing$. Ainsi, les I_i constituent chacun des sous-ensembles de $\{1, \ldots, n\}$ de deux éléments ou moins. À chaque possibilité de I correspond un ensemble I^* unique parmi ces $\binom{n}{2} + n + 1$ sous-ensembles, et les nouvelles probabilités de sélection sont conditionnées sur ce I^* plutôt que sur I. Ainsi, les nouvelles probabilités de sélection sont conditionnés sur $\binom{n}{2} + n + 1$ événements plutôt que sur un nombre possible de 2^n événements, ce qui explique la réduction de la taille. Le paramètre I^* est le premier I_i pour lequel $I_i \subset I$. C'est-à-dire que si I consiste en au moins deux nombres entiers, I^* correspondra à la première paire de l'ordonnancement contenue dans I, tandis que si I est un ensemble à un seul élément ou un ensemble vide, on obtiendra alors $I^* = I$.

Le problème de transport à taille réduite cherche à conserver les UPÉ correspondant aux éléments de l'ensemble I^* dans le nouvel échantillon, mais n'utilise pas l'information sur les éléments dans $I \sim I^*$. Ce problème de transport à taille réduite fondé sur l'ensemble de I_i prend la forme décrite ci-après. Désignons par p_i^* la probabilité que $I^* = I_i$, $i = 1, \ldots, \binom{n}{2} + n + 1$, et abrégeons $\pi_j^* = P(S_j)$, $j = 1, \ldots, \binom{n}{2}$. Pour chaque i,j, la variable x_{ij} correspond à la probabilité conjointe que $I^* = I_i$ et que $N = S_j$, alors que c_{ij} désigne le nombre attendu d'éléments dans $I \cap S_j$ étant donné $I^* = I_i$. Le problème consiste à déterminer le $x_{ij} \geq 0$ qui maximise

$$\sum_{i=1}^{\binom{n}{2}+n+1} \sum_{j=1}^{\binom{n}{2}} c_{ij}\, x_{ij}, \tag{3.1}$$

sous réserve que

$$\sum_{j=1}^{\binom{n}{2}} x_{ij} = p_i^*, \quad i = 1, \ldots, \binom{n}{2} + n + 1, \tag{3.2}$$

$$\sum_{i=1}^{\binom{n}{2}+n+1} x_{ij} = \pi_j^*, \quad j = 1, \ldots, \binom{n}{2}. \tag{3.3}$$

Une fois obtenues les valeurs optimales de x_{ij}, les probabilités conditionnelles de nouvelle sélection pour S_j, $j = 1, \ldots, \binom{n}{2}$, étant donné $I^* = I_i$, sont x_{ij}/p_i^*. À noter que le nombre de variables x_{ij} dans les formules (3.1) à (3.3) est $\left(\binom{n}{2} + n + 1\right) \times \binom{n}{2}$, comparativement à un maximum de $2^n \times \binom{n}{2}$ dans les formules (2.1) à (2.3).

Il nous reste à expliquer la méthode générale utilisée pour ordonner les $\binom{n}{2}$ paires ainsi que les méthodes de calcul des valeurs de p_i^* et de c_{ij}. Avant cela, nous procéderons à une démonstration de la méthode à taille réduite avec l'exemple à deux UPÉ par strate utilisé dans la section 2 afin d'illustrer la formulation du problème de transport pour la méthode optimale.

L'ordonnancement des paires de cet exemple, comme il sera démontré plus tard, est $\{2,3\}, \{1,2\}, \{1,3\}$. En conséquence, les valeurs I_i sont présentées dans le tableau 3. Noter que si $I = \{1,2,3\}$ ou $I = \{2,3\}$, l'ensemble associé devient alors $I_1 = \{2,3\}$. Pour les six autres possibilités pour I l'ensemble associé est I lui-même.

Par conséquent, en utilisant les valeurs du tableau 1, nous obtenons

$$p_1^* = P(I = \{1,2,3\}) + P(I = \{2,3\}) = .525, \tag{3.4}$$

$p_i^* = P(J_i)$, $i = 2,3$, et $p_i^* = P(J_{i+1})$, $i = 4, \ldots, 7$, donnant les valeurs présentées au tableau 3. Puisque $\pi_j^* = P(S_j)$, nous obtenons $\pi_1^* = .30$, $\pi_2^* = .20$, $\pi_3^* = .50$.

Tableau 3

Probabilités des ensembles associés: méthode à taille réduite

	i						
	1	2	3	4	5	6	7
I_i	{2,3}	{1,2}	{1,3}	{1}	{2}	{3}	∅
p_i^*	.525	.135	.105	.045	.09	.07	.03

Les valeurs de c_{ij} pour cet exemple sont présentées au tableau 4. Pour les obtenir, nous avons simplifié les calculs en posant que

$$b_{it} = P(t \in I \mid I^* = I_i),$$

$$i = 1, \ldots, \binom{n}{2} + n + 1, \quad t = 1, \ldots, n, \tag{3.5}$$

et en notant que si $S_j = \{s,t\}$, alors

$$c_{ij} = b_{is} + b_{it}. \tag{3.6}$$

Ainsi, le nombre attendu d'éléments dans $I \cap S_j$, sous réserve que $I^* = I_i$, est simplement la somme des probabilités que chacun des deux éléments de S_j soit dans I, étant donné $I^* = I_i$. Observer en outre que si le problème de transport pour la méthode optimale connaît la valeur exacte de I et connaît donc avec certitude si chaque élément de S_j est également dans I, tel n'est pas le cas en ce qui concerne la méthode à taille réduite puisque dans ce cas, seul l'ensemble associé I_i est connu. À titre d'illustration, examinons le premier rang du tableau 4. Puisque $I_1 = \{2,3\}$, nous savons que $2 \in I$ et $3 \in I$, et donc que $b_{12} = b_{13} = 1$. Toutefois, nous ne savons pas avec certitude si $1 \in I$ puisque I_1 est l'ensemble associé à la fois pour $I = \{1,2,3\}$ et $I = \{2,3\}$. En fait, en se reportant au tableau 1,

$$b_{11} = \frac{P(I = \{1,2,3\})}{P(I = \{1,2,3\}) + P(I = \{2,3\})} = .6.$$

Tableau 4

Valeurs de c_{ij} et de x_{ij} qui maximisent le chevauchement pour la méthode à taille réduite

		c_{ij}			x_{ij}		
		j			j		
i	I_i	1	2	3	1	2	3
1	{2,3}	1.6	1.6	2.0	0.000	0.025	0.500
2	{1,2}	2.0	1.0	1.0	0.135	0.000	0.000
3	{1,3}	1.0	2.0	1.0	0.000	0.105	0.000
4	{1}	1.0	1.0	0.0	0.045	0.000	0.000
5	{2}	1.0	0.0	1.0	0.090	0.000	0.000
6	{3}	0.0	1.0	1.0	0.000	0.070	0.000
7	∅	0.0	0.0	0.0	0.030	0.000	0.000

Ainsi, $c_{11} = b_{11} + b_{12} = 1.6$, et c_{12}, et c_{13} se calculent de la même façon. Pour les six autres rangs du tableau 4, $I_i = I$ et nous savons donc avec certitude quels sont les nombres entiers de I. En conséquence, les c_{ij} de ces six rangs se calculent facilement.

Finalement, nous maximisons le chevauchement attendu (3.1) sous réserve de (3.2) et (3.3), obtenant ainsi les valeurs x_{ij} du tableau 4. Les probabilités conditionnelles $P(N = S_j \mid I^* = I_i)$ du tableau 5 sont ensuite calculées en divisant chacune des entrées x_{ij} dans le i-ième rang du tableau 4 par p_i^*.

Tableau 5

Probabilités conditionnelles pour la méthode à taille réduite

		j		
i	I_i	1	2	3
1	{2,3}	0	1/21	20/21
2	{1,2}	1	0	0
3	{1,3}	0	1	0
4	{1}	1	0	0
5	{2}	1	0	0
6	{3}	0	1	0
7	∅	1	0	0

Le chevauchement attendu pour la méthode à taille réduite est inférieur de .01 à la valeur optimale, c'est-à-dire 1.725 UPÉ. Cet écart par rapport à l'optimalité est dû uniquement au fait que le chevauchement attendu est de 1.6 pour l'événement conjoint $I^* = \{2,3\}$ et $N = \{1,3\}$. Puisque la probabilité de cet événement conjoint est .025 et que la méthode optimale pour cet exemple donne toujours un chevauchement de 2 lorsqu'il existait au moins 2 UPÉ dans l'échantillon initial, l'écart par rapport à l'optimalité est de $.025(2 - 1.6) = .01$.

La méthode à taille réduite ne permet pas de réaliser l'optimalité parce que la paire {2,3} est assortie d'une probabilité de sélection moindre dans le nouvel échantillon que dans l'échantillon initial. Ainsi, tant pour la méthode optimale que pour la méthode à taille réduite, il convient parfois de sélectionner une autre paire (toujours {1,3} pour les deux méthodes de cet exemple) lorsque l'échantillon initial était {2,3}. Les deux méthodes se distinguent en ce que la méthode optimale choisit uniquement {1,3} lorsque $1 \in I$. Avec la méthode à taille réduite, il n'est pas possible d'utiliser l'information concernant $1 \in I$. Ainsi, lorsque $\{2,3\} \subset I$, $1 \in N$ indépendamment de $1 \in I$. C'est là la cause de l'écart par rapport au chevauchement optimal.

3.1.2 Ordonnancement des paires

Nous examinons ci-après comment on parvient généralement à ordonner les paires. Nous désignons à cette fin par p_{st}, π_{st}, s, $t = 1, \ldots, n$, $s \neq t$, la probabilité conjointe que s, $t \in I$ et s, $t \in N$, respectivement.

Nous procédons à l'ordonnancement des paires pour les motifs suivants. Si la i-ième paire de l'ordonnancement est $\{s,t\}$ le problème de transport pourra alors conserver cette paire dans le nouvel échantillon lorsque $I^* = I_i$ avec la probabilité conditionnelle $\min\{1, \pi_{st}/p_i^*\}$. (La probabilité de conservation conditionnelle ne peut dépasser ce seuil puisqu'une valeur plus élevée donnerait une probabilité de sélection inconditionnelle supérieure à π_{st} pour la paire du nouveau plan). Ainsi, l'objectif général de l'ordonnancement est de rendre ces probabilités conditionnelles aussi grandes que possible en moyenne pour toutes les paires.

Pour montrer comment l'ordonnancement des paires influe sur le chevauchement attendu, nous examinons l'exemple du tableau 3. Notre méthode d'ordonnancement, comme nous le montrerons plus tard, donne les résultats indiqués et un chevauchement attendu de 1.725 UPÉ. Examinons maintenant une autre solution d'ordonnancement pour cet exemple. Soit {1,3} la première paire de l'ordonnancement, {1,2} la seconde et {2,3} la dernière, on obtient $I^* = \{1,3\}$ dans tous les cas où $I = \{1,2,3\}$ ou $I = \{1,3\}$. Ainsi, pour cet ordonnancement, p_1^* est la probabilité que $I^* = \{1,3\}$, et sa valeur est .42. En outre, dans ce cas, $p_3^* = P(I^* = \{2,3\}) = P(I = \{2,3\}) = .21$, alors que les 5 autres colonnes du tableau 3 restent inchangées. L'ordonnancement de rechange donne un tableau de probabilités conditionnelles semblable au tableau 5 sauf dans le rang 1 où les colonnes I_i, $j = 2$ et $j = 3$ deviennent maintenant {1,3}, 10/21 et 11/21 respectivement, et dans le rang 3 où les colonnes correspondantes deviennent {2,3}, 0 et 1 respectivement.

En utilisant la méthode de l'ordonnancement original, il est possible de calculer que le chevauchement attendu pour l'ordonnancement de rechange est inférieur de .055 à la valeur optimale, c'est-à-dire qu'il est de 1.68 UPÉ. La raison pour laquelle cet ordonnancement de rechange donne un chevauchement attendu plus faible est donnée ci-après. En général, l'entrée plus tardive d'une paire dans l'ordonnancement entraîne une valeur p_i^* correspondante plus faible et donc une probabilité de rétention conditionnelle plus élevée lorsque $I^* = I_i$. Ainsi, avec {1,3} au premier rang de l'ordonnancement, $\pi_{13}/p_1^* = 10/21$, ce qui correspond à la probabilité conditionnelle de rétention pour cette paire lorsque $I^* = \{1,3\}$. Par contre, lorsque {1,3} est au troisième rang de l'ordonnancement, $\pi_{13}/p_3^* > 1$ et cette paire est retenue avec certitude. Par ailleurs, la probabilité de rétention conditionnelle de la paire {2,3} lorsque $I^* = \{2,3\}$ augmente aussi pour atteindre 1 lorsque {2,3} passe du premier au troisième rang de l'ordonnancement, mais l'augmentation n'est que de 20/21 et l'ordonnancement original du tableau 3 produit donc une augmentation plus grande que prévue du chevauchement attendu que l'ordonnancement de rechange.

Ainsi, comme le montre cet exemple, l'objectif de l'ordonnancement consiste à placer plus tôt dans l'ordonnancement les paires assorties d'une probabilité de rétention conditionnelle relativement élevée même avec un placement précoce. Pour obtenir l'ordonnancement voulu des paires de nombres entiers, on obtiendra d'abord par récursion

un ordonnancement $f(1), \ldots, f(n)$ de $\{1, \ldots, n\}$. Ensuite, pour chaque $k = 1, \ldots, n-1$, on établira par récursion un ordonnancement $g_k(1), \ldots, g_k(n-k)$ de $\{1, \ldots, n\} \sim \{f(1), \ldots, f(k)\}$. Un ordonnancement linéaire des paires distinctes de $\{1, \ldots, n\}$ pourra alors être déterminé comme suit. Chacune de ces paires peut être représentée uniquement sous la forme d'une paire ordonnée $(f(k), g_k(\ell))$ pour un certain $k \in \{1, \ldots, n-1\}$, $\ell \in \{1, \ldots, n-k\}$. Une seconde paire pouvant être représentée sous la forme $(f(k'), g_{k'}(\ell'))$ précède $(f(k), g_k(\ell))$ si et seulement si $k' < k$, ou $k' = k$ et $\ell' < \ell$. À titre d'explication pour l'exemple que nous venons d'examiner, il sera montré plus tard que $f(1) = 2$, $f(2) = 3, f(3) = 1, g_1(1) = 3, g_1(2) = 1$ et $g_2(1) = 1$, et que l'ordonnancement des paires est donc le suivant: $\{2,3\}, \{2,1\}, \{3,1\}$. L'ordonnancement de f et l'ordonnancement de g_k seront tous les deux effectués pour répondre à l'objectif énoncé au début du présent paragraphe.

Pour obtenir l'ordonnancement $f(1), \ldots, f(n)$, on définit de façon récursive $f(k), k = 1, \ldots, n$, en choisissant un $f(k) \in T_k$ qui satisfait à

$$\pi_{f(k)}/p_{f(k)}^{(k)} = \max\{\pi_i/p_i^{(k)} : i \in T_k\},$$

où

$$T_1 = \{1, \ldots, n\}, \quad T_k = T_{k-1} \sim \{f(k-1)\},$$
$$k = 2, \ldots, n, \quad p_i^{(k)} = P(i \in I \text{ et } I \subset T_k),$$
$$k = 1, \ldots, n, \quad i \in T_k. \qquad (3.7)$$

Puisque $p_i^{(1)} = p_i$, l'ordonnancement que nous venons de définir équivaut à placer d'abord une UPÉ dont la valeur de π_i/p_i^*, est la plus élevée. Pour tous les k, $p_{f(k)}^{(k)}$ est la probabilité que $f(k)$ faisait partie de I et que aucun des éléments $k-1$ précédant $f(k)$ dans l'ordonnancement f ne faisaient partie de I. Ainsi, $p_{f(k)}^{(k)}$ est la probabilité qu'une tentative soit faite de retenir $A_{f(k)}$ dans le nouvel échantillon soit comme premier membre d'une paire ordonnée d'UPÉ de l'échantillon initial, soit comme UPÉ unique de l'échantillon initial dans S. En règle générale, plus $\pi_{f(k)}/p_{f(k)}^{(k)}$ est grand et plus cette tentative risque de porter fruit. Ainsi, la motivation de l'ordonnancement f des UPÉ individuelles est l'analogue de la motivation de l'ordonnancement des paires d'UPÉ dont nous avons déjà parlé.

Il nous reste à expliquer comment calculer $p_i^{(k)}$ pour $k \geq 2$. À cette fin, désignons par r le nombre de strates initiales qui ont des UPÉ en commun avec S, et par F_α, $\alpha = 1, \ldots, r$, une partition de $\{1, \ldots, n\}$ telle que i et j font partie du même F_α si et seulement si A_i et A_j faisaient partie de la même strate initiale. Soit

$$p'_\alpha(T) = P(I \cap F_\alpha \subset T), \quad \alpha = 1, \ldots, r,$$
$$T \subset \{1, \ldots, n\}, \qquad (3.8)$$

$$p''_{i\alpha}(T) = P(i \in I \text{ et } I \cap F_\alpha \subset T), \quad \alpha = 1, \ldots, r,$$
$$T \subset \{1, \ldots, n\}, \quad i \in F_\alpha \cap T. \qquad (3.9)$$

On observe que

$$p'_\alpha(T) = 1 - \sum_{i \in F_\alpha \sim T} p_i + \sum_{\substack{i,j \in F_\alpha \sim T \\ i<j}} p_{ij}, \qquad (3.10)$$

$$p''_{i\alpha}(T) = p_i - \sum_{j \in F_\alpha \sim T} p_{ij}, \qquad (3.11)$$

et finalement, tel que l'ont établi Ernst et Ikeda (1994),

$$p_i^{(k)} = p''_{i\alpha}(T_k) \prod_{\substack{\ell=1 \\ \ell \neq \alpha}}^{r} p'_\ell(T_k), \quad k = 1, \ldots, n,$$
$$i \in F_\alpha \cap T_k. \qquad (3.12)$$

Ensuite, pour chaque $k = 1, \ldots, n-1$, l'ordonnancement $g_k(\ell)$, $\ell = 1, \ldots, n-k$, est défini de façon récursive par le choix de $g_k(\ell) \in T_{k\ell}$ qui satisfait à

$$\pi_{f(k),g_k(\ell)}/p_{f(k),g_k(\ell)}^{(\ell)} = \max\{\pi_{f(k),j}/p_{f(k),j}^{(\ell)} : j \in T_{k\ell}\},$$

où

$$T_{k1} = \{1, \ldots, n\} \sim \{f(1), \ldots, f(k)\},$$
$$T_{k\ell} = T_{k(\ell-1)} \sim \{g_k(\ell-1)\}, \quad \ell = 2, \ldots, n-k,$$
$$T_{k\ell}^* = T_{k\ell} \cup \{f(k)\}, \quad \ell = 1, \ldots, n-k,$$
$$p_{f(k),j}^{(\ell)} = P(f(k), \quad j \in I \text{ et } I \subset T_{k\ell}^*),$$
$$\ell = 1, \ldots, n-k, \quad j \in T_{k\ell}. \qquad (3.13)$$

Noter que $p_{f(k),j}^{(\ell)}$ désigne ainsi la probabilité conjointe que $f(k)$ soit le premier nombre entier de l'ordonnancement f dans I, qu'aucun des premiers nombres entiers $\ell - 1$ dans l'ordonnancement g_k ne soit dans I et que $j \in I$. En conséquence, $p_{f(k),g_k(\ell)}^{(\ell)}$ désigne la probabilité que $I^* = \{f(k), g_k(\ell)\}$. Par ailleurs, si $I_i = \{f(k), g_k(\ell)\}$, alors $p_i^* = p_{f(k),g_k(\ell)}^{(\ell)}$, et le choix de $g_k(\ell)$ donne donc la valeur de $\pi_{f(k),g_k(\ell)}/p_i^*$ la plus grande parmi les éléments de $T_{k\ell}$, conformément à l'objectif déjà mentionné de l'ordonnancement des paires d'UPÉ.

Pour calculer $p_{f(k),j}^{(\ell)}$, Ernst et Ikeda (1994) établissent que si $f(k) \in F_\alpha, j \in F_\beta$ alors

$$p_{f(k),j}^{(\ell)} = p_{f(k),j} \prod_{\substack{t=1 \\ t \neq \alpha}}^{r} p'_t(T_{k\ell}^*) \text{ si } \alpha = \beta,$$
$$= p''_{f(k),\alpha}(T_{k\ell}^*) p''_{j\beta}(T_{k\ell}^*) \prod_{\substack{t=1 \\ t \neq \alpha,\beta}}^{r} p'_t(T_{k\ell}^*) \text{ si } \alpha \neq \beta. \qquad (3.14)$$

Nous présentons ci-après les calculs utilisés pour obtenir l'ordonnancement de l'exemple sur lequel nous nous

somme penchés. Notons d'abord que $f(1) = 2$ puisque la valeur de π_i/p_i la plus grande s'obtient quand $i = 2$. Trouvons ensuite $g_1(1)$ qui, puisque $f(1) = 2$, est le $j \in \{1,3\}$ qui donne la valeur maximale de $\pi_{2j}/p_{2j}^{\{1\}}$. Pour déterminer ce j, posons d'abord $F_\alpha = \{\alpha\}$, $\alpha = 1,2,3$, et notons que $T_{11}^* = \{1,2,3\}$. Selon (3.14) avec $\alpha = 2$, $\beta = 1$, il s'ensuit que

$$p_{21}^{\{1\}} = p_{22}''\{1,2,3\}p_{11}''\{1,2,3\}p_3'\{1,2,3\} = p_2p_1 \cdot 1 = .45,$$

et on peut de la même façon obtenir que $p_{23}^{\{1\}} = .525$. Ainsi, $g_1(1) = 3$, puisque $.5/.525 > .3/.45$. Par conséquent, la première paire de l'ordonnancement est $\{f(1), g_1(1)\} = \{2,3\}$. Ensuite, $g_1(2) = 1$, puisque 1 est le seul nombre entier qu'il nous reste à utiliser dans l'ordonnancement g_1, et la seconde paire est de ce fait $\{f(1), g_1(2)\} = \{2,1\}$. Il n'est pas vraiment nécessaire de déterminer $f(2)$ puisque $\{1,3\}$ est la seule et donc la dernière paire qui reste, mais à des fins d'illustration, on peut observer que $T_2 = \{1,3\}$, $p_1^{\{2\}} = p_{11}''\{1,3\}p_2'\{1,3\}$ $p_3'\{1,3\} = p_1(1 - p_2) \cdot 1 = .15$ selon (3.12), et que de la même façon $p_3^{\{2\}} = p_3(1 - p_2) \cdot 1 = .175$. Donc, $f(2) = 3$ puisque $.7/.175 > .5/.15$. En conséquence, $g_2(1) = 1, f(3) = 1$.

3.1.3 Calcul de p_i^* et de c_{ij}

Nous expliquons ensuite le calcul des valeurs de p_i^*. Si I_i désigne la paire de nombre entiers $I_i = \{f(k), g_k(\ell)\}$ alors, comme noté antérieurement, $p_i^* = p_{f(k),g_k(\ell)}^{(\ell)}$. En conséquence, p_i^* peut être calculé à l'aide de la formule (3.14) avec $j = g_k(\ell)$.

Si I_i est un ensemble à un seul élément $\{t\}$ pour un certain $t \in F_\alpha$, alors, comme l'ont établi Ernst et Ikeda (1994),

$$p_i^* = p_{t\alpha}''(\{t\}) \prod_{\substack{u=1 \\ u \neq \alpha}}^{r} p_u'(\varnothing). \tag{3.15}$$

Finalement, si $I_i = \varnothing$, alors

$$p_i^* = \prod_{u=1}^{r} p_u'(\varnothing).$$

Il reste simplement à expliquer le calcul des c_{ij} qui, selon (3.5) et (3.6), revient à calculer b_{it}, $i = 1, \ldots, \binom{n}{2} + n + 1$, $t = 1, \ldots, n$.

Pour le calcul de b_{it}, on observe que

$$b_{it} = 0 \quad \text{si} \quad I_i = \varnothing,$$

$$= 1 \quad \text{si} \quad I_i = \{v\} \quad \text{et} \quad t = v,$$

$$= 0 \quad \text{si} \quad I_i = \{v\} \quad \text{et} \quad t \neq v,$$

tandis que si $I_i = \{f(k), g_k(\ell)\}$ et $f(k) \in F_\alpha, g_k(\ell) \in F_\beta$, $t \in F_\gamma$, alors

$$b_{it} = 1 \quad \text{si} \quad t = f(k) \quad \text{ou} \quad t = g_k(\ell), \tag{3.16}$$

$$= 0 \quad \text{si} \quad t \notin T_{k\ell}^*, \tag{3.17}$$

$$= 0 \quad \text{si} \quad t \in T_{k\ell} \sim \{g_k(\ell)\} \quad \text{et} \quad \gamma = \alpha = \beta, \tag{3.18}$$

$$= \frac{p_{f(k),t}}{p_{f(k),\alpha}''(T_{k\ell}^*)} \quad \text{si} \quad t \in T_{k\ell} \sim \{g_k(\ell)\} \quad \text{et} \quad \gamma = \alpha \neq \beta, \tag{3.19}$$

$$= \frac{p_{g_k(\ell),t}}{p_{g_k(\ell),\beta}''(T_{k\ell}^*)} \quad \text{si} \quad t \in T_{k\ell} \sim \{g_k(\ell)\} \quad \text{et} \quad \gamma = \beta \neq \alpha, \tag{3.20}$$

$$= \frac{p_{t\gamma}''(T_{k\ell}^*)}{p_\gamma'(T_{k\ell}^*)} \quad \text{si} \quad t \in T_{k\ell} \sim \{g_k(\ell)\} \quad \text{et} \quad \gamma \neq \alpha, \gamma \neq \beta. \tag{3.21}$$

Ernst et Ikeda (1994) démontrent comment les formules (3.16) à (3.21) ont été obtenues.

Lors de la mise en oeuvre du SIPP en conditions réelles, il a fallu modifier la méthode à taille réduite pour réaliser le chevauchement du plan d'échantillonnage du SIPP des années 1990 et de celui des années 1980. Ces modifications se sont avérées nécessaires par suite de changements intervenus dans les définitions des UPÉ d'une décennie à l'autre. En effet, certaines UPÉ du plan des années 1990 pouvaient recouper plus d'une UPÉ du plan des années 1980. Ces modifications sont décrites en détails dans l'article de Ernst et Ikeda (1994).

3.2 Modifications de la méthode à taille réduite pour d'autres plans

En règle générale, on considère n'importe quel plan d'échantillonnage initial sans remise de la forme m'-UPÉ par strate et n'importe quel plan d'échantillonnage final sans remise de la forme m-UPÉ par strate, où m', m sont des nombres entiers positifs quelconques. Même si la méthode à taille réduite de la section 3.1 n'a été présentée que pour le cas où $m = m' = 2$, elle peut en fait s'appliquer pour n'importe quelles valeurs de m, m'. Nous présentons brièvement ci-après les modifications nécessaires lorsque $m \neq 2$ ou $m' \neq 2$.

L'utilisation d'une valeur différente de m' ne demande le changement que d'une partie seulement des calculs. Par exemple, si $m = 2$ mais $m' \neq 2$, les calculs de $p_i^{(k)}$, $p_{f(k),j}^{(\ell)}$ et c_{ij} seront différents, mais leurs définitions resteront les mêmes.

Si $m = 3$, peu importe la valeur de m', on procède à l'ordonnancement de triplets de nombres entiers au lieu de paires dans $\{1, \ldots, n\}$. Si I consiste en au moins trois nombres entiers, les nouvelles probabilités de sélection sont conditionnées seulement sur le premier triplet de la

liste ordonnée de I. Autrement, les nouvelles probabilités de sélection sont conditionnées sur I lui-même. Ainsi, les nouvelles probabilités de sélection sont conditionnées sur $\binom{n}{3} + \binom{n}{2} + n + 1$ événements.

Pour obtenir l'ordonnancement voulu des triplets de nombres entiers, on établit d'abord l'ordonnancement des $f(1), \ldots, f(n)$ et des $g_k(1), \ldots, g_k(n-k)$ comme dans le cas où $m = 2$. Ensuite, pour chaque $k = 1, \ldots, n-2$, $\ell = 1, \ldots, n-k-1$, un ordonnancement $h_{k\ell}(1), \ldots, h_{k\ell}(n-k-\ell)$ de $\{1, \ldots, n\} \sim \{f(1), \ldots, f(k), g_k(1), \ldots, g_k(\ell)\}$ est établi d'une manière semblable à celle utilisée pour $g_k(1), \ldots, g_k(n-k)$. Par exemple, en définissant $h_{k\ell}(v)$ pour $v \geq 2$, $p^{(\ell)}_{f(k),j}$ de la définition de $g_k(\ell)$ est remplacé par

$$P\big(f(k), g_k(\ell), j \in I \quad \text{et} \quad I \subset (T^*_{k\ell} \cup g_k(\ell)) \sim \{h_{k\ell}(1), \ldots, h_{k\ell}(v-1)\}\big).$$

Un ordonnancement linéaire des triplets distincts de $\{1, \ldots, n\}$ est alors déterminé en représentant chaque triplet par un triplet ordonné unique ayant la forme $(f(k), g_k(\ell), h_{k\ell}(v))$. Un deuxième triplet $(f(k'), g_{k'}(\ell'), h_{k'\ell'}(v'))$ précède le premier si et seulement si $k' < k$, ou $k' = k$ et $\ell' < \ell$, ou $k' = k$ et $\ell' = \ell$ et $v' < v$.

Pour les valeurs de $m \geq 4$, des ordonnancements de m-uplets seraient définis selon une procédure semblable, et les nouvelles probabilités de sélections seraient conditionnées sur $\binom{n}{m} + \binom{n}{m-1} \ldots + n + 1$ événements.

Pour $m = 1$, les nouvelles probabilités de sélection sont conditionnées sur le premier membre de l'ordonnancement $f(1), \ldots, f(n)$ dans I si $I \neq \emptyset$, ou sur $\emptyset$ si $I = \emptyset$.

À noter que si $m > m'$, il est possible qu'au moins une partie des m-uplets ordonnancés ne soient pas des sous-ensembles de I, auquel cas ces sous-ensembles seraient exclus de l'ordonnancement et de l'ensemble d'événements sur lesquels les nouvelles probabilités de sélection sont conditionnées. Si aucun m-uplet ne peut être un sous-ensemble de I, alors les nouvelles probabilités de sélection sont conditionnées sur I lui-même.

Il n'est pas nécessaire de limiter les événements initiaux utilisés dans le problème de transport uniquement aux sous-ensembles de I de taille m ou moindre. Par exemple, si $m = 2$ et si $\binom{n}{3} + \binom{n}{2} + n + 1$ est suffisamment petit, on peut alors utiliser une méthode conditionnée sur des sous-ensembles de trois ou moins, ce qui donne un chevauchement attendu généralement plus grand. À l'inverse, si $\binom{n}{m} + \binom{n}{m-1} \ldots + n + 1$ est trop grand, les nouvelles probabilités de sélection peuvent être conditionnées sur des sous-ensembles de I de taille m'' ou moins, où $m'' < m$, en donnant par contre un chevauchement attendu généralement plus petit.

3.3 Rapports entre les chevauchements attendus de la méthode à taille réduite, de la méthode optimale et de la sélection indépendante

Désignons par Ω_I, Ω_R, Ω_O le chevauchement attendu avec la sélection indépendante, la méthode à taille réduite et la méthode optimale respectivement. Ernst et Ikeda (1994) examinent les rapports qui existent entre ces paramètres. Nous résumons brièvement ci-après les résultats de leur étude.

Il est admis au départ que $\Omega_I \leq \Omega_R \leq \Omega_O$ pour tout m, m', où m, m' sont tels que décrit à la section 3.2. En outre, pour le cas qui nous intéresse, $m = m' = 2$, les limites inférieures sont établies sur Ω_R, et les limites supérieures sont établies sur Ω_O et $\Omega_O - \Omega_R$.

Par exemple, désignons par μ_2 la probabilité qu'il existe au moins deux éléments dans I, et désignons par μ_1 la probabilité que I soit un ensemble à un seul élément. Soit

$$\lambda = \min\big\{\min\{\pi_i/p_i : i = 1, \ldots, n\}, \min\{\pi_{ij}/p_{ij} : i, j = 1, \ldots, n, i \neq j\}, 1\big\}.$$

Alors $\Omega_O \leq 2\mu_2 + \mu_1$, $\Omega_R \geq \lambda(2\mu_2 + \mu_1/2)$, et $\Omega_O - \Omega_R \leq 2(1-\lambda)\mu_2 + (1 - \lambda/2)\mu_1$.

Malheureusement, ces limites ne sont pas toujours très strictes. Toutefois, dans certaines circonstances, elles peuvent être utiles. Par exemple, si $\pi_{ij} \geq p_{ij}$ pour tous les i,j et s'il existe une probabilité 1 que I contienne au moins deux éléments, il s'ensuit, compte tenu de ces limites, que $\Omega_R = \Omega_O = 2$.

Finalement nous présentons un exemple du pire cas envisageable pour Ω_R en rapport avec Ω_O lorsque m, $m' = 2$. On observe que Ω_O peut être égal à 2, tandis que Ω_R est arbitrairement près de 0. Ainsi, du moins en théorie, la méthode à taille réduite peut s'avérer inefficace. En pratique toutefois, comme nous le démontrerons dans la prochaine section, Ω_R est beaucoup plus près de Ω_O que de Ω_I, au moins en ce qui concerne le SIPP.

4. APPLICATION DE LA MÉTHODE À TAILLE RÉDUITE AU SIPP

Nous présentons ci-après les résultats des simulations du chevauchement du SIPP réalisées avant l'enquête, aux fins de la recherche et des essais, ainsi que les résultats du chevauchement du SIPP obtenus en conditions réelles. Pour en savoir plus sur cette question, consulter Ernst et Ikeda (1992b, 1994).

Pour la mise en oeuvre de la méthode de chevauchement à taille réduite, nous avons utilisé un logiciel d'optimisation du flux de coût minimal (minimum cost flow ou MCF) créé par Darwin Kingman et John Mote de la University of Texas, à Austin, qui nous a permis de résoudre le problème de transport requis. Un programme en FORTRAN a été créé pour préparer les données d'entrée et pour traiter les données de sortie du logiciel MCF.

Pour tester le logiciel avant la tenue de l'enquête, nous avons utilisé le programme pour effectuer le chevauchement de deux stratifications du SIPP de la région du Midwest fondées sur les données du recensement de 1970 avec la stratification du plan véritable utilisé dans la même région pour le SIPP au cours des années 1980. (À l'époque de la réalisation de ce test, les données du recensement de 1990 n'étaient pas encore disponibles.) Les stratifications de 1970 ont été produites en stratifiant les UPÉ sélectionnées

avec probabilité des SIPP réalisés dans la région du Midwest au cours des années 1980, en utilisant les données de 1970. Les deux stratification de 1970 ont donné une répartition des UPÉ sélectionnées avec probabilité en 31 strates, en utilisant différents groupes de variables de stratification. Les stratifications fondées sur les données de 1980 et de 1970 ont été assimilées à des stratifications "initiales" et "finales" aux fins de l'algorithme de chevauchement.

Au cours de l'enquête véritable, tel qu'il est mentionné dans la section 3.1 et comme l'expliquent en détails Ernst et Ikeda (1994), on a utilisé une modification de la méthode à taille réduite pour le chevauchement du plan du SIPP des années 1990 et de celui des années 1980 à cause d'une différence dans la définition des UPÉ entre les deux décennies. La méthode à taille réduite modifiée a été utilisée pour le chevauchement de 103 strates non autoreprésentatives finales (plan des années 1990) du SIPP.

Le chevauchement attendu a été calculé pour l'algorithme de chevauchement maximum à taille réduite, pour la sélection indépendante des UPÉ finales et pour une limite supérieure du chevauchement attendu avec la méthode optimale. On a calculé une limite supérieure au lieu du chevauchement optimal réel parce qu'il est impossible de calculer un chevauchement optimal pour la strate la plus grande. Aux fins de la simulation, la limite supérieure utilisée à été celle mentionnée dans la section 3.3, $\mu_2 + 2\mu_1$, tandis qu'aux fins de l'enquête SIPP, il a fallu utiliser une limite supérieure différente, décrite par Ernst et Ikeda (1994), à cause de la différence des définitions des UPÉ entre les années 1980 et 1990.

Les résultats des deux stratifications finales de la simulation étaient généralement semblables l'un à l'autre. La combinaison des résultats des deux a donné un chevauchement attendu moyen pour cet ensemble de 62 strates de 1.552, 1.569 et .480 UPÉ/strate pour la méthode à taille réduite, la limite supérieure du chevauchement optimal et la sélection indépendante respectivement. Pour la mise en oeuvre du SIPP en conditions réelles, les données correspondantes ont été de 1.523, 1.647 et .582 respectivement, alors que les chevauchements attendus correspondant d'UPÉ pour les 103 strates étaient 156.9, 169.6 et 59.9 respectivement. Ainsi, tant dans les simulations que dans l'enquête SIPP véritable, la méthode à taille réduite a donné des résultats raisonnablement proches de la limite supérieure de la méthode optimale.

Tableau 6

Temps machine pour la méthode à taille réduite

Nombre d'UPÉ	Temps machine (hh:mm:ss)
18	0:36
37	5:44
49	24:05
68	2:23:43

Le temps d'exécution de l'algorithme à taille réduite a été relativement court pour la plupart des strates. Nous présentons ci-après les temps machine de la simulation pour la strate finale avec différents nombres d'UPÉ. Nous avons utilisé un ordinateur Solbourne 5/605. Le nombre médian d'UPÉ dans une strate, pour le groupe entier de 62 strates, était de 17. La strate la plus grande contenait 68 UPÉ.

Nous avons également déterminé, pour l'enquête SIPP en conditions réelles, que 41 des 103 strates finales du chevauchement de la méthode modifiée à taille réduite auraient été incompatibles avec la méthode optimale. En effet, selon nos estimations, la taille maximale du problème de transport, en termes de nombre de variables, pour l'exécution du programme, était fixée à 4×10^6. Le nombre de variables pour la méthode optimale était inférieur à 4×10^6 pour toutes les strates (56) à $n \leq 14$, mais dépassait la limite pour 41 des 47 strates à $n \geq 15$, y compris deux à $n = 15$. La taille maximale du problème de transport avec la méthode optimale pour les 103 strates a été atteinte avec une strate à $n = 46$, pour laquelle il y avait 3.61×10^{12} variables. Par contre, il y avait 1.03×10^6 variables pour la même strate avec la méthode à taille réduite modifiée.

L'efficacité relative du chevauchement de la méthode à taille réduite comparativement à la méthode de Ernst (1986) présente également un intérêt. On pense en général que la méthode à taille réduite devrait produire un chevauchement plus grand dans les cas où les deux méthodes sont utilisables, puisque la méthode à taille réduite tire profit de l'indépendance de strate à strate du plan initial. Toutefois, même si la méthode de Ernst (1986) s'applique aux plans à deux UPÉ par strate, aucun logiciel n'a jamais été créé au Census Bureau (ni ailleurs à notre connaissance) pour en permettre l'utilisation avec ce type de plan puisqu'il n'y a pas encore eu d'application en conditions réelles pour un tel programme. En conséquence, il n'est pas possible de comparer directement ces deux méthodes avec les mêmes données. On peut toutefois faire une comparaison sommaire à partir des résultats de la méthode de chevauchement à taille réduite pour l'enquête SIPP et des résultats du chevauchement obtenus avec la méthode de Ernst (1986) pour le chevauchement des plans du CPS et du NCVS pour les années 1990 avec leurs contreparties respectives des années 1980. (Les plans des enquêtes CPS et NCVS des années 1980 et 1990 comptent tous une UPÉ par strate.)

Dans le cas du CPS, la méthode de chevauchement a donné une augmentation moyenne du chevauchement attendu de .26 UPÉ par strate comparativement au résultat de la sélection indépendante; dans le cas du NCVS, elle a donné une augmentation moyenne du chevauchement attendu de .30 UPÉ par strate. À titre comparatif, on a obtenu une augmentation de .94 UPÉ par strate avec la méthode à taille réduite par rapport à la sélection indépendante pour le SIPP. Si les deux méthodes de chevauchement sont également efficaces, on pourrait alors s'attendre à une augmentation du chevauchement par strate à peu près deux fois plus importante pour le SIPP que pour le

CPS ou le NCVS puisque le SIPP utilise un plan à deux UPÉ par strate. Vue sous cet angle, la méthode à taille réduite fonctionne mieux que celle proposée par Ernst (1986). Toutefois, comme les stratifications ont été passablement différentes dans ces trois enquêtes, il est permis de douter de la validité d'une telle comparaison.

Pour l'exemple considéré dans les sections 2 et 3, il est possible de procéder à une comparaison valide des différentes méthodes de chevauchement puisque la valeur du chevauchement attendu dans Ernst (1986), soit 1.625, a été facilement calculée à la main. Les valeurs correspondantes du chevauchement pour la méthode à taille réduite et pour la méthode optimale sont 1.725 et 1.735 respectivement.

CONCLUSIONS

La méthode de chevauchement à taille réduite présentée dans le présent article répond en pratique à ses deux objectifs principaux. Elle réduit suffisamment la taille du problème de transport, comme le démontrent d'une part la taille du problème de transport dans les formules (3.1) à (3.3), et d'autre part le fait qu'elle a effectivement été utilisée dans le remaniement d'une enquête importante. En outre, cette méthode réalise la réduction de taille tout en donnant un chevauchement presque optimal, à tout le moins dans le cas du SIPP. Elle ne peut être utilisée que lorsque les UPÉ du plan initial sont sélectionnées indépendamment d'une strate à l'autre, mais lorsque cette condition est respectée, nous croyons qu'il s'agit de la meilleure méthode de chevauchement pour les grandes strates.

REMERCIEMENTS

Nous remercions M. Todd Williams pour son aide précieuse à la programmation. Merci également aux critiques et aux rédacteurs pour leurs commentaires constructifs. Les opinions exprimées dans le présent article sont celles des auteurs et ne sont pas nécessairement partagées par le Bureau of Labor Statistics ni par le Census Bureau.

BIBLIOGRAPHIE

ARAGON, J., et PATHAK, P.K. (1990). An algorithm for optimal integration of two surveys. *Sankhyā: The Indian Journal of Statistics*, 52, 198-203.

ARTHANARI, T.S., et DODGE, Y. (1981). *Mathematical Programming in Statistics*. New York: John Wiley and Sons.

CAUSEY, B.D., COX, L.H., et ERNST, L.R. (1985). Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80, 903-909.

ERNST, L.R. (1986). Maximizing the overlap between surveys when information is incomplete. *European Journal of Operational Research*, 27, 192-200.

ERNST, L.R. (1989). Further Applications of Linear Programming to Sampling Problems. Bureau of the Census, Statistical Research Division, Research Report Series, No. RR-89/05.

ERNST, L.R., et IKEDA, M. (1992a). Modification of the Reduced-Size Transportation Problem for Maximizing Overlap When Primary Sampling Units Are Redefined in the New Design. Bureau of the Census, Statistical Research Division, Technical Note Series, No. TN-91/01.

ERNST, L.R., et IKEDA, M. (1992b). Summary of the Performance of the Maximum Overlap Algorithms for the 1990's Redesign of the Demographic Surveys. Bureau of the Census, Statistical Research Division, Technical Note Series, No. TN-92/01.

ERNST, L.R., et IKEDA, M. (1994). A Reduced-Size Transportation Algorithm for Maximizing the Overlap Between Surveys. Bureau of the Census, Statistical Research Division, Research Report Series, No. RR-93/02.

GLOVER, F., KARNEY, D., KLINGMAN, D., et NAPIER, A. (1974). A computation study on start procedures, basic change criteria and solution algorithms for transportation problems. *Management Sciences*, 20, 793-813.

KEYFITZ, N. (1951). Sampling with probabilities proportional to size: Adjustment for changes in probabilities. *Journal of the American Statistical Association*, 46, 105-109.

KISH, L., et SCOTT, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.

PATHAK, P.K., et FAHIMI, M. (1992). Optimal integration of surveys. In *Essays in Honor of D. Basu*. Éds. M. Ghosh, et P.K. Pathak. Hayward, California: Institute of Mathematical Statistics, 208-224.

PERKINS, W.M. (1970). 1970 CPS Redesign: Proposed Method for Deriving Sample PSU Selection Probabilities Within 1970 NSR Stata. Note de service à Joseph Waksberg, Bureau of the Census.

RAJ, D. (1968). *Sampling Theory*. New York: McGraw Hill.

Techniques d'enquête, décembre 1995
Vol. 21, n° 2, pp. 173–179
Statistique Canada

Incidence des lettres de préavis, enveloppes-réponse affranchies et cartes de rappel sur les taux de réponse par la poste lors du recensement

DON A. DILLMAN, JON R. CLARK et MICHAEL D. SINCLAIR[1]

RÉSUMÉ

Lors d'un recensement national pilote effectué en 1992, la séquence d'envoi d'une lettre de préavis, d'un formulaire de recensement, d'une carte de rappel et d'un questionnaire de remplacement a engendré un taux de réponse global de 63.4%. Ce taux de réponse était considérablement plus élevé que le taux de 49.2% obtenu lors du National Content Test Census de 1986, dans le cadre duquel on avait également utilisé un questionnaire de remplacement. L'écart a semblé principalement attribuable à la séquence "préavis - formulaire de recensement - rappel" des envois, mais on n'a pas su dans quelle mesure chacun des principaux effets et leurs interactions avaient influé sur les résultats globaux. Le présent article rend compte des résultats du test de mise en oeuvre du Recensement de 1992, qui vérifiait l'influence individuelle et combinée, sur le taux de réponse, de la lettre de préavis, de l'enveloppe-réponse affranchie ainsi que de la carte de rappel. Il s'agissait d'un échantillon national de ménages (n = 50,000) interrogés à l'automne 1992. On a utilisé un plan factoriel pour vérifier les huit combinaisons possibles des principaux effets, ainsi que leurs interactions. On a également eu recours à la régression logistique et à des comparaisons multiples pour analyser les résultats du test.

MOTS CLÉS: Enquête par la poste; taux de réponse; comparaisons multiples; régression logistique.

1. INTRODUCTION

Une diminution de 10 points du taux de réponse postale, lequel a glissé de 75% à 65% pour le recensement décennal américain de 1990, a conduit à des recherches visant à améliorer le taux de réponse dans ce genre d'enquête. Chaque point gagné peut faire économiser environ 16 millions de dollars en coûts de dénombrement à domicile (Miskura 1992). Une démarche antérieure nous a appris que des questions simples et quelque peu moins nombreuses que celles qui figuraient dans le questionnaire abrégé de 1990 s'étaient traduites par une amélioration de 8 points du taux de réponse postale (Dillman, Clark et Sinclair 1993). En effet, un formulaire de recensement expérimental comportant ces caractéristiques a été retourné par 71.4% des ménages, alors que 63.4% seulement des ménages d'un groupe de contrôle ayant reçu le formulaire abrégé du Recensement de 1990 l'avaient retourné. Les taux de réponse obtenus avec ces deux formulaires étaient toutefois beaucoup plus élevés que l'avaient été ceux enregistrés à l'occasion de tests similaires effectués des années où il n'y avait pas de recensement. Par exemple, dans le National Content Test de 1986, qui reposait sur un questionnaire équivalant à la version abrégée du formulaire de recensement de 1990, un taux de réponse de 49.2% a été obtenu. On a posé l'hypothèse qu'une partie du taux de réponse élevé qu'on avait constaté dans cette récente expérience était attribuable à une stratégie privilégiant des contacts multiples: lettre de préavis, carte de rappel et questionnaire de remplacement.

Notre article a pour but de rendre compte des résultats du test de mise en oeuvre de 1992, qui vise à déterminer l'influence relative et combinée, sur le taux de réponse, de la lettre de préavis et de la carte de rappel utilisés dans la démarche exposée auparavant (Dillman et coll. 1993). Le test cherchait également à déterminer l'effet produit par l'inclusion d'une enveloppe-réponse affranchie (plutôt qu'une enveloppe-réponse d'affaires) avec le formulaire de recensement envoyé par la poste.

Le recensement décennal américain de 1990 nous a obligés à enquêter auprès de plus de 100 millions de ménages. Les seuls coûts d'un tel recensement montrent à quel point il est important de déterminer la mesure dans laquelle ces trois techniques d'incitation à répondre pourraient contribuer à améliorer le taux de réponse des ménages. Si des recherches antérieures laissent deviner l'importance de chacune de ces techniques à cet égard, on dispose de peu de renseignements sur leurs éventuelles interactions. L'étude nous permet de déterminer dans quelle mesure ces trois techniques interagissent ou se complètent lorsqu'elles sont utilisées ensemble.

1.1 Travaux de recherche antérieurs

Maintes études ont confirmé que le principal déterminant du taux de réponse global aux enquêtes postales est le nombre de contacts (p. ex., Scott 1961; Heberlein et Baumgartner 1978). Les préavis et les rappels se sont tous deux révélés efficaces pour susciter une réponse (p. ex., Kanuke et Berenson 1975; Linsky 1975; Fox et coll. 1988).

[1] Don A. Dillman, Washington State University, Pullman, WA, U.S.A.; Jon R. Clark, U.S. Bureau of the Census, Washington DC 20233, U.S.A.; et Michael D. Sinclair, Response Analysis Corp., Princeton, NJ, U.S.A.

Cependant, les données de ces recherches sont fort peu éclairantes sur l'importance relative des préavis et des rappels comme mesures d'incitation.

En général, les résultats des travaux de recherche antérieurs portent à croire que l'insertion d'une enveloppe-réponse affranchie (par comparaison avec une enveloppe-réponse d'affaires) améliore le taux de réponse (Scott 1961; Kanuk et Berenson 1975; Duncan 1979; Harvey 1987; Fox et coll. 1988). Rappelons cependant l'exception digne de mention que constitue une analyse de régression qui porte sur des études de Heberlein et Baumgartner, et qui a montré que l'insertion d'enveloppes-réponse affranchies n'exerçait pas d'effet significatif (1979). Une revue de la documentation effectuée par Armstrong et Luske signale 20 études dans lesquelles des solutions de remplacement aux enveloppes-réponse d'affaires (1987) ont été mises à l'essai. Pour chacune des comparaisons, le taux de réponse absolu suscité par la solution de remplacement était beaucoup plus élevé que celui découlant de la technique habituelle dans 15 cas sur 20, l'augmentation moyenne étant de 9.2 points. On a signalé six études comparant l'utilisation d'enveloppes affranchies à la machine et d'enveloppes affranchies avec un timbre. En moyenne, le taux de réponse découlant de ces dernières était supérieur de 3.4 points à celui des enveloppes afranchies à la machine. Enfin, quatre études reposant sur une pléiade de mesures d'incitation ont montré que les enveloppes affranchies engendraient un taux de réponse supérieur de 2 à 4 points à celui obtenu avec des enveloppes-réponse d'affaires (Dillman 1978).

Les trois mesures d'incitation qui sont testées dans notre enquête font partie des huit principales techniques uniformément reconnues dans les rapports de recherche comme des déterminants de l'amélioration du taux de réponse postale. Parmi les autres facteurs, citons les stimulants financiers, les services postaux spéciaux, l'identité de l'organisme enquêteur, la personnalisation de la correspondance et l'intérêt du répondant (ou l'importance des questions) (Dillman 1991).

On a jugé que deux de ces huit facteurs, les stimulants financiers et les services postaux spéciaux (p. ex., courrier certifié ou courrier prioritaire avec délai de deux jours), n'étaient pas pratiques dans le cadre d'un recensement touchant plus de 100 millions de ménages. On a estimé qu'un troisième facteur, soit le parrainage du U.S. Bureau of the Census, était souhaitable sur le plan de la stimulation du taux de réponse. Un quatrième facteur, soit l'intérêt du répondant ou l'importance des questions, ne se prêtait pas à une intervention, les questions de l'enquête étant définies à l'avance dans des lois fédérales. Un cinquième facteur, la personnalisation de la correspondance, avait une valeur limitée du fait que les formulaires de recensement s'adressent strictement aux ménages et non à des individus. En examinant les effets individuels et combinés, sur les réponses, des préavis, des rappels et des enveloppes-réponse affranchies, nous espérions pouvoir déterminer si l'une ou plusieurs de ces composantes pouvaient se substituer à une autre composante, ce qui nous aurait permis d'améliorer le taux de réponse en réduisant le coût.

1.2 Conception et intégration des éléments d'intervention

La trousse postale renfermant le formulaire de recensement présentait certaines caractéristiques nous donnant à penser que les destinataires pourraient ne pas l'avoir repérée ou ne pas en avoir tenu compte. Par nécessité, cette trousse est seulement adressée à des ménages anonymes; on ne peut utiliser de nom. Or, pour assurer le traitement approprié des questionnaires retournés, il faut que l'adresse exacte du ménage figure sur le questionnaire proprement dit. Dans un recensement de grande envergure, cependant, l'adressage distinct d'une enveloppe extérieure, d'une lettre et d'un questionnaire, et la procédure visant à s'assurer que les bons objets sont insérés dans la bonne enveloppe, posent de sérieux problèmes en matière de contrôle de la qualité. C'est pourquoi il importe d'imprimer l'adresse sur un seul des objets devant être rassemblés en une trousse postale. Nous utilisons donc pour la livraison de la trousse postale une enveloppe à fenêtre laissant voir l'adresse imprimée sur le questionnaire.

L'impossibilité de recourir au nom des répondants ainsi que la taille et l'apparence de l'enveloppe à fenêtre donnent malheureusement l'impression que celle-ci ne contient rien d'important ou renferme peut-être même de la publicité-rebut. De plus, des recherches sur les cas de non-réponse au Recensement de 1990 ont révélé que certaines personnes ne se rappelaient pas avoir reçu un questionnaire de recensement par la poste, ou l'avaient vu, mais n'avaient pas ouvert l'enveloppe, deux phénomènes qui peuvent être attribuables à l'apparence de publipostage (Kulka et coll. 1991).

Dans ce test, la lettre de préavis et la carte de rappel avaient pour but d'attirer l'attention sur l'enveloppe renfermant le formulaire de recensement. L'objectif a été atteint de cinq manières. Premièrement, on a élaboré le préavis sous forme de lettre et le rappel sous forme de carte postale. On a pensé que les gens seraient plus susceptibles d'examiner deux objets de correspondance s'ils paraissaient différents l'un de l'autre. On a choisi la lettre en guise de préavis, préférant conserver la carte postale pour le rappel, en raison de la commodité de sa présentation.

Deuxièmement, le préavis consistait en une lettre du directeur du Census Bureau, avec la mention "Aux résidents du", suivie de l'adresse imprimée à l'emplacement habituel sur le papier à lettre. Nous voulions faire savoir aux destinataires que le questionnaire du recensement qui leur arriverait sous peu était spécialement destiné au ménage résidant à cette adresse. L'adresse était également visible par la fenêtre de l'enveloppe, ce qui réglait l'éventuel problème de contrôle de la qualité qu'aurait soulevé l'envoi par la poste du formulaire de recensement regroupant des objets de correspondance adressés séparément.

Troisièmement, le préavis était censé être livré quelques jours avant l'enveloppe contenant le formulaire de recensement proprement dit, et le rappel devait suivre à peine quelques jours plus tard. Les dates fixées pour ces envois étaient les 21, 24 et 29 septembre respectivement. On avait pensé que, pour être efficace, un rappel (sans questionnaire de remplacement) devait arriver dans les quelques jours suivant la réception du questionnaire, avant que l'on ait jeté à la poubelle le courrier non décacheté dans le cadre des travaux normaux d'entretien ménager.

Quatrièmement, le libellé du préavis, "D'ici quelques jours, vous devriez recevoir..." et du rappel, "Vous auriez dû recevoir ces jours derniers..." avait pour objectif d'encourager les destinataires à être à l'affût du formulaire de recensement. Cinquièmement, l'utilisation du papier à en-tête du directeur du Census Bureau et de la carte postale blanche affichant le sceau du Department of Commerce au-dessus du message de rappel avait pour but d'informer le destinataire que le questionnaire provenait du gouvernement et non d'un quelconque groupe tentant de se faire passer pour un organisme gouvernemental, comme cela se fait parfois au moyen d'une mention du genre "Nous vous informons officiellement que...".

L'influence favorable - à supposer qu'il y en ait eu une - de l'enveloppe-réponse affranchie sur le taux de réponse pourrait être attribuable au fait qu'on réussit à persuader le destinataire de la légitimité et de l'importance de la demande qui lui est faite. (Autrement, pourquoi l'expéditeur "gaspillerait-il" un timbre qu'on peut facilement retirer de l'enveloppe et utiliser ailleurs? Le destinataire hésiterait peut-être à jeter un objet qui a de la valeur (p. ex., un timbre non oblitéré)). Le préavis et, dans une certaine mesure, le rappel, peuvent accroître l'effet du timbre en incitant le destinataire à ouvrir l'enveloppe contenant le formulaire de recensement. De plus, le fait de s'apercevoir, une fois qu'on a ouvert l'enveloppe, qu'elle renferme un timbre non utilisé peut inciter à en conserver le contenu, ce qui renforce l'effet du rappel.

Pour que le préavis, l'enveloppe-réponse affranchie et le rappel s'appuient réciproquement, on a jugé important d'utiliser le courrier de première classe. Si l'on avait eu recours à des envois en nombre au tarif réduit et si ces envois avaient été étroitement espacés, il est probable que, dans certains cas, ces derniers ne seraient pas parvenus à destination dans le bon ordre.

En somme, notre test comportait plus que la simple juxtaposition de trois composantes distinctes dont il est fait mention dans la documentation. Ces composantes ont été opérationnalisées de façon à accroître la probabilité que chacune renforce l'effet des autres et à ce qu'elles s'appliquent aux envois postaux à grande échelle. En réalité, nous espérions apprendre si l'une ou plus d'une de ces composantes pouvaient être éliminées sans que le taux de réponse diminue exagérément, ce qui nous montrerait comment réduire le coût des envois postaux liés au recensement.

2. PLAN EXPÉRIMENTAL

Un plan factoriel, formé des huit combinaisons possibles des trois principaux effets, a été utilisé pour l'expérience. Les types d'intervention ont été les suivants:

1) Néant (groupe de contrôle),
2) lettre de préavis seulement,
3) enveloppe-réponse affranchie seulement,
4) carte de rappel seulement,
5) lettre et enveloppe-réponse affranchie,
6) enveloppe-réponse affranchie et rappel,
7) lettre et rappel, et
8) lettre, enveloppe-réponse affranchie et rappel.

2.1 Plan d'échantillonnage

L'univers d'échantillonnage correspondait à l'ensemble des unités de logement situées dans les régions visées par les questionnaires postaux et établies à partir du fichier d'adresses du Census Bureau. Les 449 régions correspondant aux bureaux de district (BD) du Recensement de 1990 ont été assimilées aux unités géographiques devant servir à définir les strates aux fins du test. Deux strates ont été définies. Compte tenu de la forte corrélation entre le taux relatif aux minorités (on entend par "minorités" toutes les classifications relatives aux populations de race noire et d'origine hispanique) et le taux de réponse par la poste au Recensement de 1990, on a pu réaliser la stratification en classant les BD selon le pourcentage de membres des minorités que l'on y trouvait. Les BD dont une forte proportion de la population se composait de minorités (noire ou d'origine hispanique) et dont le taux de réponse au Recensement de 1990 était faible ont été définis comme des "régions à taux de réponse faible" (RTRF) et formaient la première strate. Les BD restants ont été classés comme des "régions à taux de réponse élevé" (RTRE) et constituaient la seconde strate.

La première strate se composait de 67 BD, correspondait à une population dont environ 64% des membres appartenaient à une minorité, et englobait approximativement 11% des unités de logement situées dans les régions soumises au recensement par la poste. La seconde strate se composait de 382 BD et correspondait à une population dont environ 15% des membres appartenaient à une minorité. La strate des RTRE affichait un taux de réponse cumulatif au Recensement de 1990 supérieur de près de dix points à celui de la strate des RTRF.

Un échantillon de 50,000 unités de logement a été sélectionné, puis divisé en deux strates de 25,000 unités chacune. On a suréchantillonné la strate des RTRF afin d'étudier simultanément divers facteurs liés au sous-dénombrement différentiel, thème que nous n'abordons pas dans le présent article. Chaque strate a été divisée en huit panels de même taille afin de mettre à l'essai les huit types d'intervention. Un échantillon systématique de 3,125 unités de logement a été sélectionné dans chaque combinaison panel-strate. Une fois qu'une unité de logement était sélectionnée, les sept unités suivantes l'étaient également. Les ménages de chacun des huit groupes ont été assignés au hasard à un panel. Par conséquent, les huit voisins ont été soumis à un type d'intervention différent. L'échantillon a été regroupé, ce qui a permis de réduire la variance d'échantillonnage lors de la comparaison des panels.

La taille de l'échantillon sélectionné pour cette étude a été définie grâce à une vaste opération de simulation de données, laquelle a indiqué que l'échantillon de 50,000 unités était suffisant pour qu'on décèle un écart d'un minimum de 3% dans l'ensemble des comparaisons par paire des types d'intervention.

Tableau 1

Taux finals - Test de mise en oeuvre - Estimations à l'échelle nationale et des strates

Type d'intervention	Taux de réponse (%) - Estimations et erreurs-types (%)					
	Échelle nationale		Régions à taux de réponse élevé (RTRE) - 1990		Régions à taux de réponse faible (RTRF) - 1990	
	Estimation	Erreur-type	Estimation	Erreur-type	Estimation	Erreur-type
1. Groupe de contrôle	50.0	0.8	51.9	0.9	36.3	0.9
2. Lettre de préavis seulement	56.4	0.8	58.6	0.9	40.5	0.9
3. Enveloppe-réponse affranchie seulement	52.6	0.8	54.5	0.9	37.9	0.9
4. Carte de rappel seulement	58.0	0.8	60.2	0.9	42.0	0.9
5. Lettre et enveloppe-réponse affranchie	59.8	0.8	62.1	0.9	43.0	0.9
6. Enveloppe-réponse affranchie et carte de rappel	59.5	0.8	61.8	0.9	42.6	0.9
7. Lettre de préavis et carte de rappel	62.7	0.8	65.0	0.9	45.4	0.9
8. Lettre de préavis, enveloppe-réponse affranchie et carte de rappel	64.3	0.8	66.5	0.9	47.8	0.9

Tableau 2

Écarts dans les taux de réponse - Chaque composante en présence d'une autre composante

Comparaisons expérimentales	Écarts dans les taux de réponse (%) et intervalles de confiance (i.c.) de 90%					
	Échelle nationale		Régions à taux de réponse faible (RTRF) - 1990		Régions à taux de réponse élevé (RTRE) - 1990	
	Écart	I.C. de 90%	Écart	I.C. de 90%	Écart	I.C. de 90%
1. L - C	6.4	3.3 à 9.5*	4.2	0.9 à 7.5*	6.7	3.2 à 10.2*
2. S - C	2.5	−0.5 à 5.6	1.7	−1.7 à 5.0	2.7	−0.8 à 6.1
3. R - C	8.0	4.9 à 11.1*	5.7	2.4 à 9.1*	8.3	4.9 à 11.7*
4. LS - C	9.8	6.7 à 12.9*	6.8	3.4 à 10.1*	10.2	6.7 à 13.7*
5. SR - C	9.5	6.4 à 12.5*	6.4	3.0 à 9.7*	9.9	6.5 à 13.3*
6. LR - C	12.7	9.6 à 15.7*	9.2	5.8 à 12.5*	13.2	9.7 à 16.6*
7. LSR - C	14.2	11.2 à 17.2*	11.5	8.2 à 14.8*	14.6	11.3 à 18.0*
8. L - S	3.8	0.8 à 6.9*	2.5	−0.9 à 5.9	4.1	0.6 à 7.5*
9. R - L	1.6	−1.5 à 4.8	1.5	−1.9 à 5.0	1.6	−1.96 à 5.10
10. R - S	5.5	2.4 à 8.5*	4.1	0.7 à 7.5*	5.6	2.2 à 9.0*
11. LS - L	3.4	0.3 à 6.5*	2.6	−0.9 à 6.0	3.5	0.03 à 7.0*
12. SR - L	3.1	0.03 à 6.2*	2.2	−1.3 à 5.6	3.2	−0.3 à 6.6
13. LR - L	6.3	3.2 à 9.3*	5.0	1.5 à 8.4*	6.4	3.0 à 9.9*
14. LS - S	7.3	4.2 à 10.3*	5.1	1.7 à 8.5*	7.6	4.1 à 11.0*
15. SR - S	6.9	3.8 à 10.1*	4.7	1.2 à 8.2*	7.2	3.8 à 10.7*
16. LR - S	10.1	7.1 à 13.2*	7.5	4.1 à 11.0*	10.5	7.0 à 13.9*
17. LS - R	1.8	−1.3 à 4.9	1.1	−2.4 à 4.5	1.9	−1.6 à 5.4
18. SR - R	1.5	−1.6 à 4.5	0.7	−2.8 à 4.1	1.6	−1.8 à 5.0
19. LR - R	4.7	1.6 à 7.7*	3.5	−0.02 à 6.9	4.9	1.5 à 8.3*
20. LSR - L	7.9	4.8 à 10.9*	7.3	3.9 à 10.7*	7.9	4.5 à 11.4*
21. LSR - S	11.7	8.7 à 14.7*	9.8	6.4 à 13.3*	12.0	8.6 à 15.4*
22. LSR - R	6.2	3.2 à 9.3*	5.8	2.3 à 9.3*	6.3	2.9 à 9.7*
23. LSR - LS	4.4	1.4 à 7.5*	4.7	1.2 à 8.2*	4.4	1.0 à 7.8*
24. LSR - SR	4.8	1.7 à 7.8*	5.1	1.7 à 8.6*	4.7	1.3 à 8.2*
25. LSR - LR	1.6	−1.4 à 4.5	2.3	−1.1 à 5.8	1.5	−1.8 à 4.8
26. SR - LS	−0.3	−3.3 à 2.7	−0.4	−3.8 à 3.1	−0.3	−3.7 à 3.1
27. LR - LS	2.9	−0.2 à 6.0	2.4	−1.1 à 5.9	2.9	−0.6 à 6.4
28. LR - SR	3.2	0.2 à 6.2*	2.8	−0.6 à 6.2	3.3	−0.1 à 6.6

Un intervalle de confiance marqué d'un astérisque (*) indique que l'écart est statistiquement significatif à $\alpha = 0.10$ (9 chances sur 10 que l'i.c. comprenne l'écart actuel).

3. RÉSULTATS

Deux méthodes analytiques servent à présenter les principales constatations de cette étude. La première fait appel à des comparaisons multiples, par paire, des moyennes des types d'intervention; la deuxième recourt à la régression logistique. Les méthodes d'estimation sont précisées en annexe. Les deux méthodes donnent des résultats cohérents. Les taux de réponse globaux et les erreurs-types globales pour chacun des types d'intervention à l'échelle nationale et des strates sont présentés au tableau 1. Ils varient entre 50.0% pour le groupe de contrôle et 64.3% lorsque les trois principaux effets sont combinés.

3.1 Comparaisons multiples des taux de réponse par la poste

Le tableau 2 présente 28 comparaisons correspondant à tous les appariements possibles des huit types d'intervention. Compte tenu du peu d'espace qu'offre le tableau, nous avons utilisé les abréviations suivantes: C = contrôle, L = lettre de préavis, E = Enveloppe-réponse affranchie, R = Carte de rappel.

Les trois premières comparaisons figurant au tableau 2 indiquent une amélioration du taux de réponse attribuable à chacune des trois principales composantes s'ajoutant au groupe de contrôle, prises séparément. L'amélioration estimée du taux de réponse attribuable à la lettre de préavis s'établissait à 4.2% dans la strate des RTRF, à 6.7% dans la strate des RTRE, et à 6.4% à l'échelle nationale. L'amélioration estimée attribuable à la carte de rappel était de 5.7% dans la strate des RTRF, de 8.3% dans la strate des RTRE, et de 8.0% à l'échelle nationale. Toutes ces améliorations sont significatives. Par conséquent, la principale constatation de cette étude est que tant la lettre de préavis que la carte de rappel ont accru le taux de réponse à l'échelle nationale et des strates. On n'a enregistré aucune amélioration significative avec l'enveloppe-réponse affranchie, ni à l'échelle nationale ni dans les strates.

3.2 Analyse de régression logistique

Nous avons évalué un modèle comportant une composante pour les strates, la lettre de préavis, l'enveloppe-réponse affranchie et la carte de rappel, y compris tous les termes de l'interaction. Il y a également eu modélisation à l'échelle des strates, avec pour seuls paramètres les effets des composantes et leurs interactions.

Les résultats de l'analyse du modèle global indiquent que seuls les principaux effets de la lettre et de la carte de rappel, ainsi que l'élément stratification et les coordonnées à l'origine sont statistiquement significatifs. Compte tenu de ces résultats, une autre modélisation à l'échelle nationale a été effectuée au moyen d'un modèle réduit comportant seulement les principaux effets de la stratification, des composantes individuelles et des interactions entre ces dernières. Les résultats sont présentés au tableau 3 ci-après.

Tableau 3

Analyse de régression logistique des moindres carrés pondérés, modèle interaction réduit, sans strate composante

Paramètres du modèle	Estimations et intervalles de confiance (i.c.) de Bonferroni de 90%	
	Estimation	I.C. de 90%
Coordonnées à l'origine, β_0	−.61	−.686 à −.545*
Stratification, β_1	.738	.689 à .789*
Lettre de préavis, β_2	.227	.130 à .324*
Enveloppe-réponse affranchie, β_3	.090	−.006 à .186
Carte de rappel, β_4	.291	.194 à .387*
Lettre de préavis - enveloppe-réponse affranchie, β_5	.036	−.101 à .173
Lettre de préavis - carte de rappel, β_6	−.054	−.192 à .083
Carte de rappel - enveloppe-réponse affranchie, β_7	−.043	−.179 à .093
Lettre de préavis - carte de rappel - enveloppe affranchie, β_8	−.003	−.197 à .191

Un i.c. marqué d'un astérisque (*) indique que l'écart est statistiquement significatif à $\alpha = .10$.

Les deux modélisations montrent, tant pour le modèle national que pour le modèle des strates, une amélioration significative du taux de réponse résultant de la lettre et du rappel, mais non de l'enveloppe-réponse affranchie. Ces résultats correspondent à ceux qui ont été obtenus pour les comparaisons multiples présentées ci-dessus. Aucune interaction n'était statistiquement significative, signe que les effets des composantes sont essentiellement additifs.

4. DISCUSSION ET CONCLUSIONS

La lettre de préavis, l'enveloppe-réponse affranchie et la carte de rappel ont chacune amélioré le taux de réponse de 6.4, 2.5 et 8.0 points respectivement. L'accroissement de 2.5 points n'est pas statistiquement significatif. On a également établi que les effets des composantes étaient essentiellement additifs ou indépendants les uns des autres. Par rapport au groupe de contrôle, la combinaison lettre de préavis-enveloppe affranchie a amélioré le taux de réponse de 9.8%, tandis que la combinaison enveloppe affranchie-rappel a accru le taux de réponse de 9.5%, et la combinaison lettre-rappel l'a accru de 12.7%. Ensemble, les trois composantes ont amélioré le taux de réponse de 14.3%. Chaque utilisation de la lettre et du rappel a amélioré considérablement le taux de réponse; quant à l'enveloppe-réponse affranchie, elle n'a eu d'effet significatif que lorsqu'elle était utilisée avec une lettre de préavis, sans rappel. La conclusion la plus importante à tirer de cette expérience est que la lettre de préavis et la carte de rappel sont essentielles à l'atteinte d'un taux de réponse élevé et qu'aucune de ces composantes n'élimine l'effet de l'autre.

Bien que l'effet individuel (2.5% globalement) de l'enveloppe-réponse affranchie soit légèrement inférieur au taux nécessaire pour présenter une valeur significative, son ordre de grandeur concorde avec la valeur jugée significative dans des recherches antérieures (Armstrong et Luske 1987; Dillman 1978 et 1991). Face aux nombreux résultats de recherche qui ont démontré son utilité, cette technique ne devrait probablement pas être complètement écartée pour raison d'inefficacité. Il semble aussi que l'on puisse établir un lien d'une autre nature entre, d'une part, l'enveloppe affranchie et, d'autre part, la lettre de préavis et le rappel. Lorsque l'enveloppe affranchie est utilisée seule avec la lettre de préavis, son effet est significatif (3.4 points), mais à l'évidence il ne l'est pas (1.6 point) lorsqu'un rappel est inclus dans l'envoi. Le rappel compense l'absence d'enveloppe affranchie, alors que le préavis semble accroître l'effet de l'enveloppe lorsqu'elle y est. C'est peut-être qu'un préavis signale aux gens d'être à l'affût de la trousse postale sur le recensement et que, une fois celle-ci ouverte, la présence de l'enveloppe-réponse affranchie incite les gens à répondre. Ce lien différentiel avec les envois qui précèdent et ceux qui suivent ne semble pas avoir été examiné dans les recherches antérieures. Il en découle qu'en cas de réception d'une lettre de préavis sans rappel ultérieur, la présence d'une enveloppe-réponse affranchie pourrait augmenter le taux de réponse de manière statistiquement significative, mais revêtir une importance moindre lorsque l'expéditeur utilise aussi une carte de rappel, comme c'était le cas lors du dernier recensement.

Il existe au moins deux obstacles importants à l'application directe des résultats de notre recherche au recensement de l'an 2000. D'abord, il faut reconnaître que les tests actuels sont effectués hors d'une période de recensement. Par le passé, le Census Bureau a obtenu des taux de réponse beaucoup plus faibles dans les années de non-recensement que dans les années de recensement. Par exemple, le National Content Test de 1986 a obtenu seulement un taux de réponse de 49.2% avec un questionnaire de remplacement, tandis que le Recensement de 1990, sans questionnaire de remplacement, a obtenu un taux de réponse de 65%. Un écart aussi prononcé tiendrait à un climat dit "favorable au recensement", explication succincte de l'association de divers éléments, comme l'attention médiatique, la publicité et le sens culturel de la participation qui semblent faire surface chaque décennie au cours de l'année du recensement.

Les taux de réponse obtenus dans nos tests au moyen des cinq composantes qui augmentent le taux de réponse sont beaucoup plus élevés que ceux qu'on obtient généralement en période de non-recensement; ils sont néanmoins à peu près égaux ou peut-être mieux légèrement inférieurs aux résultats obtenus au cours du dernier recensement décennal, où aucune de ces composantes n'a été utilisée. Nous ignorons si l'existence d'un "climat favorable au recensement" peut remplacer les effets de ces composantes ou améliorer le taux de réponse susceptible d'être atteint l'année d'un recensement. Nous ne pensons certainement pas qu'une augmentation de 30 points du taux de réponse au Recensement de l'an 2000 soit possible, car cela signifierait qu'à peu près 100% du public-cible a répondu. Par conséquent, il reste passablement d'incertitude quant aux répercussions exactes des résultats actuels sur le Recensement de l'an 2000.

ANNEXE
Méthodes d'estimation

Les résultats de l'analyse découlent de deux méthodes distinctes: les comparaisons multiples réalisées entre les taux de réponse postale selon le type d'intervention, et la régression logistique. Chaque méthode a ses avantages pour ce qui est de l'interprétation des données dans un cas, et de l'inférence statistique dans l'autre; c'est pourquoi nous avons utilisé une approche combinée, recourant aux deux.

Dans notre étude, nous avons calculé le taux de réponse national estimé d'un panel donné en divisant le total pondéré du nombre de questionnaires retournés par le total pondéré des formulaires de recensement postés, moins le total pondéré des envois retournés par le maître de poste (généralement des unités vacantes).

Nous avons examiné les résultats des comparaisons multiples des taux de réponse relatifs aux huit types d'intervention afin de déterminer l'importance de l'augmentation du taux de réponse pour chacun d'eux. Ces comparaisons comportaient l'évaluation par paire de tous les types d'intervention, les uns avec les autres ainsi qu'avec le groupe de contrôle.

La méthode de régression logistique est un moyen rapide et efficace de déterminer si l'augmentation du taux de réponse attribuable à chaque composante (mais surtout à leur interaction) résulte d'une variation de l'échantillonnage ou est bien réelle, et si l'augmentation observée est influencée par la présence d'autres variables. Toutefois, les paramètres estimés ne peuvent être facilement assimilés aux taux de réponse postale. Pour plus de précisions, voir la méthode de régression logistique dans Thompson 1993.

Les taux de réponse ont été calculés pour chaque type d'intervention à l'intérieur de chacune des strates et à l'échelle nationale (strate 1 et strate 2 réunies). Les erreurs-types relatives aux estimations nationales ont été obtenues au moyen de la méthode jackknife du calcul de la variance pour un échantillonnage stratifié (Wolter 1985). Nos estimations ont été produites au moyen du progiciel statistique VPLX. Les erreurs-types pour les estimations relatives aux strates ont été obtenues avec la méthode jackknife de calcul de la variance pour un échantillonnage aléatoire simple.

L'analyse principale a fait appel à la comparaison par paire des écarts observés entre les taux de réponse obtenus pour huit types d'intervention, tant à l'échelle nationale qu'à l'échelle des strates (RTRF et RTRE).

En raison de la diversité des hypothèses testées, l'analyse porte sur la comparaison de toutes les combinaisons possibles (28 au total) entre les huit types d'intervention. Au moyen de la régression logistique, nous avons testé huit paramètres de modèle ou plus afin de déterminer si les résultats étaient statistiquement significatifs. Plus il y a de

comparaisons, plus il y a de chances que certaines d'entre elles soient déclarées à tort statistiquement significatives. Dans ce cas, nous employons d'autres mesures statistiques pour contrôler le taux global d'erreur du processus décisionnel.

Nous avons procédé à l'analyse de manière à ce que les constatations sur la "famille" complète des 28 combinaisons ou sur les paramètres de régression logistique maintiennent l'intervalle de confiance de 90% (une norme du Census Bureau) dans tous les cas. L'intervalle de confiance de 90% des comparaisons par paire a été ajusté au moyen de la méthode C de Dunnett permettant la comparaison par paire contrastante des estimations du panel (Hockberg et Tamhane 1987). La méthode d'inférence simultanée de Bonferroni a servi à évaluer la signification statistique des paramètres de régression logistique.

BIBLIOGRAPHIE

ARMSTRONG, J.S., et LUSKE, E.J. (1987). Return postage in mail surveys: A meta analysis. *Public Opinion Quarterly*, 51 (1) 233-248.

DILLMAN, D.A., CLARK, J., et SINCLAIR, M. (1993). Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. *Public Opinion Quarterly*.

DILLMAN, D.A., SINCLAIR, M., et CLARK, J. (1992). Mail-back response rates for simplified decennial census questionnaires. *Proceeding of the Section on Survey Research Methods, American Statistical Association*, 776-783.

DILLMAN, D.A. (1991). The design and administration of mail surveys. *Annual Review of Sociology*, 17, 225-249.

DILLMAN, D.A. (1978). *Mail and Telephone Surveys: The Total Design Method*. New York: Wiley-Interscience.

DUNCAN, W.J. (1979). Mail questionnaires in survey research: A review of response inducement techniques. *Journal of Management*, 5, 39-55.

FOX, R.J., CRASK, M.R., et KIM, J. (1988). Mail survey response rate: A meta-analysis of selected techniques for inducing response. *Public Opinion Quarterly*, 52, 467-491.

HARVEY, L. (1987). Factors affecting response rates to mailed questionnaires: A comprehensive literature review. *Journal of the Market Research Society*, 29, 3, 342-353.

HEBERLEIN, T., et BAUMGARTNER, R. (1978). Factors affecting response rates to mailed questionnaires: A quantitative analysis of the published literature. *American Sociological Review*, 43, 447-462.

HOCHBERG, Y., et TAMHANE, A.C. (1987). *Multiple Comparison Procedures*. New York: John Wiley and Sons.

KANUK, L., et BERENSON, C. (1975). Mail surveys and response rates: A literature review. *Journal of Marketing Research*, 12, 440-453.

KULKA, R.A., HOLT, N.A., CARTER, W., et DOWD, K.L. (1991). Self reports of time pressures, concerns for privacy and participation in the 1990 Mail Census. *Proceedings of the 1991 Annual Research Conference*. U.S. Bureau of the Census, 33-54.

LINSKY, A.S. (1975). Stimulating responses to mailed questionnaires: A review. *Public Opinion Quarterly*, 39, 82-101.

MISKURA, S.M. (1992). Estimating the Full Cycle Costs for the Simplified Questionnaire Test (SQT), 2KS Memorandum Series, Design 2000, Book I, Chapter 30, #6.

SCOTT, C. (1961). Research in mail surveys. *Journal of Royal Statistical Society*, 143-205.

THOMPSON, J.H. (1993). Final Results of the Mail Response Evaluation for the Implementation Test (IT), DSSD 2000 Census Memorandum Series, #E-32.

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

Techniques d'enquête, décembre 1995
Vol. 21, n° 2, pp. 181–192
Statistique Canada

Concordance des données censitaires et des registres de l'état civil concernant les aînés aux États-Unis: 1970-1990

LAURA B. SHRESTHA et SAMUEL H. PRESTON[1]

RÉSUMÉ

Les lacunes et les erreurs relevées dans les données censitaires et les registres de l'état civil aux États-Unis laissent planer un doute sérieux sur la qualité des données concernant les aînés et le recensement des décès dans ce pays. Dans le présent article, nous évaluons la concordance des données provenant de ces deux sources pour les populations blanche et afro-américaine. Nous étudions en particulier les aînés (60 ans et plus), où les tendances de la mortalité ont la plus grande incidence sur les programmes sociaux et où les problèmes de données sont les plus graves. Au moyen de l'analyse intercensitaire des cohortes, nous déterminons les discordances en fonction de l'âge entre les deux sources pour deux périodes: 1970-1980 et 1980-1990. Les anomalies relevées quant à la nature de la couverture et aux erreurs de contenu sont analysées à la lumière des données de la documentation spécialisée. Les données sur la population afro-américaine montrent de très nombreuses discordances pour la période 1970-1990, qui sont vraisemblablement attribuables à une exagération de l'âge déclaré lors des recensements, par rapport aux données des registres de l'état civil. On relève également des anomalies pour la population blanche dans la période intercensitaire 1970-1980. Selon nous, ces anomalies sont principalement dues à un sous-dénombrement dans le recensement de 1970, par rapport à celui de 1980 et au recensement des décès. Par contre, les données de 1980-1990 concernant les blancs, et en particulier les femmes, sont très cohérentes, beaucoup plus même que dans la plupart des pays européens.

MOTS CLÉS: Erreurs de déclaration au sujet de l'âge; couverture; mortalité; évaluation des recensements; recensement des décès; qualité des données; croisement des courbes du taux de mortalité, États-Unis.

1. INTRODUCTION

Les méthodes classiques d'estimation de l'ampleur de la mortalité dans les pays développés utilisent des données provenant de deux sources distinctes. Les données sur les décès tirées des registres de l'état civil servent habituellement de numérateurs des taux de mortalité. Les dénominateurs correspondent pour leur part habituellement au dénombrement des personnes vivantes réalisé lors des recensements. L'exactitude des taux ainsi calculés dépend donc de la qualité des données provenant de ces deux sources.

Nous présentons ci-après les résultats d'un test de la qualité des données américaines correspondant à deux périodes intercensitaires: 1970-1980 et 1980-1990. Nous étudions en particulier la concordance des changements relevés dans la taille d'une cohorte d'un recensement à l'autre et du nombre de décès recensés pendant cette période, en procédant aux ajustements voulus pour tenir compte des migrations. Toutes les données sont exprimées en années d'âge, et des tests séparés sont réalisés pour les populations blanche et noire.

Nous nous intéressons essentiellement aux aînés (60 ans et plus), où les tendances de la mortalité ont la plus grande incidence sur les programmes sociaux (Preston 1993) et où les problèmes de données sont les plus graves. La population blanche des États-Unis semble afficher des taux de mortalité plus bas, chez les plus de 80 ans, que dans tous les autres pays industrialisés (Vaupel 1993). Si elle s'avérait valide, cette comparaison pourrait avoir une grande incidence sur l'évaluation de la qualité relative des services de santé. Cependant, la population afro-américaine présente des taux de mortalité encore plus bas que ceux de la population blanche chez les plus de 80 ans, reflétant ainsi le phénomène bien connu du croisement des courbes du taux de mortalité des deux races qui s'observe entre 75 et 85 ans. La fiabilité de ces deux ensembles de données sur la mortalité dépend bien sûr de la qualité des données. Or, on a déjà émis des doutes très sérieux sur la qualité des données concernant la population de race noire (voir à ce propos Zelnik 1969; Coale et Kisker 1990), même si la plupart des observateurs semblent admettre d'emblée la validité du croisement des tendances (Manton et coll. 1986; McCord et Freeman 1990).

Dans le cadre de leurs travaux de détermination de nouveaux profils modèles du taux de mortalité dans les pays à faible mortalité, Condran, Himes et Preston (1991) ont fait état de tests semblables de la qualité des données pour 68 périodes intercensitaires dans 18 pays industrialisés. En général, la concordance des données s'est avérée très bonne pour les cohortes de 65 ans au deuxième recensement (66 des 68 ensembles de données ont passé le test de cohérence). La concordance a cependant diminué avec l'âge: la moitié seulement des ensembles de données montraient une concordance à 85 ans et la proportion des données concordantes passait à 15%, à 95 ans (Condran et coll. 1991: tableau 7). Les États-Unis ne faisaient pas partie des pays qui ont fait l'objet de ces tests parce que leurs données publiées sur la mortalité ne sont pas classées

[1] Laura B. Shrestha, The World Bank, Human Development Department, 1818 H Street, N.W., Washington, DC 20433, U.S.A.; Samuel H. Preston, Population Studies Center, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104, U.S.A.

par années d'âge. Nous sommes maintenant en mesure de combler cette grave lacune puisque nous avons traité les bandes de données pour chacun des décès enregistrés aux États-Unis de 1970 à 1988. (Les chiffres pour 1989 (année complète) et pour 1990 (janvier à mars seulement) ont été estimés à partir des données par groupes d'âge publiées par le National Center for Health Statistics et des données de l'année 1988. Voir l'annexe A pour plus de détails.) Ces bandes sont produites par le National Center for Health Statistics (NCHS) et sont diffusées par le Inter-University Consortium for Political and Social Research (ICPSR). Pour les années que nous avons examinées, elles portent sur un nombre approximatif de 50 millions de décès.

2. POPULATIONS ET DONNÉES

2.1 Contexte général

Trois sources principales de données ont été utilisées: 1) recensements nationaux du U.S. Bureau of the Census pour les années 1970, 1980 et 1990; 2) données annuelles sur les décès compilées par le NCHS; 3) estimations inédites sur l'immigration nette obtenues auprès du U.S. Bureau of the Census. La description détaillée de nos sources de données est jointe au présent article dont elle constitue l'annexe A. Il nous paraît cependant utile de décrire ci-après brièvement ces données et les ajustements importants que nous leur avons apportés.

2.2 Recensements

Nous utilisons des données censitaires classées en fonction de la race (noire/blanche) et du sexe, et par années d'âge (avec une classe ouverte pour les 100 ans et plus). Les compilations englobent l'ensemble des résidents des 50 états et du district de Columbia; elles comprennent les personnes placées en établissements, les Américains en déplacement temporaire à l'étranger et les étrangers dont la résidence habituelle (légale ou non) est située aux États-Unis (sauf les militaires et le personnel diplomatique). Sont par contre exclus les Américains qui résident à l'étranger pour une période prolongée et les ressortissants étrangers en visite temporaire aux États-Unis. Les statistiques officielles ne sont pas ajustées pour tenir compte du sous-dénombrement (p. ex., résidents légaux et immigrants sans document non recensés).

Le choix de l'expression "population résidente" signifie que les tableaux de recensement comprennent à la fois les résidents légaux et les immigrants sans document. Sur l'ensemble des personnes sans document qui résidaient aux États-Unis à l'époque du recensement de 1970, il semble qu'une proportion négligeable ait été dénombrée. Ainsi, la population des résidents légaux était approximativement comparable à la population résidente totale lors du recensement de 1970. Par contre, lors du recensement de 1980, le U.S. Bureau of the Census a estimé que, pour la toute première fois, un nombre significatif de personnes sans document avaient été dénombrées. Cette population a été évaluée à 2.06 millions de personnes. De ce nombre, dans le groupe d'âge des 60 ans et plus, on a compté 10,000 hommes et 19,000 femmes de race blanche, et 3,000 hommes et 6,000 femmes de race noire (U.S. Bureau of the Census 1988).

On reconnaît que les tableaux officiels du recensement de 1970 contiennent des erreurs parmi lesquelles la plus évidente est le surdénombrement marqué du nombre de personnes âgées de 100 ans ou plus. Même si le recensement a dénombré 106,000 personnes appartenant à ce groupe d'âge, les estimations démographiques indirectes laisseraient plutôt conclure à un nombre oscillant entre 3,000 et 8,000 personnes, le chiffre estimatif le plus plausible étant fixé à 4,800 (Siegel 1974; Siegel et Passel 1976; U.S. Bureau of the Census 1974). Nous utilisons les données inédites du U.S. Bureau of the Census pour le recensement de 1970, où cette surestimation des centenaires a été corrigée. Cette utilisation des estimations corrigées est motivée par deux raisons: premièrement, à défaut d'une correction, l'excédent serait suffisamment important pour biaiser les résultats pour les groupes plus âgés. Deuxièmement, il semble que cette surestimation n'ait pas été due à des déclarations erronées de l'âge par les intéressés, mais plutôt à un malentendu qui a porté certaines personnes à confondre les colonnes destinées à l'inscription du mois et de l'année de la naissance (Siegel et Passel 1976).

Au cours des recensements de 1980 et de 1990, un grand nombre de personnes ont choisi de fournir une réponse de leur cru à la question concernant la race, au lieu de choisir une des catégories préétablies. Sur la population totale, 6.8 millions de personnes, en majeure partie d'origine hispanique, ont choisi cette option en 1980, et 9.3 millions on fait de même en 1990. Les tableaux officiels de ces recensements ne sont donc pas directement comparables à d'autres sources de données puisqu'ils sont seuls à prévoir une catégorie raciale résiduelle. Pour autoriser une comparaison avec d'autres sources de données, le Census Bureau a modifié les données censitaires de 1980 et de 1990 pour qu'elles soient conformes aux catégories historiques de groupes raciaux. Le Bureau a en outre procédé, en 1990, à une "correction" du problème lié à l'âge (pour plus de détails, voir Word et Spencer 1991). Nous avons décidé d'utiliser les statistiques du Bureau de 1980 et de 1990 modifiées pour la race aux fins de notre étude. Ce choix a été motivé par le nombre considérable de personnes qui auraient été exclues autrement, en particulier dans la population blanche.

2.3 Registres des décès

Les registres américains des décès portent sur l'ensemble des décès recensés dans les 50 états et dans le district de Columbia, classés selon la race et le sexe et par années d'âge (jusqu'à 125+). Pour permettre une comparaison avec les données des recensements, les décès des non résidents (ressortissants étrangers et citoyens américains résidant à l'étranger) ont été exclus.

Aucune correction n'est faite pour tenir compte du sous-dénombrement des décès ou des décès dont la description sur le certificat comporte des erreurs. Nous avons relevé deux problèmes qui influent sur l'utilisation de notre

méthode d'analyse intercensitaire. La période intercensitaire s'étale du 1er avril au 31 mars, alors que les registres des décès utilisent l'année civile. En outre, les registres des décès et les recensements américains utilisent l'âge au dernier anniversaire au lieu de l'année de la naissance. Nous avons corrigé les deux problèmes en répartissant les décès sur des graphes triangulés du temps et de l'âge correspondant aux "années de recensement" débutant le 1er avril. Par exemple, les décès déclarés dans l'intervalle d'un an compris entre la date de recensement du 1er avril 1970 et celle du 1er avril 1971 et concernant des personnes âgées de 60 ans (à leur dernier anniversaire) au moment du recensement peuvent être répartis en quatre catégories: 1) personnes de 60 ans décédées au cours de l'année civile 1970; 2) personnes de 60 ans décédées au cours de l'année civile 1971; 3) personnes de 61 ans décédées au cours de l'année civile 1970; 4) personnes de 61 ans décédées au cours de l'année civile 1971. À l'aide des données sur les dates des décès tirées des bandes du NCHS, nous avons répartis les décès sur des triangles du temps et de l'âge correspondant à l'année de recensement débutant le 1er avril 1970. Ce faisant, nous avons présumé que les décès compris dans chaque triangle étaient répartis uniformément. Cette supposition est rendue nécessaire par l'absence de données fiables sur la naissance pour la plupart des cohortes visées par notre étude, données grâce auxquelles il nous aurait été possible de répartir les décès plus précisément entre les cohortes de naissance adjacentes. Pour une description plus détaillée de cette méthode, voir Shrestha (1993).

2.4 Statistiques sur l'immigration nette

Nous avons utilisé des statistiques inédites sur l'immigration nette obtenues auprès du U.S. Bureau of the Census. Si la qualité des statistiques américaines sur l'immigration est généralement mise en doute (Hill 1985), l'estimation de la taille de la population des aînés est passablement indépendante des variations dans les estimations de la migration intercensitaire. Cette robustesse découle à la fois du nombre moindre de migrants nets chez les personnes plus âgées, comparativement aux plus jeunes, et du rôle plus important de la mortalité dans la diminution du nombre de personnes âgées, comparativement à la migration nette. Par exemple, les données sur l'immigration nette font état d'un apport de 64 hommes de race noire pour la cohorte des personnes âgées de 75 ans et plus (en 1970) pendant la décennie de 1970 à 1980. À titre de comparaison, plus de 141,000 décès ont été enregistrés dans cette cohorte au cours de la même période.

Les estimations concernant les migrations des résidents sans document sont exclues des ensembles de données sur l'immigration nette, mais elles seront prises en compte dans l'interprétation des résultats. Cette exclusion a été motivée par un certain nombre de facteurs. Les estimations de la taille et de la répartition selon l'âge et le sexe des immigrants illégaux (illegal aliens) varient largement par suite de l'insuffisance d'instruments de collecte de données aux États-Unis. Cependant, même les estimations les plus exagérées du nombre de migrants sans document sont négligeables par rapport aux décès en ce qui concerne les aînés.

Nous avons décrit un certain nombre d'ajustements apportés aux données de base: utilisation de compilations de données censitaires inédites de 1970 à cause d'un surdénombrement marqué de la population des centenaires dans les statistiques officielles; utilisation de compilations modifiées pour la race des recensements de 1980 et de 1990 et exclusion des estimations concernant les populations d'étrangers sans document. Pour juger des effets de ces ajustements sur nos résultats, nous avons procédé à plusieurs analyses de sensibilité utilisant des données non corrigées. Même si nous n'avons observé que de légères différences entre les résultats obtenus avec les données officielles et les données corrigées (sauf pour les 100 ans et plus), les analyses intercensitaires de cohortes utilisant des données non corrigées ont généralement laissé voir des écarts plus grands dans les résultats définitifs, confirmant dans l'ensemble la pertinence des corrections apportées.

3. SOURCES D'ERREURS DANS LES RECENSEMENTS ET LES REGISTRES DE DÉCÈS

Les erreurs dans les données démographiques ont été classées en deux catégories: erreurs de couverture et erreurs de contenu. La couverture s'entend de la mesure dans laquelle les personnes ou les événements faisant partie d'un ensemble défini dans un système de données particulières sont pris en compte. Le contenu s'entend de la qualité de l'information effectivement compilée sur les personnes ou les événements. L'un ou l'autre de ces deux types d'erreurs peut créer des discordances entre l'évolution intercensitaire de la taille d'une cohorte et les décès intercensitaires. Toutefois, si les deux sources de données sont marquées par le même taux net d'omissions, la concordance entre les deux sera maintenue. En outre, dans ces conditions, les taux de mortalité resteront également exacts.

Même si les cas d'erreurs dans la déclaration de l'âge suivent la même tendance dans les recensements et dans les certificats de décès, cela ne permettra pas, en règle générale, de maintenir la concordance des changements de taille des cohortes entre les deux sources de données. En effet, comme les taux de mortalité augmentent avec l'âge, la distribution par âge des décès chez les aînés est "plus vieille" que la distribution par âge de la population. Par exemple, si on place par erreur 10% des personnes et des décès du groupe d'âge 75-79 dans le groupe d'âge 80-84, l'incidence proportionnelle de cette erreur sur les données démographiques sera plus grande que l'incidence proportionnelle sur les taux de mortalité. Cette tendance à déclarer un âge erroné aurait pour effet de biaiser les taux de mortalité et influerait également sur la cohérence des tests que nous utilisons.

Le Census Bureau a utilisé des méthodes démographiques et statistiques pour estimer la qualité de la couverture des recensements. Les méthodes démographiques comparent les estimations du nombre réel de naissances,

diminuées des nombres estimés de décès dans la cohorte et des migrations, aux données censitaires (voir à ce propos le résumé dans Robinson et coll. 1993, et Himes et Clogg 1992). Les méthodes statistiques comparent un groupe de personnes tiré d'une source de données de rechange (p. ex., le Current Population Survey) aux données individuelles du recensement. Une troisième méthode consiste à comparer les données censitaires concernant les aînés aux dénombrements des personnes inscrites dans les dossiers du programme fédéral d'assurance-maladie (Medicare).

Les évaluations du recensement de 1970 réalisées par le U.S. Bureau of the Census (1973, 1974, 1975) ont conduit à un certain nombre de conclusions générales concernant la population des aînés. Premièrement, l'erreur nette (combinaison des erreurs de couverture et de contenu) est plus importante dans le cas des aînés que des personnes plus jeunes. Deuxièmement, les taux d'erreur nets ont été plus élevés dans le cas des femmes que dans celui des hommes, par suite du plus grand nombre d'erreurs commises dans la déclaration de l'âge. Cependant, les taux bruts d'omission (qui ne constituent qu'un élément de l'erreur nette) étaient plus élevés dans le cas des hommes. Troisièmement, les taux d'erreur nets, d'omission bruts et d'erreur dans la déclaration des caractéristiques démographiques sont beaucoup plus élevés dans le cas des américains de race noire que de ceux de race blanche. Quatrièmement, il semble que les statistiques officielles contiennent une très grande quantité d'erreurs dans la déclaration de l'âge. Par exemple, il est intéressant de noter que pour les quatre groupes de race-sexe des personnes âgées de 65 à 69 ans en 1970, les estimations tirées de l'analyse démographique laissent conclure à une surestimation nette dans les données censitaires tandis que l'étude fondée sur les données du Medicare laisse transparaître d'importantes omissions dans les données censitaires (de 2.1% dans le cas des femmes blanches à 12.6% dans le cas des hommes noirs et de ceux des autres catégories raciales). Cette comparaison donne à conclure qu'en plus des taux d'omission bruts dans le dénombrement des personnes âgées de 65 à 69 ans, d'autres erreurs encore plus grandes (surtout, semble-t-il, une exagération de l'âge chez les personnes de moins de 65 ans) agissent dans le sens contraire pour augmenter les estimations du dénombrement net des personnes appartenant à ces groupes d'âge. Il découle notamment de cette situation que les caractéristiques d'une portion substantielle de la population classée dans le groupe des 65 ans et plus lors du recensement se rapportent en fait à des personnes âgées de moins de 65 ans (U.S. Bureau of the Census 1976).

Comparativement aux données censitaires de 1970, les taux d'erreurs nets observés en 1980 dans la plupart des groupes d'âge-race-sexe ont été sensiblement moins élevés. Toutefois, comme le faisait observer le Bureau of the Census (1988), les résultats du programme postcensitaire (Post Enumeration Program, ou PEP) et de la Housing Unit Enumeration Duplication study de 1980 donnent à conclure qu'une part considérable du total des dénombrements censitaires, soit plus de 1.1% selon toute vraisemblance, contient des personnes déjà recensées. Les données laissent supposer que le taux de duplication était beaucoup moindre au cours des recensements antérieurs. Ainsi, les améliorations relevées dans la couverture nette du recensement de 1980 semble malheureusement en partie attribuables à un problème de duplication des données (U.S. Bureau of the Census 1988:10).

Le Census Bureau compte procéder à des évaluations complètes de la qualité du recensement de 1990, mais la diffusion des résultats obtenus est demeurée jusqu'à maintenant fragmentaire. Il semble que le problème de sous-dénombrement des données brutes ait été moins grave en 1980 qu'en 1990 (Robinson et coll. 1993), mais cette différence pourrait être due à un taux de duplication plus élevé dans le recensement de 1980. On peut formuler un certain nombre de généralisations concernant la tendance au sous-dénombrement net dans le recensement de 1990, dans le cas des aînés. Premièrement, confirmant la tendance historique, les estimations de l'erreur nette pour les Afro-américains dépassent largement celles des blancs. La différence la plus importante s'observe dans le cas des hommes âgés de 60 à 64 ans. Le taux de sous-dénombrement net pour les hommes de race noire atteint 10.3%, alors qu'il n'est que de 2.6% chez les hommes de race blanche (différence de 7.7 points). Deuxièmement, alors qu'on observe un sous-dénombrement pour tous les groupes d'âge chez les hommes, il existe des problèmes de surdénombrement dans plusieurs des groupes de femmes. Finalement, comme le soulignent Robinson et coll. (*ibid*), les tendances de la couverture nette sont généralement comparables dans les trois derniers recensements pour chaque groupe de race-sexe.

Les statistiques officielles sur les décès produites par le National Center for Health Statistics constituent la source de base de données sur la mortalité annuelle aux États-Unis. Ces chiffres sont généralement utilisés sans ajustement pour tenir compte du sous-dénombrement et des erreurs de déclaration qui se glissent dans les certificats de décès. Toutefois, on présume en général que le système de recensement des décès est pratiquement complet (Wilkin 1981; U.S. Bureau of the Census 1984a; National Center for Health Statistics 1968) même si aucun test national n'a été fait pour confirmer ce diagnostic depuis la mise en place de la Death Registration Area en 1933. Cette supposition s'appuie sur les exigences légales strictes en vigueur concernant le recensement des décès ainsi que sur la nécessité, pour les survivants, de prouver le décès aux fins des arrangements funéraires, du règlement des droits successoraux et de la collecte des prestations d'assurance (U.S. Bureau of the Census 1984a; Wilkin 1981). Les calculs réalisés par Coale et Kisker (1990) donnent toutefois à penser qu'il existe un problème de sous-dénombrement des décès, en particulier chez les aînés. Dans le cas des personnes de couleur, par exemple, le total des décès portés aux registres de l'état civil est inférieur de 7% à celui des décès inscrits aux dossiers du Medicare dans le cas des hommes âgés de plus de 80 ans en 1980, et de 10% dans le cas des femmes. Toutefois, il est possible que l'écart relevé soit dû à des différences dans l'âge déclaré entre les deux sources plutôt qu'à un problème de sous-dénombrement.

Une étude au cours de laquelle on a comparé un échantillon de certificats de décès datant de mai à août 1960 aux données du recensement de 1960 constitue la meilleure évaluation de la concordance des données sur l'âge déclaré contenues dans les tableaux de recensements et dans les registres de l'état civil – sans doute une des sources les plus importantes d'erreurs de contenu influant sur notre test de cohérence (NCHS 1968; Hambright 1969). Même si les données ont été recueillies avant la période visée par notre projet, les résultats de cette étude mettent en lumière un problème qui risque d'exister encore aujourd'hui. Les auteurs de l'étude ont constaté ce qui suit: 1) dans le cas des blancs, les données des deux sources concordent assez bien même lorsque l'âge augmente, mais pour les personnes de couleur, la concordance est moins nette; 2) lorsqu'il y a discordance entre les deux sources de données, la différence d'âge dans le cas des blancs est généralement inférieure à un an, mais cette différence est typiquement supérieure à un an dans le cas des personnes de couleur, en particulier pour celles âgées de 45 ans et plus; 3) pour les blancs de tous âges et pour les personnes de couleur âgées de moins de 45 ans, l'âge indiqué sur le certificat de décès est typiquement plus avancé que celui déclaré au moment du recensement. Par contre, pour les personnes de couleur âgées de 45 ans et plus, l'âge indiqué sur le certificat de décès est en moyenne moins avancé que celui déclaré lors du recensement.

Les auteurs de cette étude n'ont pas été en mesure de déterminer laquelle des deux sources donnait l'âge "réel". Rosenwaike et Logue (1983) ont cherché à vérifier les données sur l'âge indiqué dans les certificats de décès des personnes âgées de 85 ans et plus pour la période de 1968 à 1972. Ils ont à cette fin tiré un échantillon de certificats de personnes décédées à un âge très avancé en Pennsylvanie et aux New Jersey. Ils ont ensuite comparé ces données à celles du recensement manuscrit de l'année 1900. Au total, 1,429 comparaisons ont été établies: 960 concernant des personnes de race blanche et 496 concernant des personnes de couleur.

Ils ont constaté que la concordance des renseignements sur l'âge entre les deux sources diminuait à mesure que l'âge augmentait pour les deux groupes raciaux. Des différences frappantes ont été observées entre les groupes raciaux. La concordance était élevée dans le cas des blancs, sauf pour les personnes âgées de 100 ans et plus. Toutefois, pour les personnes de couleur, la concordance était sensiblement moins bonne. Les auteurs ont par ailleurs relevé qu'à l'intérieur de chaque groupe racial, le sexe avait peu d'incidence sur la concordance des données.

4. MÉTHODE INTERCENSITAIRE D'ÉVALUATION DE LA QUALITÉ DES STATISTIQUES SUR LES AÎNÉS

La présente analyse permet de déterminer l'ampleur de la discordance des sources de données sur les aînés à l'aide d'une méthode intercensitaire fondée sur les cohortes. La taille prévisible d'une cohorte d'âge ouverte dans le second recensement peut être estimée à partir de sa taille dans le premier recensement et des décès survenus au sein de cette cohorte au cours de la période intercensitaire, après l'ajustement tenant compte des migrations (Condran, Himes et Preston 1991). Le recours à une catégorie ouverte permet l'observation de la tendance du rapport tout en atténuant l'effet des valeurs extrêmes dues à des erreurs dans des groupes d'âge particuliers. Cette méthode est insensible aux erreurs commises dans la déclaration de l'âge des personnes décédées ou des personnes recensées dont l'âge est supérieur à celui qui marque le début de la catégorie ouverte.

En utilisant les données censitaires et celles portant sur les décès et les migrations pour une période intercensitaire donnée, l'analyse intercensitaire des cohortes nous permet d'estimer la taille prévisible de chaque cohorte d'âge ouverte du recensement suivant. Les statistiques susmentionnées, classées par années d'âge, par sexe et par race (blanche et noire), ont servi à calculer l'équation suivante pour la population prévue à l'époque du second recensement:

$$\hat{N}_x(2) = N_{x-10}(1) - D_{x-10}(1) + M_{x-10}(1) \qquad (1)$$

où

$\hat{N}_x(2)$ = la population prévue d'âge x et plus, lors du second recensement, réalisé 10 ans après le premier;

$N_{x-10}(1)$ = la population recensée d'âge $x - 10$ et plus au temps 1, soit au premier recensement;

$D_{x-10}(1)$ = les décès survenus pendant la période intercensitaire dans la cohorte des personnes d'âge $x - 10$ et plus (au moment du premier recensement);

$M_{x-10}(1)$ = l'immigration légale nette intercensitaire dans la cohorte des personnes d'âge $x - 10$ et plus (au moment du premier recensement).

On peut également calculer à l'aide d'une méthode analogue la population prévue d'un âge donné (par opposition à celle d'âge x et plus). Dans l'un ou l'autre de ces cas, il est ensuite possible de calculer le rapport de la population observée, dénombrée lors du recensement suivant, sur la population prévue (après simplification et en présumant que la migration nette est nulle) à l'aide de la formule suivante:

$$R_x = \frac{N_x(2)}{N_{x-10}(1) - D}. \qquad (2)$$

Le changement de taille de la cohorte mesuré entre deux recensements successifs ne peut être dû qu'aux décès et aux migrations. Un rapport de 1.00 indiquerait donc une concordance complète entre les sources de données. (Noter toutefois qu'un tel rapport n'est pas nécessairement un signe d'exactitude des données. Ainsi, si l'âge d'une personne donnée a systématiquement été augmenté de n années, la méthode ne permettra pas de relever cette déclaration erronée). Dans les faits, toutefois, le dénombrement

indiqué pourra également subir l'influence: 1) des erreurs de couverture dans l'un ou l'autre ou les deux recensements; 2) des erreurs (en plus ou en moins) dans les registres des décès ou les statistiques de l'immigration; 3) des déclarations erronées (concernant l'âge, la race, *etc.*) dans l'une ou l'autre ou l'ensemble des sources de données (Ewbank 1981; Shryock et Siegel 1976; Condran et coll. 1991). Le rapport des populations observées sur les populations prévues est un outil de diagnostic utile lorsqu'il est possible d'attribuer les écarts par rapport à 1.00 à l'effet de ces erreurs sous-jacentes. Toutefois, il ne constitue pas un outil très précis puisque différents types d'erreurs peuvent produire le même genre d'écarts. Il peut néanmoins permettre un choix plus facile entre diverses solutions concurrentes.

5. EFFETS DES ERREURS SUR LES RAPPORTS DES POPULATIONS OBSERVÉES SUR LES POPULATIONS PRÉVUES

Certains types d'erreurs ont un effet direct visible sur la formule du rapport lui-même (effets qui ont été confirmés par les simulations que nous avons réalisées). Pour simplifier la démonstration, désignons par R_x dans l'équation (2) le rapport de la population observée sur la population prévue pour l'âge x et plus au deuxième recensement. On peut distinguer les possibilités principales suivantes d'erreurs de couverture et leurs répercussions sur la configuration des rapports en fonction de l'âge:

1) Si $N_{x-10}(1)$ et D sont également complets et que $N_x(2)$ est assorti d'un degré de complétude relative de $C(2)$, alors la configuration des rapports en fonction de l'âge sera constante et son niveau sera $C(2)$.

2) Si $N_x(2)$ et D sont également complets et que $N_{x-10}(1)$ est assorti d'un degré de complétude relative de $C(1)$, alors la configuration des rapports en fonction de l'âge sera:

   a) supérieure à 1.00 et augmentera en fonction de l'âge si $C(1) < 1.00$;

   b) inférieure à 1.00 et diminuera en fonction de l'âge si $C(1) > 1.00$.

La raison pour laquelle ces erreurs induisent un effet relatif à l'âge sur le rapport R_x est qu'une erreur proportionnelle donnée dans $N_{x-10}(1)$ engendre des erreurs proportionnelles de plus en plus grandes dans le dénominateur à mesure que les deux termes compensatoires du dénominateur (l'un positif et l'autre négatif) tendent l'un vers l'autre en valeur absolue. Cette égalisation est due au fait qu'une proportion plus grande des membres de chaque cohorte meurent pendant la période intercensitaire à mesure que l'âge avance.

3) Si $N_{x-10}(1)$ et $N_x(2)$ sont également complets et que D est assorti d'un degré de complétude relative de $C(D)$, alors la configuration des rapports en fonction de l'âge sera:

   a) supérieure à 1.00 et augmentera en fonction de l'âge si $C(D) > 1.00$ (c.-à-d., si les certificats de décès sont plus complets que les dénombrements des deux recensements);

   b) inférieure à 1.00 et diminuera en fonction de l'âge si $C(D) < 1.00$.

Ici encore, on observe une tendance relative à l'âge puisqu'une erreur proportionnelle égale dans D engendrera des erreurs proportionnelles plus grandes dans le dénominateur à mesure que les deux éléments de ce dernier tendent l'un vers l'autre en termes absolus.

On peut mieux comprendre les effets des déclarations erronées de l'âge en examinant les éléments qui composent cette formule. Shrestha (1993) et Condran et coll. (1991) ont introduit diverses erreurs dans des ensembles simulés, dépourvus d'erreurs, typiques des conditions démographiques actuelles existant aux États-Unis et aux Pays-Bas respectivement. Ils ont démontré qu'une tendance nette à exagérer l'âge qui se limiterait à deux recensements produirait des rapports oscillant aux alentours de 1.00 jusqu'à un âge avancé, pour chuter ensuite à des valeurs très faibles. La diminution du rapport à des valeurs inférieures à 1.00 s'explique ici encore par le fait qu'une erreur dans un des éléments du dénominateur (dans ce cas, l'inflation de $N_{x-10}(1)$ par une exagération de l'âge) engendre des effets disproportionnés dans le dénominateur. Même si la chute rapide de la distribution par âge peut entraîner une inflation plus grande de $N_x(2)$ par rapport à $N_{x-10}(1)$, l'inflation du dénominateur finira tôt ou tard par excéder celle du numérateur et les rapports diminueront. (À titre d'exemple, voir la figure 1 de Condran et coll. 1991).

Un problème d'exagération de l'âge limité aux recensements des décès donnera un rapport supérieur à 1.00 et qui augmentera avec l'âge. Dans ce cas, le dénominateur est trop petit (sa composante négative est trop grande) et le déficit proportionnel augmente avec l'âge.

L'existence d'une tendance identique à exagérer l'âge dans les certificats de décès et les recensements donnera également des rapports qui finiront par augmenter avec l'âge. Ce résultat important est robuste dans la mesure de l'erreur introduite (Condran et coll. 1991). Il découle du fait que les distributions par âge chutent de plus en plus rapidement à mesure que l'âge avance, de sorte qu'un pourcentage identique de personnes qui exagèrent leur âge véritable entraînera un pourcentage plus grand d'erreurs dans les distributions par âge aux âges très avancés. Autrement dit, $N_x(2)$ est assorti d'un facteur d'inflation plus grand que $N_{x-10}(1)$. Dans ce cas, une certaine inflation de $N_{x-10}(1)$ verra ses effets sur le dénominateur compensés par l'inflation de D.

6. RÉSULTATS

L'analyse intercensitaire des cohortes a porté sur quatre groupes de sexe-race des États-Unis, au cours des périodes 1970-1980 et 1980-1990. Les figures 1 et 2 illustrent les rapports calculés de la population observée sur la population prévue pour des groupes d'âge précis en fonction de la race, du sexe et de la période intercensitaire.

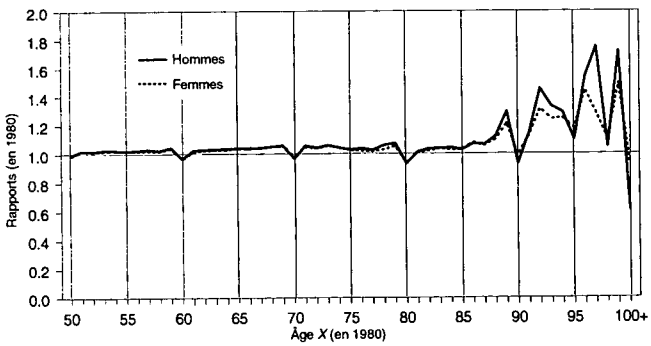


Figure 1A. Rapports intercensitaires de la population observée sur la population prévue: blancs, 1970-1980.

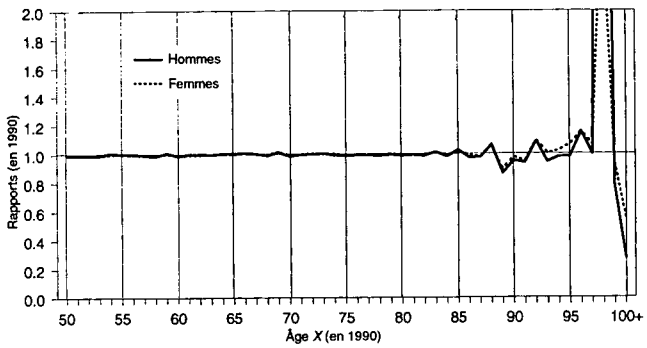


Figure 1B. Rapports intercensitaires de la population observée sur la population prévue: blancs, 1980-1990.

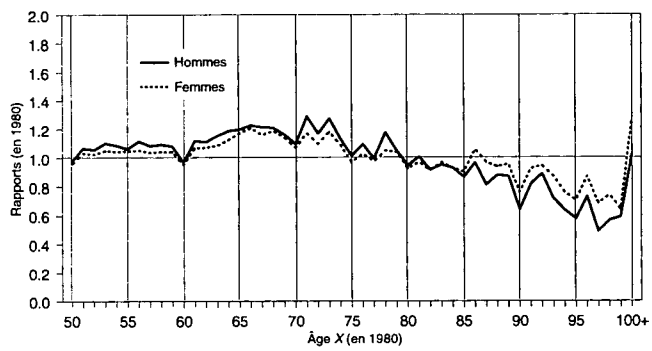


Figure 2A. Rapports intercensitaires de la population observée sur la population prévue: noirs, 1970-1980.

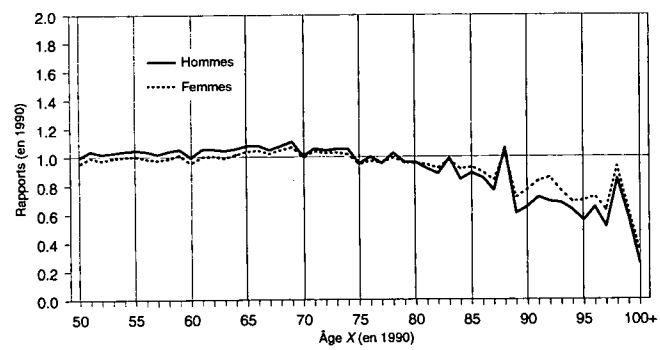


Figure 2B. Rapports intercensitaires de la population observée sur la population prévue: noirs, 1980-1990.

Pour toutes les combinaisons de races et de périodes, l'effet des tendances relatives à l'âge sur les rapports est pratiquement le même pour les hommes et pour les femmes. Dans tous les cas, le degré de discordance augmente avec l'âge, même si on ne commence à observer un écart systématique et significatif par rapport à 1.00 qu'à partir de 95 ans dans le cas des blancs en 1980-1990. On observe une discontinuité nette dans beaucoup de ces séries à l'âge de 100 ans, ce qui reflète le caractère idiosyncratique de la déclaration de l'âge et les méthodes d'ajustement utilisées par le Census Bureau dans le cas des centenaires.

6.1 Résultats pour les blancs

6.1.1 Période intercensitaire 1970-1980

Comme l'indique la figure 1A, les rapports de la population blanche ont tendance à se maintenir au-dessus de 1.00 au cours de cette période et ils augmentent avec l'âge (jusqu'à 100 ans). Cette configuration pourrait être le résultat de divers types d'erreurs de données dont les deux plus plausibles sont:

1) un sous-dénombrement dans le recensement de 1970 par rapport au recensement de 1980 et au recensement des décès;

2) des probabilités à peu près égales d'exagération de l'âge dans les certificats de décès et lors des deux recensements.

Nous croyons que la première explication risque davantage d'être la bonne. Si l'évolution des rapports en fonction du vieillissement découlait de tendances semblables à l'exagération de l'âge déclaré dans les certificats de décès et les recensements, on pourrait s'attendre à une situation comparable au cours de la décennie 1980-1990, notamment du fait que le recensement de 1980 figure dans les deux comparaisons. On ne s'attendrait pas non plus à voir les propensions culturelles à déclarer un âge erroné disparaître soudainement. Pourtant, le graphique de l'évolution des rapports pour les blancs au cours de la décennie 1980-1990 (figure 1B) laisse constater une remarquable concordance, bien meilleure que dans la plupart des pays européens et équivalant à celle observée en Suède et dans les Pays-Bas, deux pays dotés de registres de la population extrêmement efficaces (Condran et coll. 1991). La concordance des données observée pour la décennie 1980-1990 est également beaucoup plus grande que celle obtenue dans d'autres pays de langue anglaise: Angleterre et pays de Galles, Canada, Australie et Nouvelle-Zélande.

Notre préférence pour la première explication est également motivée par le fait que le Census Bureau a conclu que le recensement de 1980 était plus complet que celui de 1970 (U.S. Bureau of the Census 1988; Robinson et coll. 1993). Cette conclusion est fondée en partie sur l'analyse démographique et n'est donc pas entièrement indépendante du genre de preuve que nous examinons. Toutefois, l'analyse démographique du Bureau fait grand usage de groupes d'âge plus jeunes que ceux auxquels nous nous intéressons ici. En outre, la conclusion selon laquelle la couverture des recensements a été améliorée est également confirmée par

le programme postcensitaire au cours duquel des données individuelles du recensement sont comparées à des données provenant d'autres systèmes.

6.1.2 Période intercensitaire 1980-1990

Comme mentionné précédemment, l'évolution des rapports dans le cas des blancs (en particulier les femmes) laisse constater une très grande concordance des données, de loin meilleure que dans la plupart des pays européens. Notre étude semble confirmer l'hypothèse de Vaupel (1993) selon laquelle la population blanche des États-Unis présenterait des taux de mortalité plus bas, au-delà de l'âge de 80 ans, que dans tous les autres pays industrialisés. Toutefois, la prudence est de mise. Si nos méthodes laissent voir une grande concordance des données entre les recensements et les certificats de décès pour la période 1980-1990, rien n'indique que ces données soient nécessairement exactes. Condran et coll. (1991) décrit une situation dans laquelle une tendance à déclarer des âges erronés donne une série de rapports de 1.00 pour tous les âges. Par ailleurs, la méthode intercensitaire ne permet pas de dépister les sujets qui font systématiquement des déclarations erronées de leur âge. Comme le soulignait Horiuchi (1993), une personne qui exagère une première fois son âge - pour être admise à une école ou être autorisée à travailler plus jeune, pour éviter le service militaire lorsqu'elle approche de l'âge limite supérieur ou pour bénéficier plus tôt de la sécurité sociale, du programme Medicare ou d'une pension - pourra avoir tendance à continuer d'exagérer son âge par la suite. Une telle possibilité ne peut être ignorée, même si nous sommes incapables de la mesurer directement.

6.2 Résultats pour les noirs

L'évolution des rapports en fonction de l'âge est beaucoup plus régulière dans le cas des Afro-américains (voir les figures 2A et 2B). Ils commencent à diminuer vers 70 ans pour les deux sexes dans les deux périodes et poursuivent cette chute jusqu'aux âges avancés (jusqu'à 100 ans en 1970-1980). Avant 70 ans, les rapports se maintiennent typiquement au-dessus de 1.00 en 1970-1980, et légèrement au-dessus de 1.00 pour les hommes en 1980-1990.

L'écart observé entre les rapports pour des groupes d'âge donnés entre les résultats de 1970-1980 et de 1980-1990 n'est pas étranger au sous-dénombrement relatif qui a caractérisé le recensement de 1970. Comme nous l'avons déjà mentionné, il est vraisemblable qu'un tel sous-dénombrement se soit produit également dans le cas des blancs. Ce sous-dénombrement ne suffit pas toutefois à expliquer la tendance à la baisse persistante des rapports au-delà de 70 ans pour les deux périodes. Deux raisons principales peuvent permettre d'expliquer ce déclin des séries de rapports pour les Afro-américains:

1) L'inscription des décès aux registres d'état civil n'est pas faite avec la même efficacité que les recensements;

2) L'exagération de l'âge est plus importante dans les recensements que dans les certificats de décès.

Coale et Kisker (1990) préfèrent la première explication. Ils relèvent que les populations reconstituées à partir des données sur les décès à l'aide de méthodes à r variables (Preston et Coale 1982) sont trop petites par rapport aux données censitaires de 1980 pour les plus de 65 ans, ce qui porte à conclure que les décès n'ont pas tous été portés aux registres. Ils relèvent également que chez les Afro-américains d'âge avancé, on observe un déficit dans les décès déclarés par rapport aux dossiers Medicare.

Toutefois, ces deux observations pourraient également s'expliquer par une exagération de l'âge lors des recensements (et dans les dossiers Medicare) par rapport aux données des certificats de décès. Une comparaison directe des certificats de décès de 1960 et des données se rapportant aux mêmes sujets dans le recensement de 1960 (NCHS 1968; Hambright 1969) tend à confirmer l'existence d'une telle tendance. Pour les hommes comme pour les femmes, le total des décès chez les plus de 50 ans, lorsque ces décès sont classés selon l'âge déclaré au recensement, ne s'écarte pas de plus de 1% du total des décès classés selon l'âge inscrit dans les certificats de décès. Toutefois, pour l'ensemble des personnes de 65 ans et plus, le total des décès classés selon l'âge au recensement dépasse de 15.4% celui des décès classés selon l'âge inscrit dans le certificat de décès dans le cas des femmes, et de 7.1% dans le cas des hommes. Pour les personnes de 75 ans et plus, les écarts atteignent 23.3 et 17.8% respectivement et pour le groupe des 85 ans et plus, ils sont de 39.2 et 17.6% respectivement.

Ces écarts importants de l'âge déclaré entre les recensements et les certificats de décès peuvent expliquer la tendance à la baisse des rapports observée dans les graphiques des figures 2A et 2B. Elo et Preston (1994) ont calculé que les valeurs de R_x pour les Afro-américains, entre les périodes 1950-1960 et 1960-1970, périodes qui encadrent la comparaison des données censitaires et celles des certificats de 1960. Ils démontrent que si les âges au décès sont "corrigés" pour correspondre à ceux déclarés lors des recensements, la tendance à la baisse des rapports est éliminée.

Les raisons qui poussent les Afro-américains à déclarer un âge plus avancé lors des recensements, comparativement à l'âge inscrit dans le certificat de décès, ne sont pas claires. Cette tendance n'apparaît pas avant le recensement de 1940, c'est-à-dire le premier réalisé après l'adoption de la Loi sur la sécurité sociale. Ce recensement laisse constater un large excédent d'Afro-américains appartenant aux groupes d'âge 65-69 ans et 70-74 ans, et un déficit des personnes du groupe d'âge 50-64 ans (Elo et Preston 1994). Comme le rappelle Wolfenden (1954:56), "Les distorsions dans les données sur les noirs étaient si marquées qu'il a fallu procéder à une nouvelle répartition préliminaire de ces populations (et de ces décès) entre 55 et 69 ans, aux fins de la préparation des tables de survie." Cet excédent persiste, même s'il s'atténue graduellement, jusque dans les recensements les plus récents (comme le montre la figure 2). Quelle qu'en soit la raison, nous croyons que l'explication principale des écarts importants observés entre les données censitaires et les données des certificats de décès pour la population noire américaine s'expliquent

principalement par une tendance, chez les personnes de ce groupe, à exagérer leur âge au moment des recensements par rapport aux données inscrites dans les certificats de décès. Il s'ensuit de cette tendance que les taux de mortalité des groupes d'âge supérieurs à 65 ans pour les Afro-américains donneront vraisemblablement lieu à de graves sous-estimations. Il est effectivement possible que les courbes des taux de mortalité des blancs et des noirs se croisent à un âge avancé, mais il serait dangereux de fonder une telle conclusion uniquement sur les données des recensements et des registres de l'état civil. La discordance entre ces deux sources de données est tout simplement trop grande pour nous permettre de faire une estimation fiable des taux de mortalité pour les groupes d'âge avancé.

7. CONCLUSION

Les erreurs de couverture et de contenu relevées dans les systèmes de recensement des personnes et des décès aux États-Unis laissent planer un doute sérieux sur la qualité de ces données en ce qui concerne les aînés. Nous avons évalué la concordance des données provenant de ces deux sources pour les populations blanche et afro-américaine, en concentrant notre attention sur les personnes âgées de 60 ans et plus, où les tendances affichées par les taux de mortalité ont la plus grande incidence sur les programmes sociaux et où les données sont les plus problématiques. L'analyse intercensitaire des cohortes laisse constater des discordances liées à l'âge entre les deux sources de données pour les deux périodes: 1970-1980 et 1980-1990.

Afin de déterminer quelles combinaisons de degré de couverture et de tendance à la déclaration d'un âge erroné produirait les résultats empiriques, nous avons réalisé une série de simulations. Les discordances observées dans les données américaines sont examinées à la lumière des résultats de ces simulations et des données de la documentation spécialisée concernant la nature des erreurs de couverture et de contenu dans les sources de données.

Les données concernant les blancs pour la période intercensitaire 1980-1990 laissent constater une concordance remarquable. Leur qualité, jusqu'à 95 ans, s'approche de celle des données de la Suède et des Pays-Bas, deux pays dotés de registres de la population extrêmement efficaces. Les données portant sur les blancs pour la décennie 1970-1980 laissent voir plus de discordances. L'explication la plus plausible de cette différence est le sous-dénombrement relatif net du recensement de 1970 combiné à la tenue de statistiques plus complètes sur les décès. En conséquence, les estimations de la mortalité chez les aînés qui utilisent des numérateurs tirés des certificats de décès et des dénominateurs tirés du recensement de 1970 risquent de surestimer la mortalité.

Les données concernant les Afro-américains laissent voir une tendance différente. Au-delà de 70 ans, la population recensée diminue graduellement comparativement à la population prévisible tant en 1980 qu'en 1990. Il semble que cette discordance soit principalement due à une tendance à exagérer l'âge lors des recensements, par rapport aux données portées aux certificats de décès. Il s'ensuit de cette tendance que les taux de mortalité des Afro-américains plus âgés risquent d'être sérieusement sous-estimés. Il est possible que les courbes des taux de mortalité des noirs et des blancs se croisent à un âge avancé, mais il serait imprudent de fonder une telle conclusion uniquement sur les données des recensements et des registres de l'état civil.

REMERCIEMENTS

Cette recherche a été réalisée au Population Studies Center de la University of Pennsylvania, avec l'aide financière du National Institute of Aging (AG10168) et du Boettner Institute of Financial Gerontology. Nous remercions Irma Elo, Douglas Ewbank, Shiro Horiuchi et J. Gregory Robinson pour leurs commentaires utiles sur notre projet et sur le texte du présent article. Merci également à J. Gregory Robinson et au U.S. Bureau of the Census qui nous ont fournis les données censitaires inédites et les tableaux de données sur les migrations internationales.

ANNEXE A

Source: Shrestha (1993)

Trois sources principales de données ont été utilisées dans notre étude: 1) recensements de 1970, 1980 et 1990; 2) registres officiels des décès; 3) statistiques sur l'immigration nette. Nous décrivons ci-après ces sources et les ajustements que nous avons apportés aux données.

1.A Recensement de 1970

Les tableaux officiels des caractéristiques démographiques de base de la population en 1970 sont présentés dans le document Series B - U.S. Summary of the 1970 Census (U.S. Bureau of the Census 1972). Nous savons que ces données officielles contiennent des erreurs importantes qui risquent de biaiser notre enquête sur le recensement des aînés aux États-Unis. La première est un surdénombrement évident des personnes centenaires. Alors que 106,000 personnes ont été classées dans cette catégorie ouverte, des analyses démographiques indirectes donnent plutôt à conclure que le nombre correct oscillerait plutôt entre 3,000 et 8,000 (Siegel 1974; Siegel et Passel 1976). Ce surdénombrement semble avoir été causé par une interprétation erronée d'une des questions du formulaire plutôt que par une tendance systématique à se déclarer faussement dans la catégorie des centenaires. Le deuxième problème découle d'une erreur de classification de la population en groupes raciaux dans les tableaux du recensement total qui touche 21,000 personnes âgées de 65 ans et plus. Finalement, le recensement officiel a omis plus de 23,000 personnes (de tous âges) dont les données ont été retrouvées après la publication des tableaux officiels.

À cause des erreurs contenues dans les tableaux officiels, nous avons utilisé des tableaux corrigés inédits obtenus auprès du U.S. Bureau of the Census. Ces données statistiques modifiées étaient exemptes des trois types d'erreurs susmentionnés. Les données sont classées selon la race (blanche ou noire), le sexe, les années d'âge (de 0 à 94 ans) et les groupes d'âge (95-99 ans et 100 et plus). Pour répartir les données du groupe des 95-99 ans en années d'âge, nous avons utilisé des distributions de l'âge moyen par sexe et par race des recensements de 1960 et de 1980 pour les blancs, et des recensements de 1950 et 1980 pour les noirs (les données en années d'âge ne sont pas disponibles pour ce groupe d'âge chez les noirs dans le recensement de 1960).

1.B Recensement de 1980

Les tableaux officiels du recensement de 1980 sont présentés dans le document Series B – U.S. Summary of the 1980 Census (U.S. Bureau of the Census 1983). Au cours de ce recensement, un grand nombre de gens (environ 6.8 millions) ont choisi de fournir une réponse de leur cru à la question concernant leur race au lieu de choisir une des catégories préétablies. Comme le recensement de 1980 est le seul à contenir une catégorie raciale résiduelle, il n'est pas directement comparable aux autres sources de données (registres de l'état civil, recensements antérieurs, *etc.*). le Census Bureau a produit un dossier modifié conforme aux catégories raciales historiques (U.S. Bureau of the Census 1984b). La méthode d'adaptation des données comportait une redistribution des macro-données sur la race fondée sur un tableau détaillé croisant la race et l'origine hispanique et découlant des données échantillons et des données intégrales du recensement. Nous décrivons en détails ci-après les modifications apportées par le Census Bureau.

Pour les 219.8 millions de personnes qui ont choisi l'une des 14 catégories raciales préétablies, aucun ajustement n'a été fait. Deux catégories de personnes (6.7 millions au total) ont choisi de fournir une description de leur cru: les personnes d'origine hispanique (5.8 millions) et celles d'une autre origine (0.9 million). Des méthodes d'ajustement différentes ont été élaborées pour ces deux groupes.

Les personnes d'origine hispanique ont été réparties entre les catégories blanche et noire (et non dans celles des autochtones ni des asiatiques ou des insulaires du Pacifique). Toutes les personnes d'origine mexicaine ont été replacées dans la catégorie blanche. Les personnes d'origine porto-ricaine et cubaine et toutes les autres personnes d'origine hispanique ont été replacées dans des groupes modifiés, blancs ou noirs, dans des proportions identiques au choix des Hispaniques de même origine qui avaient précisé la race blanche ou noire dans leur formulaire. Les calculs ont été effectués par cellules d'âge-sexe-comté.

Les personnes d'une autre origine ont été réparties entre trois groupes raciaux modifiés (blancs, noirs et autres) dans des proportions correspondant à celles observées dans les cellules d'âge-sexe-comté de leur État de résidence. Ces proportions étaient fondées sur un échantillon de données tirées du recensement de 1980. Pour de plus amples détails sur ces modifications, voir U.S. Bureau of the Census (1984b).

Les données des tableaux modifiés sont classées selon la race, le sexe, les années d'âge (0 à 99) et le groupe d'âge (100 ans et plus). Aux fins de notre recherche, nous avons utilisé les groupes raciaux modifiés, compte tenu du nombre considérable de personnes qui ont été transférées du groupe racial résiduel aux catégories blanche ou noire.

1.C Recensement de 1990

La publication des tableaux du recensement de 1990 par le U.S. Bureau of the Census n'est pas encore terminée. Les statistiques déjà publiées présentent toutefois un certain nombre de problèmes qui compliquent leur comparaison avec les recensements antérieurs et les autres sources de données. Trois problèmes sont évidents: classification des 9.3 millions de personnes placées dans une catégorie raciale résiduelle non spécifiée; incohérences dans la déclaration de l'âge et modification des méthodes d'attribution de l'âge pour les personnes qui ont omis de fournir ce renseignement.

Un dossier modifié intitulé MARS (Modified Age and Race Statistics) a été préparé par le Census Bureau pour régler les deux premiers problèmes (Word et Spencer 1991). Les modifications du recensement de 1990 ont porté sur les micro-données. Des méthodes d'imputation spéciales (hot-deck) ont servi à attribuer une race particulière aux personnes qui avaient choisi la catégorie "autre non précisé". Cette redistribution est appliquée aux dossiers individuels des tableaux détaillés entièrement révisés du recensement de 1990 (Robinson, Word et Spencer 1991).

Nous avons encore une fois choisi d'utiliser les statistiques modifiées, classées selon la race, le sexe et les années d'âge. Notre décision d'utiliser les statistiques modifiées des recensements de 1980 et de 1990 n'allait pas de soi; pour un examen plus détaillé de la question, voir Shrestha (1993).

2. Le système d'enregistrement des décès

Les statistiques annuelles nationales sur les décès que nous avons utilisées pour notre étude sont venues du National Center for Health Statistics (NCHS). Les données de 1970 à 1988 ont été tirées des bandes de données du NCHS obtenues auprès du ICPSR (NCHS 1970-1988). Elles sont classées selon la race (noire ou blanche), le sexe et les années d'âge (0 à 124 ans; catégorie ouverte pour les 125 ans et plus). Comme les bandes de données de l'année civile 1989 et des trois premiers mois de 1990 n'ont pas encore été diffusées, nous avons élaboré une méthode pour en estimer la distribution. Les statistiques finales sur la mortalité pour 1989 classées selon la race et le sexe ont été publiées par le NCHS en 1992. Les données présentées par groupes d'âge ont été réparties en années d'âge en fonction de la répartition des données de 1988. La répartition selon le mois du décès s'est fondée sur les données des registres de l'état civil (NCHS 1989). Les estimations de la distribution des décès en 1990 sont fondées sur les rapports provisoires mensuels sur la mortalité du NCHS (1990). Ces données provisoires ont été classées en années d'âge en

fonction de la répartition des données du groupe d'âge correspondant en 1988.

Tel que mentionné dans le texte, nous avons ajusté les données disponibles pour corriger deux problèmes. Premièrement, les périodes intercensitaires s'étendent du 1er avril au 31 mars tandis que les registres des décès utilisent l'année civile. Deuxièmement, les deux sources de données indiquent l'âge au dernier anniversaire au lieu de donner l'année de la naissance. Comme la date des recensements est fixée au 1er avril, nous préférons la deuxième méthode qui identifie la cohorte de naissance utilisée dans notre analyse des cohortes. Pour corriger ces deux problèmes, nous présumons que la surface tridimensionnelle du nombre de décès en fonction de l'âge et du temps est égale sur tout l'intervalle. Nous ne tenons pas compte du sous-dénombrement ni des déclarations erronées dans les certificats de décès.

3. Statistiques sur l'immigration nette

Nous utilisons des statistiques inédites sur l'immigration nette obtenues auprès du U.S. Bureau of the Census. Les données sont classées par catégories de "facteurs de changement" pour chacune des deux décennies.

L'immigration nette en fonction de l'âge, de la race et du sexe a été calculée par cohortes à l'aide de l'équation suivante:

Immigration nette = immigration internationale légale + réfugiés et admissions conditionnelles + immigration nette de civils américains + immigration nette de Porto-ricains + immigration nette d'étudiants étrangers + déplacements nets des membres des Forces armées américaines outre-mer − émigration légale.

Compte tenu de l'imprécision des données brutes fournies par le U.S. Census Bureau, un certain nombre d'ajustements ont dû être apportés. Premièrement, la classe d'âge terminale des données fournies était trop précoce (75 ans et plus au début de la décennie). Pour répartir ces données en groupes quinquennaux (75-79, . . . , 95-99; 100 et plus), nous avons présumé que les taux d'immigration nette en fonction de l'âge, de la race et du sexe demeuraient constants pendant toute la période ouverte débutant à 75 ans. Cette estimation plutôt imprécise est adéquate à cause du petit nombre d'immigrants appartenant à ce groupe d'âge. Deuxièmement, pour convertir les groupes quinquennaux en années d'âge, nous avons eu recours aux multiplicateurs ou à l'interpolation osculatrice de Sprague (Sprague, 1880-81).

BIBLIOGRAPHIE

COALE, A.J., et KISKER, E.E. (1990). Defects in data on old-age mortality in the United States: new procedures for calculating mortality schedules and life tables at the highest ages. *Asian and Pacific Population Forum*, 4(1), 1-31.

CONDRAN, G.A., HIMES, C.L., et PRESTON, S.H. (1991). Old-age mortality patterns in low-mortality countries: an evaluation of population and death data at advanced ages, 1950 to present. *Population Bulletin of the United Nations*, 30, 23-60.

ELO, I.T., et PRESTON, S.H. (1994). New estimates of old-age mortality among African Americans, 1930-1990. *Demography*, 31(3), 427-58.

EWBANK, D.C. (1981). *Age Misreporting and Age-Selective Underenumeration: Sources, Patterns, and Consequences for Demographic Analysis*. National Academy of Sciences, Committee on Population and Demography. Report No. 4. Washington, DC: National Academy Press.

HAMBRIGHT, T.Z. (1969). Comparison of information on death certificates and matching 1960 census records: age, marital status, nativity, and country of origin. *Demography*, 6(4), 413-24.

HILL, K. (1985). Illegal Aliens: an assessment. In: Panel on Immigration Statistics. *Immigration Statistics, A Story of Neglect*. Washington, DC: National Academy Press.

HIMES, C.L., et CLOGG, C.C. (1992). An overview of demographic analysis as a method for evaluating census coverage in the United States. *Population Index*, 58(4), 587-607.

HORIUCHI, S. (1993). Personal communication dated April 20, 1993.

MANTON, K.G., STALLARD, E., et VAUPEL, J.W. (1986). Alternative models for the heterogeneity of mortality risks among the aged. *Journal of the American Statistical Association*. 81, 635-644.

MC CORD, C., et FREEMAN, H.P. (1990). Excess mortality in Harlem. *New England Journal of Medicine*. 322, 172-177.

NATIONAL CENTER FOR HEALTH STATISTICS (1968). *Comparability of age on the death certificate and matching census record: United States - May - August 1960*; Vital and Health Statistics: Data Evaluation and Methods Research. By Thea Zelman Hambright. Series 2, No. 29.

NATIONAL CENTER FOR HEALTH STATISTICS (1970-1988). *Mortality Detail Files* (data and codebooks). Data were made available by the Inter-University Consortium for Political and Social Research, University of Michigan.

NATIONAL CENTER FOR HEALTH STATISTICS (1989). Births, marriages, divorces, and deaths for January-December 1989. *Monthly* vital statistics report. Vol. 38, Nos. 1-12. Hyattsville, Maryland: Public Health Service.

NATIONAL CENTER FOR HEALTH STATISTICS (1990). Births, marriages, divorces, and deaths for January-March 1990. *Monthly* vital statistics report. Vol. 39, Nos. 1-3. Hyattsville, Maryland: Public Health Service.

NATIONAL CENTER FOR HEALTH STATISTICS (1992). Advance report of final mortality statistics, 1989. *Monthly* vital statistics report. Vol. 40, No. 8, supp. 2. Hyattsville, Maryland: Public Health Service.

PRESTON, S.H. (1993). Demographic change in the United States, 1970-2050. *Demography and Retirement: The 21st Century*, (Sylvester Scheiber, éd.). New York: Praeger Press.

PRESTON, S.H., et COALE, A.J. (1982). Age structure, growth, attrition, and accession: a new synthesis. *Population Index*. 48(2), 217-59.

ROBINSON, J.G., AHMED, B., DAS GUPTA, P., et WOODROW, K.A. (1993). Estimation of population coverage in the 1990 United States census based on demographic analysis. *Journal of the American Statistical Association*, 88, 1061-1079.

ROBINSON, J.G., WORD, D.L., et SPENCER, G. (1991). Uncertainty for models to translate 1990 census concepts into historical racial classifications. 1990 Decennial Census, Preliminary Research and Evaluation Memorandum (PREM) No. 81, Demographic Analysis Evaluation Project D8.

ROSENWAIKE, I., et LOGUE, B. 1983. Accuracy of death certificate ages for the extreme aged. *Demography*, 20(4), 569-585.

SHRESTHA, L.B. (1993). Age Misreporting and its Effects on Old-Age Population and Death Registration Estimates: United States, 1970-1990. Unpublished doctoral dissertation, University of Pennsylvania, Population Studies Center.

SHRYOCK, H.S., et SIEGEL, J.S. (1976). *The Methods and Material of Demography*. Orlando, Florida: Academic Press, Inc. (Harcourt Brace Jovanovich, Publishers): Studies in Population Series.

SIEGEL, J.S. (1974). Estimates of coverage of the population by sex, race, and age in the 1970 census. *Demography*, 11(1), 1-23.

SIEGEL, J.S., et PASSEL, J.S. (1976). New estimates of the number of centenarians in the United States. *Journal of the American Statistical Association*. 71, 559-566.

SPRAGUE, T.B. (1880-81). Explanation of a new formula for interpolation. *Journal of the Institute of Actuaries*. 22, 270.

UNITED STATES BUREAU OF THE CENSUS (1972). *General Population Characteristics*. 1970 Census of Population and Housing. Final Report PC(1)-B1. U.S. Summary. Washington, DC: U.S. Government Printing Office.

UNITED STATES BUREAU OF THE CENSUS (1973). *The Medicare Record Check: an Evaluation of the Coverage of Persons 65 Years of Age and Over in the 1970 Census*. 1970 Census of Population: Evaluation and Research Program, PHC(E)-7. Washington, DC: U.S. Government Printing Office.

UNITED STATES BUREAU OF THE CENSUS (1974). *Estimates of Coverage of Population by Sex, Race, and Age: Demographic Analysis*. 1970 Census of Population: Evaluation and Research Program, PHC(E)-4. By Jacob S. Siegel. Washington, DC: U.S. Government Printing Office.

UNITED STATES BUREAU OF THE CENSUS (1975). *Accuracy of Data for Selected Population Characteristics as Measured by the 1970 CPS-Census Match*. 1970 Census of Population: Evaluation and Research Program, PHC(E)-11. Washington, DC: U.S. Government Printing Office.

UNITED STATES BUREAU OF THE CENSUS (1976). *Demographic Aspects of Aging and the Older Population in the United States*. Current Population Reports. Series P-23, No. 59. By J.S. Siegel. Washington, DC: U.S. Government Printing Office.

UNITED STATES BUREAU OF THE CENSUS (1983). *General Population Characteristics*. 1980 Census of Population and Housing. Final Report PC80-1-B1. United States Summary. Washington, DC: U.S. Government Printing Office.

UNITED STATES BUREAU OF THE CENSUS (1984a). *Demographic and Socioeconomic Aspects of Aging in the United States*. Current Population Reports. Series P-23, No. 138. By J.S. Siegel and M. Davidson. Washington, DC: US Government Printing Office.

UNITED STATES BUREAU OF THE CENSUS (1984b). Census of Population: 1980. Race detail file. 100% count. Table IV: modified counts (OMB-consistent) by age, race, and sex. Unpublished tabulations.

UNITED STATES BUREAU OF THE CENSUS (1988). *The Coverage of Population in the 1980 Census*. 1980 Census of Population and Housing: Evaluation and Research Reports, PHC80-E4. By R.E. Fay, J.S. Passel and J.G. Robinson. Washington, DC: U.S. Government Printing Office.

VAUPEL, J.W. (1993). Verbal presentation at Research Workshop on Oldest Old Mortality, Duke University, Durham, North Carolina. March, 1993.

WILKIN, J.C. (1981). Recent trends in the mortality of the aged. *Transactions of the Society of Actuaries*. Vol. XXXIII, 11-62.

WOLFENDEN, H.H. (1954). *Population Statistics and their Compilation*. Revised Edition. The University of Chicago Press: published for the Society of Actuaries.

WORD, D.L., et SPENCER, G. (1991). Age, sex, race, and Hispanic origin information from the 1990 census: a comparison of census results with results where age and race have been modified. 1990 CHS-L-74. Draft dated August 1991.

ZELNIK, M. (1969). Age patterns of mortality of American Negroes: 1900-02 to 1959-61. *Journal of the American Statistical Association*. 64, 433-451.

Techniques d'enquête, décembre 1995
Vol. 21, n° 2, pp. 193–199
Statistique Canada

Évaluation de l'utilisation des ordinateurs de poche pour la réalisation d'enquêtes démographiques dans les pays en développement

D. FORSTER et R.W. SNOW[1]

RÉSUMÉ

Les enquêtes à grande échelle réalisées dans les pays en développement peuvent fournir un instantané extrêmement utile de l'état de santé d'une collectivité, mais les résultats produits reflètent rarement la réalité actuelle puisqu'ils ne sont souvent diffusés que plusieurs mois ou même des années après la collecte des données. Ces retards sont en partie attribuables au temps nécessaire à la saisie, au codage et au traitement des données. Or, il est désormais possible de faire la saisie sur les lieux mêmes de l'enquête, grâce aux ordinateurs de poche. Les erreurs sont également moins nombreuses puisqu'une vérification automatique permet de rejeter les entrées illogiques ou incohérentes dès l'étape de l'administration du questionnaire. Le présent article porte sur l'utilisation d'une telle méthode d'enquête assistée par ordinateur pour la collecte de données démographiques au Kenya. Même si les coûts initiaux de la mise sur pied d'un tel système sont élevés, ses avantages sont évidents: les erreurs qui peuvent se glisser dans les données recueillies, malgré l'expérience du personnel technique, peuvent être réduites à des niveaux négligeables. Dans les situations où la rapidité d'exécution est essentielle, où les employés sont nombreux et où on a recours au précodage des réponses pour la collecte systématique de données sur une période de temps prolongée, l'entrevue assistée par ordinateur pourrait s'avérer économique à long terme, tout en améliorant radicalement la qualité des données à court terme.

MOTS CLÉS: Ordinateurs de poche; enquêtes démographiques; Psion.

1. INTRODUCTION

On procède régulièrement dans les pays en développement à des enquêtes à grande échelle qui peuvent rejoindre des dizaines de milliers de répondants (p. ex., recensements nationaux, enquêtes démographiques et sondages sur l'état de santé des populations). Ces études ont pour objectif de fournir rapidement des informations à jour sur les populations et leur état de santé, aux fins de l'évaluation et de la planification. Leur très grande portée exige la participation d'un grand nombre de formateurs, d'intervieweurs, de surveillants, de préposés à la saisie et de gestionnaires. L'Enquête mondiale sur la fécondité (EMF 1986) et les enquêtes nationales sur la démographie et la santé (DHS Kenya 1989) sont des exemples de telles enquêtes fondées sur des questionnaires. La comparaison des dates publiées du début des enquêtes EMF dans 12 pays africains et des dates de publication des premiers rapports nationaux (tableau 1) montre le temps qui peut s'écouler avant que les données ne deviennent accessibles aux planificateurs intéressés (EMF 1986). En moyenne, il faut patienter pendant 45.6 mois avant qu'un rapport final ne soit diffusé. Les retards accusés sont sans doute dus en partie à des problèmes logistiques inhérents aux pays en développement, mais la mécanique du traitement des données a aussi sa part de responsabilités. L'enquête sur la démographie et la santé récemment réalisée au Kenya a employé cinq préposés à la saisie, deux surveillants et un commis au contrôle pour traiter les données de 8,343 entrevues dans les ménages. La collecte des données a commencé en février 1989 et la première ébauche du rapport final a été diffusée sept mois plus tard (DHS Kenya 1989).

Tableau 1

Résumé chronologique des enquêtes mondiales sur la fécondité réalisées dans 12 pays africains (source: EMF 1986)

Pays	Nombre d'entrevues	Date du début de l'enquête	Date du premier rapport	Nombre de mois écoulés du début de l'enquête à la publication du rapport
Bénin	4,018	12/1981	06/1984	30
Cameroun	8,219	01/1978	04/1983	63
Ghana	6,125	02/1979	06/1983	52
Côte d'Ivoire	6,270	08/1980	12/1984	52
Kenya	8,100	08/1977	06/1980	34
Lesotho	3,603	08/1977	12/1981	52
Mauritanie	3,500	01/1981	06/1984	41
Maroc	5,800	04/1980	05/1984	49
Nigéria	9,727	10/1981	09/1984	35
Sénégal	3,985	05/1978	07/1981	38
Soudan (Nord)	3,115	12/1978	04/1982	40
Tunisie	4,123	05/1978	06/1983	61

[1] D. Forster, Department of Tropical Medicine, University of Oxford, John Radcliffe Hospital, Headington OX3 9DU, England; R.W. Snow, CRC - Research Unit, Kenyan Medical Research Institute, P.O. Box 230, Kilifi, Kenya.

Les enquêtes de cette ampleur comportent de nombreuses étapes de vérification et de codage des données brutes, qui constituent une autre source de retards. Or, comme la rapidité d'exécution est essentielle à l'évaluation de l'état de santé des populations (Anker 1991; Vlassoff et Tanner 1992), la réduction du temps consacré à ces activités de vérification et de codage peut constituer un avantage important pour les enquêtes. Les progrès réalisés dans le domaine du matériel informatique ont conduit à la mise au point de micro-ordinateurs utilisables sur le terrain. Par ailleurs, grâce aux nouveaux logiciels améliorés spécialement conçus pour l'administration des questionnaires, l'entrevue assistée par ordinateur est maintenant devenue une option envisageable. Les bureaux nationaux des statistiques de certains pays industrialisés ont évalué le recours à cette technique et certains l'utilisent régulièrement (Nicholls et Groves 1986; Lyberg 1985; Denteneer et coll. 1987; Bench et coll. 1994). Ces systèmes présentent l'avantage de réduire les risques d'erreurs en simplifiant le cheminement et en rejetant les entrées inexactes, illogiques ou incohérentes. En outre, ils permettent de stocker les données de très nombreuses entrevues et de les transférer ensuite simplement dans un ordinateur central à la fin de chaque période de travail, éliminant ainsi l'étape de la saisie.

Malgré les avantages apparents de cette technologie, son adoption se butte à des réticences surprenantes dans les pays en développement. Plusieurs raisons peuvent être invoquées pour expliquer ce problème. Premièrement, le coût initial peut paraître prohibitif pour des pays aux ressources limitées. Deuxièmement, il n'y a eu jusqu'à maintenant que peu de tentatives de validation de cette méthode en conditions réelles, et les données quantitatives sur ses avantages et ses inconvénients, qui pourraient en permettre la comparaison avec les méthodes de collecte classiques, sont donc encore limitées (Reitmaier 1985; Ferry et Cantrelle 1988; Forster et coll. 1991). Nous présentons ci-après les résultats d'une étude comparative des deux méthodes de collecte et de traitement des données utilisées au cours d'une étude démographique réalisée sur la côte kényane.

2. ENQUÊTE SUR LA MORTALITÉ DES ADULTES

Cette enquête a été réalisée dans le cadre d'une série d'études démographiques et épidémiologiques portant sur 60,000 habitants de la côte du Kenya. La population étudiée et les méthodes d'enquête démographique utilisées ont été décrites ailleurs (Snow et coll. 1994). En résumé, un recensement initial de la population est suivi d'un contrôle des données de l'état civil réalisé au moyen de visites des ménages individuels effectuées aux six semaines et de recensements semestriels de la population entière. Au cours d'un nouveau recensement de la population effectué en novembre 1993, on a procédé à une estimation de la mortalité des adultes à l'aide de méthodes démographiques indirectes (Timaeus 1991). Toutes les femmes âgées de 25 à 44 ans ont été interviewées à l'aide du questionnaire structuré reproduit à la figure 1. Les questions posées étaient fermées et précodées; le questionnaire comportait des branchements par sauts logiques ainsi qu'une question de vérification de la cohérence.

Vingt-quatre enquêteurs, tous diplômés du niveau secondaire, ont participé à l'étude. Ils étaient tous familiers avec les méthodes d'enquête et de recensement, ayant déjà reçu une formation officielle aux techniques d'enquête sur le terrain et justifiant tous de une à cinq années d'expérience pratique. Ils ont en plus reçu deux jours supplémentaires de formation à l'administration du questionnaire sur la mortalité des adultes. Pendant l'étude, les enquêteurs ont été répartis en deux équipes, chacune placée sous la direction d'un enquêteur principal. Les questionnaires remplis étaient vérifiés à la fin de chaque journée par les chefs d'équipe, puis transmis aux préposés à la saisie. On utilisait un logiciel FoxPro (version 2.0) qui reproduisait à l'écran le questionnaire utilisé. Les données ont été entrées deux fois par deux préposés différents et les dossiers ainsi complétés ont été comparés pour dépister les erreurs de saisie qui ont par la suite été corrigées. Les dossiers ont ensuite été soumis à une vérification logique et à des contrôles de vraisemblance et de cohérence; il s'agissait par exemple de relever les omissions, les codes erronés (p. ex., emploi de lettres autres que "Y" ou "N") et les discordances entre les dates indiquées, l'âge des répondantes et la date de l'enquête (figure 1; questions 5 et 6), et de s'assurer que la somme des réponses aux questions 7, 9, 11 et 13 concordait avec la réponse à la question 15.

3. TEST DE COLLECTE DES DONNÉES ASSISTÉE PAR ORDINATEUR

3.1 Matériel et logiciel

Une version antérieure du logiciel d'enquête avait été élaborée pour le Psion Organiser II (Forster et coll. 1991). Ce modèle comportait un écran aux dimensions limitées à 16 caractères sur deux lignes, mais utilisait un clavier complet. L'ordinateur Psion Series 3 utilisé pour la présente étude offre de nouvelles possibilités: l'écran est beaucoup plus grand (40 caractères sur 8 lignes) et autorise l'affichage graphique. L'appareil reste tout de même petit (165 × 85 × 22 mm) et ne pèse que 265 g, piles comprises (2 piles AA). Les dispositifs de stockage peuvent contenir jusqu'à un mégaoctet. Le clavier QWERTY comporte 58 touches. Le transfert des données entre le Psion Series 3 et un ordinateur personnel classique s'effectue à l'aide d'une simple opération de copie.

Le logiciel a été mis au point à l'aide du langage de programmation intégré du Psion: OPL. Le questionnaire-papier est reproduit sous une forme structurée, dans un fichier texte. La définition du questionnaire comprend un mélange de questions et de commandes (p. ex., pour le branchement par saut et le contrôle de vraisemblance). Les contrôles internes de vraisemblance comprennent ceux mis

Figure 1. Questionnaire sur la mortalité des adultes

Questionnaire sur la survie des parents
(pour toutes les femmes âgées de 25 à 44 ans)

Noms
Dates ID ⌊_⌋⌊_⌋-⌊_⌋⌊_⌋⌊_⌋-⌊_⌋⌊_⌋

J'aimerais vous poser quelques questions sur vos parents naturels ainsi que sur vos frères et soeurs qui ont la même mère que vous.

1. Votre mère est-elle vivante? (1 = oui, 2 = non) ⌊_⌋
2. Votre père est-il vivant? (1 = oui, 2 = non) ⌊_⌋
   INTERVIEWEUR: Si les deux parents sont vivants, (Q1 et Q2 = 1), passer à la question 7.
3. Avez-vous déjà accouché? (1 = oui, 2 = non) ⌊_⌋
   INTERVIEWEUR: Si la répondante n'a jamais accouché, (Q3 = 2), passer à la question 6.
4. Est-ce que (*MENTIONNER TOUS LES PARENTS QUI NE SONT PLUS VIVANTS*) était toujours vivant lors de la naissance de votre premier enfant?

	Oui	Non	Ne sais pas
Mère de la répondante	1	2	9
Père de la répondante	1	2	9

5. En quelle année votre premier enfant est-il né? ⌊_⌋⌊_⌋
6. En quelle année (*MENTIONNER TOUS LES PARENTS QUI NE SONT PLUS VIVANTS*) est-il mort?
   Mère de la répondante ⌊_⌋⌊_⌋
   Père de la répondante ⌊_⌋⌊_⌋
7. Combien avez-vous de soeurs nées de votre mère? (TOUJOURS VIVANTES) ⌊_⌋⌊_⌋
   INTERVIEWEUR: S'il n'y a pas de soeur vivante, (Q7 = 0), passer à la question 9.
8. Combien de ces soeurs toujours vivantes ont moins de 15 ans? ⌊_⌋⌊_⌋
9. Combien de vos soeurs, nées de votre mère, sont maintenant décédées? ⌊_⌋⌊_⌋
   INTERVIEWEUR: Si aucune soeur n'est décédée, (Q9 = 0), passer à la question 11.
10. Combien de ces soeurs aujourd'hui décédées sont mortes avant d'avoir 15 ans? ⌊_⌋⌊_⌋
11. Combien avez-vous de frères nés de votre mère? (TOUJOURS VIVANTS) ⌊_⌋⌊_⌋
    INTERVIEWEUR: S'il n'y a pas de frère vivant, (Q11 = 0), passer à la question 13.
12. Combien de ces frères toujours vivants ont moins de 15 ans? ⌊_⌋⌊_⌋
13. Combien de vos frères, nés de votre mère, sont maintenant décédés? ⌊_⌋⌊_⌋
    INTERVIEWEUR: Si aucun frère n'est décédé, (Q13 = 0), passer à la question 15.
14. Combien de ces frères aujourd'hui décédés sont morts avant d'avoir 15 ans? ⌊_⌋⌊_⌋
    INTERVIEWEUR: Faire la somme des questions 7, 9, 11 et 13: Q7 = ⌊_⌋⌊_⌋
    Q9 = ⌊_⌋⌊_⌋
    Q11 = ⌊_⌋⌊_⌋
    Q13 = ⌊_⌋⌊_⌋
15. Je veux m'assurer que j'ai bien compris. Exception faite de vous-même, votre mère a eu . . . . . ⌊_⌋⌊_⌋ enfants au total. Est-ce bien exact?
    INTERVIEWEUR: En cas de discordance, reprendre les questions 7 à 14 et corriger le cas échéant.

INTERVIEWEUR: Prière de remercier la répondante de sa coopération.

Code de l'intervieweur ⌊_⌋⌊_⌋

au point pour la vérification de la concordance des données entrées à l'aide du logiciel FoxPro décris plus haut. Les données entrées sur le Psion sont stockées dans un fichier séparé, à raison d'une ligne par entrevue.

Pour être indiquée correctement, la question doit inclure un numéro, le texte de la question et la réponse (choix d'options, caractère ou numéro). La définition devrait en outre préciser à quelle position sur la ligne l'entrée correspondante doit être stockée et quelle est sa longueur. Les réponses numériques peuvent également inclure un nombre préétabli de signes décimaux. La gamme d'entrées acceptables est une caractéristique optionnelle pour les réponses utilisant des chiffres ou des lettres; elle précisera un minimum, un maximum ou les deux à la fois. Les listes d'options peuvent servir à préciser les codes et leurs valeurs.

Les commandes peuvent être intégrées dans le texte des questions afin de permettre l'évaluation au moment de l'administration du questionnaire. Par exemple, la question de contre-vérification finale de la figure 1 requiert une addition. La syntaxe permet d'inclure cette instruction dans le corps du texte de la question. D'autres commandes peuvent contenir des instructions pour sauter une question, effectuer une contre-vérification entre deux réponses ou passer à une question différente.

Le logiciel utilise ainsi une méthode souple et d'application générale de définition du questionnaire. Il incorpore la manipulation des informations saisies et intègre les fonctions arithmétiques dans les questions ou les lignes de commande. La prochaine étape consisterait à élaborer une interface pour la détermination des caractéristiques du questionnaire qui nous éviterait d'avoir à élaborer une définition du questionnaire correcte au plan syntactique. On peut se procurer ce logiciel auprès des auteurs.

3.2 Conception du test

On a organisé une journée supplémentaire de formation à l'utilisation du Psion pour les deux chefs d'équipe. Cette formation comportait une explication du matériel et du logiciel ainsi que des séances d'exercice en conditions réelles. Ni l'un ni l'autre des deux chefs d'équipe ne possédait une expérience préalable en informatique. Les deux ont effectué des entrevues en alternant, chaque jour, l'utilisation du Psion et des questionnaires-papier. Ces entrevues ont servi de base à la comparaison des deux méthodes.

Les erreurs commises par les 22 intervieweurs utilisant le questionnaire-papier ont été comptées et totalisées aux fins de l'estimation du taux d'erreurs inhérent à cette méthode de collecte des données. Le temps consacré à la vérification des formulaires remplis, à la saisie, à la vérification et à la correction des données après les contrôles de vraisemblance et de concordance a été mesuré tout au long de l'enquête. Une évaluation semblable du temps d'exécution a été faite dans le cas de la collecte assistée par l'ordinateur Psion.

4. RÉSULTATS DES TESTS

4.1 Temps

Les entrevues effectuées avec les questionnaires-papier ont duré en moyenne 5.1 minutes, comparativement à 5.0 minutes pour celles réalisées à l'aide de l'ordinateur de poche; les deux méthodes n'étaient donc pas différentes sur ce plan (tableau 2; noter que 215 des 234 entrevues ont été chronométrées). La durée des entrevues variait considérablement (de 1 à 18 minutes); elle dépendait non seulement du nombre de questions sautées, mais également de la clarté et de la cohérence des réponses fournies.

Tableau 2

Comparaison de la méthode d'enquête classique (questionnaire-papier) et de la méthode d'enquête assistée par ordinateur (Psion Series 3)

	Durée de l'entrevue en minutes				
	Minimum général	Maximum général	Moyenne générale	Moyenne Équipe A	Moyenne Équipe B
Papier	1	16	5.1 (215)*	5.2 (128)	4.9 (87)
Ordinateur	1	18	5.0 (363)	5.5 (190)	4.5 (173)

* Nombre d'entrevues chronométrées.

Les chefs d'équipe ont pour leur part consacré 2 à 3 heures par jour à la vérification des questionnaires. Par ailleurs, il a fallu en moyenne à chaque préposé 3 heures et 40 minutes pour saisir les données de 500 dossiers (nombre approximatif d'entrevues réalisées par semaine); l'entrée des données en double a donc demandé 7 heures et 20 minutes (tableau 3). La vérification de ces questionnaires a demandé 2 heures et 23 minutes de plus. Les dossiers complétés ont été révisés deux fois pour entrer les corrections, puis vérifiés à nouveau; cette dernière opération a demandé en moyenne 2 heures et 30 minutes pour 500 dossiers.

Tableau 3

Temps nécessaire au traitement des données de 500 questionnaires

Activité	Temps moyen
Vérification des données	4 heures 8 minutes
Première saisie des données	3 heures 40 minutes
Seconde saisie des données	3 heures 40 minutes
Vérification	2 heures 23 minutes
Révision	2 heures 33 minutes
Total	18 heures 24 minutes

4.2 Erreurs

Les erreurs commises ont été comptabilisées sur deux périodes de deux semaines chacune, pour évaluer l'incidence de la familiarisation graduelle des employés avec leur travail. Les 22 intervieweurs (à l'exclusion des deux chefs d'équipe) ont commis 1,704 erreurs sur 1,427 questionnaires au cours de la première période, et 1,049 erreurs sur 1,158 questionnaires au cours de la seconde période. Le taux moyen d'erreurs par questionnaire a donc atteint 1.19 au cours de la première quinzaine, et 0.90 au cours de la seconde. En outre, sur l'ensemble de la période, il a fallu retourner 37 questionnaires (1.2% de toutes les entrevues) pour les faire reprendre parce que les erreurs qu'ils contenaient ne pouvaient être corrigées au bureau. Ces questionnaires contenaient entre 1 et 6 erreurs chacun, pour un total de 61. C'est la question 5 qui a donné lieu au plus grand nombre d'erreurs (17), suivie par la question 6b (15). La compilation des erreurs commises à chaque question est présentée au tableau 4. Quatorze des 22 intervieweurs ont dû reprendre au moins un questionnaire. L'un d'eux a dû en reprendre 8.

Tableau 4

Compilation des erreurs commises par les 22 intervieweurs utilisant des questionnaires-papier (pour le libellé des question, voir figure 1)

	Période 1 (première quinzaine)	Période 2 (seconde quinzaine)
Identification	163	48
Question 1	6	1
Question 2	8	2
Question 3	125	92
Question 4a	201	138
Question 4b	151	93
Question 5	105	61
Question 6a	94	57
Question 6b	65	41
Question 7	14	0
Question 8	109	63
Question 9	51	10
Question 10	178	134
Question 11	13	1
Question 12	108	71
Question 13	53	3
Question 14	204	149
Question 15	19	76
Code de l'intervieweur	37	9
Total des erreurs	**1,704**	**1,049**
Total des questionnaires	**1,427**	**1,158**

Les erreurs ont été détectées manuellement, lors de la vérification finale effectuée par un des enquêteurs (D. Forster), ou par le biais des contrôles de vraisemblance et de cohérence effectuées dans FoxPro sur les données saisies. Le chef de l'équipe A a commis 8 erreurs sur 144 questionnaires (0.06 erreur par questionnaire), et celui de l'équipe B en a commis 18 sur 90 questionnaires (0.20 erreur par questionnaire). La plupart de ces erreurs ont été commises à la question 10 (4 erreurs) et à la question 15 (5 erreurs). Les seules erreurs trouvées dans les données entrées sur ordinateur étaient des erreurs d'identification des répondants. On en a relevé 7 : 2 pour le chef de l'équipe A et 5 pour le chef de l'équipe B (taux d'erreurs de 0.01 et 0.03 par questionnaire respectivement). Ces erreurs auraient pu être évitées en entrant à l'avance, dans l'ordinateur de poche, la liste des personnes à interviewer.

4.3 Coût

Nous présentons au tableau 5 les coûts comparatifs d'une enquête de cette envergure effectuée à l'aide de l'ordinateur Psion et de questionnaires-papier. Les prix indiqués pour les ordinateurs de poche sont les prix de détail recommandés. Le jeu de la concurrence entre les détaillants pourrait entraîner une réduction des prix de ces ordinateurs atteignant 20% des prix indiqués dans ce tableau. Les prix du matériel informatique sont aussi en baisse. Compte tenu des prix en vigueur actuellement, le coût d'achat d'un système Psion pourrait être amorti après 12 à 15 enquêtes portant sur environ 7,000 répondants.

Tableau 5

Étude comparative de l'enquête assistée par ordinateur et de la méthode classique avec questionnaire-papier (coûts en £ sterling du Royaume-Uni)

	Équipement requis	Coût
Enquête assistée par ordinateur	20 ordinateurs Psion Series 3	2,539.00
	20 dispositifs de stockage de 1 Mo	2,039.00
	1 dispositif de communication en série	59.45
	80 piles rechargeables	146.20
	1 chargeur à piles	15.95
	Coût total	**4,799.59**
Enquête classique	14 rames de papier pour 7,000 entrevues	42.00
	Reproduction de 7,000 questionnaires	70.00
	20 stylos, gommes et liquide correcteur	27.40
	20 planchettes à pince	100.00
	2 préposés à la saisie des données (2 semaines)	70.00
	2 surveillants* (un mois plus surtemps)	85.00
	Coût total	**394.40**

* Nécessaires pour la vérification manuelle des formulaires à la fin de chaque journée d'entrevues.

5. DISCUSSION

Le taux d'erreurs le plus bas enregistré avec un questionnaire classique utilisé par des intervieweurs expérimentés justifiant de 5 années d'expérience dans la collecte des données était en moyenne de 0.11 erreur par questionnaire comportant 17 champs. Ces erreurs ont été pratiquement éliminées à l'aide du logiciel d'enquête mis au point pour un ordinateur de poche Psion Series 3. Cette méthode a éliminé la plupart des erreurs de cheminement commises par les intervieweurs dans l'administration du questionnaire (tableau 4) grâce à des modules préétablis de branchement par saut qui ont réduit le taux d'erreurs d'au moins 90%. Si on ajoutait une liste préétablie de répondants dans le logiciel, on pourrait par surcroît réduire le taux d'erreurs d'identification des répondants.

Les chefs d'équipe étaient heureux d'utiliser l'ordinateur; ils ont maîtrisé le clavier QWERTY et appris le fonctionnement du logiciel assez rapidement pour être en mesure de se débrouiller seuls au bout de deux jours. Même si aucune enquête officielle n'a été réalisée pour évaluer et quantifier les réactions des personnes interviewées à l'utilisation du Psion, le nombre de commentaires recueillis a été remarquablement faible et personne n'a refusé d'être interviewé.

Le traitement des données de l'enquête complète a employé deux préposés à la saisie qui ont travaillé à temps plein sur deux ordinateurs IBM pendant 92 heures. Un gestionnaire était sur place pour offrir son aide au besoin et déterminer le format de saisie des données. Il convient de souligner que les deux préposés à la saisie connaissaient bien les procédures de saisie des données ainsi que le matériel et le logiciel utilisés. Dans d'autres situations, il faudrait prévoir une surveillance plus étroite et les services d'un gestionnaire des données, ce qui augmenterait les coûts de l'opération. Un système informatisé de collecte des données Psion permettrait au gestionnaire de passer beaucoup moins de temps à transférer les données recueillies chaque jour, ce qui réduirait cet élément du coût du personnel. Néanmoins, le coût initial des ordinateurs Psion Series 3 risque d'être prohibitif, comparativement au coût prévisible du papier et de la reproduction des questionnaires, si on n'envisage pas de les utiliser dans le cadre d'activités futures de collecte des données.

QUESTOR (Ferry et Cantrelle 1988) offre un logiciel propice à la réalisation d'entrevues assistées par ordinateur. Toutefois, ce logiciel nécessite un ordinateur personnel beaucoup plus cher que l'ordinateur de poche Psion. Nous avons été à même de confirmer qu'il sera profitable de poursuivre la mise au point d'un ensemble d'outils appropriés intégrant la technologie des ordinateurs de poche compatibles avec les ordinateurs personnels et qui seront à la fois plus faciles à utiliser sur le terrain, plus robustes et moins énergivores.

Il faut toujours chercher un compromis entre la réduction des taux d'erreurs, le temps et le coût d'une enquête. Le recours aux entrevues assistées par ordinateur peut permettre de réduire à la fois les taux d'erreurs et le temps consacré à la préparation des données et ce, par une marge considérable. Le recours à un tel système de collecte devrait réduire les retards inacceptables dans la présentation des données observés lors d'enquêtes telles que l'Enquête mondiale sur la fécondité (tableau 1). La présente étude comparative s'est déroulée dans un contexte qui différait de celui de beaucoup des vastes enquêtes démographiques où le personnel recruté n'est pas familier avec l'administration des questionnaires. Nous croyons donc que les résultats dont nous faisons état représentent l'amélioration minimale à laquelle il est permis de s'attendre dans la qualité des données. Le coût initial de la mise en place d'un tel système d'enquêtes est peut-être impressionnant, mais il pourra être amorti si on peut l'utiliser pour des enquêtes répétées ou pour une série d'enquêtes de diverses natures.

REMERCIEMENTS

Les auteurs tiennent à remercier toutes les personnes qui ont participé à l'enquête et notamment les deux chefs d'équipe, Rodgers Chisengwa et Lewis Mitsanze, et les préposés à la saisie, Robert Mutai et Monica Omondi. Merci également à M. Ian Timaeus, qui a conçu le questionnaire sur la mortalité des adultes, et à M. Chris Nevill, qui a réalisé le nouveau recensement. Nous remercions M. Kevin Marsh pour son aide dans la réalisation des enquêtes assistées par ordinateur à Kilifi, le Wellcome Trust (Royaume-Uni) pour son aide financière, la société Psion PLC (Royaume-Uni) qui nous a donné un ordinateur de poche Psion Series 3 et le directeur du KEMRI qui nous a autorisés à publier ses résultats. Les travaux de M. Bob Snow sont financés en partie par le programme de bourses pour la recherche fondamentale en sciences biomédicales Wellcome Trust Senior Fellowship.

BIBLIOGRAPHIE

ANKER, M. (1991). Epidemiological and statistical methods for rapid health assessment. *World Health Statistics Quarterly*, 44, 94-97.

BENCH, J., CLARK, C., DUFOUR, J., et KAUSHAL, R. (1994). Computer-assisted interviewing for the labour force survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

DHS, KENYA (1989). Report on Kenyan Demographic and Health Survey. Institute for Resource Development/Macro Systems Inc., Columbia, Maryland, USA.

DENTENEER, D., BETHLEHEM, J.G., HUNDEPOOL, A.J., et KELLER, W.J. (1987). The BLAISE System for Computer-Assisted Processing, Automation in Survey Processing. Netherlands Bureau of Statistics, CBS Select No. 4, 67-76.

ENQUÊTE MONDIALE SUR LA FÉCONDITÉ (1986). Rapport final. Institut international de statistique, Pays-Bas.

FERRY, B., et CANTRELLE, P. (1988). L'utilisation de micro-ordinateurs de terrain pour la collecte en démographie. Dans *Congrès africain sur la population*, Dakar, Sénégal. Liege, Belgique: International Union for the Scientific Study of Population (IUSSP), 15-30.

FORSTER, D., BEHRENS, R.H., CAMPBELL, H., et BYASS, P. (1991). Evaluation of a computerized field data collection system for health surveys. *Bulletin of the World Health Organisation*, 69, 107-111.

FORSTER, D., et SNOW, R.W. (1992). Using microcomputers for rapid data collection in developing countries. *Health Policy and Planning*, 7, 667-71.

LYBERG, L. (1985). Plans for computer-assisted data collection at Statistics Sweden. *Bulletin de l'Institut International de Statistique*, actes de la 45ième session, communications demandées, tome LI, livre 3, section 18.2.

NICHOLLS, W.L., et GROVES, R.M. (1986). The status of computer-assisted telephone interviewing: Part I - Introduction and impact on cost and timeliness of survey data. *Journal of Official Statistics*, 2, 93-115.

REITMAIER, P., DUPRET, A., et CUTTING, W.A.M. (1987). Better health data with a portable microcomputer at the periphery: An anthropometric survey in Cape Verde. *Bulletin de l'Organisation mondiale de la santé*, 65, 651-657.

SNOW, R.W., MUNG'ALA, V.O., FORSTER, D., et MARSH, K. (1994). The role of the district hospital in child survival on the Kenyan Coast. *African Journal of Health Sciences*, 1, 71-75.

TIMAEUS, I.M. (1991). Measurement of adult mortality in less developed countries: a comparative review. *Population Index*, 57, 552-568.

VLASSOFF, C., et TANNER, M. (1992). The relevance of rapid assessment to health research and interventions. *Health Policy and Planning*, 7, 1-9.

Techniques d'enquête, décembre 1995
Vol. 21, n° 2, pp. 201–207
Statistique Canada

Contrôle statistique du processus relatif aux bases de sondage

A.W. SPISAK[1]

RÉSUMÉ

Le contrôle statistique du processus est un instrument qui peut servir à vérifier l'exactitude des bases de sondage construites périodiquement. La taille des bases de sondage est reportée sur un graphique de contrôle afin de déceler les causes spéciales de variation. On décrit les méthodes qui permettent de choisir le modèle chronologique (ARMMI) approprié pour des observations liées. On examine aussi les applications de l'analyse des séries chronologiques à la construction de graphiques de contrôle. Les données du programme de contrôle de la qualité des prestations d'assurance-chômage du United States Department of Labor servent à illustrer la méthode.

MOTS CLÉS: Autocorrélation; modèles ARMMI; graphiques de contrôle; contrôle de la qualité.

1. INTRODUCTION

L'intégrité de la base de sondage est d'une importance capitale dans le domaine de la recherche par sondage. Les imperfections de la base de sondage incluent les éléments manquants (base incomplète), les agrégats d'éléments (plus d'un élément sur une seule liste), les blancs ou les éléments étrangers, et les listes répétées. Ces imperfections peuvent causer plusieurs difficultés, car elles contribuent à l'erreur non due à l'échantillonnage, limitent le nombre d'unités d'échantillonnage tirées des sous-classes de la population, et obligent à utiliser des coefficients de pondération complexes pour estimer les caractéristiques de la population. Les méthodes qui visent à réduire au minimum les problèmes liés aux bases de sondage, ou à limiter leurs répercussions sur l'enquête, sont exposées en détail dans la plupart des traités sur les enquêtes statistiques.

Le présent article met l'accent sur le contrôle statistique du processus relatif aux bases de sondage construites périodiquement (quotidiennement, hebdomadairement ou mensuellement, par exemple) et constituées d'éléments générés selon un processus continu. Étant donné les fluctuations inhérentes à tout processus dynamique, la taille des bases de sondage varie inévitablement. Or, comment sait-on si les modifications de la taille des bases de sondage reflètent les variations aléatoires du processus, plutôt que des erreurs de construction? Le contrôle statistique du processus permet aux chargés d'enquête de faire la distinction entre les variations inhérentes au processus (causes ordinaires) et celles qui sont indicatrices d'un problème éventuel de construction de la base de sondage (causes spéciales).

2. VARIATIONS DU PROCESSUS ET CONTRÔLE STATISTIQUE DU PROCESSUS

Ces dernières années, les gestionnaires des secteurs privés de la fabrication et des services et ceux du secteur public de l'économie ont été de plus en plus nombreux à adopter les théories sur le contrôle de la qualité proposées par W. Edwards Deming, J.M. Juran, Philip B. Crosby, Kaoru Ishikawa et d'autres. La gestion de la qualité englobe tout un éventail d'instruments et de méthodes, dont le recours aux graphiques de contrôle pour déterminer si un processus est sous contrôle statistique. Selon Deming (1982), on arrive à ce dernier en éliminant les causes spéciales de variation, de sorte que seules subsistent les variations aléatoires d'un processus stable. L'évolution d'un processus sous contrôle statistique est prévisible.

Distinguer entre les causes ordinaires et spéciales de variation est un principe essentiel du contrôle statistique des processus. Deming (1982) attribue à Walter A. Shewhart, qui a défini nombre des principes du contrôle statistique durant les années 20 et 30, l'énoncication du concept des causes spéciales ou assignables. Les causes spéciales sont ordinairement attribuables à un élément du processus, comme un travailleur, une machine ou un bureau, et se manifestent répétitivement à moins d'être repérées et éliminées. Elles sont signalées par des points de données qui tombent en dehors des limites de contrôle, par des points consécutifs situés au-dessus ou au-dessous de la moyenne, ou par des séries d'observations de valeurs croissantes ou décroissantes.

Les causes ordinaires de variation, quant à elles, sont inhérentes au processus. Présentes en tout temps, elles ont une incidence sur l'ensemble du processus. On les limite ou on les élimine grâce à des mesures de gestion visant à modifier ce dernier.

3. APPLICATION DU CONTRÔLE STATISTIQUE DU PROCESSUS À LA CONSTRUCTION DE BASES DE SONDAGE POUR LES ENQUÊTES PÉRIODIQUES

3.1 Programme américain de contrôle de la qualité des prestations d'assurance-chômage

Le programme de contrôle de la qualité des prestations d'assurance-chômage du *United States Department of Labor* est un exemple de l'utilisation du contrôle statistique

[1] A.W. Spisak, Mathematical Statistician, Unemployment Insurance Service, U.S. Department of Labor, Washington, DC 20210, U.S.A.

du processus comme instrument de gestion de la qualité des bases de sondage. Depuis 1987, les 50 États, le district de Columbia et Porto Rico ont mis en oeuvre le programme de contrôle de la qualité des prestations d'assurance-chômage en collaboration avec le *United States Department of Labor*. L'objectif du programme consiste à limiter les paiements excédentaires ou insuffisants de prestations d'assurance-chômage, grâce à la détermination des causes d'erreur de paiement et à la prise de mesures visant à améliorer le processus de paiement des prestations.

Quand une personne présente une demande de prestation, le personnel de l'assurance-chômage détermine si elle satisfait tous les critères d'admissibilité - par exemple, le requérant a tiré de son emploi précédent des gains suffisants pour avoir droit à des prestations; le requérant est au chômage contre son gré; et, enfin, le requérant est capable et libre de travailler, et cherche activement un emploi. Si tous les critères d'admissibilité sont satisfaits, le Bureau d'assurance-chômage de l'État émet un chèque de prestation pour la semaine de chômage faisant l'objet de la demande.

3.2 Méthodes d'échantillonnage relatives au contrôle de la qualité des prestations et sources d'erreur

Chaque État tire hebdomadairement des échantillons aléatoires de versements au titre de l'assurance-chômage qu'on examine afin de déterminer si le requérant a reçu le bon montant. Si ce dernier est incorrect, l'enquêteur détermine les types d'erreurs et leurs causes, afin que les gestionnaires du programme puissent prendre des mesures correctives. Les bases de sondage sont construites chaque semaine en se fondant sur l'ensemble des prestations d'assurance-chômage versées entre le dimanche à 12h et le samedi suivant à 23h59. Un programme informatique permet de vérifier la base de données de l'État pour s'assurer que seuls les versements conformes à la définition opérationnelle de la population cible du programme soient inclus dans la base de sondage. Par exemple, on exclut de celle-ci les paiements effectués au titre de certains programmes d'assurance-chômage temporaires ou peu importants.

Le nombre de chèques d'assurance-chômage émis chaque semaine (et, par conséquent, la taille de la base de sondage) varie selon le nombre de personnes qui demandent des prestations et reçoivent ces dernières durant la semaine en question. Cependant, il existe plusieurs sources d'erreurs potentielles, susceptibles d'avoir une incidence sur l'intégrité de la base de sondage. Voici quelques-unes des erreurs les plus graves.

- Les paiements faits par certains bureaux locaux d'assurance-chômage risquent de ne pas être repris dans la base de données centrale de l'État, en raison de problèmes de télécommunication ou de TAD.
- Si l'État crée un fichier distinct pour chaque journée de transaction, les transactions d'un ou de plusieurs jours peuvent être omises par erreur dans le fichier cumulatif final.
- Le codage incorrect des transactions peut entraîner soit l'inclusion d'éléments étrangers à la base de sondage, ou l'élimination de transactions qui devraient y être incluses.

4. ANALYSE DES DONNÉES ET DÉVELOPPEMENT DU MODÈLE

La figure 1 représente un graphique chronologique de la taille des bases de sondage pendant une période de 52 semaines. La base de sondage de chaque semaine comprend les bénéficiaires de l'assurance-chômage de la semaine précédente qui continuent à recevoir une prestation, moins les bénéficiaires de l'assurance-chômage de la semaine précédente qui sont retournés au travail, n'ont plus droit à une prestation ou ont omis de présenter une demande, plus les requérants nouvellement admissibles ainsi que les requérants admissibles qui n'ont pas présenté de demande ou reçu de prestation pour une demande la semaine précédente.

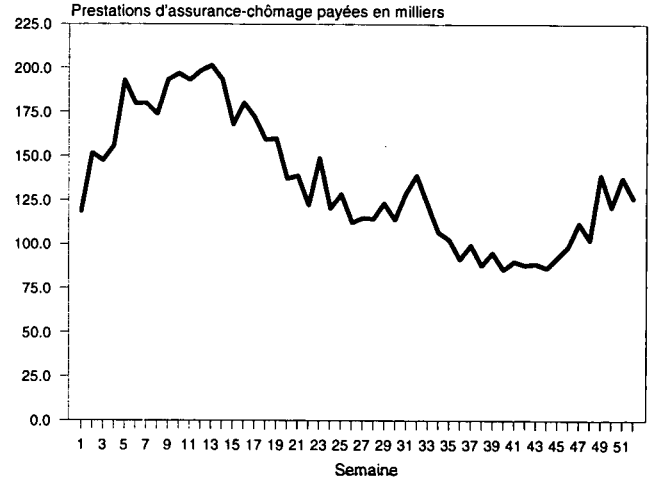


Figure 1. Nombre de prestations d'assurance-chômage payées par semaine.

On établit les graphiques de contrôle pour les observations individuelles en supposant que les données sont indépendantes et identiquement distribuées (i.i.d.). Toutefois, l'existence d'une corrélation sériale des données risque de fausser considérablement les estimations de la variance du processus (et, donc, les limites de contrôle). Aussi doit-on déterminer si les observations sont liées avant de pouvoir construire les graphiques de contrôle relatifs aux bases de sondage de l'assurance-chômage.

Le graphique de la série chronologique présenté à la figure 1 fournit une preuve *visuelle* que les observations ne sont pas indépendantes. Les données sur les bases de sondage affichent des tendances distinctes, caractérisées par l'augmentation des valeurs durant les 13 semaines du premier trimestre, puis leur diminution au cours des deux trimestres suivants, et enfin leur augmentation durant les 13 semaines du dernier trimestre. On peut tester l'existence de la corrélation sériale évoquée par la représentation graphique de la figure 1 au moyen des méthodes établies pour analyser les séries chronologiques. Bien qu'une discussion approfondie de l'analyse des séries chronologiques dépasse le cadre du présent article, on examinera

les concepts de stationnarité et d'autocorrélation, afin d'expliquer les méthodes utilisées pour choisir le modèle approprié. Les lecteurs qui connaissent mal les principes fondamentaux de l'analyse des séries chronologiques devraient consulter un des nombreux traités sur le sujet, en particulier Box et Jenkins (1976).

4.1 Stationnarité

On peut imaginer les observations individuelles qui constituent une série chronologique comme une collection de variables aléatoires à plusieurs dimensions - $p(z_1, \ldots, z_n)$ - où p est une densité de probabilité et où $z_1, \ldots, z_n$ sont des variables aléatoires. Si la distribution conjointe des variables aléatoires ne varie pas en fonction du temps, c'est-à-dire si $p(z_t, \ldots, z_{t+n}) = p(z_{t+m}, \ldots, z_{t+n+m})$, le processus est dit *strictement* stationnaire. En pratique, il est difficile d'établir une stationnarité stricte. Dans le cas de la présente application, on présume que la série chronologique est *faiblement* stationnaire. On parle aussi de stationnarité d'ordre 2, car les premier et deuxième moments du processus sont invariants en fonction du temps - $E(z_t) = E(z_{t+m})$, $\mathrm{VAR}(z_t) = \mathrm{VAR}(z_{t+m})$, et $\mathrm{COV}(z_t, z_{t+k}) = \mathrm{COV}(z_{t+m}, z_{t+k+m})$.

Jusqu'à la fin du présent article, les termes *stationnaire* ou *stationnarité* feront allusion à un processus qui répond aux critères de faible stationnarité.

4.2 Autocorrélation

Dans le cas d'une série chronologique stationnaire, la covariance entre deux observations dépend uniquement du nombre de périodes (décalages) qui les séparent - $\mathrm{COV}(z_t, z_{t+k}) = \mathrm{COV}(z_{t+m}, z_{t+k+m})$. La corrélation de z_t et de z_{t+k} est égale à $\mathrm{COV}(z_t, z_{t+k})/\mathrm{VAR}(z_t)$ et est représentée par la notation ρ_k, où k correspond au nombre de périodes entre les observations. Par exemple, ρ_1 représente la corrélation des observations de la série chronologique séparées par une période, et est égal à $\mathrm{COV}(z_t, z_{t+1})/\mathrm{VAR}(z_t)$. Une corrélation pour la période k reçoit le nom d'autocorrélation, parce qu'il s'agit de la corrélation entre des observations qui forment une série chronologique. Les autocorrélations pour les divers décalages peuvent être représentées sur un graphique, appelé corrélogramme, qui facilite le choix du modèle approprié pour une série chronologique.

4.3 Choix du modèle chronologique

La figure 2 représente le corrélogramme de la série chronologique, couvrant 52 semaines, du nombre de prestations hebdomadaires d'assurance-chômage payées dans les bases de sondage. Les autocorrélations diminuent ou "disparaissent" très lentement, phénomène qui est caractéristique d'un processus non stationnaire. (À nouveau, le lecteur trouvera dans Box (1976) et dans d'autres traités sur les séries chronologiques une discussion approfondie du choix du modèle.)

Le *calcul de différences* est une méthode qui permet de transformer une série non stationnaire en une série stationnaire. Le symbole B est l'opérateur de décalage qui, appliqué à z_t, fait reculer l'indice inférieur d'une période. Donc, la différence d'ordre 1 de z_t est $(1 - B)z_t = z_t - z_{t-1}$.

Décalage	Auto-corr.	Erreur-type	-1 -.75 -.5 -.25 0 .25 .5 .75 1	Box-Ljung	Prob.
1	.915	.135		46.086	.000
2	.900	.133		91.592	.000
3	.830	.132		131.116	.000
4	.778	.131		166.570	.000
5	.695	.129		195.749	.000
6	.631	.128		220.036	.000
7	.557	.127		239.375	.000
8	.468	.125		253.366	.000
9	.381	.124		262.822	.000
10	.298	.122		268.774	.000
11	.216	.121		271.979	.000
12	.157	.119		273.714	.000
13	.084	.118		274.220	.000
14	.056	.116		274.452	.000
15	-.012	.115		274.464	.000
16	-.040	.113		274.586	.000
17	-.084	.112		275.156	.000
18	-.099	.110		275.967	.000
19	-.144	.108		277.737	.000
20	-.178	.107		280.516	.000
21	-.221	.105		284.959	.000
22	-.263	.103		291.419	.000
23	-.297	.102		299.936	.000
24	-.328	.100		310.739	.000
25	-.366	.098		324.658	.000
26	-.371	.096		339.530	.000

Symboles du tracé: Autocorrélations * Limites fixées à deux erreurs-type.

Nombres total de cas: 52 Décalages de premier ordre calculables: 51

Figure 2. Autocorrélations de la série chronologique des prestations hebdomadaires d'assurance-chômage payées.

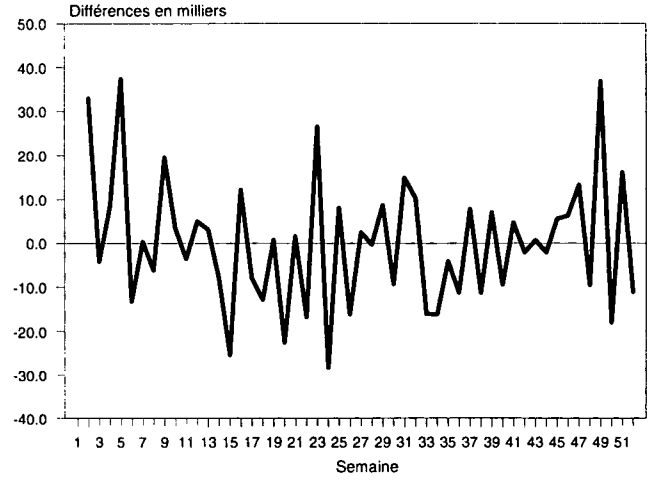


Figure 3. Différences de premier ordre des prestations d'assurance-chômage payées.

La figure 3 représente la série chronologique des différences $z_t - z_{t-1}$ pour les données sur les bases de sondage de l'assurance-chômage. Cette série paraît stationnaire, autour d'une moyenne égale à zéro. (L'estimation de la moyenne d'échantillon pour les différences est 150.8, avec une erreur-type de 2,064.0. La variable à tester $t = (150.8 - 0)/2{,}064$ est égale à .07, et on ne peut donc rejeter l'hypothèse selon laquelle $\mu = 0$). Pour d'autres séries chronologiques, le calcul de différences d'ordre 1 pourrait ne pas permettre d'atteindre la stationnarité. Le cas échéant, il est nécessaire de recourir à des transformations telles que le calcul de différences d'ordre 2 -

$(1 - B)^2 z_t = (z_t - z_{t-1}) - (z_{t-1} - z_{t-2})$, le calcul de différences saisonnières, ou encore, à des transformations logarithmiques ou à d'autres procédés de stabilisation de la variance.

Les autocorrélations des différences d'ordre 1 de la série chronologique, que l'on trouve à la figure 4, sont compatibles avec un processus stationnaire. Les autocorrélations diminuent rapidement, tandis que les autocorrélations partielles (non représentées) disparaissent après le premier décalage. Ces observations donnent à penser qu'on peut modéliser les données au moyen d'un processus autorégressif intégré d'ordre 1, ARI (1,1). Le terme d'autorégression AR indique qu'on n'estimera qu'un seul paramètre autorégressif, et le caractère d'intégration (I) montre qu'on a tranformé la série chronologique originale en calculant les différences d'ordre 1.

Décalage	Auto-corr.	Erreur-type	Box-Ljung	Prob.
1	-.367	.136	7.293	.007
2	.314	.135	12.731	.002
3	-.080	.133	13.089	.004
4	.196	.132	15.294	.004
5	-.082	.130	15.685	.008
6	.027	.129	15.728	.015
7	.074	.128	16.065	.025
8	.022	.126	16.095	.041
9	-.025	.125	16.135	.064
10	-.005	.123	16.137	.096
11	-.129	.122	17.254	.101
12	.089	.120	17.807	.122
13	-.280	.119	23.401	.037
14	.204	.117	26.443	.023
15	-.222	.115	30.133	.011
16	.089	.114	30.743	.015
17	-.191	.112	33.658	.009
18	.189	.110	36.597	.006
19	-.090	.109	37.287	.007
20	.075	.107	37.776	.009
21	-.048	.105	37.983	.013
22	-.016	.104	38.007	.018
23	-.049	.102	38.241	.024
24	.042	.100	38.415	.031
25	-.221	.098	43.481	.012
26	.184	.096	47.147	.007

Symboles du tracé: Autocorrélations * Limites fixées à deux erreurs-type.

Nombres total de cas: 52 Décalages de premier ordre calculables après le calcul de différences: 50

Figure 4. Autocorrélations des différences de premier ordre des prestations hebdomadaires d'assurance-chômage payées.

4.4 Estimation du modèle

On a estimé le modèle grâce à la procédure ARMMI du logiciel SPSS Trends (version 4.0), qui est basée sur les travaux de Box et Jenkins.

Le modèle provisoire est:

$$z_t = (1 + \phi_1)z_{t-1} - \phi_1 z_{t-2} + e_t, \quad \text{ou}$$

$$z_t - z_{t-1} = \phi_1(z_{t-1} - z_{t-2}) + e_t,$$

où ϕ_1 est le paramètre autorégressif d'ordre 1, et e est le terme d'erreur, qu'on suppose distribué normalement, avec une moyenne égale à 0 et une variance σ_e^2. L'estimation du paramètre autorégressif, ϕ_1', est égale à $-.4045$, et l'estimation de la variance résiduelle, $\sigma_e^{2\prime}$, est 184,275,853 (avec 50 degrés de liberté). Le signe négatif du paramètre AR est compatible avec les signes alternants des autocorrélations à la figure 4. Le modèle n'inclut pas de constante, car l'estimation de la moyenne du processus ne diffère pas de zéro de façon significative.

4.5 Vérifications du modèle

On peut déterminer si le modèle estimé convient aux données observées en examinant les résidus. Si le modèle est bien ajusté aux données, les résidus (e_t) devraient être des "bruits blancs", c'est-à-dire non corrélés. La figure 5 montre les autocorrélations des résidus du modèle. Bien que l'autocorrélation pour le décalage 13 soit significative, la valeur de la variable statistique Q de Box-Ljung n'est pas significative jusqu'au décalage 13. (La variable statistique Q teste la signification des autocorrélations pour les décalages 1 à k. Pour une discussion approfondie, consulter Box et Pierce (1970)). En outre, aucune autocorrélation partielle (non représentée dans la figure) n'est significative. Ces résultats indiquent qu'il n'existe pas de corrélation sériale des résidus.

Décalage	Auto-corr.	Erreur-type	Box-Ljung	Prob.
1	.118	.136	.757	.384
2	.218	.135	3.382	.184
3	.130	.133	4.333	.228
4	.205	.132	6.753	.150
5	-.007	.130	6.755	.239
6	.041	.129	6.858	.334
7	.128	.128	7.858	.345
8	.056	.126	8.054	.428
9	-.025	.125	8.095	.525
10	-.077	.123	8.491	.581
11	-.132	.122	9.676	.560
12	-.069	.120	10.009	.615
13	-.242	.119	14.175	.362
14	.042	.117	14.303	.427
15	-.164	.115	16.331	.360
16	-.071	.114	16.721	.404
17	-.133	.112	18.129	.381
18	.119	.110	19.283	.375
19	-.009	.109	19.290	.438
20	.030	.107	19.370	.498
21	-.029	.105	19.447	.556
22	-.061	.104	19.799	.596
23	-.057	.102	20.113	.635
24	-.066	.100	20.549	.665
25	-.191	.098	24.352	.499
26	.096	.096	25.356	.499

Symboles du tracé: Autocorrélations * Limites fixées à deux erreurs-type.

Nombres total de cas: 52 Décalages de premier ordre calculables: 50

Figure 5. Autocorrélations des résidus du modèle chronologique.

Pour tester l'hypothèse selon laquelle la distribution des résidus du modèle est normale, $N(0,\sigma_e^2)$, on a effectué le test de validité de l'ajustement de Kolmogorov-Smirnov (K-S). Pour l'estimation de la variance égale à 184,275,853, la variable à tester est égale à .591 ($p = .876$) et on ne peut donc rejeter l'hypothèse selon laquelle la distribution des différences est normale.

Pour que le processus AR (1) soit stationnaire, la valeur absolue du paramètre autorégressif doit être inférieure à un. Afin de tester l'hypothèse que $|\phi_1| \geq 1$ pour le modèle, on calcule: $t = (|\phi_1'| - 1)/SE(\phi_1')$, où $|\phi_1'|$ est la valeur absolue de l'estimation du paramètre autorégressif, et $SE(\phi_1')$, l'erreur-type de ϕ_1'. Le modèle statistique donne $t = (.4045 - 1)/.1295$, ou $t = -4.6$. La probabilité que la valeur absolue de ϕ_1' soit aussi petite que .4045, si la valeur absolue réelle de $\phi_1 \geq 1$, est très faible (< 0.00001). On rejette donc l'hypothèse que $|\phi_1| \geq 1$, et on conclut que la série des différences d'ordre 1 est stationnaire.

5. UTILISATION DU MODÈLE ARMMI DANS UN GRAPHIQUE DE CONTRÔLE

5.1 Graphiques de contrôle pour des observations individuelles

On fixe les limites de contrôle pour le graphique d'observations individuelles à $\bar{x} \pm 3\sigma'$, où $\bar{x}$ représente la moyenne des observations et σ', l'écart-type estimé du processus. Ryan (1989) examine d'autres méthodes d'estimation de l'écart-type du processus en calculant soit la moyenne des étendues mobiles (la moyenne des écarts absolus entre des observations successives) ou l'écart-type (s) des observations d'échantillon, $\sigma' = s/c$, où c est une constante d'ajustement qui dépend de la taille de l'échantillon.

Quand il existe une corrélation sériale des données, l'utilisation de l'écart-type de l'échantillon ou de l'étendue mobile moyenne aboutit parfois à de mauvaises estimations de σ. Les limites de contrôle construites pour ces estimations peuvent donner des résultats très trompeurs, soit parce qu'elles produisent des signaux erronés donnant à penser que le processus est hors de contrôle, ou qu'elles empêchent de déceler certaines causes spéciales de la variation du processus. Le calcul de l'étendue moyenne risque de donner une sous-estimation de σ, car les écarts entre les observations successives sont généralement faibles quand ces dernières sont fortement corrélées. La sous-estimation de σ se traduira par des limites de contrôle trop rapprochées et par une augmentation de la fréquence des signaux indicateurs de causes spéciales. Selon Ryan, quand les données sont corrélées, le calcul de l'écart-type de l'échantillon pour estimer l'écart-type du processus fournit une meilleure estimation de σ que celui de l'étendue mobile moyenne, à condition que l'échantillon comprenne au moins 50 observations. Cependant, l'écart-type de l'échantillon n'est un estimateur non biaisé de σ que quand les observations sont indépendantes.

Vasilopoulos et Stamboulis (1978) ont analysé l'effet de la corrélation sériale des données sur les limites de contrôle des graphiques de $\bar{x}$ et s (écart-type), et développé des équations définissant certains facteurs qui permettent d'ajuster les limites de contrôle des données produites par un processus autorégressif. Autrement, on peut définir un modèle chronologique pour les données corrélées et construire un graphique de contrôle en se servant des résidus du modèle pour surveiller le processus. Cette méthode est décrite par Berthouex, Hunter et Pallesen (1978) pour des sous-groupes de mesures de variables environnementales collectées par des stations d'épuration de l'eau. Alwan et Roberts (1988) se servent des résidus de modèles à moyenne mobile pondérée exponentiellement (MMPE) pour les séries chronologiques tant stationnaires que non stationnaires. Montgomery et Mastrangelo (1991) utilisent les résidus d'un modèle autorégressif dans un graphique MMPE et prétendent que ce type de graphique peut servir d'approximation à bon nombre de modèles autocorrélés, particulièrement si la corrélation des observations est positive et que la moyenne ne dévie pas trop rapidement. En outre, le lecteur trouvera dans Maragah et Woodall (1992) et dans Woodall et Faltin (1993) une discussion supplémentaire des effets de l'autocorrélation sur les méthodes de contrôle statistique des processus.

5.2 Graphiques de contrôle des données sur l'assurance-chômage

La figure 6 représente un graphique de contrôle des résidus ($e_t = z_t - z_t'$) du modèle ARI (1,1) choisi pour les données sur les bases de sondage relatives à l'assurance-chômage. Puisque les vérifications du modèle confirment que les résidus sont indépendants et identiquement distribués (i.i.d.) $N(0,\sigma_e^2)$, on a normalisé ces derniers, de sorte que la ligne centrale du graphique corresponde à la valeur 0, et fixé les limites de contrôle à ± 3. Le graphique inclut les résidus du modèle pour les tailles de la base de sondage observées durant la période de référence de 52 semaines et durant le trimestre suivant. Pour la semaine 56, la différence entre la taille de la base de sondage et la valeur prédite par le modèle tombe au-delà de la limite de contrôle supérieure, ce qui indique l'existence d'une cause spéciale de variation.

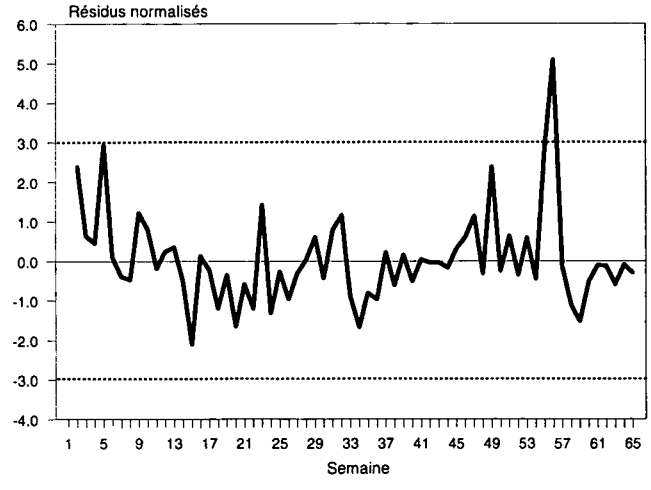


Figure 6. Graphique de contrôle des résidus du modèle (données de référence + trimestre suivant).

Au lieu de représenter graphiquement les résidus du modèle, on peut construire un graphique de contrôle de la taille des bases de sondage. Au besoin, on doit transformer les observations originales pour atteindre la stationnarité. On se sert des estimations des paramètres du modèle chronologique pour déterminer la moyenne et les limites de contrôle du graphique. La variance d'un processus AR(1) est $\sigma^2 = \sigma_e^2/(1 - \phi_1^2)$. Pour le modèle chronologique des différences d'ordre 1, ϕ_1' est égal à $-.4045$, et l'estimation de la variance résiduelle, $\sigma_e'^2$, est 184,275,853. L'estimation de la variance du processus est 184,275,853/(1 − .1636), ou 220,325,579.4, et l'écart-type du processus est égal à 14,843.4. On fixe les limites de contrôle inférieure et supérieure à $\pm 3\sigma'$, soit $\pm 44{,}530.2$, par rapport à la différence moyenne estimée, égale à zéro. Le graphique de contrôle (figure 7) indique qu'il existe une cause spéciale de variation pour l'observation 56, tout comme le graphique de contrôle des résidus présenté à la figure 6.

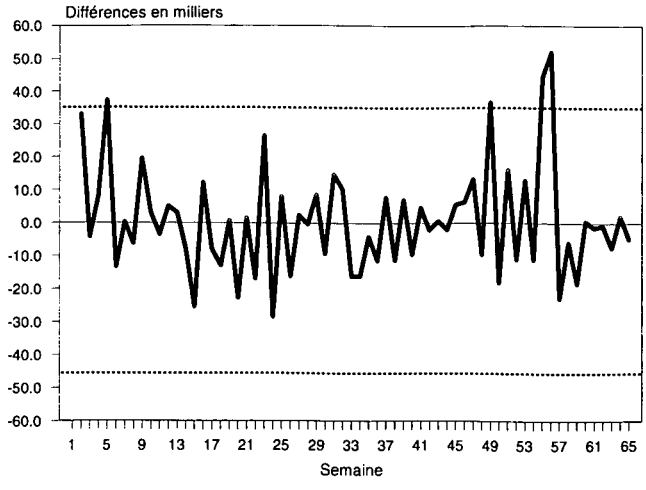


Figure 7. Graphique de contrôle des prestations d'assurance-chômage payées (différences de premier ordre – données de référence + trimestre suivant).

6. CONCLUSIONS

Le contrôle statistique du processus est un instrument de contrôle de la qualité utile pour les enquêtes dont les échantillons sont tirés de bases de sondage construites pour des périodes particulières, par un procédé continu. Comme les tailles des bases de sondage constituent une série chronologique, les données sont parfois liées et doivent alors être transformées pour atteindre la stationnarité. Si les observations sont corrélées, il convient de choisir le modèle chronologique (ARMMI) approprié pour estimer la variance du processus utilisée pour fixer les limites de contrôle. Dans le cas de l'exemple mentionné plus haut, la série chronologique s'accorde avec un modèle autorégressif intégré (à calcul de différences) d'ordre 1 – ARI (1,1). De façon plus générale, on peut décrire les séries chronologiques au moyen d'autres modèles ARMMI (p,d,q), où p est le nombre de termes autorégressifs dans le modèle, d est le degré de calcul de différences nécessaire pour atteindre la stationnarité, et q est le nombre de termes à moyenne mobile dans le modèle. Les modèles chronologiques saisonniers incluent des paramètres AR, MM et de calcul de différences pour le ou les décalage(s) approprié(s).

Une fois le modèle déterminé à partir des données de référence, on peut reporter sur le graphique de contrôle les observations faites pour les périodes subséquentes. Les graphiques de contrôle des figures 6 et 7 comportent les observations de la période de référence de 52 semaines, et celles faites durant le trimestre suivant (13 semaines). Le modèle chronologique devrait être vérifié périodiquement, en fonction de l'intervalle de collecte des données, afin de s'assurer que les paramètres n'ont pas changé.

Quand les procédures de contrôle statistique du processus signalent l'existence d'une cause spéciale de variation, les chargés d'enquête doivent se servir d'autres instruments de gestion de la qualité pour déterminer la cause originelle des problèmes liés à la base de sondage, puis appliquer des mesures correctives pour améliorer les procédures d'enquête. L'élimination systématique des causes assignables de variation mises en évidence grâce au contrôle statistique du processus permettent aux chargés d'enquête de passer du diagnostic des anomalies et de la correction des erreurs à l'amélioration continue du processus d'enquête.

Dans le cas des données sur les bases de sondage relatives à l'assurance-chômage, la cause spéciale de variation était inévitable: la semaine où le nombre de versements au titre de l'assurance-chômage a augmenté brusquement suivait une semaine de travail écourtée en raison d'un congé et coïncidait avec une mise à pied dans un grand établissement. La grande base de sondage observée cette semaine-là n'était donc pas le résultat d'un problème technique de construction. Dans d'autres États, à d'autres périodes, le contrôle statistique du processus a permis de déceler des erreurs très diverses, dont une erreur de saisie de données (déclaration d'une base de sondage égale à 558,432 au lieu de 5,558,432), l'omission des transactions effectuées au titre de l'assurance-chômage durant une des cinq journées de travail, ce qui a réduit d'environ 20% la taille de la base de sondage, et, enfin, l'oubli d'entrer les révisions dans le logiciel de sélection de l'échantillon, avec pour conséquence l'insertion d'éléments étrangers dans la base de sondage.

Outre la vérification de l'intégrité des bases de sondage, la méthode décrite dans le présent article s'applique à d'autres domaines de la gestion des enquêtes et de l'information. On peut l'utiliser pour diminuer l'erreur non due à l'échantillonnage attribuable à l'enregistrement ou à la saisie des données dans le cas des enquêtes effectuées quotidiennement, mensuellement, *etc.* De façon plus générale, le contrôle statistique du processus permet de confirmer l'intégrité des bases de données ou des systèmes de gestion de l'information, dans toutes les situations où l'information est collectée ou enregistrée en sous-groupes, notamment quand les données sont collectées à des sites multiples ou par plusieurs chercheurs ou vérificateurs.

REMERCIEMENTS

L'auteur remercie les réviseurs pour leur suggestions et leurs commentaires judicieux.

BIBLIOGRAPHIE

ALWAN, L.C., et ROBERTS, H.V. (1988). Time series modeling for statistical process control. *Journal of Business and Economic Statistics*, 6, 87-95.

BERTHOUEX, P.M., HUNTER, W.G., et PALLESEN, L. (1978). Monitoring sewage treatment plants: some quality control aspects. *Journal of Quality Technology*, 10, 139-149.

BOX, G.E.P., et PIERCE, D.A. (1970). Distribution of residual autocorrelations in autoregressive moving average time series models. *Journal of the American Statistical Association*, 65, 1509-1526.

BOX, G.E.P., et JENKINS, G.M. (1976). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

DEMING, W.E. (1982). *Quality, Productivity, and Competitive Position*. Cambridge: Massachusetts Institute of Technology Center for Advanced Engineering Study.

MARAGAH, H.D., et WOODALL, W.H. (1992). The effect of autocorrelation on the retrospective X-chart. *Journal of Statistical Computation and Simulation*, 40, 29-42.

MONTGOMERY, D.C., et MASTRANGELO C.M. (1991). Some statistical process control methods for autocorrelated data. *Journal of Quality Technology*, 23, 179-204.

RYAN, T.P. (1989). *Statistical Methods for Quality Improvement*, New York: John Wiley and Sons.

VASILOPOULOS, A.V., et STAMBOULIS, A.P. (1978). Modification of control chart limits in the presence of data correlation. *Journal of Quality Technology*, 10, 20-30.

WOODALL, W.H., et FALTIN, F.W. (1993). Autocorrelated data and SPC. *American Society for Quality Control Statistics Division Newsletter*, 13, 18-21.

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont accepté de faire la critique d'un article durant l'année 1995. Un astérisque indique que la personne a participé plus d'une fois.

C. Alexander, *U.S. Bureau of the Census*
R. Bell, *The Rand Corporation*
*D.R. Bellhouse, *University of Western Ontario*
N. Bennett, *Yale University*
*D.A. Binder, *Statistique Canada*
J.-R. Boudreau, *Statistique Canada*
J. Brehm, *Duke University*
F.J. Breidt, *Iowa State University*
S.J. Butani, *U.S. Bureau of Labor Statistics*
B.D. Causey, *U.S. Bureau of the Census*
R.L. Chambers, *Australian National University*
G. Chen, *University of Regina*
J. Chen, *University of Waterloo*
G.H. Choudhry, *Statistique Canada*
M.L. Cohen, *National Academy of Sciences*
M.J. Colledge, *Australian Bureau of Statistics*
J.L. Czajka, *Mathematica Policy Research*
T. DeMaio, *U.S. Bureau of the Census*
J. Denis, *Statistique Canada*
J.-C. Deville, *INSEE*
J.D. Drew, *Statistique Canada*
J.-J. Droesbeke, *Université Libre de Bruxelles*
D. Findley, *U.S. Bureau of the Census*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistique Canada*
M. Gonzalez, *U.S. Office of Management and Budget*
R.M. Groves, *University of Maryland*
K.P. Hapuarachchi, *Statistique Canada*
M.A. Hidiroglou, *Statistique Canada*
D. Hill, *University of Michigan*
*D. Holt, *Central Statistical Office, U.K.*
C.T. Ireland, *U.S. National Security Agency*
S. Itzhaki, *Hebrew University*
G. Kalton, *Westat, Inc.*
P.N. Kokic, *Australian Bureau of Agricultural and Resource Economics*
P.S. Kott, *National Agricultural Statistical Service*
M. Kovacevic, *Statistique Canada*
R.A. Kulka, *Research Triangle Institute*
S. Kumar, *Statistique Canada*
*N. Laniel, *Statistique Canada*
M. Latouche, *Statistique Canada*
*P. Lavallée, *Statistique Canada*
L. Lazzeroni, *Stanford University*
*H. Lee, *Statistique Canada*
N. Luther, *East-West Center*
*L.Mach, *Statistique Canada*
T.K. Mak, *Concordia University*
*H. Mantel, *Statistique Canada*
A. Mason, *East-West Center*
P. Merkouris, *Statistique Canada*
*D. Pfeffermann, *Hebrew University*
H. Pold, *Statistique Canada*
R.F. Potthoff, *Duke University*
B. Quenneville, *Statistique Canada*
T.E. Raghunathan, *University of Michigan*
É. Rancourt, *Statistique Canada*
*J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
K. Rust, *Westat, Inc.*
I. Sande, *Bell Communications Research, U.S.A.*
C.-E. Särndal, *Université de Montréal*
J. Schafer, *Pennsylvania State University*
*W.L. Schaible, *U.S. Bureau of Labor Statistics*
*F.J. Scheuren, *George Washington University*
*J. Sedransk, *State University of New York - Albany*
R. Sigman, *U.S. Bureau of the Census*
M. Simard, *Statistique Canada*
*A. Singh, *Statistique Canada*
*M.P. Singh, *Statistique Canada*
R.P. Singh, *U.S. Bureau of the Census*
*C.J. Skinner, *University of Southampton*
*D. Stukel, *Statistique Canada*
*A. Théberge, *Statistique Canada*
Y. Tillé, *Université Libre de Bruxelles*
M. Traugott, *University of Michigan*
J. Trépanier, *Statistique Canada*
R. Valliant, *U.S. Bureau of Labor Statistics*
J. Waite, *U.S. Bureau of the Census*
*J. Waksberg, *Westat, Inc.*
S. Wing, *Synetics*
W.E. Winkler, *U.S. Bureau of the Census*
*K.M. Wolter, *National Opinion Research Center*
*A. Zaslavsky, *Harvard University*

On remercie également ceux qui ont contribué à la production des numéros de la revue pour 1995: S. Beauchamp (photocomposition), L. Rousseau et S. Cadieux (Langues officielles et traduction). Finalement on désire exprimer notre reconnaissance à S. DiLoreto, S.F. Bertrand, C. Larabie et D. Lemire de la Division des méthodes d'enquêtes-ménages, pour leur apport à la coordination, la dactylographie et la rédaction.

The Canadian Journal of Statistics **La Revue Canadienne de Statistique**

CONTENTS TABLE DES MATIÈRES

Volume 23, No. 3, September/Septembre 1995

Research papers/Articles

Case study in data analysis
Étude de cas en analyse des données

CONTENTS TABLE DES MATIÈRES

Volume 23, No. 4, December/Décembre 1995

JOURNAL OF OFFICIAL STATISTICS

An International Quarterly Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey Methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 11, Number 1, 1995

Contents Volume 11, Number 2, 1995

Contents Volume 11, Number 3, 1995

Lars Lyberg, U/LEDN, Statistics Sweden, S-115 81 Stockholm, Sweden

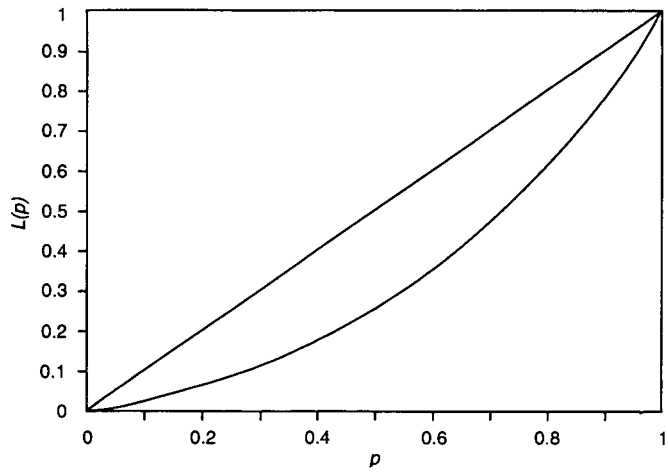


Figure 1. Courbe de Lorenz pour la distribution de Weibull avec $\alpha = 1.6$ comme paramètre de forme.

Le coefficient de Gini détermine le degré d'inégalité de la distribution du revenu. On le définit notamment comme une fonction linéaire de la surface située entre la courbe de Lorenz et l'axe de 45°, normalisée pour rester entre 0 et 1. À la figure 1, le coefficient de Gini est égal à 0.35. La définition formelle du coefficient de Gini (Nygård et Sandström 1981) est la suivante:

$$G = 1 - 2\int_0^1 L(p)dp = \frac{1}{\mu}\int_0^\infty [2F(y) - 1]ydF(y).$$

Nygård et Sandström (1981) donnent le groupe plus général de coefficients de Gini suivant:

$$G_J = \frac{1}{\mu_Y}\int_0^\infty J[F(y)]ydF(y), \tag{1.2}$$

où J est une fonction continue et liée. Avec le coefficient de Gini habituel, $J(p) = 2p - 1$.

Certains économistes recourent à une autre mesure de l'inégalité du revenu, soit la mesure du faible revenu qui représente la part de la population dont le revenu est inférieur à la moitié du revenu médian de la population. On l'exprime formellement comme suit:

$$\Theta = \int_0^{M/2} dF(y), \tag{1.3a}$$

où M est la médiane définie par

$$\int_0^M dF(y) = \frac{1}{2}. \tag{1.3b}$$

Le paramètre Θ dont nous avons besoin pour ces différentes mesures est la solution à l'équation suivante:

$$\int u(y,\Theta)dF(y) = 0,$$

où $u(y,\Theta)$ est le noyau de l'équation d'estimation. Nous reviendrons à la formulation de cette dernière à la partie 2. Les équations d'estimation des mesures qui précèdent et la valeur approximative de leur erreur quadratique moyenne apparaissent aux parties 3, 4 et 5. La partie 6 présente les estimateurs des mesures pour un plan d'échantillonnage complexe. Enfin, la partie 7 illustre la méthode au moyen des données de l'Enquête sur les finances des consommateurs du Canada.

2. APPLICATION DES ÉQUATIONS D'ESTIMATION À UNE POPULATION FINIE

La théorie permettant d'estimer les moyennes et les totaux d'une population finie est très bien développée dans les ouvrages de statistique contemporains. Särndal, Swensson et Wretman (1992) donnent une formule qui englobe la plupart des estimateurs utilisés dans la pratique. Nous passerons ici brièvement en revue cette théorie et montrerons comment il est possible de l'appliquer à des statistiques plus complexes grâce à des équations d'estimation, comme le décrivent Binder (1991) et Binder et Patak (1994).

Pour exposer l'idée principale, commençons par examiner l'estimation de la population totale T_Y et la fonction de distribution de la population finie $F(y)$. L'estimation de la population totale se trouve à la base même de l'approche des équations d'estimation de Binder (1991) et de Binder et Patak (1994). Supposons que la population totale de la variable Y soit décrite par

$$T_Y = N\int ydF(y).$$

Soulignons que $F(y)$ est une fonction en escalier correspondant à la fonction de distribution de la population finie. Nous utiliserons des estimateurs semblables à

$$\hat{T}_Y = \sum_{i\in s} w_i(s)y_i = \sum_{i=1}^{N} w_i(s)Y_i, \tag{2.1}$$

où $w_i(s)$ est égal à zéro lorsque la i-ème unité ne fait pas partie de l'échantillon. Par exemple, l'expression (2.1) donne l'estimateur non biaisé d'Horvitz-Thompson (HT) si

$$w_i(s) = \begin{cases} 1/\pi_i, & i\in s, \\ 0, & i\notin s, \end{cases}$$

ou l'estimateur de régression général si

$$w_i(s) = \begin{cases} [1 + (T_X - \hat{T}_X)x_i/\hat{T}_{X2}]/\pi_i, & i\in s, \\ 0, & i\notin s, \end{cases}$$

Techniques d'enquête, décembre 1995
Vol. 21, n° 2, pp. 151–159
Statistique Canada

Estimation de l'inégalité du revenu d'après les données d'enquête: application de la méthode des équations d'estimation

DAVID A. BINDER et MILORAD S. KOVACEVIC[1]

RÉSUMÉ

Les auteurs exposent brièvement quelques-uns des principaux aspects de la théorie des fonctions d'estimation pour les populations finies. Plus précisément, ils abordent la question de l'estimation des moyennes et des totaux et étendent la théorie pertinente aux fonctions d'estimation qu'ils appliquent ensuite au problème de l'estimation de l'inégalité du revenu. Les statistiques qui résultent de l'exercice expriment des fonctions non linéaires des observations, certaines d'entre elles dépendant de l'ordre des observations ou quantiles. Par conséquent, l'erreur quadratique moyenne des estimations ne peut être représentée par une formule simple ni être estimée par les méthodes classiques d'estimation de la variance. Les auteurs montrent que, pour les fonctions d'estimation, il est possible de résoudre ce problème par la méthode de linéarisation de Taylor. Enfin, ils font la démonstration de la méthode proposée en utilisant les données relatives au revenu provenant de l'Enquête sur les finances des consommateurs du Canada et comparent cette méthode à celle du jackknife avec suppression d'une grappe.

MOTS CLÉS: Plan d'enquête complexe; coefficient de la famille de Gini; ordonnée de la courbe de Lorenz; mesure du faible revenu, part du quantile.

1. INTRODUCTION

Les ouvrages d'économétrie traitent abondamment des façons de mesurer et d'analyser l'inégalité économique, tant sur le plan théorique que du point de vue des applications, quoiqu'on accorde la préférence aux questions théoriques. On s'est moins intéressé, par contre, à l'estimation de l'inégalité et à l'incidence de la structure des enquêtes par sondage. Les ouvrages d'économétrie pertinents mentionnent rarement l'estimation de la variance, pourtant inévitable dans les inférences statistiques issues de l'estimation de l'inégalité. On contourne habituellement la difficulté en formulant de très fortes hypothèses et en simplifiant à outrance la structure de l'enquête ou les formules servant à obtenir la variance approximative. Nous proposons ici une méthode qui se prête à l'estimation de l'inégalité du revenu et à l'estimation de la variance des statistiques non linéaires qui en dérivent. Cette méthode peut s'appliquer à divers plans d'échantillonnage.

En règle générale, la répartition de la population peut être décrite par sa fonction de distribution cumulative $F(y) = \Pr\{Y \leq y\}$, où Y est la variable aléatoire représentant la sélection d'un sujet au hasard. Nous supposerons ici que Y n'a pas de valeur négative. Si Y représente le revenu, nous nous intéressons à ses propriétés de distribution, c'est-à-dire à la concentration du revenu, à la part du revenu détenue par certains groupes de la population, à la proportion du faible revenu, *etc.* La fonction de quantile $\xi(p) = F^{-1}(p) = \inf\{y \mid F(y) \geq p\}$ suscite aussi notre intérêt.

La courbe de Lorenz, par exemple, compare le revenu cumulatif à la part de la population. La définition officielle de l'ordonnée de la courbe de Lorenz correspondant au $100p$-ième percentile de la population est

$$L(p) = \frac{\int_0^{\xi_p} y dF(y)}{\mu_Y}, \qquad (1.1)$$

où

$$\int_0^{\xi_p} dF(y) = p, \quad \text{et} \quad \int_0^{\infty} y dF(y) = \mu_Y.$$

La population finie représentée par (1.1) peut aussi être exprimée sous une autre forme, que les statisticiens jugeront plus familière:

$$L(p) = \sum_U Y_i I\{Y_i \leq \xi_p\} \Big/ \sum_U Y_i,$$

où U correspond à la population finie et $I\{.\}$ est une fonction indicatrice.

La part du revenu (quantile) est égale au pourcentage du revenu total que partage la population associée au quantile de revenu $[\xi_{p_1}, \xi_{p_2}]$, $p_1 \leq p_2$. Cette valeur est égale à la différence entre deux ordonnées de la courbe de Lorenz

$$Q(p_1,p_2) = L(p_2) - L(p_1).$$

La figure 1 illustre graphiquement la courbe de Lorenz pour la distribution de Weibull, avec pour paramètre de forme $\alpha = 1.6$, par rapport à l'axe de 45°. On remarque sur le graphique que la couche défavorisée, qui représente la moitié de la population, reçoit au maximum 25% du revenu total alors que la couche aisée (10% de la population) touche 20% du revenu total disponible.

[1] David A. Binder, Directeur, Division des méthodes d'enquêtes-entreprises, et Milorad S. Kovacevic, méthodologiste principal, Division des méthodes d'enquêtes-ménages, Statistique Canada, immeuble R.H. Coats, Parc Tunney, Ottawa (Ontario), Canada, K1A 0T6.

8. CONCLUSION

Dans un sondage en deux phases, lorsque deux informations différentes sont disponibles pour l'ensemble de la population d'une part et l'échantillon issu de la première phase d'autre part, plusieurs stratégies sont possibles lorsqu'on souhaite utiliser l'information auxiliaire pour améliorer l'estimation de totaux.

Deux approches naturelles différentes ont été utilisées pour dériver les estimateurs: une approche assistée par modèle de régression qui cherche à adapter l'idée de l'estimateur par régression, et une approche calage qui tente d'adapter l'idée du calage. Les estimateurs obtenus par les deux approches peuvent être reliés. On a fait apparaître 3 modélisations sous jacentes alternatives auxquelles peuvent se rattacher l'ensemble des estimateurs obtenus. On obtient ainsi trois classes d'estimateurs. Quelques stratégies de calage auxquelles on pouvait penser ont été éliminées d'emblée comme non pertinentes.

On a montré que les estimateurs d'une même classe, c'est-à-dire rattachés à un même modèle sont asymptotiquement équivalents et on donne la forme des variances dérivées dans le cas d'une fonction de calage linéaire, ces résultats restent valables compte tenu des équivalences asymptotiques pour une fonction de calage quelconque.

La forme des variances indique, conformément à l'intuition qu'une des classes d'estimateurs (estimateurs 3, 6 et 7 (stratégies de calage 3 et 4)) domine l'autre du point de vue de la variance, lorsque l'on se fonde uniquement sur le plan de sondage. Lorsqu'elle s'appuie sur une modélisation de la variable d'intérêt, l'évaluation des stratégies, conduit à privilégier la classe d'estimateurs associée à la modélisation adoptée.

Dans un contexte où l'on souhaite effectuer le redressement d'une enquête, et que l'on souhaite simultanément corriger les biais qu'induiraient l'utilisation des pondérations brutes et réduire la variance, les conclusions doivent être adaptées. Les modifications introduites dans les pondérations pour corriger les biais sont plus importantes que les corrections pour réduction de variance. Dès lors, on intégrera dans le calage les variables, dès que l'on pense qu'elles interviennent dans la probabilité de sélection et donc participent à la création du biais.

Lorsque les variables auxiliaires sont qualitatives, l'arbitrage entre utilisation a priori et a posteriori de l'information auxiliaire, c'est-à-dire entre, utilisation au niveau de l'échantillonnage ou au niveau du redressement, repose encore sur la distinction entre les deux modélisations de la variable d'intérêt.

Ces résultats se généralisent sans difficulté au cas d'un sondage à plus de deux phases.

REMERCIEMENTS

Je remercie chaleureusement Jean-Claude Deville, Louis Meuric et Carl-Erik Särndal pour toutes les suggestions dont ce texte a pu bénéficier.

BIBLIOGRAPHIE

CASSEL, C.M., SÄRNDAL, C.-E., et WRETMAN, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite population. *Biometrica*, 63, 615-620.

DEVILLE, J.-C. (1992). Constrained samples, conditional inference, weighting: three aspects of the utilisation of auxiliary information. *Proceedings of the Workshop on Uses of Auxiliary Information in Surveys*, October 1992, Örebro.

DEVILLE, J.-C., et SÄRNDAL, C.-E. (1992). Calibration estimators and generalized raking techniques in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DEVILLE, J.-C., SÄRNDAL, C.-E., et SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

DEVILLE, J.-C., et DUPONT, F. (1993). Calage et redressement de la non-réponse totale. Journées de Méthodologie.

DUPONT, F. (1994). Redressements alternatifs en présence de plusieurs niveaux d'information auxiliaire. Document de travail de la Direction des Statistiques Démographiques et Sociales, F9409.

FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā* C, 37, 117-132.

GOURIEROUX, C. (1981). *Théorie des sondages*. Edition Economica Paris.

ISAKI, C.T., et FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

SÄRNDAL, C.-E. (1980). On π inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.

SÄRNDAL, C.-E., et SWENSSON, B. (1987). A general view of estimation for two phases of selection with applications to two-phases sampling and nonresponse. *International Statistical Review*, 55, 279-294.

SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN, J. (1991). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

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

en apportant une réduction de variance (*cf.* Deville et Dupont 1993). Cependant, asymptotiquement, les corrections de biais à apporter aux poids sont plus importantes que les modifications à apporter pour améliorer les estimateurs. C'est donc le modèle de réponse implicite qui va orienter le choix entre les différents estimateurs:

Le modèle de réponse implicite de la première classe d'estimateurs est $p_i = 1/F(x_{i2}'B_2)$

Le modèle de réponse implicite de la deuxième et la troisième classe d'estimateurs est $p_i = 1/F(x_{i2}'A_2 + x_{i1}'A_1)$.

- Quelque soit le modèle de réponse, l'évaluation des trois classes d'estimateurs sur la base du seul plan de sondage donne encore la préférence à la troisième puisqu'elle convient pour tous les modèles de réponse.
- Si on évalue les stratégies sur la base d'une modélisation par régression, on n'utilisera la première classe d'estimateurs que si le mécanisme de réponse est bien expliqué par x_2 c'est-à-dire: $p_i = 1/F(x_{i2}'B_2)$. Or, on a vu que la modélisation associée à la première classe d'estimateurs prend son sens lorsque les variables x_1 et x_2 sont très corrélées. Il est donc assez probable dans ce contexte, que la variable x_2 suffise à expliquer le mécanisme de réponse. Dans le cas contraire, il faudrait s'orienter vers la troisième classe d'estimateurs.

La comparaison entre les trois stratégies peut donc être adaptée dans un contexte où on souhaite corriger les biais introduits par les sélections non contrôlées. Les conclusions restent sensiblement les mêmes.

Il est bien sûr possible d'établir selon le même principe, des comparaisons entre stratégies de redressement alternatives dans un contexte de sondages comportant plus de deux phases, et une ou plusieurs sélections non contrôlées.

7. UTILISATION A PRIORI ET A POSTERIORI DE L'INFORMATION AUXILIAIRE

L'estimateur par calage permet d'améliorer a posteriori l'estimation, en réduisant la variance et en corrigeant le biais, comme cela vient d'être mentionné. Toutefois, on peut souhaiter intégrer l'information auxiliaire a priori, au stade du sondage plutôt qu'a posteriori au niveau de l'estimation. On retrouve alors, dans un contexte plus complexe, l'opposition classique entre stratification ou poststratification, bien connue dans le cas d'un sondage en une phase, lorsque toutes les variables auxiliaires sont qualitatives.

Il est possible de transposer les termes de l'arbitrage entre, utilisation de l'information a priori et a posteriori, dans la configuration de sondage et d'information auxiliaire étudiée, lorsque les variables auxiliaires sont qualitatives. En effet, lorsque les variables auxiliaires sont qualitatives, un calage correspond exactement à une poststratification.

Nous avons vu précédemment, que pour déterminer la procédure de redressement adéquate, il fallait distinguer deux modélisations de la variable d'intérêt possibles, selon que les informations contenues dans x_1 et x_2, sont jugées substituables ou complémentaires. Chacune de ces deux modélisations conduisait alors à une ou des procédures de redressement différentes. De la même façon, ces deux modélisations apparaissent lorsqu'on s'interroge sur la stratégie d'échantillonnage qui permet d'intégrer au mieux l'information auxiliaire:

- Lorsque l'information contenue dans x_1 et celle contenue dans x_2 sont substituables, la modélisation de la variable d'intérêt y est:

  (1) $y_i = x_{i1}b_1 + u_{i1}$ et

  (2) $y_i = x_{i2}b_2 + u_{i2}$ où le second modèle est de meilleure qualité pour prédire la valeur de y_i.

  Nous avons vu que l'utilisation de l'information auxiliaire a posteriori conduit à la stratégie de calage n° 1, soit à la première classe d'estimateurs. Si l'on souhaite tenir compte de l'information auxiliaire au niveau du sondage, il est naturel de proposer, un tirage stratifié en x_1 pour la première phase et un tirage stratifié en x_2 pour la deuxième phase.

  Toutefois, le parallèle entre procédure de redressement et procédure d'échantillonnage n'est pas complet: dans un calage, on peut n'utiliser que l'information marginale en x_1. On réalise alors une poststratification incomplète (Särndal et Deville 1992). En revanche, dans la procédure d'échantillonnage proposée comme alternative a priori, on est amené à utiliser tous les croisements des variables x_1. L'équivalent a priori d'un calage serait alors un sondage équilibré sur les marges du vecteur de variables x_1.

- Lorsque l'information contenue dans x_1 et celle contenue dans x_2 sont complémentaires, la modélisation de la variable d'intérêt y est $y_i = x_{i1}b_1 + x_{i2}b_2 + u_i$. Nous avons vu que dans ce cas l'utilisation de l'information auxiliaire a posteriori conduisait aux stratégies de calage 2, 3 et 4 aux classes d'estimateurs 2 et 3. Si l'on souhaite tenir compte de l'information auxiliaire au niveau du sondage, il est naturel de proposer, un tirage stratifié en $x1$ pour la première phase et un tirage stratifié en x_1 et x_2 pour la deuxième phase.

  De la même façon que précédemment, il n'y a pas un parallèle exact entre les procédures a priori et a posteriori puisque l'utilisation de l'information a priori mobilise tous les croisements entre les variables x_1 et x_2.

Ainsi, il est possible d'effectuer un arbitrage entre intégrer l'information a priori ou a posteriori, voire d'optimiser le plan de sondage, lorsque les variables auxiliaires sont qualitatives. Les termes de l'arbitrage sont les mêmes que dans un sondage en une phase avec un seul niveau d'information. Il s'y ajoute la multiplicité des strates crées par les croisements de x_1 et x_2 dans le cas où la modélisation utilisée est $y_i = x_{i1}b_1 + x_{i2}b_2 + u_i$ qui renforce les avantages de l'utilisation a posteriori de l'information.

Lorsque les variables auxiliaires sont quantitatives, l'arbitrage passe par leur transformation en variables qualitatives, l'extension correcte ne pouvant être réalisée qu'en utilisant le parallèle entre calage et sondage équilibré (*cf.* Deville 1992).

$$\hat{V}(\hat{Y}_4) =$$

$$\left(\sum_{i,j\in s}\sum_{j\in s}\frac{\Delta^1_{ij}}{\pi_{ij}\pi_{aij}}\frac{\tilde{u}_{1i}\tilde{u}_{1j}}{\pi_i\pi_j}\right)+\left(\sum_{i,j\in s}\frac{\Delta^2_{ij}}{\pi_{aij}}\frac{\tilde{u}_{2i}\tilde{u}_{2j}}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

avec:
$$\tilde{u}_{1i} = y_i - x'_{i1}\hat{B}_1,$$
$$\tilde{u}_{2i} = y_i - x'_{i2}\hat{B}_2.$$

5.2 Estimateurs $\hat{Y}_2 = \hat{Y}_5$: modèle $y_i = x'_{i1}a_1 + x'_{i2}a_2 + u_i$

On montre (voir Dupont 1994) facilement que:

$$\mathrm{VA}(\hat{Y}_2) = \mathrm{VA}(\hat{Y}_5) \cong$$

$$\left(\sum_{i,j\in U}\Delta^1_{ij}\frac{v_iv_j}{\pi_i\pi_j}\right)+\left(E_{s_a}\sum_{i,j\in s_a}\Delta^2_{ij}\frac{u_iu_j}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

avec:
$$v_i = y_i - x'_{i1}a_1,$$
$$u_i = y_i - x'_{i1}a_1 - x'_{i2}a_2$$

$$a = \begin{pmatrix}a_1\\a_2\end{pmatrix} = \begin{pmatrix}\sum_{i\in U}x_{i1}x'_{i1} & \sum_{i\in U}x_{i1}x'_{i2}\\ \sum_{i\in U}x_{i2}x'_{i1} & \sum_{i\in U}x_{i2}x'_{i2}\end{pmatrix}\begin{pmatrix}\sum_{i\in U}x_{i1}y_i\\ \sum_{i\in U}x_{i2}y_i\end{pmatrix}.$$

On en déduit que:

$$\hat{V}(\hat{Y}_2) = \hat{V}(\hat{Y}_5) =$$

$$\left(\sum_{i,j\in s}\frac{\Delta^1_{ij}}{\pi_{ij}\pi_{aij}}\frac{\hat{v}_i\hat{v}_j}{\pi_i\pi_j}\right)+\left(\sum_{i,j\in s}\frac{\Delta^2_{ij}}{\pi_{aij}}\frac{\hat{u}_i\hat{u}_j}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

avec:
$$\hat{v}_i = y_i - x'_{i1}\hat{a}_1,$$
$$\hat{u}_i = y_i - x'_{i1}\hat{a}_1 - x'_{i2}\hat{a}_2.$$

On retrouve dans cette écriture que par construction de $\hat{Y}_2 - \hat{Y}_5$, x_1 réduit la variance apportée par la première phase et x_1 et x_2 sont utilisés simultanément pour réduire la variance apportée par la deuxième phase.

5.3 Estimateurs $\hat{Y}_3$, $\hat{Y}_6$ et $\hat{Y}_7$: modèle $y_i = x'_{i1}c_1 + M_{x_1}(x_{i2})'c_2 + u_i$

On montre que $\mathrm{VA}(\hat{Y}_6) = \mathrm{VA}(\hat{Y}_7) = \mathrm{VA}(\hat{Y}_3)$. Ainsi:

$$\mathrm{VA}(\hat{Y}_3) = \mathrm{VA}(\hat{Y}_6) = \mathrm{VA}(\hat{Y}_7) \cong$$

$$\left(\sum_{i,j\in U}\Delta^1_{ij}\frac{u_{1i}u_{1j}}{\pi_i\pi_j}\right)+\left(E_{s_a}\sum_{i,j\in s_a}\Delta^2_{ij}\frac{u_iu_j}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

$$u_{1i} = y_i - x'_{i1}c_1 = y_i - x'_{i1}b_1,$$
$$u_i = y_i - x'_{i1}c_1 - M_{x_1}x'_{i2}c_2 = y_i - x'_{i1}a_1 - x'_{i2}a_2.$$

On en déduit les trois estimateurs de variance qui diffèrent en raison de coefficients estimés différents:

$$\hat{V}(\hat{Y}_3) =$$

$$\left(\sum_{i,j\in s}\frac{\Delta^1_{ij}}{\pi_{ij}\pi_{aij}}\frac{\hat{u}_{1i}\hat{u}_{1j}}{\pi_i\pi_j}\right)+\left(\sum_{i,j\in s}\frac{\Delta^2_{ij}}{\pi_{aij}}\frac{\hat{u}_i\hat{u}_j}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

$$\hat{u}_{1i} = y_i - x'_{i1}\hat{c}_1,$$
$$\hat{u}_i = y_i - x'_{i1}\hat{c}_1 - M_{x_1}x'_{i2}\hat{c}_2,$$

$$\hat{V}(\hat{Y}_6) =$$

$$\left(\sum_{i,j\in s}\frac{\Delta^1_{ij}}{\pi_{ij}\pi_{aij}}\frac{\tilde{u}_{1i}\tilde{u}_{1j}}{\pi_i\pi_j}\right)+\left(\sum_{i,j\in s}\frac{\Delta^2_{ij}}{\pi_{aij}}\frac{\tilde{u}_i\tilde{u}_j}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

$$\tilde{u}_{1i} = y_i - x'_{i1}\hat{C}_1,$$
$$\tilde{u}_i = y_i - x'_{i1}\hat{a}_1 - x'_{i2}\hat{a}_2,$$

$$\hat{V}(\hat{Y}_7) =$$

$$\left(\sum_{i,j\in s}\frac{\Delta^1_{ij}}{\pi_{ij}\pi_{aij}}\frac{\tilde{u}_{1i}\tilde{u}_{1j}}{\pi_i\pi_j}\right)+\left(\sum_{i,j\in s}\frac{\Delta^2_{ij}}{\pi_{aij}}\frac{\tilde{u}_i\tilde{u}_j}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

$$\tilde{u}_{1i} = y_i - x'_{i1}\hat{C}^*_1,$$
$$\tilde{u}_i = y_i - x'_{i1}\hat{a}^*_1 - x'_{i2}\hat{a}^*_2,$$
$$\hat{C}_1 = \hat{a}^*_1 + \left(\sum_{s_a}\frac{x_1x'_1}{\pi_a}\right)^{-1}\left(\sum_{s_a}\frac{x_1x'_2}{\pi_a}\right)\hat{a}^*_2.$$

On retrouve que par construction de $\hat{Y}_3$, $\hat{Y}_6$ et $\hat{Y}_7$, x_1 est utilisée pour réduire *au maximum* la variance apportée par la première phase et x_2 permet de réduire la variance apportée par la deuxième phase.

6. CHOIX D'ESTIMATEURS EN PRÉSENCE DE BIAIS DE SÉLECTION

En pratique, lorqu'on effectue le redressement d'une enquête, il est fréquent que l'on souhaite non seulement améliorer l'estimation, mais aussi, et surtout corriger les biais qu'introduisent les sélections non contrôlées d'individus, comme la non réponse.

Étudions le cas d'un sondage en deux phases, dont la deuxième phase correspond à la non réponse totale. Les poids π_i du sondage $2^{\text{ième}}$ phase sont alors inconnus. Le calage de s va permettre d'estimer ces probabilités, tout

En effet, posons $w = y - x_1' \hat{a}_1^* - x_2' \hat{a}_2^*$. Alors $\hat{Y}_7 = \hat{Y}_6 + [\hat{\hat{W}}^* - \hat{\hat{W}}]$. Or, asymptotiquement $[\hat{\hat{W}}^* - \hat{\hat{W}}]$ est un infiniment petit négligeable devant $\hat{Y}_6$:

$$[\hat{\hat{W}}^* - \hat{\hat{W}}] = \left(\sum_s \frac{wx_1'}{\pi_a \pi}\right)\left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a}\right)^{-1}[X_1 - \hat{\hat{X}}_1],$$

et

$\left(\sum_s \frac{wx_1'}{\pi_a \pi}\right)$ tend vers zéro et $[X_1 - \hat{\hat{X}}_1] = O\left(\frac{1}{\sqrt{m}}\right)$.

On obtient finalement: $\hat{Y}_7 \cong \hat{Y}_6$.

En conclusion, lorsque la fonction de calage est exponentielle l'estimateur $\hat{Y}_7$ coïncide exactement avec le précédent. Lorsque F est linéaire, $\hat{Y}_7$ est proche du précédent et correspond donc encore à l'approche modèle de régression dans le cas où l'information contenue dans x_1 est jugée complémentaire à l'information contenue dans x_2 pour estimer y et où on utilise la décomposition $y_i = x_{i1}' c_1 + M_{x_1}(x_{i2})' c_2 + u_i$.

Conclusion: les trois classes d'estimateurs

On vient de voir que les quatre stratégies dérivées d'une approche par calage pouvaient être associées à une modélisation par régression. On obtient ainsi trois classes d'estimateur qui sont:

$Y_4 \cong Y_1$ associés aux modèles

$$(1) \quad y_i = x_{i1}' b_1 + u_{i1},$$

et

$$(2) \quad y_i = x_{i2}' b_2 + u_{i2}$$

$\hat{Y}_5 = \hat{Y}_2$ associé au modèle

$$y_i = x_{i1}' a_1 + x_{i2}' a_2 + u_i$$

$\hat{Y}_6 \cong \hat{Y}_3$ et $\hat{Y}_7 \cong \hat{Y}_3$ associés au modèle

$$y_i = x_{i1}' c_1 + M_{x_1}(x_{i2})' c_2 + u_1.$$

L'approximation $\cong$ qui indique que les estimateurs sont rattachés au même modèle de régression prend tout son sens lorsqu'on s'intéresse au calcul de variance de ces différents estimateurs. Les estimateurs qui sont rattachés au même modèle de régression ont en effet la même variance asymptotique.

5. ESTIMATION DE VARIANCES

Étudions les variances des différents estimateurs $\hat{Y}_1$, ..., $\hat{Y}_7$ définis précédemment. VA désigne la variance asymptotique d'un estimateur qui est obtenue lorsque N, n et m tendent vers l'infini dans un rapport constant.

5.1 Estimateur $\hat{Y}_1$ et $\hat{Y}_4$: modèle $y_i = x_{i1}' b_1 + u_{i1}$ et (2) $y_i = x_{i2}' b_2 + u_{i2}$.

• Estimateur $\hat{Y}_1$

La variance de cette estimateur et son estimation sont données dans l'ouvrage de Särndal, Swensson et Wretman (1991). La variance se décompose en deux termes qui mesurent la part de variance due respectivement à la première et à la deuxième phase du sondage.

$$\mathrm{VA}(\hat{Y}_1) = \left(\sum_{i,j \in U} \Delta_{ij}^1 \frac{u_{1i} u_{1j}}{\pi_i \pi_j}\right) + \left(E_{s_a} \sum_{i,j \in s_a} \Delta_{ij}^2 \frac{u_{2i} u_{2j}}{\pi_i \pi_{ai} \pi_j \pi_{aj}}\right),$$

avec: $\Delta_{ij}^2 = \pi_{ij} - \pi_i \pi_j$,

$\Delta_{ij}^1 = \pi_{aij} - \pi_{ai} \pi_{aj}$,

$u_{1i} = y_i - x_{i1}' b_1$,

$u_{2i} = y_i - x_{i2}' b_2$,

$$b_1 = \left(\sum_{i \in U} x_{i1} x_{i1}'\right)^{-1}\left(\sum_{i \in U} x_{i1} y_i\right),$$

$$b_2 = \left(\sum_{i \in U} x_{i2} x_{i2}'\right)^{-1}\left(\sum_{i \in U} x_{i2} y_i\right).$$

Ainsi l'estimateur de variance se décompose aussi en deux termes qui estiment la part de variance relative à chacune des phases de tirage. On retrouve que par construction de $\hat{Y}_1$, x_1 permet de réduire la variance apportée par la première phase et x_2 permet de réduire la variance apportée par la deuxième phase

$$\hat{V}(\hat{Y}_1) = \left(\sum_{i \in s}\sum_{j \in s} \frac{\Delta_{ij}^1}{\pi_{ij} \pi_{aij}} \frac{\hat{u}_{1i} \hat{u}_{1j}}{\pi_i \pi_j}\right) + \left(\sum_{i,j \in s} \frac{\Delta_{ij}^2}{\pi_{aij}} \frac{\hat{u}_{2i} \hat{u}_{2j}}{\pi_i \pi_{ai} \pi_j \pi_{aj}}\right),$$

1ière phase 2ième phase

avec: $\hat{u}_{1i} = y_i - x_{i1}' \hat{b}_1$,

$\hat{u}_{2i} = y_i - x_{i2}' \hat{b}_2$.

Une telle décomposition repose sur l'écriture $V(\hat{Y}_1) = V(E[\hat{Y}_1 \mid s_a]) + E(V[\hat{Y}_1 \mid s_a])$ qui interviendra pour tous les autres estimateurs.

• Estimateur $\hat{Y}_4$

Les termes du développement au premier ordre en $1/\sqrt{m}$ de $\hat{Y}_1$ et $\hat{Y}_4$ coïncident exactement. On peut donc donner un sens plus précis à l'écriture $\hat{Y}_4 \cong \hat{Y}_1$. On en déduit que $\mathrm{VA}(\hat{Y}_1) = \mathrm{VA}(\hat{Y}_4)$. Ainsi:

Or $\hat{Y}_1$ se réécrit:

$$\hat{Y}_1 = [X_1' \hat{b}_1] + [\hat{X}_2^{*\prime} \hat{b}_2 - \hat{X}_1' \hat{b}_1] + [\hat{\hat{Y}} - \hat{\hat{X}}_2^{*\prime} \hat{b}_2].$$

On retrouve donc un estimateur semblable à l'estimateur $\hat{Y}_1$ issu de l'approche modèle dans le cas où l'information contenue dans x_1 est jugée substituable à l'information contenue dans x_2 pour estimer y et de moins bonne qualité. Les différences entre $\hat{Y}_1$ et $\hat{Y}_4$ portent sur les points suivants:

1. $\hat{B}_2$ est estimé en intégrant les modifications du calage en x_1 contrairement à $\hat{b}_2$.
2. L'estimation $\hat{B}_1 = \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a}\right)^{-1}\left(\sum_{s} \frac{x_1 y}{\pi_a \pi}\right)$ de B_1, est effectuée en partie sur s_a contrairement à $\hat{b}_1$.
3. Enfin on utilise les poids redressés $F(x_1\beta_1)/\pi_a\pi$ dans les sommes en x_2 sur s et sur S_a dans $\hat{Y}_4$ contrairement à ce que l'on fait pour $\hat{Y}_1$: l'estimation en x_2 est améliorée par la connaissance de x_1.

La modélisation sous-jacente est donc bien ici: (1) $y_i = x_{i1}' b_1 + u_{i1}$ et (2) $y_i = x_{i2}' b_2 + u_{i2}$ dont on pense que le second est a priori meilleur pour prédire la valeur de y_i.

Stratégie 2

On obtient des poids

$$w_i^5 = \frac{F(x_{i1}'\alpha_1 + x_{i2}'\alpha_2)}{\pi_{ai}\pi_i},$$

l'estimateur associé se réécrit:

$$\hat{Y}_5 = [X_1' \hat{a}_1] + [\hat{X}_2' \hat{a}_2] + [\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{a}_1 - \hat{\hat{X}}_2' \hat{a}_2].$$

On retrouve donc exactement l'estimateur $\hat{Y}_2$ proposé dans l'approche modèle de régression dans le cas où l'information contenue dans x_1 est jugée complémentaire à l'information contenue dans x_2 pour estimer y. Le modèle sous-jacent est bien ici $y_i = x_{i1} a_1 + x_{i2} a_2 + u_i$.

Stratégie 3

On obtient des poids

$$w_i^6 = \frac{F(x_{i1}'\gamma_1 + x_{i2}'\gamma_2)}{\pi_{ai}\pi_i},$$

l'estimateur associé se réécrit:

$$\hat{Y}_6 = \hat{\hat{Y}} + [X_1 - \hat{\hat{X}}_1]' \hat{a}_1 + [\hat{X}_2^* - \hat{\hat{X}}_2]' \hat{a}_2$$

soit:

$$\hat{Y}_6 = [X_1' \hat{a}_1] + [\hat{X}_2^{*\prime} \hat{a}_2] + [\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{a}_1 - \hat{\hat{X}}_2' \hat{a}_2].$$

Or,

$$\hat{X}_2^* = \sum_{s_a} \frac{x_2}{\pi_a} + \left(\sum_{s_a} \frac{x_2 x_1'}{\pi_a}\right)^{-1}\left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a}\right)\left[X - \sum_{s_a} \frac{x_1}{\pi_a}\right].$$

On en déduit en remplaçant dans $\hat{Y}_6$ que:

$$\hat{Y}_6 = [X_1' \hat{C}_1] + [M_{x_1} \hat{X}_2' \hat{a}_2] + [\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{a}_1 - \hat{\hat{X}}_2' \hat{a}_2],$$

avec

$$\hat{C}_1 = \hat{a}_1 + \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a}\right)^{-1}\left(\sum_{s_a} \frac{x_1 x_2'}{\pi_a}\right)\hat{a}_2.$$

On retrouve donc un estimateur proche de l'estimateur $\hat{Y}_3$ proposé dans l'approche modèle de régression dans le cas où l'information contenue dans x_1 est jugée complémentaire à l'information contenue dans x_2 pour estimer y. Le modèle sous-jacent est ici $y_i = x_{i1}' c_1 + M_{x_1}(x_{i2})' c_2 + u_i$. Les différences entre $\hat{Y}_3$ et $\hat{Y}_6$ portent sur les coefficients estimés: $\begin{pmatrix}\hat{C}_1\\ \hat{a}_2\end{pmatrix}$ diffère légèrement de $\begin{pmatrix}\hat{c}_1\\ \hat{c}_2\end{pmatrix}$ et $[\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{a}_1 - \hat{\hat{X}}_2' \hat{a}_2]$ diffère légèrement de $[\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{c}_1 - M_{x_1} \hat{\hat{X}}_2' \hat{c}_2]$. En revanche ces quantités sont asymptotiquement équivalentes.

Stratégie 4

On obtient des poids

$$w_i^7 = \frac{F(x_{i1}'\beta_1)F(x_{i1}'\delta_1 + x_{i2}'\delta_2)}{\pi_{ai}\pi_i},$$

l'estimateur associé se réécrit:

$$\hat{Y}_7 = \hat{\hat{Y}}^* + [X_1 - \hat{\hat{X}}_1^*]' \hat{a}_1^* + [\hat{X}_2^* - \hat{\hat{X}}_2^*]' \hat{a}_2^*.$$

En modifiant les poids initiaux en $d_i = F(x_{i1}\beta_1)/\pi_{ai}\pi_i$, on obtient de la même façon:

$$\hat{Y}_7 = [X_1' \hat{a}_1^*] + [\hat{X}_2^{*\prime} \hat{a}_2^*] + [\hat{\hat{Y}}^* - \hat{\hat{X}}_1^{*\prime} \hat{a}_1^* - \hat{\hat{X}}_2^{*\prime} \hat{a}_2^*].$$

En remplaçant $\hat{X}_2^*$ par son expression trouvée plus haut on obtient:

$$\hat{Y}_7 = [X_1' \hat{C}_1^*] + [M_{x_1} \hat{X}_2' \hat{a}_2^*] + [\hat{\hat{Y}}^* - \hat{\hat{X}}_1'^* \hat{a}_1^* - \hat{\hat{X}}_2^{*\prime} \hat{a}_2^*],$$

avec

$$\hat{C}_1 = \hat{a}_1^* + \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a}\right)^{-1}\left(\sum_{s_a} \frac{x_1 x_2'}{\pi_a}\right)\hat{a}_2^*.$$

$\hat{Y}_7$ et $\hat{Y}_6$ sont asymptotiquement équivalents.

Cette stratégie conduit aux équations de calage suivantes:

étape a:

$$\sum_{i\in s_a} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}} x_{i1} = \sum_{i\in U} x_{i1}$$

qui détermine β_1, puis

étape b:

$$\sum_{i\in s} \frac{F(x_{i1}'\gamma_1 + x_{i2}'\gamma_2)}{\pi_{ai}\pi_i} x_{i1} = \sum_{i\in U} x_{i1} = X_1, \quad \text{et}$$

$$\sum_{i\in s} \frac{F(x_{i1}'\gamma_1 + x_{i2}'\gamma_2)}{\pi_{ai}\pi_i} x_{i2} = \sum_{i\in s_a} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}} x_{i2} = \hat{X}_2^*,$$

qui déterminent γ_1 et γ_2.

Enfin une quatrième stratégie peut être proposée qui peut être vue comme une variante de la stratégie précédente:

Stratégie 4

a) caler la structure de l'échantillon 1ière phase s_a sur celle de la population totale U en terme de variable x_1, puis,

b) caler la structure de l'échantillon 2ième phase s simultanément en terme de variables x_1 et x_2, en partant des poids modifiés par le calage précédent, soit:
   - sur la structure de la population totale U en ce qui concerne x_1
   - sur la structure de s_a modifiée en tenant compte du calage précédent pour x_2.

Cette stratégie conduit aux équations de calage suivantes:

étape a:

$$\sum_{i\in s_a} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}} x_{i1} = \sum_{i\in U} x_{i1},$$

qui détermine β_1, puis

étape b:

$$\sum_{i\in s} \frac{F(x_{i1}'\beta_1)F(x_{i1}'\delta_1 + x_{i2}'\delta_2)}{\pi_{ai}\pi_i} x_{i1} = \sum_{i\in U} x_{i1} = X_1, \quad \text{et}$$

$$\sum_{i\in s} \frac{F(x_{i1}'\beta_1)F(x_{i1}'\delta_1 + x_{i2}'\delta_2)}{\pi_{ai}\pi_i} x_{i2} = \sum_{i\in s_a} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}} x_{i2} = \hat{X}_2^*,$$

qui déterminent δ_1 et δ_2.

Lorsque la fonction de calage est exponentielle, il est clair que les stratégies 3 et 4 coïncident.

Dans cette approche en termes de calage, le point de vue adopté est celui de la réduction de variance basée sur les caractéristiques du plan de sondage, sans considération de modèle. Deux questions se posent alors naturellement:

- Peut-on relier chacune de ces quatre stratégies à une approche en terme de modèle?
- Peut-on comparer ces quatre stratégies en termes de variance?

Nous étudierons d'abord le lien entre les trois stratégies définies par une approche calage et les stratégies définies par une approche modèle ou régression, puis dans un deuxième temps nous nous intéresserons au calcul des variances des estimateurs associés à chacune des stratégies.

4.2 Lien entre les différentes stratégies possibles et l'approche régression

Lorsque F est linéaire, chacun des estimateurs associés aux quatre stratégies peut se réécrire simplement.

Notations

Dans toute la suite on utilisera les notations suivantes pour un vecteur de variable z quelconque:

$$\hat{\hat{Z}}^* = \sum_{i\in s} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}\pi_i} z_i \qquad \hat{Z}^* = \sum_{i\in s_a} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}} z_i.$$

On omettra également les indices i pour alléger la présentation lorsqu'il n'y a pas d'ambiguïté.

Stratégie 1

Les pondérations sont de la forme:

$$w_i^4 = \frac{F(x_{i1}'\beta_1)}{\pi_{ai}\pi_i} F(x_{i2}'\beta_2),$$

l'estimateur associé $\hat{Y}_4$ peut se réécrire en traduisant l'effet du deuxième calage en x_2:

$$\hat{Y}_4 = \hat{\hat{Y}}^* + [\hat{X}_2^* - \hat{\hat{X}}_2^*]'\hat{B}_2 \qquad \text{avec}$$

$$\hat{B}_2 = \left(\sum_s \frac{F(x_1'\beta_1)}{\pi_a\pi} x_2 x_2'\right)^{-1} \left(\sum_s \frac{F(x_1'\beta_1)}{\pi_a\pi} x_2 y\right),$$

puis en traduisant l'effet du premier calage en x_1:

$$\hat{Y}_4 = \hat{\hat{Y}} + [X_1 - \hat{X}_1]'\hat{B}_1 + [\hat{X}_2^* - \hat{\hat{X}}_2^*]'\hat{B}_2,$$

soit encore:

$$\hat{Y}_4 = [X_1'\hat{B}_1] + [\hat{X}_2^{*\prime}\hat{B}_2 - \hat{X}_1'\hat{B}_1] + [\hat{\hat{Y}} - \hat{\hat{X}}_2^{*\prime}\hat{B}_2],$$

$$\text{avec} \quad \hat{B}_1 = \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a}\right)^{-1} \left(\sum_s \frac{x_1 y}{\pi_a\pi}\right).$$

Avec ces notations les trois estimateurs se réécrivent:

$$\hat{Y}_1 = [X_1' \hat{b}_1] + [\hat{X}_2' \hat{b}_2 - \hat{X}_1' \hat{b}_1] + [\hat{\hat{Y}} - \hat{\hat{X}}_2' \hat{b}_2]$$

associé aux modèles:

$$\text{(1)} \quad y_i = x_{i1}' b_1 + u_{i1}$$

et

$$\text{(2)} \quad y_i = x_{i2}' b_2 + u_{i2}$$

$$\hat{Y}_2 = [X_1' \hat{a}_1] + [\hat{X}_2' \hat{a}_2] + [\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{a}_1 - \hat{\hat{X}}_2' \hat{a}_2]$$

associé au modèle

$$y_i = x_{i1}' a_1 + x_{i2}' a_2 + u_i$$

$$\hat{Y}_3 = [X_1' \hat{c}_1] + [M_{x_1} \hat{X}_2' \hat{c}_2] + [\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{c}_1 - M_{x_1} \hat{\hat{X}}_2' \hat{c}_2]$$

associé au modèle

$$y_i = x_{i1}' c_1 + M_{x_1}(x_{i2})' c_2 + u_i .$$

De la même façon que l'on généralise l'estimateur par la régression par les estimateurs par calage, on peut aborder le problème de l'utilisation de l'information auxiliaire à plusieurs niveaux à travers la théorie du calage, en essayant de construire des stratégies de calage adaptées à la configuration d'information auxiliaire étudiée dans cette article.

4. APPROCHE PAR CALAGE

4.1 Différentes stratégies possibles

Lorsque l'on cherche à généraliser l'estimation par calage proposée dans un contexte où l'information auxiliaire est présente à un seul niveau: celui de la population totale, plusieurs stratégies apparaissent naturellement:

Stratégie 1

a) caler la structure de l'échantillon 1ière phase s_a sur celle de la population totale U en termes de variable x_1, puis,

b) caler la structure de l'échantillon 2ième phase s sur celle de l'échantillon 1ière phase s_a en terme de variable x_2.

NB: pour cette dernière opération, mieux vaut tenir compte du calage précédent en x_1 pour déterminer la valeur de référence dans le calage en x_2 sur s_a. Si l'on ne tient pas compte du calage précédent, seules les estimations effectuées au niveau de s_a profiteront de l'amélioration apportée par l'étape a). Une bonne façon de s'en convaincre consiste à considérer le cas particulier extrême où $x_1 = x_2$.

Cette stratégie correspond aux équations de calage suivantes:

étape a:

$$\sum_{i \in s_a} \frac{F(x_{i1}' \beta_1)}{\pi_{ai}} x_{i1} = \sum_{i \in U} x_{i1} = X_1$$

qui détermine β_1, puis

étape b:

$$\sum_{i \in s} \frac{F(x_{i1}' \beta_1)}{\pi_{ai} \pi_i} F(x_{i2}' \beta_2) x_{i2} = \sum_{i \in s_a} \frac{F(x_{i1}' \beta_1)}{\pi_{ai}} x_{i2} = \hat{X}_2^*$$

qui détermine β_2,

où F désigne tout au long de l'article la fonction utilisée dans le calage qui peut être linéaire, exponentielle, linéaire tronquée, logit (*cf.* Deville, Särndal 1993).

Stratégie 2

Caler la structure de l'échantillon 2ième phase s simultanément en terme de variables x_1 et x_2, c'est-à-dire:
- sur la structure de la population totale U en ce qui concerne x_1
- sur la structure de s_a pour x_2.

Cette deuxième stratégie nous conduit aux équations de calage suivantes:

$$\sum_{i \in s} \frac{F(x_{i1} \alpha_1 + x_{i2} \alpha_2)}{\pi_{ai} \pi_i} x_{i1} = \sum_{i \in U} x_{i1} = X_1, \quad \text{et}$$

$$\sum_{i \in s} \frac{F(x_{i1}' \alpha_1 + x_{i2}' \alpha_2)}{\pi_{ai} \pi_i} x_{i2} = \sum_{i \in s_a} \frac{x_{i2}}{\pi_{ai}} = \hat{X}_2,$$

qui déterminent α_1 et α_2.

La première stratégie présente l'avantage de corriger les pondérations 1ière phase, c'est-à-dire d'intégrer l'information auxiliaire au niveau le plus élevé. La deuxième stratégie permet quant à elle, de corriger les pondérations qui seront effectivement utilisées dans l'estimation, et notamment d'obtenir une estimation parfaite du total de x_1.

Une troisième stratégie peut être proposée qui réunit les avantages des deux stratégies précédentes, et semble donc pouvoir les dominer:

Stratégie 3

a) caler la structure de l'échantillon 1ière phase s_a sur celle de la population totale U en termes de variables x_1, puis,

b) caler la structure de l'échantillon 2ième phase s simultanément en termes de variables x_1 et x_2, c'est-à-dire:
   - sur la structure de la population totale U en ce qui concerne x_1
   - sur la structure de s_a modifiée en tenant compte du calage précédent pour x_2.

dont on pense que le second est a priori meilleur pour prédire la valeur de y_i. x_1 fonctionne donc dans cette optique modèle, comme une "*proxy*" de x_2. Une situation de ce genre correspond par exemple au cas où x_2 représente la mise à jour c'est-à-dire à la date de l'enquête de la variable extraite de la base de sondage x_1. Autrement dit si x_2 était disponible au niveau de la population toute entière, l'estimateur utilisé serait

$$\sum_{i\in U} x'_{i2}\hat{b}_1 + \sum_{i\in s} \frac{(y_i - x'_{i2}\hat{b}_2)}{\pi_i \pi_{ai}}.$$

Imaginons maintenant le cas d'une enquête par sondage en deux phases effectuée auprès des ménages. Supposons que la base de sondage soit constituée de logements pour lesquels on dispose d'une information constituée de la taille du logement notée x_1, connue donc pour tous les individus de la population cible . Si tous les individus issus de la première phase de sondage sont interrogés sur la composition du ménage notée x_2, notamment sur le nombre d'enfants du ménage, les deux informations apparaissent plutôt complémentaires que substituables, lorsqu'il s'agit d'étudier le budget du ménage. Ceci est encore renforcé si au lieu de la composition du ménage on recueille l'âge du chef de ménage ou bien sa profession.

Dans une optique modèle, la situation alternative où l'information contenue dans x_1 est jugée complémentaire de celle contenue dans x_2 pour estimer y, apparaît donc naturellement.

3.2 L'information contenue dans x_1 est jugée complémentaire de celle contenue dans x_2 pour estimer y

3.2.1 Décomposition $y_i = x'_{i1}a_1 + x'_{i2}a_2 + u_i$

Le modèle sous-jacent est alors:

$$y_i = x'_{i1}a_1 + x'_{i2}a_2 + u_i \quad \text{avec} \quad E(u_i) = 0 \quad \text{et} \quad V(u_i) = \sigma_1^2.$$

L'estimateur à utiliser est alors:

$$\hat{Y}_2 = \sum_{i\in U} x'_{i1}\hat{a}_1 + \sum_{i\in s_a} \frac{x'_{i2}\hat{a}_2}{\pi_{ai}} + \sum_{i\in s} \frac{(y_i - x'_{i1}\hat{a}_1 - x'_{i2}\hat{a}_2)}{\pi_i \pi_{ai}}$$

avec:

$$\begin{pmatrix}\hat{a}_1\\ \hat{a}_2\end{pmatrix} = \begin{pmatrix} \sum_{i\in s} \frac{x_{i1}x'_{i1}}{\pi_{ai}\pi_i} & \sum_{i\in s} \frac{x_{i1}x'_{i2}}{\pi_{ai}\pi_i} \\ \sum_{i\in s} \frac{x_{i2}x'_{i1}}{\pi_{ai}\pi_i} & \sum_{i\in s} \frac{x_{i2}x'_{i2}}{\pi_{ai}\pi_i} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i\in s} \frac{x_{i1}y_i}{\pi_{ai}\pi_i} \\ \sum_{i\in s} \frac{x_{i2}y_i}{\pi_{ai}\pi_i} \end{pmatrix}.$$

En effet, la variable y se décompose en trois composantes $y_i = x'_{i1}\hat{a}_1 + x'_{i2}\hat{a}_2 + \hat{u}_i$. Le total de y se décompose donc en trois composantes qui sont chacune estimée au niveau le plus haut, c'est-à-dire avec le maximum de précision possible:

- U pour $x'_{i1}\hat{a}_1$,
- s_a pour $x'_{i2}\hat{a}_2$, et
- s pour $\hat{u}_i$.

3.2.2 Décomposition $y_i = x'_{i1}c_1 + M_{x_1}(x_{i2})'c_2 + u_i$

Si on souhaite utiliser au maximum l'information contenue dans x_1 disponible sur U, il est naturel d'introduire une autre écriture du même modèle $y_i = x'_{i1}a_1 + x'_{i2}a_2 + u_i$ qui isole tout ce qui dans y peut être pris en compte à travers x_1. Elle s'écrit:

$$y_i = x'_{i1}c_1 + M_{x_1}(x_{i2})'c_2 + u_i \quad \text{avec}$$
$$E(u_i) = 0 \quad \text{et} \quad V(u_i) = \sigma_1^2,$$

où M_{x_1} représente la projection orthogonale, dans la métrique associée aux poids $1/\pi_{ai}$, sur l'orthogonal de l'espace vectoriel engendré dans s_a (assimilé à $\Re^n$) par le groupe de variables x_1.

$M_{x_1}(x_{i2})$ est définie par:

$$M_{x_1}(x_{i2}) = x'_{i2} - \left(\sum_{i\in s_a} \frac{x_{i2}x'_{i1}}{\pi_{ai}}\right)\left(\sum_{i\in s_a} \frac{x_{i1}x'_{i1}}{\pi_{ai}}\right)^{-1} x_{i1}.$$

L'estimateur naturel associé est alors:

$$\hat{Y}_3 = \sum_{i\in U} x'_{i1}\hat{c}_1 + \sum_{i\in s_a} \frac{(M_{x_1}x'_{i2}\hat{c}_2)}{\pi_{ai}} + \sum_{i\in s} \frac{(y_i - x'_{i1}\hat{c}_1 - M_{x_1}x'_{i2}\hat{c}_2)}{\pi_i \pi_{ai}},$$

où $\hat{c} = \begin{pmatrix}\hat{c}_1\\ \hat{c}_2\end{pmatrix}$ est le coefficient de la régression $y = x'c_1 + (M_{x_1}x_2)'c_2 + u$ estimée sur s avec des poids $1/\pi_{ai}\pi_i$ (qui diffère légèrement de $\begin{pmatrix}\hat{b}_1\\ \hat{a}_2\end{pmatrix}$).

3.3 Les trois estimateurs dérivés de l'approche modèles

L'approche par la modélisation nous a permis de construire 3 estimateurs que l'on peut réécrire synthétiquement en introduisant de nouvelles notations. On écrira dans toute la suite pour un vecteur de variable z quelconque:

$$\hat{\hat{Z}} = \sum_{i\in s} \frac{1}{\pi_{ai}\pi_i} z$$

$$\hat{Z} = \sum_{i\in s_a} \frac{1}{\pi_{ai}} z.$$

Ainsi (section 4), les deux approches peuvent être reliées et débouchent sur trois classes d'estimateurs associées chacune à une décomposition de la variable d'intérêt. Les estimateurs d'une même classe ont la même variance asymptotique.

L'évaluation des stratégies fondée sur le seul plan de sondage oriente le choix vers la troisième classe d'estimateurs qui domine les deux autres du point de vue de la variance.

L'évaluation des stratégies, lorsqu'elle s'appuie sur une modélisation de la variable d'intérêt conduit à privilégier la classe d'estimateurs associée à la modélisation adoptée.

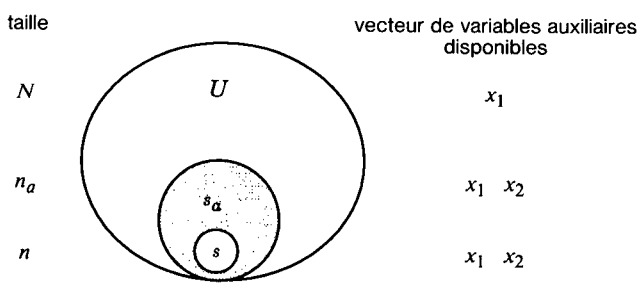
Dans un contexte où l'on souhaite effectuer le redressement d'une enquête, et que l'on souhaite simultanément corriger les biais qu'induiraient l'utilisation des pondérations brutes et réduire la variance, les conclusions doivent être adaptées: les modifications introduites dans les pondérations pour corriger les biais sont plus importantes que les corrections pour réduction de variance. Dès lors, on intégrera dans le calage les variables, dès que l'on pense qu'elles interviennent dans la probabilité de sélection et donc participent à la création du biais.

Lorsque les variables auxiliaires sont qualitatives, l'arbitrage entre utilisation a priori et a posteriori de l'information auxiliaire, c'est-à-dire entre, utilisation au niveau de l'échantillonnage ou au niveau du redressement, repose encore sur la distinction entre les deux modélisations de la variable d'intérêt.

Il est possible d'étendre ces résultats au cas de sondages en plus de deux phases.

2. NOTATIONS

Le cadre est celui d'un sondage en deux phases. On suppose qu'on dispose d'information auxiliaire à deux niveaux différents: population cible et population issue du tirage 1ière phase. La situation peut être schématisée de la manière suivante:



Où U représente la population cible pour laquelle les valeurs du vecteur de variable x_1 sont connues ou à défaut le total $X_1 = \sum_{i\in U} x_{i1}$. s_a représente un niveau intermédiaire de sondage pour lequel les valeurs des vecteurs de k_1 variables x_1 et k_2 variables x_2 sont connues pour tous les individus. On note π_{ia}, la probabilité de sélection du tirage associée à la première phase du sondage. s représente l'échantillon final pour lequel les valeurs de la variable y dont on cherche à estimer le total sont disponibles, ainsi que les valeurs des vecteurs de variables auxiliaires x_1 et x_2. On note $\pi_i = P(i \mid s_a)$.

On souhaite utiliser au mieux toute cette information auxiliaire pour améliorer les estimations qui seront effectuées à partir des données recueillies sur l'échantillon issu de la deuxième phase de sondage s.

Une première idée naturelle consiste à chercher la généralisation de l'estimateur par régression dans ce contexte.

3. APPROCHE ESTIMATION PAR RÉGRESSION

3.1 L'information contenue dans x_1 est jugée substituable à l'information contenue dans x_2 pour estimer y et de moins bonne qualité que x_2

Särndal, Swensson et Wretman proposent dans leur ouvrage l'estimateur par régression suivant pour l'estimation du total de y:

$$\hat{Y}_1 = \sum_{i\in s} \frac{y_i}{\pi_i \pi_{ai}} + \left(\sum_{i\in s_a} \frac{x'_{i2}\,\hat{b}_2}{\pi_i} - \sum_{i\in s} \frac{x'_{i2}\,\hat{b}_2}{\pi_i \pi_{ai}} \right) + \left(\sum_{i\in U} x'_{i1}\,\hat{b}_1 - \sum_{i\in s_a} \frac{x'_{i1}\,\hat{b}_1}{\pi_{ai}} \right)$$

où le deuxième terme est la correction pour mauvaise estimation sur s_a et le troisième est la correction pour mauvaise estimation sur s.

L'estimateur peut aussi s'écrire:

$$\hat{Y}_1 = \sum_{i\in U} x'_{i1}\,\hat{b}_1 + \sum_{i\in s_a} \frac{(x'_{i1}\,\hat{b}_1 - x'_{i2}\,\hat{b}_2)}{\pi_{ai}} + \sum_{i\in s} \frac{(y_i - x'_{i2}\,\hat{b}_2)}{\pi_i \pi_{ai}}$$

où le deuxième terme est la correction pour mauvaise approximation de y_i par $x'_{i1}\,\hat{b}_1$ et le troisième est la correction pour mauvaise approximation de y_i par $x'_{i2}\,\hat{b}_2$;

avec $$\hat{b}_1 = \left(\sum_{i\in s_a} \frac{x_{i1} x'_{i1}}{\pi_{ai}} \right)^{-1} \left(\sum_{i\in s} \frac{x_{i1} y_i}{\pi_i \pi_{ai}} \right)$$

et $$\hat{b}_2 = \left(\sum_{i\in s} \frac{x_{i2} x'_{i2}}{\pi_i \pi_{ai}} \right)^{-1} \left(\sum_{i\in s} \frac{x_{i2} y_i}{\pi_i \pi_{ai}} \right).$$

L'idée sous-jacente est que l'on a deux modèles concurrents pour y qui sont:

(1) $y_i = x'_{i1} b_1 + u_{i1}$ avec $E(u_{i1}) = 0$ et $V(u_{i1}) = \sigma_1^2$

et

(2) $y_i = x'_{i2} b_2 + u_{i2}$ avec $E(u_{i2}) = 0$ et $V(u_{i2}) = \sigma_2^2$

Techniques d'enquête, décembre 1995
Vol. 21, n° 2, pp. 141–150
Statistique Canada

Redressements alternatifs en présence de plusieurs niveaux d'information auxiliaire

F. DUPONT[1]

RÉSUMÉ

L'estimation par régression et sa généralisation, l'estimation par calage introduite par Deville et Särndal en 1993 permet de réduire a posteriori la variance des estimateur, grâce à l'utilisation de l'information auxiliaire. Dans les enquêtes réalisées par sondage, on dispose souvent d'information additionnelle exploitable répartie selon un schéma complexe notamment lorsque le sondage est réalisé en plusieurs phases. Une adaptation de l'estimation par régression a été proposée avec ses variantes dans le cadre du sondage à deux phases par Särndal et Swensson en 1987. Cet article propose d'étudier les stratégies d'estimation alternatives selon deux configurations alternatives pour l'information auxiliaire, en reliant les deux approches possibles pour aborder le problème: modèle de régression et estimation par calage.

MOTS CLÉS: Information auxiliaire; estimateur par régression; estimateur par calage; sondage en deux phases.

1. INTRODUCTION

L'estimateur par régression étudié par Fuller (1975), Cassel, Särndal et Wretman (1976), Särndal (1980), Gourieroux (1981), Isaki et Fuller (1982), et Wright (1983) permet d'améliorer, a posteriori, c'est-à-dire une fois le sondage réalisé, l'estimation d'un total d'une variable d'intérêt sur la base de variables auxiliaires $x_1, \ldots, x_k$ pour lesquelles on dispose d'information additionnelle. On réduit en effet la variance par rapport à l'estimateur d'Horwitz-Thompson en utilisant l'estimateur par régression à condition de connaître la valeur sans aléa, des totaux sur la population cible des variables auxiliaires, qui constitueront l'information additionnelle appelée information auxiliaire. Deville et Särndal ont proposé en 1992 une classe d'estimateurs dérivée d'une approche par repondération qui répond à la même problématique de réduction de variance: les estimateurs par calage. Le calage des poids de sondage réalisé permet en effet d'intégrer a posteriori une information auxiliaire du type totaux $X_1, \ldots, X_k$ de k variables $x_1, \ldots, x_k$ dans l'estimation réalisée à partir des nouvelles pondérations et donc de l'améliorer. Cette approche généralise l'estimation par régression qui constitue l'un des éléments de la classe.

Toutefois, dans les enquêtes réalisées par sondage, on dispose souvent d'information additionnelle exploitable répartie selon un schéma plus complexe que ce qui vient d'être décrit, notamment lorsque le sondage est réalisé en plusieurs phases. Cet article étudie différentes stratégies d'utilisation de cette information auxiliaire complexe dans le cadre d'un sondage en deux phases, la généralisation à plus de deux phases étant possible.

Lorsque le plan de sondage comporte deux phases, l'information auxiliaire est constituée de l'information connue pour la population toute entière, mais aussi de l'information connue pour l'échantillon issu de la première phase de sondage. Or, ces deux informations peuvent porter sur des variables différentes.

Särndal et Swensson proposent dans leur article de 1987, un estimateur utilisant toute l'information auxiliaire disponible pour un tirage en deux phases, avec une information auxiliaire différente pour la population toute entière et la population issu du tirage de la première phase. Il s'agit d'un estimateur adaptant le principe de l'estimateur par régression lorsque l'information connue pour les individus issus du tirage première phase est considérée comme substituable à l'information agrégée et de meilleure qualité que l'information disponible pour la population cible toute entière, vis-à-vis de l'estimation de la variable d'intérêt. Or, Il arrive en pratique que ces deux informations soient complémentaires plutôt que substituables. On a donc cherché dans cette étude à développer l'estimation par régression, dans un contexte où les informations auxiliaires sont complémentaires.

Par ailleurs, dans la mesure où l'estimation par calage généralise l'estimation par régression lorsque l'information auxiliaire n'est constituée que d'un seul niveau, on a cherché à adapter l'estimation par calage dans ce contexte. On passe en revue les stratégies de calages afin de proposer les plus adaptées, en cherchant à les relier aux généralisations de l'estimation par régression possibles dans ce contexte.

On montre (section 2) que l'utilisation conjointe de deux informations auxiliaires différentes conduit à deux modèles de régression et trois décompositions de la variable d'intérêt associées. L'approche assistée par modèle de régression (AAMR) permet donc de dériver 3 estimateurs alternatifs.

L'approche calage (AC) (section 3) permet quant à elle de dériver 4 estimateurs. Chacun de ces estimateurs peut être relié (associé) aux trois estimateurs dérivés de l'approche modèle de régression.

[1] F. Dupont, Unité Méthodes Statistiques, Institut National de la Statistique et des Études Économiques, (INSEE), 18 Blvd. Adolphe Pinard, 75675 Paris Cedex.

LÊ, T., et VERMA, V. (1995). *Sample Designs and Sampling Errors for the DHS*. Calverton MD: MACRO International.

McNEMAR, Q. (1949). *Psychological Statistics*. New York: John Wiley and Sons.

MOSTELLER, F. (1952). Some statistical problems in measuring the subjective responses to drugs. *Biometrika*, 8, 220-226.

RAO, J.N.K., et SCOTT, A.J. (1987). On simple adjustments to Chi-square tests with sample survey data. *Annals of Statistics*, 15, 385-397.

RAO, J.N.K., et WU, C.F.S. (1985). Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.

SCOTT, A.J., et HOLT, D. (1982) The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77, 848-54.

SCOTT, A.J., et SEBER, G.A.F. (1983). Difference of proportions from the same survey. *The American Statistician*, 37, 319-20.

SKINNER, C.J., HOLT, D., et SMITH, T.M.F. (1989). *Analysis of Complex Surveys*. New York: John Wiley and Sons.

VERMA, V., et LÊ, T. (1995). Sampling errors for the DHS survey. 50ième Session de l'Institut International de Statistique, Beijing.

VERMA, V., SCOTT, C., et O'MUIRCHEARTAIGH, C. (1980). Sample designs and sampling errors for the World Fertility Surveys. *Journal of the Royal Statistical Society (A)*, 143, 431-473.

WILD, C.J., et SEBER, G.A.F. (1993). Comparing two proportions for the same survey. *The American Statistician*, 47, 178-181.

tableaux 1, 2 et 3. La figure 1 présente les données provenant de trois pays (Maroc, Niger et Colombie); il y a donc six populations puisque les eps en zone urbaine sont assez différents de ceux observés en zone rurale. La figure 2 présente les résultats selon le sexe; on distingue deux populations passablement distinctes en ce qui concerne l'emploi, mais la différence est moins marquée en ce qui concerne le niveau d'éducation.

Nous attendions les données empiriques des tableaux des études 4, 5 et 6 avec anxiété. Il est vrai que les cinq ensembles précédents avaient conduit à des conclusions similaires, même s'ils traitaient de onze populations différentes, de pointages et de variables. Pourtant, les études 1 à 5 portaient sur des paires de catégories issues de polytomies, les plans 1 et 2 du type A. Dorénavant, nous cherchions des données pour des comparaisons de type B provenant d'enquêtes par panel, de manière à pouvoir explorer les conjectures des plans test-retest et avant-après. Du point de vue mathématique, on peut facilement montrer une ressemblance avec les polytomies (c.-à-d., avec les tétratomies), mais les rapports entre ces valeurs et les valeurs empiriques des eps ne sont pas du tout évidentes. C'est ce qui explique pourquoi ces valeurs empiriques sont si utiles et si remarquables. Nous avons ici observé un effet du plan de sondage extrêmement important pour les tests de chi carré des comparaisons analytiques.

5. SIGNIFICATION DES RÉSULTATS POUR LES RECHERCHES CONNEXES

Nos travaux antérieurs et ceux d'autres chercheurs fournissent déjà une masse considérable de données empiriques sur les effets du plan de sondage sur l'échantillon total, les sous-classes et les différences, pour des variables et des plans différents. Il serait donc utile d'examiner les résultats décrits plus haut à la lumière de ces données antérieures.

On a déterminé que la nature des variables d'enquête qui font l'objet d'une estimation est un facteur important (souvent le plus important) pour déterminer l'ampleur des eps. On peut observer des eps extrêmement différents selon le type de variables, même au sein d'un seul échantillon ou avec des plans de sondage très semblables. C'est la raison pour laquelle nous avons toujours recommandé que les eps soient calculés pour plusieurs variables différentes, alors qu'il est généralement moins important de les calculer pour plusieurs sous-classes différentes, en particulier lorsque ces sous-classes sont définies en fonction des mêmes caractéristiques.

Nos résultats montrent que les eps peuvent également varier énormément entre les catégories différentes de la même variable d'enquête, lorsque l'échantillon total sert de base commune à l'estimation. Ainsi, il convient dans un certain sens de considérer chaque catégorie individuelle et chaque différence entre les paires de catégories, même lorsqu'on a affaire aux variables d'une seule et même enquête, comme s'il s'agissait de variables séparées aux fins du calcul et de l'analyse des effets du plan de sondage.

Pour ce qui est du rapport existant entre les eps pour les sous-classes et pour les différences entre sous-classes, les recherches antérieures ont surtout porté sur la situation décrite ci-après. Compte tenu d'un échantillon total n réparti en sous-classes i de taille $n_i = p_i.n$, les valeurs d'eps(r_i) pour les statistiques r_i (comme le rapport m_i/n_i, la moyenne $\sum y_i/n_i$ ou le rapport $\sum y_i/\sum x_i$), calculées sur la base des éléments de sous-classe n_i, se rapportent à la valeur eps(r) pour la même variable, calculée sur la base de l'échantillon total. De même, les valeurs de eps$(r_i - r_j)$ pour les différences de sous-classes se rapportent à l'eps(r_i), à l'eps(r_j) fondé sur les sous-classes individuelles et à l'eps(r) fondé sur l'échantillon total. De nombreux calculs confirment que ces rapports s'accordent avec notre conjecture (2) de la section 3:

$$\text{eps}(r) > \text{eps}(r_i); \quad \text{et} \quad \text{eps}(r_i) > \text{eps}(r_i - r_j) > 1.$$

Ces effets des covariances sur les effets du plan de sondage pour des échantillons par grappes sont essentiellement empiriques (même sociologiques, au sens large); ils doivent être vérifiés comme tels.

Ainsi en est-il du nouveau rapport que nous avons découvert pour $(p_i - p_j)$ pour deux catégories, lesquelles sont si différentes de ce qui précède. Les rapports $\text{eps}(p_i - p_j) \cong [\text{eps}(p_i) + \text{eps}(p_j)]/2$ sont également empiriques et approximatifs, et doivent être constamment vérifiés. Pourtant, ils semblent s'appliquer largement à nos données et sont nettement préférables aux autres hypothèses, telles que $\text{eps}(p_i - p_j) = 1$, qu'on a souvent utilisées jusqu'ici.

REMERCIEMENTS

Les auteurs désirent remercier l'éditeur associé et les arbitres dont les commentaires ont permis d'abréger et d'améliorer cet article.

BIBLIOGRAPHIE

COCHRAN, W.G. (1950). The comparison of percentages in matched samples. *Biometrika*, 37, 256-66.

DEMING, W.E. (1953). On the distinction between enumerative and analytic studies. *Journal of the American Statistical Association*, 48, 244-45.

KISH, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.

KISH, L. (1987). *Statistical Research Design*. New York: John Wiley and Sons.

KISH, L. (1995). Methods for design effects. *Journal of Official Statistics*, 11, 55-77.

KISH, L., et FRANKEL, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society*, (B), 36, 1-37.

KISH, L., GROVES, R.M., et KROTKI, K. (1976). *Sampling Errors for Fertility Surveys*. Document hors série n° 17, *Enquête mondiale de la fécondité*. Institut International de Statistique: The Hague.

des quelques cas où $\text{eps}(p_i - p_j)$ ne se situe pas entre $\text{eps}(p_i)$ et $\text{eps}(p_j)$, et où on obtient $\text{eps}(p_i) < \text{eps}(p_i - p_j) > \text{eps}(p_j)$ ou $\text{eps}(p_i) > \text{eps}(p_i - p_j) < \text{eps}(p_j)$. Ces cas montrent en passant que nos résultats ne sont pas des conséquences mathématiques, mais qu'ils ont plutôt un caractère empirique.

Tableau 4

Les National Election Studies Panels de 1990 et de 1992, réalisés par le Survey Research Center, Institute for Social Research, Ann Arbor

Catégories avant/après (90/92)	Eps pour			
	P_i	P_j	Moyenne	$(P_i - P_j)$
Entièrement d'accord avec Bush	1.14	.93	1.04	1.02
D'accord avec la politique étrangère de Bush	.92	1.05	.99	1.00
Désapprouve entièrement la politique étrangère de Bush	1.23	1.24	1.24	1.32
D'accord avec la politique économique de Bush	.97	.94	.96	.96
Entièrement d'accord avec la politique économique de Bush	1.14	1.04	1.09	1.10
D'accord avec Bush	1.00	1.00	1.00	1.00
Désapprouve entièrement Bush	1.16	1.10	1.13	1.12
A suivi la campagne électorale à la télé	.89	1.55	1.22	1.40
Moyenne	*1.06*	*1.11*	*1.08*	*1.11*

Tableau 5

Les Panel Study of Income Dynamics de 1983 et de 1987, réalisés par le Survey Research Center, Ann Arbor

Catégories* avant/après (83/87)	Eps pour			
	P_i	P_j	Moyenne	$(P_i - P_j)$
Vit dans le sud	1.22	1.23	1.23	1.11
Âge du chef de famille	1.28	1.33	1.31	1.37
Taille de la famille	1.29	1.43	1.36	1.47
Nombre d'enfants dans la famille	1.23	1.43	1.33	1.49
Heures de travail du chef de famille	1.12	.84	.98	1.03
Âge de l'enfant le plus jeune	.93	.91	.92	.87
Moyenne	*1.18*	*1.20*	*1.19*	*1.22*

* Toutes les variables sont classées dans deux catégories.

Les résultats empiriques présentés aux figures 1 et 2 viennent encore confirmer ceux présentés aux tableaux 1, 2 et 3. Nous y constatons également que: 1) $\text{eps}(p_i - p_j) \cong [\text{eps}(p_i) + \text{eps}(p_j)]/2$ approximativement, le long de la ligne de 45°; 2) ces égalités se vérifient pour une vaste gamme d'eps; 3) la variation entre les variables est en effet très importante. Cette variation est particulièrement évidente pour l'Indonésie rurale, où les valeurs d'eps dépassent 4 et où les valeurs d'eps^2 sont donc supérieures à 16. Ces importants effets de groupement sont dus à la grande taille des grappes: avec $\bar{b} = 133$ et 137, les valeurs de $roh = 0.12$ sont suffisantes pour un grand eps. À noter que ces résultats empiriques viennent à la fois de populations et de variables très diversifiées; elles sont différentes l'une de l'autre et différentes également des données des

Tableau 6

Les études Americans' Changing Lives de 1986 et de 1989 réalisées par le Survey Research Center, Ann Arbor

Catégories avant/après	Eps pour				Catégories avant/après	Eps pour			
	P_i	P_j	Moyenne	$(P_i - P_j)$		P_i	P_j	Moyenne	$(P_i - P_j)$
A. Rencontre des amis					**B. Fait de l'exercice physique**				
Une fois par semaine	1.30	1.26	1.28	1.28	Souvent	1.51	1.67	1.59	1.26
2 ou 3 fois par mois	.88	1.00	.94	1.02	Jamais	1.62	1.97	1.80	1.41
Moyenne	*1.09*	*1.13*	*1.11*	*1.15*	*Moyenne*	*1.56*	*1.82*	*1.70*	*1.34*
C. Satisfaction de soi					**D. Aime sa maison**				
Très satisfait	1.28	1.21	1.25	1.33	Beaucoup	1.24	.90	1.07	.91
Pas satisfait	1.04	1.16	1.10	1.00	Pas beaucoup	1.33	.98	1.16	1.12
Moyenne	*1.16*	*1.19*	*1.18*	*1.17*	*Moyenne*	*1.29*	*.94*	*1.12*	*1.02*
E. Pratique le jardinage					**F. A une attitude positive**				
Souvent	1.40	1.16	1.28	1.19	Oui	1.10	1.33	1.22	1.19
Rarement	.91	1.11	1.01	1.18	Non	1.05	1.28	1.17	1.21
Jamais	1.66	1.17	1.42	1.26	*Moyenne*	*1.08*	*1.31*	*1.20*	*1.20*
Moyenne	*1.32*	*1.15*	*1.24*	*1.21*					
Moyenne globale	*1.25*	*1.26*	*1.26*	*1.18*					

Tableau 2

La National Education Longitudinal Study (NELS) de 1988, réalisée par le National Opinion Research Center de la University of Chicago (n = 24,355)

Catégories	Eps pour				Catégories	Eps pour			
$i - j$	P_i	P_j	Moyenne	$(P_i - P_j)$	$i - j$	P_i	P_j	Moyenne	$(P_i - P_j)$
	A. Scolarité					**B. Trouve les cours ennuyants**			
1-2	1.38	1.22	1.30	1.11	1-2	.99	1.11	1.05	1.04
1-3	1.38	1.14	1.26	1.16	1-3	.99	1.12	1.06	1.07
1-4	1.38	1.19	1.29	1.30	2-3	1.11	1.12	1.12	1.13
1-5	1.38	1.42	1.40	1.54	*Moyenne*	*1.03*	*1.12*	*1.08*	*1.08*
2-3	1.22	1.14	1.18	1.11		**C. Libre de choisir sa voie**			
2-4	1.22	1.19	1.21	1.24	1-2	1.28	1.10	1.19	1.21
2-5	1.22	1.42	1.32	1.45	1-3	1.28	1.08	1.18	1.28
3-4	1.14	1.19	1.17	1.18	2-3	1.10	1.08	1.09	.97
3-5	1.14	1.42	1.28	1.37	*Moyenne*	*1.22*	*1.09*	*1.15*	*1.15*
4-5	1.19	1.42	1.31	1.20		**D. L'école donne accès à de bons emplois**			
Moyenne	*1.27*	*1.28*	*1.27*	*1.25*	1-2	1.24	1.07	1.16	1.17
	E. Religion				1-3	1.24	1.11	1.18	1.24
1-2	2.48	2.83	2.65	2.74	2-3	1.07	1.11	1.09	1.01
1-3	2.48	2.02	2.25	2.09	*Moyenne*	*1.18*	*1.10*	*1.14*	*1.14*
2-3	2.83	2.02	2.42	2.59		**F. Scolarité du père**			
Moyenne	*2.60*	*2.29*	*2.44*	*2.47*	1-2	1.61	1.76	1.69	1.83
	I. Assurance				1-3	1.61	1.68	1.65	1.65
1-2	1.42	1.28	1.35	1.37	2-3	1.76	1.68	1.72	2.48
					Moyenne	*1.65*	*1.71*	*1.69*	*1.99*
Moyenne globale						*1.48*	*1.41*	*1.45*	*1.49*

Tableau 3

La National Longitudinal Study of Labor Market Experience of Youth, réalisée par le National Opinion Research Center de la University of Chicago (n = 5,857)

Catégories	Eps pour				Catégories	Eps pour			
$i - j$	P_i	P_j	Moyenne	$(P_i - P_j)$	$i - j$	P_i	P_j	Moyenne	$(P_i - P_j)$
	A. La chance joue un rôle important dans ma vie					**B. Quelque chose m'empêche de progresser**			
1-2	1.26	1.20	1.23	1.06	1-2	1.07	1.22	1.14	1.04
1-3	1.26	1.18	1.22	1.30	1-3	1.07	1.12	1.10	1.14
1-4	1.26	1.16	1.21	1.28	1-4	1.07	1.09	1.08	1.09
2-3	1.20	1.18	1.19	1.22	2-3	1.22	1.12	1.17	1.28
2-4	1.20	1.16	1.18	1.25	2-4	1.22	1.09	1.16	1.14
3-4	1.18	1.16	1.17	1.05	3-4	1.12	1.09	1.11	1.07
Moyenne	*1.23*	*1.17*	*1.20*	*1.19*	*Moyenne*	*1.13*	*1.12*	*1.13*	*1.13*
	C. Je fais ce que je veux dans la vie					**D. Je suis aussi capable que les autres**			
1-2	1.13	1.06	1.10	1.09	1-2	1.17	1.13	1.15	1.16
1-3	1.13	1.10	1.12	1.13	1-3	1.17	1.07	1.12	1.16
2-3	1.06	1.10	1.08	1.07	2-3	1.13	1.07	1.10	1.08
Moyenne	*1.11*	*1.09*	*1.10*	*1.10*	*Moyenne*	*1.16*	*1.09*	*1.12*	*1.13*
	E. Mes projets aboutissent rarement					**F. Je suis satisfait**			
1-2	1.19	1.07	1.13	1.12	1-2	1.19	1.12	1.16	1.16
1-3	1.19	1.13	1.16	1.20	1-3	1.19	1.13	1.16	1.20
2-3	1.07	1.13	1.10	1.08	2-3	1.12	1.13	1.13	1.09
Moyenne	*1.15*	*1.11*	*1.13*	*1.13*	*Moyenne*	*1.17*	*1.13*	*1.15*	*1.15*
	I. Travail de la mère								
1-2	1.49	1.36	1.43	1.41					
1-3	1.49	1.52	1.51	1.53					
2-3	1.36	1.52	1.44	1.44					
Moyenne	*1.45*	*1.47*	*1.46*	*1.47*					
Moyenne globale	*1.20*	*1.17*	*1.18*	*1.19*					

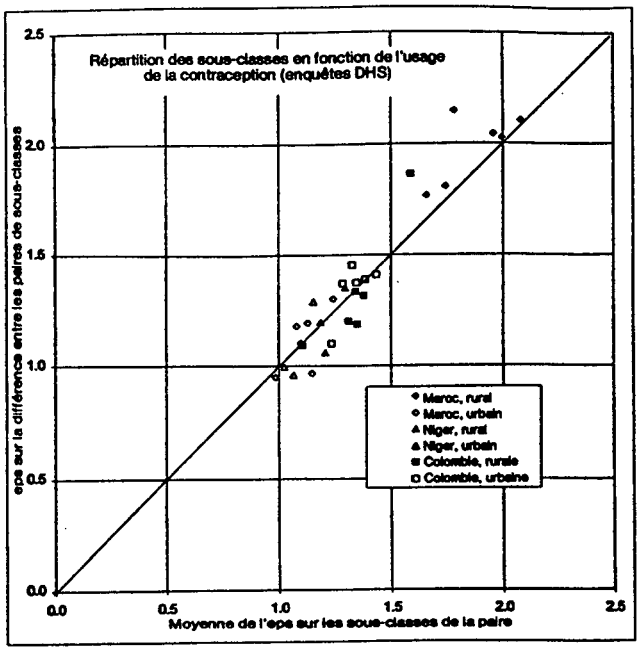


Figure 1. Comparaison de eps$(p_i - p_j)$ à la moyenne de eps(p_i) et de eps(p_j) pour diverses catégories d'usage de la contraception*. Illustration de six populations tirées des enquêtes sur la démographie et la santé des populations (DHS).

* 1 = aucune méthode de contraception
2 = usage des méthodes traditionnelles uniquement
3 = utilisation d'une méthode moderne "réversible"
4 = stérilisation.

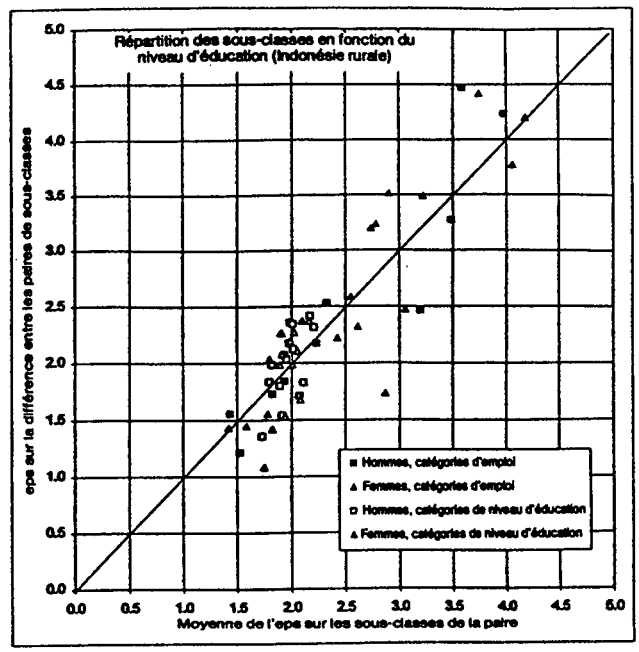


Figure 2. Comparaison de eps$(p_i - p_j)$ à la moyenne de eps(p_i) et de eps(p_j) pour les diverses catégories d'emploi et de niveau d'éducation par sexe. Illustration d'un recensement.

Tableau 1

La National Election Study de 1986 réalisée par le Institute for Social Research de la University of Michigan (n = 2,135)

Catégories	Eps pour			
$i - j$	P_i	P_j	Moyenne	$(P_i - P_j)$
	A. Religion			
1-2	1.21	1.42	1.32	1.10
1-3	1.21	2.32	1.77	2.02
1-4	1.21	1.50	1.36	1.18
1-5	1.21	1.18	1.19	1.17
2-3	1.42	2.32	1.87	1.93
2-4	1.42	1.50	1.46	1.57
2-5	1.42	1.18	1.30	1.27
3-4	2.32	1.50	1.91	2.03
3-5	2.32	1.18	1.75	2.04
4-5	1.50	1.18	1.34	1.19
Moyenne	*1.56*	*1.53*	*1.54*	*1.55*
	C. Appui à Reagan			
1-2	1.32	1.10	1.21	1.07
1-3	1.32	.86	1.09	1.26
1-4	1.32	1.48	1.40	1.50
2-3	1.10	.86	.98	.96
2-4	1.10	1.48	1.29	1.38
3-4	.86	1.48	1.17	1.09
Moyenne	*1.17*	*1.21*	*1.19*	*1.21*
Moyenne globale	*1.23*	*1.24*	*1.23*	*1.24*

Catégories	Eps pour			
$i - j$	P_i	P_j	Moyenne	$(P_i - P_j)$
	B. Position sur l'avortement			
1-2	1.27	.97	1.12	.97
1-3	1.27	1.28	1.28	1.32
1-4	1.27	1.31	1.29	1.36
2-3	.97	1.28	1.12	1.08
2-4	.97	1.31	1.14	1.16
3-4	1.28	1.31	1.30	1.32
Moyenne	*1.17*	*1.24*	*1.21*	*1.20*
	D. Problèmes dans le pays			
1-2	1.07	.94	1.00	.98
1-3	1.07	1.04	1.05	1.09
1-4	1.07	.93	1.00	1.12
2-3	.94	1.04	.99	1.01
2-4	.94	.93	.93	.85
3-4	1.04	.93	.98	.82
Moyenne	*1.02*	*.97*	*.99*	*.98*

4. RÉSULTATS EMPIRIQUES POUR EPS$(P_i - P_j)$

À défaut d'un solide fondement théorique ou mathématique permettant d'accorder une préférence à l'une ou l'autre des conjectures énumérées plus haut, les résultats empiriques sur les valeurs d'eps$(p_i - p_j)$ deviennent essentiels puisqu'ils permettent d'établir un rapport entre ces valeurs et les valeurs calculées d'eps(p_i). Ceci rappelle nos hypothèses plus familières concernant eps$(p_i) = \sqrt{1 + roh[\bar{b} - 1]}$; leur valeur dépend de plusieurs facteurs influant sur *roh*, le coefficient de corrélation intraclasse, ainsi que de la taille moyenne de la grappe $\bar{b}$ (Kish 1965, 5.4, 8.2). Les valeurs d'eps(p_i) varient beaucoup d'un sondage à l'autre, ainsi que d'une variable à l'autre dans un sondage particulier (Kish, Groves et Krotki 1976; Verma, Scott et O'Muircheartaigh 1980; Verma et Lê 1995). Toutefois, les études empiriques portant sur les erreurs d'échantillonnage observées dans diverses enquêtes constituent, pour les statisticiens d'enquête, une source utile de connaissances qui leur permettent également d'établir des rapports entre les valeurs d'eps de statistiques complexes et celles d'eps(p_i) plus simples (Kish 1995; Rao et Wu 1985; Rao et Scott 1987). De même, pour en savoir plus sur le rapport entre eps$(p_i - p_j)$ et eps(p_i), nous pouvons utiliser les résultats empiriques issus de plusieurs variables et de plusieurs enquêtes.

Dans un premier essai portant sur cette question, nous présentons les données tirées de quatorze enquêtes réalisées dans une vaste gamme de situations. Onze de ces enquêtes présentent cinq ensembles de données (figures 1 et 2 et tableaux 1 à 3) et portent sur les différences de catégories appariées tirées d'enquêtes uniques (type A). Trois ensembles de résultats (tableaux 1 à 3) proviennent d'enquêtes sociales, et les deux autres (figures 1 et 2) proviennent d'enquêtes sur la démographie et la santé des populations. Finalement, trois autres ensembles comportant chacun deux vagues de données sont fondés sur deux nouvelles séries d'entrevues menées auprès des mêmes répondants (tableaux 4, 5 et 6), et représentent les plans de comparaison du type B.

Tableaux:

1. La National Election Study de 1986, réalisée par le Institute for Social Research de la University of Michigan, $n = 2{,}135$.
2. La National Education Longitudinal Study (NELS) de 1988, réalisée par le National Opinion Research Center de la University of Chicago, $n = 24{,}355$.
3. La National Longitudinal Study of Labor Market Experience of Youth, réalisée par le National Opinion Research Center de la University of Chicago, $n = 5{,}857$.
4. Les National Election Studies Panels de 1990 et de 1992, réalisés par le Survey Research Center, Institute for Social Research, Ann Arbor, Michigan 48106.
5. Les Panel Study of Income Dynamics de 1983 et de 1987, réalisés par le Survey Research Center.
6. Les études Americans' Changing Lives de 1986 et de 1989 réalisées par le Survey Research Center.

Figures:

1. Les études sur la démographie et la santé réalisées au Maroc, au Niger et en Colombie par MACRO International.
2. La portion du recensement indonésien portant sur la strate rurale de Java (données inédites).

Nous relevons en particulier dans ces études les résultats EMPIRIQUES suivants:

1) D'abord et avant tout: les effets du plan de sondage sur les différences eps$(p_i - p_j)$ ne sont habituellement PAS MOINS IMPORTANTS que les eps(p_i) sur les proportions elles-mêmes, et eps$(p_i - p_j) \simeq 0.5$ [eps(p_i) + eps(p_j)] dans tous les cas. Ils varient de concert avec l'importante variation des valeurs d'eps entre les variables, et avec les variations moins importantes observées entre les paires de catégories d'une même variable. Les chercheurs qui négligent les eps font l'erreur commune de sous-estimer les erreurs d'échantillonnage propres aux statistiques d'enquêtes par grappes. Outre l'intérêt qu'elle présente, cette observation fournit un modèle utile de déduction puisque les trois autres sources de variation - entre les variables, entre les catégories de chaque variable et erreurs d'échantillonnage des statistiques individuelles - sont toutes plus importantes.
2) On peut obtenir ces résultats avec les 14 ensembles de données d'enquêtes des tableaux et des graphiques, et en donner une illustration au tableau 1. Noter que les eps varient essentiellement entre 1.00 pour la variable D (problèmes dans le pays) et 2.32 pour la variable A (religion), ce qui nous donne une valeur d'eps^2 = $2.32^2 = 5.38$. Il est rassurant de constater que notre règle empirique (1) se vérifie pour toute la gamme des valeurs. On observe fréquemment une telle variation entre les variables d'un même échantillon; ceci devrait nous convaincre de cesser d'utiliser une moyenne commune pour tous les eps d'un échantillon (Verma et Lê 1995; Kish 1995).

   Il convient par ailleurs de souligner la grande variation des valeurs d'eps pour les cinq catégories de la même variable, soit de 1.21 à 2.32 (n° 3 pour les protestants "intégristes"). Il en découle que la valeur d'eps$(p_i - p_j)$ n'est grande que lorsque i ou j appartiennent à la catégorie 3 pour cette variable. Ces variations des eps entre les catégories de la même variable signifient qu'il vaudrait mieux les calculer pour toutes les catégories plutôt que pour une seule catégorie "représentative". La possibilité d'une grande variation entre les catégories pour une seule et même variable est une conclusion importante de notre étude qui était semble-t-il passée inaperçue jusqu'à maintenant.
3) Des erreurs d'échantillonnage sont également présentes dans le calcul des eps. Il semble que seuls les statisticiens qui ont une bonne expérience du traitement des erreurs d'échantillonnage et des effets du plan de sondage puissent avoir une idée de l'importance de ces erreurs. Ces erreurs risquent d'être les principales responsables

3. ERREURS D'ÉCHANTILLONNAGE ET EFFETS DU PLAN DE SONDAGE

Pour les échantillons aléatoires simples de taille n, on peut facilement montrer (Kish 1965, 12.10) que

$$\text{var}(p_2 - p_0) = \left[\frac{(1-f)n}{(n-1)}\right][p_2 + p_0 - (p_2 - p_0)^2]/n.$$

La plupart des exemples trouvés et utilisés sont tirés de grands échantillons d'enquêtes où la valeur de $(1 - f)$ peut être ignorée. Il convient de noter que pour la variance des éléments

$$p_2 + p_0 - (p_2 - p_0)^2 = p_2q_2 + p_0q_0 + 2p_2p_0,$$

où le dernier terme $\text{cov}(p_2,p_0) = -p_2p_0$ désigne la covariance découlant de ce que p_2 et p_0 sont des parties concurrentes du même échantillon, plutôt que les proportions d'échantillons indépendants. Le carré de la différence de proportions $(p_2 - p_0)^2$ sera habituellement un terme de correction petit; en l'excluant, nous obtiendrons l'équivalent de la variance $(p_2 + p_0)/n$ de deux échantillons de Poisson indépendants. On se rappellera en outre (Kish 1965, 12.10) que:

> Le test de chi carré a été utilisé pour certains de ces problèmes traités séparément (Cochran 1950; Mosteller 1952; McNemar 1962, p. 225). Il s'agit essentiellement de $(n_2 - n_0)^2/(n_2 + n_0)$, c'est-à-dire du carré de la différence divisé par la variance, en vertu de l'hypothèse nulle $n_2 = n_0$. On applique exactement les théories disponibles pour les tests d'hypothèses nulles dans les petits échantillons, y compris la "correction de Yates", toutes fondées sur l'hypothèse d'un échantillonnage aléatoire simple. Toutefois, le traitement de ces problèmes dans de grands échantillons avec des moyennes estimatives et des écarts-types adéquats présente d'énormes avantages. Premièrement, au lieu de nous limiter à tester les hypothèses nulles, nous pouvons faire des déductions fondées sur les intervalles de probabilité $(p_2 - p_0) \pm t_p$ é.-t. $(p_2 - p_0)$. Deuxièmement, les formules des écarts-types d'échantillons complexes peuvent s'appliquer directement à la moyenne $(p_2 - p_0)$. Troisièmement, la structure logique de cette statistique $(p_2 - p_0)$ s'observe plus facilement dans son application à plusieurs problèmes distincts.

Les proportions corrélées proviennent habituellement de données tirées d'enquêtes complexes, et les calculs de la variance devraient être appropriés pour le plan de sondage. On peut adopter les formules de la variance propres aux échantillons complexes stratifiés, mais la formule directe comporte huit termes (Kish 1965, 12.10.3). Au lieu de cela, il est plus pratique de traduire le problème en une variable trichotomique avec des valeurs de -1, 0, $+1$, comme dans le plan 2 de la section 2. Les calculs de la section 4 utilisent cette transformation.

Les comparaisons entre les variables et entre les échantillons peuvent alors être facilitées par le recours aux effets du plan de sondage:

$$\text{eps}^2(p_2 - p_0) = \frac{\text{variance calculée de } (p_2 - p_0)}{[p_2 + p_0 - (p_2 - p_0)^2]/n}.$$

Il convient de s'attarder quelque peu sur les limites de l'utilisation des eps en guise d'outils pour les approximations robustes. Ils sont utiles pour les échantillons en grappes et les échantillons à plusieurs degrés qui utilisent les sélections primaires pour le calcul des erreurs d'échantillonnage. Toutefois, nous avons voulu éviter le problème des échantillons pondérés parce que leur traitement serait trop spécifique et peut-être trop complexe. La pondération pour l'absence de réponse ne serait pas importante pour le rapport de $\text{eps}(p_i - p_j)$ sur $\text{eps}(p_i)$. Toutefois, la pondération des inégalités majeures des probabilités de sélection appelle des traitements particuliers. Néanmoins, la déduction et l'expérience portent à conclure que les valeurs d'eps sont moins sensibles aux poids que les variances et les moyennes elles-mêmes. Par ailleurs, nous présumons que les rapports que nous avons observés entre les valeurs d'$\text{eps}(p_i - p_j)$ et d'$\text{eps}(p_i)$ se maintiendront avec les données pondérées, à condition que ces dernières ne soient pas extrêmes ni pathologiques.

Il serait utile de déterminer un rapport approximatif, mais fiable, d'$\text{eps}(p_i - p_j)$ sur $\text{eps}(p_i)$ et $\text{eps}(p_j)$ afin de pouvoir déduire la valeur du premier, peu aisée à calculer, à partir des valeurs des seconds que l'on peut obtenir plus facilement. Plusieurs conjectures de rechange peuvent paraître raisonnables à cet égard; aucune ne peut être dérivée mathématiquement ni exclue.

1. $\text{Eps}(p_i - p_j) = 1$, l'absence d'effet du plan de sondage, a été implicitement présumée dans les cinq publications citées à la section 1.

2. $\text{Eps}(p_i) > \text{eps}(p_i - p_j) > 1$, désignant des effets persistants, mais moins importants que les $\text{eps}(p_i)$ pour les proportions. C'est ce qui se produit dans le cas des "classes croisées" et de leurs comparaisons (Kish 1987, 7.1). Cette hypothèse a également paru raisonnable à plusieurs statisticiens d'expérience que nous avons consultés.

3. $\text{Eps}(p_i - p_j) = [\text{eps}(p_i) + \text{eps}(p_j)]/2$ nous a semblé être une bonne approximation pour l'ensemble de nos données pour diverses populations et divers plans de sondage. Cette hypothèse nous paraît raisonnable puisque les effets du plan de sondage dus au groupement des valeurs individuelles de p_i peuvent s'appliquer de façon similaire à la variable issue de la différence $(p_i - p_j)$ entre deux de ces valeurs.

4. Des résultats incohérents seraient également possibles, mais peu utiles puisqu'ils empêcheraient toute déduction.

Dans la section 2, nous décrivons les cinq problèmes apparentés (plans de sondage) au sujet desquels nous décrivons les erreurs d'échantillonnage à la section 3. Dans la section 4, nous examinons la constatation empirique exposée dans les tableaux. Dans la section 5, nous plaçons nos résultats dans le contexte des travaux antérieurs portant sur les valeurs d'eps pour les sous-classes et leurs différences.

2. STATISTIQUES SEMBLABLES POUR LES CINQ PLANS DE SONDAGE

On a montré que cinq plans de sondage (problèmes) appartenant à deux types distincts peuvent être traités à l'aide des mêmes statistiques simples (Kish 1965, Section 12.10). Pour notre présentation à la fois empirique et simple, nous utilisons des symboles qui dénotent les valeurs d'échantillons (eps, p_i et n_i), même si à l'occasion, l'usage de majuscules pour les valeurs de la population totale serait plus approprié.

La différence de proportions $p_2 - p_0 = n_2/n - n_0/n$ dénote l'estimation souhaitée, où $n = n_0 + n_1 + n_2 + \ldots$ n_k désigne la taille de l'échantillon, avec n unités choisies et pondérées également. Par ailleurs, en vertu des hypothèses de l'échantillonnage aléatoire simple, la variance de $(p_2 - p_0)$ est $(1 - f)[p_2 + p_0 - (p_2 - p_0)^2]/(n - 1)$.

Comparaisons de type A

1. La différence entre deux catégories $(n_2 - n_0)/n = (p_2 - p_0)$ d'une polytomie peut représenter la préférence manifestée pour deux partis parmi plusieurs (k) dans un sondage sur les intentions de vote, pour deux marques d'automobiles dans une étude de marchés, pour deux types d'attitudes, d'opinions ou de comportements concernant une variable quelconque, *etc.* Les autres choix $(k - 2)$ sont pris en compte dans p_1 et sont ignorés dans l'évaluation de la différence. (Voir à ce propos Scott et Seber 1983)

2. On peut attribuer les valeurs de rang de $-1, 0, +1$ (ou $0, 1, 2$ ou $c, c + 1, c + 2$) à une variable trichotomique ordonnée sans métrique, et les assimiler à une forme simple de la différence entre deux catégories. Cette forme est particulièrement utile pour le calcul des erreurs d'échantillonnage puisque les cinq plans de sondage peuvent par exemple utiliser $-1, 0, +1$ à titre de variable de calcul transformée.

3. La différence de proportions de deux variables différentes (x et y) peut être traitée comme en (1) et en (2). Ne définir comme positifs en x (succès) que les éléments qui sont positifs en x, mais non en y, de manière que $n_{10} = n(x_1, y_o)$. Définir de la même manière comme positifs en y les $n_{01} = n(x_0, y_1)$. Ainsi, $(n_{10} - n_{01})/n = (p_x - p_y)$ représentera la différence nette dans la proportion des éléments positifs en x et y. Les éléments qui sont positifs ou négatifs à la fois en x et y ne comptent pas dans les différences. Nous obtenons ainsi un cas de trois catégories comme en (1) et en (2). On peut citer à titre d'exemple la différence entre les proportions de gens qui "interdiraient tous les essais nucléaires" et ceux qui "souhaitent un désarmement nucléaire complet", ou entre ceux qui "obligeraient les Irakiens à quitter le Koweit" et ceux qui voudraient "remplacer Saddam" (Wild et Seber 1993). Toutefois, les deux catégories pourraient également être tirées de deux *sondages différents* portant sur les mêmes n cas, si on avait affaire par exemple à une vérification de la qualité, à des observations provenant de deux bases de sondage ou aux deux vagues d'un même échantillon. Ces situations ressemblent à celles examinées aux points (4) et (5).

Comparaisons du type B

4. Les expressions test-retest et avant-après qualifient des plans dans lesquels les mêmes sujets font l'objet de deux observations. Les réponses dichotomiques $n_2 = n_{10}$ désignent alors le nombre de changements négatifs; $n_0 = n_{01}$ le nombre de changements positifs et $n_{11} + n_{00}$ la somme des positifs et des négatifs inchangés. On dénote respectivement les réponses positives et négatives par 1 et 0, et les première et deuxième vagues par l'ordre de l'indice. La différence $(n_{10} + n_{11}) - (n_{01} + n_{11}) = n_{10} - n_{01} = n_2 - n_0$ représente une mesure du changement entre les positifs pour les deux observations, et $p_2 - p_0 = n_2/n - n_0/n$ représente la mesure du changement des proportions (McNemar 1949; Cochran 1950; Mosteller 1952).

5. Les paires appariées de n paires de sujets peuvent également être traitées comme une généralisation du plan test-retest (Mosteller 1952). Par exemple, n paires de sujets randomisés peuvent représenter le traitement expérimental par rapport au traitement témoin; ou n paires de garçons ou de filles par rapport aux variables de contrôle. Le traitement statistique $(p_{10} - p_{01})$ des n paires de sujets appariés est le même que celui des n paires de traitements du même groupe de n sujets (4).

La similitude des traitements statistiques de ces cinq plans de deux types distincts est pratique; nous présentons les résultats empiriques obtenus pour les deux types. "Elle a également une valeur heuristique qu'on a ignorée dans des publications récentes (Scott et Seber 1983; Wild et Seber 1993). Les résultats des tests de chi carré pour les types 4 et 5 ont été publiés depuis longtemps (McNemar 1949; Cochran 1950; Mosteller 1952) et la similitude des cas des catégories 1, 2 et 3 a également été démontrée" (Kish 1965, 12.10). (Kish faisait erreur lorsqu'il a parlé de "trinômes et de binômes appariés" pour désigner les "trichotomies et les dichotomies appariées" puisque ces termes s'appliquent uniquement aux échantillons IID.)

Tous ces traitements portent sur les différences des proportions p_i fondées sur les variables de dénombrement n_i. Les extensions aux différences corrélées $(y_i - y_j)$ pour d'autres variables sont envisageables, mais elles sortent du cadre de la présente étude. On pourrait citer à titre d'exemple pratique la différence entre les parts en dollars (pas seulement les nombres n_i) entre deux marques d'automobiles sur un total de $\sum y_i$ ventes.

Techniques d'enquête, décembre 1995
Vol. 21, n° 2, pp. 131–139
Statistique Canada

Effets du plan de sondage sur les $(P_i - P_j)$ corrélés

LESLIE KISH, MARTIN R. FRANKEL, VIJAY VERMA et NIKO KAĆIROTI[1]

RÉSUMÉ

En nous fondant sur 14 enquêtes menées dans six pays, nous présentons la constatation empirique de l'existence et de l'ampleur des effets du plan de sondage (eps) pour cinq plans appartenant à deux types principaux. Le premier type a trait à eps$(p_i - p_j)$, la différence de deux proportions d'une variable polytomique de trois catégories ou plus. Le deuxième type utilise les tests de chi carré pour l'analyse des différences entre deux échantillons. Nous montrons que pour toutes les variables et pour tous les plans, eps$(p_i - p_j) \cong [\text{eps}(p_i) + \text{eps}(p_j)]/2$ constituent de bonnes approximations. Ces résultats sont *empiriques*, et les exceptions prouvent qu'il ne peut s'agir de simples inégalités analytiques. Il convient de signaler que ces résultats restent valables malgré les grandes variations des valeurs d'eps entre les variables et entre les catégories d'une même variable. Ils montrent en outre la nécessité d'utiliser des méthodes de traitement adaptées aux échantillons d'enquêtes pour l'analyse des données d'enquête, même lorsqu'on a affaire à des statistiques analytiques. En outre, ils permettent d'utiliser des approximations d'eps$(p_i - p_j)$ tirées des valeurs plus facilement accessibles d'eps(p_i).

MOTS CLÉS: Effets du plan de sondage; échantillonnage d'enquête; erreurs d'échantillonnage.

1. EFFETS DU PLAN DE SONDAGE SUR LES STATISTIQUES ANALYTIQUES

Nous nous penchons sur l'existence et sur l'ampleur des effets du plan de sondage sur certaines statistiques analytiques spéciales, en nous servant de données tirées d'échantillons d'enquêtes. Notre étude s'inspire à la fois de la méthode et de l'empirisme; elle s'appuie sur des données venant de plusieurs enquêtes aux variables différentes et provenant de populations contrastantes, et s'expose donc aux risques que présentent les résultats empiriques incohérents. On dit et on écrit souvent que l'échantillonnage probabiliste, bien que nécessaire à la réalisation d'enquêtes descriptives, n'est pas nécessaire pour les enquêtes analytiques. Dans une section intitulée "Four Obstacles to Representation in Analytic Studies", l'un de nous écrit qu'en plus des quatre obstacles concrets à la représentation auxquels il est fait allusion, il en existe un autre plus artificiel: le refus d'admettre la nécessité de la représentation (Kish 1987, section 2.7). Les études sur l'échantillonnage montrent que les méthodes de sélection probabiliste complexes, en particulier l'échantillonnage par grappes, n'ont pas d'effet appréciable sur les statistiques descriptives (p. ex., les moyennes et les coefficients de régression), mais qu'elles peuvent influer d'une façon extrêmement marquée sur les statistiques déductives comme les intervalles de confiance et les tests de signification (Kish et Frankel 1974).

Les effets du plan de sondage (eps) sont définis par la formule eps^2 = variance réelle/variance aléatoire simple, pour une même taille d'échantillon n (les deux variances étant estimatives). On a observé des valeurs d'eps supérieures à 1 pour des erreurs d'échantillonnage concernant non seulement les moyennes, mais également des statistiques analytiques comme les écarts de moyennes (et les tests de chi carré), les coefficients de régression, *etc.* Il est vrai qu'on a observé des réductions et des différences considérables des valeurs d'eps pour certaines statistiques analytiques. Les variations des valeurs d'eps ne sont pas une simple conséquence mathématique inévitable du choix du plan de sondage que l'on peut déduire une fois pour toutes. Elles ont un contenu empirique et doivent donc être reproductibles dans le cadre d'études empiriques (Kish et Frankel 1974; Kish 1987, 7.1; Kish 1965, 14.1-14.2; Rao et Wu 1985; Scott et Holt 1982; Skinner, Holt et Smith 1989). Dans le présent article, nous examinons l'incidence possible et l'importance des effets du plan de sondage sur un ensemble de statistiques apparentées qui n'ont fait l'objet d'aucune étude antérieure. Il se trouve que, dans de nombreux articles, des auteurs reconnus à bon droit ont simplement présumé que le choix du plan de sondage était sans effet, sans fournir aux lecteurs les mises en garde appropriées, et que cela n'a soulevé aucune objection de la part des examinateurs. Nous vérifierons donc si les eps sont réduits ou éliminés pour cet ensemble de statistiques analytiques (Cochran 1950; Mosteller 1952; Scott et Seber 1983; Seber et Wild 1993).

Par ailleurs, nous proposerons explicitement ce à quoi nous avons déjà fait allusion auparavant, à savoir que l'existence d'effets de sondage importants justifie largement le recours à la sélection probabiliste. Il serait difficile d'imaginer un modèle de distribution de population où le plan de sondage serait sans importance (ou sans intérêt) tout en produisant des effets considérables. Toutefois, à l'inverse, l'absence d'eps est nécessaire, quoique non suffisante, pour justifier la décision de ne pas recourir à la sélection probabiliste. Cette proposition donne du poids à notre étude qui met en rapport les valeurs d'eps$(p_i - p_j)$ des statistiques analytiques d'une part, et celles d'eps(p_i) et d'eps(p_j) pour deux des nombreuses catégories de la même variable d'autre part.

[1] Leslie Kish, ISR, University of Michigan, Ann Arbor MI 48106, U.S.A.; Martin R. Frankel, NORC et City University of New York; Vijay Verma, University of Essex, Colchester, C04 3SQ, U.K.; Niko Kaćiroti, Institut de statistiques, Tirana, Albanie.

HICKS, S., et FETTER, M. (1993). An evaluation of robust estimation techniques for improving estimates of total hogs. *Proceedings of the International Conference on Establishment Surveys*. Alexandria, Virginia: American Statistical Association, 385-389.

HIDIROGLOU, M.A. (1987). The construction of a self-representing stratum of large units in survey design. *The American Statistician*. 40, 27-31.

HUBER, P.J. (1981). *Robust Statistics*. New York: John Wiley.

LEE, H. (1994). Outliers in Survey Data. Document de Statistique Canada.

RIVEST, L.-P. (1994). Some sampling properties of winsorized means for skewed distributions. *Biometrika*, 81, 373-384.

RIVEST, L.-P. (1993a). Winsorization of survey data. *Bulletin de l'Institut International de Statistique, Actes de la 49ième Session*, livre 2, 73-89.

RIVEST, L.-P. (1993b). Winsorization of survey data. *Proceedings of the International Conference on Establishment Surveys*. Alexandria, Virginia: American Statistical Association, 396-401.

SEARLS, D.T. (1966). An estimator which reduces large true observations. *Journal of the American Statistical Association*, 61, 1200-1204.

THISTED, R.A. (1988). *Elements of Statistical Computing*. New York: Chapman and Hall.

VON MISES, R. (1964). *Selected Papers of Richard von Mises*. Volume II. Providence: American Mathematical Society.

$$\frac{\psi(R) - E(Y)}{n-1} < \int_{\psi(R)}^{\infty} [1 - F_Y(x)]\,dx. \quad \text{(A.1)}$$

Selon l'inégalité de Jensen, $E(Y) = E[\psi(X)] > \psi[E(X)]$. Ainsi, en utilisant (2.3), le membre de gauche de l'équation (A.1) est inférieur ou égal à

$$\frac{R - E(X)}{n-1}\,\frac{\psi(R) - \psi[E(X)]}{R - E(X)} < \frac{R - E(X)}{n-1}\,\psi'(R) = \int_{R}^{\infty} [1 - F_X(y)]\,dy \cdot \psi'(R)$$

où ψ' est la dérivée de ψ. Comme ψ' augmente, le membre de gauche de l'inégalité ci-dessus est inférieur ou égal à:

$$\int_{R}^{\infty} \psi'(y)[1 - F_X(y)]\,dy = \int_{\psi(R)}^{\infty} [1 - F_Y(x)]\,dx.$$

ce qui prouve la validité de (A.1).

Preuve de la proposition 2 Le résultat suivant obtenu en appliquant les théorèmes 2.7.5 et 2.7.11 de Galambos (1987) à la distribution $F(z^{p+1})$ est largement mis à contribution. Si le maximum de l'échantillon de la distribution $F(x)$ tend vers $H_{3,0}(x)$, tous les moments de F existent et

$$\int_{x}^{\infty} y^p[1 - F(y)]\,dy \sim \frac{[1 - F(x)]\,x^p}{g(x)} \quad \text{(A.2)}$$

où $g(x) \sim h(x)$ signifie que $g(x)/h(x)$ tend vers 1 à mesure que x tend vers l'infini. À l'aide de la formule (A.2), on obtient $R(F,n)$ en résolvant

$$\frac{R - \mu}{n-1} = \frac{1 - F(R)}{g(R)}\,(1 + o(1)).$$

Soit $R = F^{-1}(1 - a/n)$, alors, jusqu'à la valeur $(1 + o(1))$, l'équation ci-dessus devient

$$a = g\left[F^{-1}\left(1 - \frac{a}{n}\right)\right]F^{-1}\left(1 - \frac{a}{n}\right). \quad \text{(A.3)}$$

Soit $a_0 = g[F^{-1}(1 - 1/n)]F^{-1}(1 - 1/n)$ et $a_1 = g[F^{-1}(1 - a_0/n)]F^{-1}(1 - a_0/n)$. Puisque pour les grandes valeurs de $x, xg(x)$ augmente, $a_0 > a_1$ et la solution de (A.3) tombe dans l'intervalle (a_1, a_0). Pour prouver le résultat, il suffit de montrer que a_1/a_0 tend vers 1 à mesure que n tend vers l'infini.

Puisque $g(x) = f(x)/[1 - F(x)]$, on peut poser

$$a_0 = \exp\left[\int_{F^{-1}(1-a_0/n)}^{F^{-1}(1-1/n)} g(t)\,dt\right] = \exp\left[a_0 - a_1 - \int_{F^{-1}(1-a_0/n)}^{F^{-1}(1-1/n)} tg'(t)\,dt\right],$$

où la seconde expression est obtenue en intégrant par parties. Puisque $tg'(t)/g(t)$ est inférieur à c, on obtient $a_0 > \exp(a_0 - a_1)a_0^{-c}$. Si a_1/a_0 ne tend pas vers 1, c'est-à-dire par exemple que $a_1/a_0 < 1 - \epsilon < 1$ pour une séquence infinie de tailles d'échantillons, l'inégalité précédente signifie que $a_0^{1+c} > \exp(a_0\epsilon)$. Ce résultat est contradictoire puisque a_0 tend vers l'infini à mesure que n devient plus grand. L'approximation de $\mathrm{MSE}(\bar{X}_R)$ s'obtient en utilisant la formule (A.2) avec $p = 2$.

Preuve de la proposition 3 Si le maximum de l'échantillon de la distribution $F(x)$ tend vers $H_{1,\alpha}(x)$, F satisfait alors aux propriétés suivantes (Feller 1971, p. 281):

$$\int_{x}^{\infty} y^p[1 - F(y)]\,dy \sim \frac{[1 - F(x)]x^{p+1}}{\alpha - p - 1} \quad \text{(A.4)}$$

pour toute valeur de p telle que $\alpha - p - 1 \geq 0$. Avec la formule (A.4), $R(F,n)$ s'obtient en résolvant $F(R) = 1 - [\alpha - 1 + o(1)]/n$. Ceci conduit à l'approximation de $R(F,n)$. Pour calculer l'approximation de $\mathrm{MSE}(\bar{X}_R)$, on utilise (A.4) avec $p = 1$.

BIBLIOGRAPHIE

BARLOW, R.E., et PROSCHAN, F. (1981). *Statistical Theory of Reliability and Life Testing*. Silver Spring MD: To Begin With.

CHAMBERS, R.L., et KOKIC, P.N. (1993). Outlier robust sample survey inference. *Bulletin de l'Institut International de Statistique, Actes de la 49ième Session*, livre 2, 54-72.

ERNST, L.R. (1980). Comparison of estimators of the mean which adjust for large observations. *Sankhyā* C, 42, 1-16.

FELLER, W. (1971). *An Introduction to Probability Theory and its Applications*. Volume II. Deuxième édition. New York: Wiley.

FULLER, W.A. (1991). Simple estimators for the mean of skewed populations. *Statistica Sinica*, 1, 137-158.

FULLER, W.A. (1993). Estimators for long-tailed distributions. *Bulletin de l'Institut International de Statistique, Actes de la 49ième Session*, livre 2, 39-54.

GALAMBOS, J. (1987). *The Asymptotic Theory of Extreme Order Statistics*. Deuxième édition. Malabar FL: Krieger.

GNEDENKO, B.V. (1962). *The Theory of Probability*. New York: Chelsea.

GLASSER, G.J. (1962). On the complete coverage of large units in a statistical study. *Revue de l'Institut International de Statistique*, 30, 28-32.

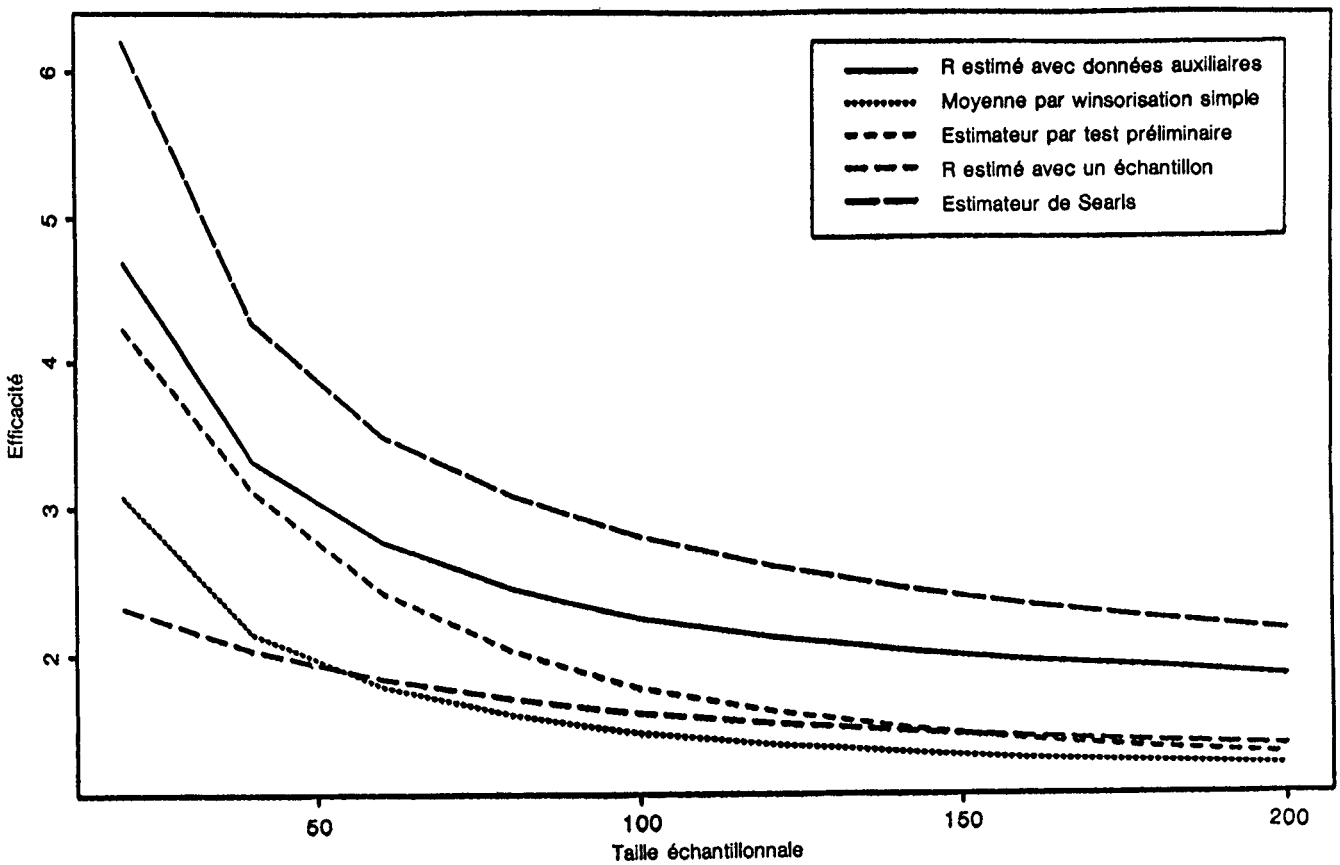


Figure 4. Efficacité de cinq estimateurs pour la moyenne du nombre de poulets.

avec les petits échantillons. Ainsi, comme nous l'avons montré dans la section 4, l'estimateur obtenu est très sensible aux valeurs aberrantes qui apparaissent parfois dans les petits échantillons. Cet estimateur n'est pas recommandé.

6. CONCLUSIONS

Un grand nombre de stratégies peuvent nous permettre de traiter les grandes valeurs qui apparaissent parfois lors des enquêtes. Si on dispose d'informations auxiliaires (par exemple, des données de recensement), on peut utiliser l'estimateur de Searls avec l'échantillonnage aléatoire simple, l'échantillonnage stratifié ou l'échantillonnage avec ppt. Comme les valeurs limites sont des constantes fixes, les estimateurs de l'erreur quadratique moyenne peuvent être dérivés à l'aide des formules (2.3) et (3.1).

En l'absence d'informations auxiliaires, on peut utiliser la moyenne de la winsorisation simple et l'estimateur de test préliminaire de Fuller. Des travaux sont en cours afin de généraliser ces estimateurs avec les plans d'échantillonnage stratifié. Rivest (1994) propose un estimateur de l'erreur quadratique moyenne de la moyenne de la winsorisation simple:

$$v(\bar{X}_1) = \frac{1}{n} S^2 - \frac{1}{n^2} (X_n + X_{n-1} - 2\bar{X}_1) (X_n - 3X_{n-1} + 2X_{n-2})$$

où S^2 désigne la variance de l'échantillon X et $X_n > X_{n-1} > X_{n-2}$ désignent les trois plus grandes valeurs de cet échantillon. Cet estimateur présente un léger biais dans les populations infinies. Toutefois, la couverture de l'intervalle de confiance normal est souvent bien en dessous du niveau nominal de $100(1 - \alpha)\%$, en particulier lorsque la distribution sous-jacente est asymétrique. De plus amples recherches sont nécessaires pour obtenir des intervalles de confiance fiables pour les estimateurs de la moyenne des populations asymétriques.

REMERCIEMENTS

La présente recherche a bénéficié de l'aide financière du Conseil de recherches en sciences naturelles et en génie et du Fonds pour la formation des chercheurs et l'aide à la recherche du Québec.

ANNEXE 1

Preuve de la proposition 1 La supposition selon laquelle Y est plus asymétrique que X signifie qu'il existe une fonction convexe ψ telle que $\psi(X)$ et Y suivent la même distribution. Désignons par R la valeur $R(F_X, n)$. Pour prouver le résultat, il suffit de montrer que $\psi(R) < R(F_Y, n)$. Ceci équivaut à dire que

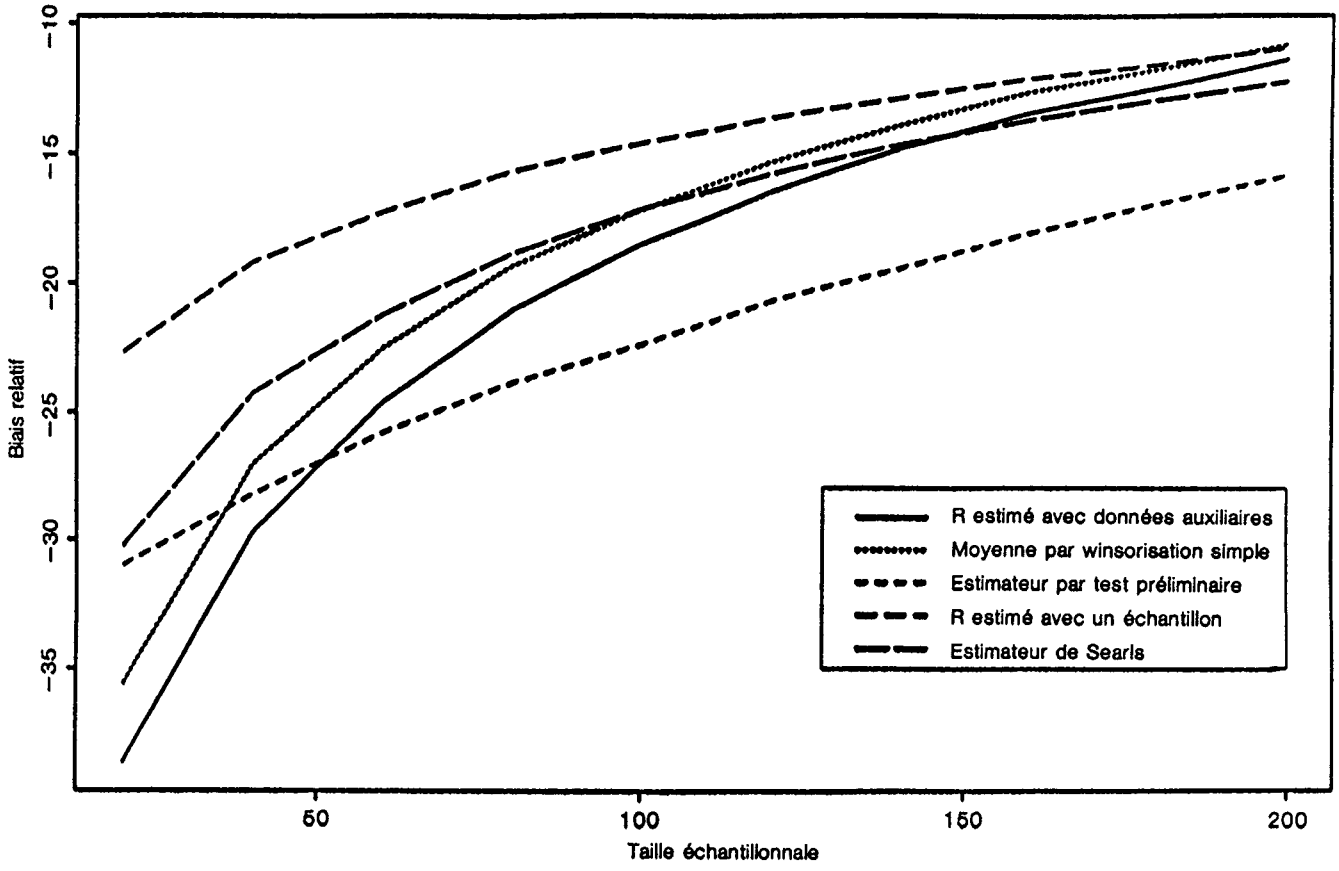


Figure 3. Biais relatif de cinq estimateurs pour la moyenne du nombre de poulets.

Les cinq estimateurs considérés sont:

- L'estimateur winsorisé de Searls, $\bar{X}_R$, calculé comme si la distribution sous-jacente était connue;
- Un estimateur winsorisé où la valeur limite correspond à la deuxième valeur en importance d'un échantillon auxiliaire de taille $2n$; il s'agit d'un cas où on dispose d'informations auxiliaires limitées sur la distribution sous-jacente F (dans les simulations de Monte Carlo, chaque échantillon simulé possède son propre échantillon auxiliaire);
- La moyenne $\bar{X}_1$ de la winsorisation simple, présentée dans la section 4;
- Un estimateur winsorisé où R est estimé à partir de l'échantillon par la résolution de l'équation (4.3);
- L'estimateur d'un test préliminaire plus complet avec $j = 3$ (c.-à-d., le numérateur du test préliminaire utilise les trois plus grandes observations), T (le nombre total de points de données utilisés dans le test préliminaire) étant égal à $[4n^{1/2} - 10]$ et K_3, la valeur limite, étant égale à 3.5. On fournit une description détaillée de cet estimateur dans Fuller (1991) et dans Rivest (1993a et b). Cet estimateur ne réduit les valeurs les plus grandes que lorsqu'un test statistique de détection des valeurs extrêmes donne des résultats significatifs.

Les biais et les valeurs de l'efficacité de $\bar{X}_R$ ont été calculés exactement. Pour les autres estimateurs, les résultats présentés dans les figures 1 et 2 ont été obtenus à l'aide des simulations de Monte Carlo, avec 100,000 répétitions.

La figure 3 montre que les biais des estimateurs winsorisés sont importants, même dans le cas des grands échantillons. On peut par ailleurs tirer plusieurs conclusions importantes de la figure 4. Premièrement, comme le laissait prévoir le tableau 2, l'estimateur de Searls est beaucoup plus efficace que la moyenne de la winsorisation simple. En outre, l'estimation de la valeur limite optimale à l'aide d'une information auxiliaire limitée est extrêmement efficace. Ceci reste vrai tant que la variable à l'étude peut être modélisée par la distribution d'une superpopulation assortie d'une variance finie (pour en savoir plus, voir Rivest 1993a). Dans le contexte d'un échantillonnage, les échantillons auxiliaires pourraient être constitués de données provenant des enquêtes antérieures normalisées pour tenir compte des changements risquant d'être survenus avec le temps dans la distribution de la variable à l'étude.

Des trois estimateurs de la figure 4 qui ne dépendent pas de l'information auxiliaire, l'estimateur de Fuller est le meilleur. Cette constatation s'accorde avec les résultats de simulation de Fuller (1991). L'estimation de la valeur limite en minimisant une estimation de l'erreur quadratique moyenne donne des résultats inadéquats, en particulier

Tableau 2

Approximations du biais et de l'erreur quadratique moyenne de la moyenne de la winsorisation simple $\bar{X}_1$ et de la moyenne winsorisée optimale de Searls, $\bar{X}_R$, pour les distributions de Weibull et de Pareto ($\Gamma(\cdot)$ représente la fonction gamma)

		WEIBULL	PARETO
$\bar{X}_R$	MSE	$\frac{\sigma^2}{n} - \frac{(\log n)^{2\alpha}}{n^2}$	$\frac{\sigma^2}{n} - \frac{\gamma}{(\gamma-2)(\gamma-1)^{2/\gamma}n^{2-2/\gamma}}$
	bias	$-\frac{(\log n)^{\alpha}}{n}$	$-\frac{1}{(\gamma-1)^{1/\gamma}n^{1-1/\gamma}}$
$\bar{X}_1$	MSE	$\frac{\sigma^2}{n} - \frac{2\alpha(\alpha-1)(\log n)^{2\alpha-2}}{n^2}$	$\frac{\sigma^2}{n} - \frac{2\Gamma(1-2/\gamma)}{\gamma(\gamma-1)n^{2-2/\gamma}}$
	bias	$-\frac{\alpha(\log n)^{\alpha-1}}{n}$	$-\frac{\Gamma(1-1/\gamma)}{\gamma n^{1-1/\gamma}}$

$$\frac{R - \bar{X}}{n-1} = \frac{1}{n}\sum_{i=1}^{n} \max(X_i - R, 0) \qquad (4.3)$$

pour fonction d'estimation de R. Cette méthode est contestable lorsque la distribution sous-jacente est fortement asymétrique, c'est-à-dire lorsque F satisfait l'hypothèse de la proposition 3. En moyenne, il n'y aura que des termes $\alpha - 1$ non nuls dans le membre de droite de l'équation (3). Ainsi, $\hat{R}$ sera déterminé, en moyenne, par les $\alpha - 1$ plus grandes valeurs et le maximum de l'échantillon aura la plus grande incidence sur $\hat{R}$. Ceci rendra la valeur $\hat{R}$ extrêmement instable et, compte tenu des constatations faites concernant la figure 1, la valeur venant au deuxième rang pour sa grandeur deviendra probablement un estimateur plus adéquat de $R(F,n)$ que la solution de (3.3). Les simulations de Monte Carlo présentées dans la section 5 sont instructives à cet égard.

Le tableau 2 compare les approximations du biais et de l'erreur quadratique moyenne de la moyenne winsorisée de Searls, $\bar{X}_R$, à celles de la moyenne de la winsorisation simple $\bar{X}_1$, obtenue en utilisant une valeur limite R égale à la deuxième plus grande observation. Rivest (1994) montre que ce choix de la valeur limite donne la moyenne winsorisée non paramétrique optimale. Il dérive également les approximations, pour grands échantillons, du biais et de l'erreur quadratique moyenne de $\bar{X}_1$ qui sont présentées dans le tableau 2. Les expressions correspondantes pour $\bar{X}_R$ sont tirées des propositions 2 et 3. Dans le tableau 2, l'erreur quadratique moyenne de $\bar{X}_R$ est beaucoup plus petite que celle de $\bar{X}_1$. En fait, pour la distribution de Weibull, l'efficacité de $\bar{X}_R$ par rapport à $\bar{X}_1$ pour les grands échantillons est égale à celle de $\bar{X}_R$ par rapport à $\bar{X}$. Ainsi, la winsorisation non paramétrique réduit l'erreur quadratique moyenne des estimateurs de la moyenne d'une population asymétrique, mais on peut obtenir une réduction supplémentaire de l'erreur quadratique moyenne si on dispose d'informations sur la distribution sous-jacente. Les comparaisons de Monte Carlo présentées dans la section suivante sont instructives à cet égard.

Les résultats de la présente section s'appliquent à l'échantillonnage stratifié. Pour ce type de plan, la solution de l'équation (3.4) pour un grand échantillon est déterminée par la strate dont la distribution est la plus asymétrique. Si F_1 est la distribution la plus asymétrique, alors $nW_1R(F_1,n_1)/n_1$ constitue une solution approximative de (3.4), l'approximation de $R(F_1,n_1)$ étant tirée de la proposition 2 ou de la proposition 3, selon l'importance de l'asymétrie de F_1. Dans ce cas, seuls les observations de la strate 1 sont winsorisées dans les grands échantillons stratifiés. La moyenne winsorisée de Searls est alors égale à W_1 fois la moyenne winsorisée optimale pour la strate 1, additionnée d'une somme pondérée des moyennes de l'échantillon des autres strates.

5. COMPARAISONS DE MONTE CARLO DES ESTIMATEURS DE LA MOYENNE D'UNE DISTRIBUTION ASYMÉTRIQUE

Nous présentons ci-après les comparaisons de Monte Carlo de l'erreur quadratique moyenne et du biais de cinq estimateurs de la moyenne du nombre de poulets par segment d'une population comportant 2,000 unités (Fuller 1991). Le coefficient de variation est de 4.46. De plus amples comparaisons numériques des cinq estimateurs examinés dans la présente section pour d'autres distributions, finies ou infinies, sont présentées par Rivest (1993a et b).

4. APPROXIMATIONS DE L'EFFICACITÉ DE LA MOYENNE WINSORISÉE POUR LES GRANDS ÉCHANTILLONS

Pour la plupart des distributions, l'équation (2.3) définissant la valeur limite optimale n'a pas de solution explicite. Dans la présente section, nous obtenons des approximations directes de cette solution en utilisant la théorie des statistiques d'ordre extrême. Cette méthode nous permettra de calculer des approximations explicites de l'efficacité de la moyenne winsorisée optimale. Nous comparerons ensuite la méthode de winsorisation optimale de Searls à une méthode simple de winsorisation non paramétrique où la variable statistique la plus grande est remplacée par celle qui vient au second rang (Rivest 1994).

La forme de l'approximation de $R(F,n)$ dépend de la distribution limitante de la variable statistique d'ordre supérieur adéquatement normalisée, lorsque la taille de l'échantillon n tend vers l'infini. Pour les distributions tracées sur un axe positif, il n'existe que deux distributions limitatives possibles qui sont données par Galambos (1987, p. 53-54):

$$H_{1,\alpha}(x) = \exp(-x^{-\alpha}) \quad \text{pour} \quad x > 0 \quad \text{et} \quad \alpha > 0$$

et

$$H_{3,0}(x) = \exp[-\exp(-x)] \quad \text{pour} \quad x \quad \text{dans} \quad R.$$

Pour beaucoup des distributions utilisées dans l'analyse statistique des variables aléatoires positives (par exemple, les familles de Weibull et log-normale), le maximum de l'échantillon, adéquatement normalisé, tend vers $H_{3,0}(x)$. Les distributions dont le maximum de l'échantillon tend vers $H_{1,\alpha}(x)$ pour un certain $\alpha > 0$ sont caractérisées par des queues importantes. Pour de telles distributions, $1 - F(x)$ tend vers 0 à un taux de $O(x^{-\alpha})$. Les distributions de Pareto et les distributions de F appartiennent à cette classe.

Les distributions dont le maximum de l'échantillon tend vers $H_{3,0}(x)$ sont examinées les premières. La caractérisation suivante est tirée de von Mises (1964): le maximum de l'échantillon d'une distribution doublement différentiable $F(x)$ tend vers $H_{3,0}(x)$ si, lorsque x tend vers l'infini,

$$\lim_x \frac{g'(x)}{g^2(x)} = 0 \tag{4.1}$$

où $f(x)$ désigne la densité de F, $g(x) = f(x)/[1 - F(x)]$ désigne le taux d'échec de F, et g' est la dérivée de g. Nous proposons ci-après une approximation de la constante de winsorisation $R(F,n)$ pour cette classe de distributions.

Proposition 2 Si $F(x)$ est telle que l'équation (4.1) est valide et si, pour les grandes valeurs de x, elle répond aux conditions suivantes:

i) $xg(x)$ augmente;
ii) $xg'(x)/g(x)$ est inférieur à une certaine constante positive c;

alors la constante de winsorisation optimale $R(F,n)$ satisfera

$$R(F,n) = F^{-1}\left(1 - \frac{g[F^{-1}(1-1/n)]F^{-1}(1-1/n)[1+o(1)]}{n}\right);$$

et $m(F,n) = g(F^{-1}(1-1/n))F^{-1}(1-1/n)[1+o(1)]$. En outre, l'erreur quadratique moyenne de la moyenne winsorisée de Searls sera approximativement égale à:

$$\mathrm{MSE}(\bar{X}_R) \approx \frac{\sigma^2}{n} - \frac{R(F,n)^2}{n^2}.$$

Dans la famille de Weibull, $F_\alpha^{-1}(1-t) = [-\log(t)]^\alpha$, $g(x) = x^{1/\alpha - 1}/\alpha$. Les hypothèses de la proposition 2 sont réalisées et $m(F_\alpha,n)$, le nombre prévu d'observations winsorisées dans un grand échantillon de Weibull, est donné par $\log(n)[1 + o(1)]/\alpha$ qui tend vers l'infini à mesure que n augmente. La figure 1 donne à penser que la convergence est très lente, en particulier pour les grands coefficients de variation.

Examinons maintenant les distributions dont le maximum de l'échantillon tend vers $H_{1,\alpha}(x)$. Cette classe de distributions a été caractérisée par Gnedenko (1962): le maximum de l'échantillon de F converge vers $H_{1,\alpha}(x)$ si on peut affirmer que

$$1 - F(x) = L(x)/x^\alpha \tag{4.2}$$

où à mesure que x tend vers l'infini, $L(x)/L(kx)$ converge vers 1 pour n'importe quelle constante k. Notons que pour que F possède un deuxième moment fini, il suffit que $\alpha > 2$ dans (4.2). La distribution de Pareto satisfait à (4.2) avec $\alpha = \gamma$.

Proposition 3 Si F satisfait à (4.2) avec le paramètre $\alpha > 2$, alors, à mesure que n tend vers l'infini, $R(F,n) = F^{-1}[1 - (\alpha - 1)/n][1 + o(1)]$, c.-à-d., $m(F,n) \approx \alpha - 1$. En outre,

$$\mathrm{MSE}(\bar{X}_R) \approx \frac{\sigma^2}{n} - \frac{\alpha R(F,n)^2}{n^2(\alpha - 2)}.$$

Pour les distributions qui satisfont à (4.2), un nombre fini d'observations sont en moyenne winsorisés à mesure que l'échantillon tend vers l'infini. Ceci s'observe dans une certaine mesure dans la Figure 1, où les courbes de $m(F_\gamma,n)$ pour la distribution de Pareto ont $m(F_{2.33},n) = 1.33$, et $m(F_{2.67},n) = 1.67$ comme asymptotes.

Les propositions 2 et 3 jettent une certaine lumière sur l'estimation de la valeur limite optimale. Lorsque F est inconnu, la valeur qui minimise un estimateur de l'erreur quadratique moyenne de $\bar{X}_R$ est un estimateur possible de $R(F,n)$. On obtient ainsi:

$$\mathrm{MSE}(\bar{X}_R) = \sum_{h=1}^{L} \frac{W_h^2}{n_h}$$

$$\left(\sigma_h^2 - 2\int_{R_h}^{\infty} (x - \mu_h)[1 - F_h(x)]dx - B^2(\bar{X}_{Rh})\right)$$

$$+ \left(\sum_{h=1}^{L} W_h B(\bar{X}_{Rh})\right)^2 \quad (3.1)$$

où $B(\bar{X}_{Rh})$ représente le biais de $\bar{X}_{Rh}$ en tant que estimateur de μ_h

$$B(\bar{X}_{Rh}) = -\int_{R_h}^{\infty} [1 - F_h(x)]dx.$$

L'utilisation des dérivées partielles en fonction de R_h, $h = 1, \ldots, L$ donne les équations suivantes pour les valeurs optimales:

$$\frac{W_h}{n_h}[R_h - \mu_h - B(\bar{X}_{Rh})] = -\sum_{h=1}^{L} W_h B(\bar{X}_{Rh}), \quad (3.2)$$

pour $h = 1, \ldots, L$.

Il n'existe pas de méthode simple pour résoudre (3.2). On peut cependant obtenir une solution approximative en notant que $B(\bar{X}_{Rh})/n_h$ est habituellement petit comparativement aux autres éléments de l'équation, pour toutes les valeurs de h. L'élimination de ces éléments conduit à l'équation suivante:

$$\frac{W_h}{n_h}(R_h - \mu_h) = -\sum_{h=1}^{L} W_h B(\bar{X}_{Rh}), \quad (3.3)$$

pour $h = 1, \ldots, L$. Les solutions de (3.3) surestiment légèrement les valeurs optimales qui satisfont à l'équation (3.2) puisqu'elles laissent conclure que les dérivées partielles de (3.1) sont toutes positives et qu'elles représentent des fonctions croissantes de R_h, pour $h = 1, \ldots, L$. Ainsi, en résolvant (3.3) pour estimer les valeurs limites, on ne court pas le risque de winsoriser trop de valeurs. L'équation (3.3) signifie que $R_h = \mu_h + n_h R/(nW_h)$, où R est une constante positive. On obtient une équation simple de R en remplaçant la variable $y = nW_h(x - \mu_h)/n_h$ dans les intégrales par $B(\bar{X}_{Rh})$, $h = 1, \ldots, L$, où $n = \sum n_h$. On obtient ainsi:

$$-\sum_{h=1}^{L} W_h B(\bar{X}_{Rh}) = \frac{R}{n} =$$

$$\int_R^{\infty} [1 - F(y)]dy = -B(\bar{X}_R), \quad (3.4)$$

où $F(y) = \sum n_h F_h[\mu_h + n_h y/(nW_h)]/n$. L'équation (3.4) est aisément résolue en utilisant l'algorithme (2.5) proposé dans la section 2 pour le cas de l'échantillon unique. On peut ainsi calculer facilement des approximations simples des valeurs limites optimales de Searls pour les échantillons stratifiés.

Comme la distribution F définie ci-dessus a une espérance de zéro, l'erreur quadratique moyenne de la moyenne stratifiée winsorisée obtenue par la résolution de l'équation (3.3) est égale à:

$$\mathrm{MSE}(\bar{X}_R) = \frac{1}{n}$$

$$\left(\sigma_F^2 - 2\int_R^{\infty} y[1 - F(y)]dy - B(\bar{X}_R)^2\right) + B(\bar{X}_R)^2$$

$$+ \left(\frac{1}{n}B(\bar{X}_R)^2 - \sum_{h=1}^{L} \frac{W_h^2 B^2(\bar{X}_{Rh})}{n_h}\right) \quad (3.5)$$

où σ_F^2 est la variance de F. Il est facile de démontrer que le dernier terme de (3.5) est négatif ou nul; il est nul lorsque $B(\bar{X}_R) = nW_h B(\bar{X}_{Rh})/n_h$ pour $h = 1, \ldots, L$. La variance de la moyenne stratifiée, $\bar{X} = \sum W_h \bar{X}_h$, est égale à σ_F^2/n. Ainsi, une approximation prudente de l'efficacité de $\bar{X}_R$ par rapport à $\bar{X}$ dans un échantillon stratifié est égale à l'efficacité correspondante pour un échantillon aléatoire de taille n tiré de F. Remarquons également que $n[1 - F(R)]$ représente le nombre total prévu de points de données winsorisés dans la strate L.

Le plan de winsorisation optimale obtenu par la résolution de (3.3) possède une forme simple pour beaucoup de règles de répartition. En vertu de la règle de répartition proportionnelle, c.-à-d., $n_h = nW_h$ pour $h = 1, \ldots, L$, on obtient $R_h = \mu_h + R$. En vertu de la répartition optimale de Neyman, avec $n_h = nW_h\sigma_h/(\sum W_h\sigma_h)$, où σ_h est l'écart-type de la strate h, on obtient $R_h = \mu_h + \sigma_h R/(\sum W_h\sigma_h)$. Si, en plus, les distributions de X à l'intérieur des strates sont les mêmes sauf pour un changement dans l'emplacement et dans l'échelle, c.-à-d., $F_h = F_0[(x - \mu_h)/\sigma_h]$ pour une certaine distribution F_0, alors $F(x) = F_0[x/(\sum W_h\sigma_h)]$. Dans ce cas, les caractéristiques des moyennes winsorisées optimales dans l'échantillonnage stratifié et dans l'échantillonnage aléatoire simple sont les mêmes. Ainsi, la figure 1 présente le nombre total prévu de points de données winsorisés dans la strate L en fonction de la taille n de l'échantillon total, en vertu de la répartition de Neyman, lorsque F_0 correspond à une des distributions du tableau 1. La figure 2 donne les valeurs correspondantes de l'efficacité.

Les résultats de cette section sont faciles à généraliser à l'échantillonnage stratifié sans remise en remplaçant n_h par $n_h/(1 - f_h)$ tout au long des calculs. On dérive aisément les valeurs limites optimales pour l'échantillonnage stratifié avec ppt en posant $F_h(x) = \sum p_{hi} I(y_{hi}/(N_h p_{hi}) \leq x)$, où p_{hi} désigne la probabilité de sélection de la i-ième unité de la strate h.

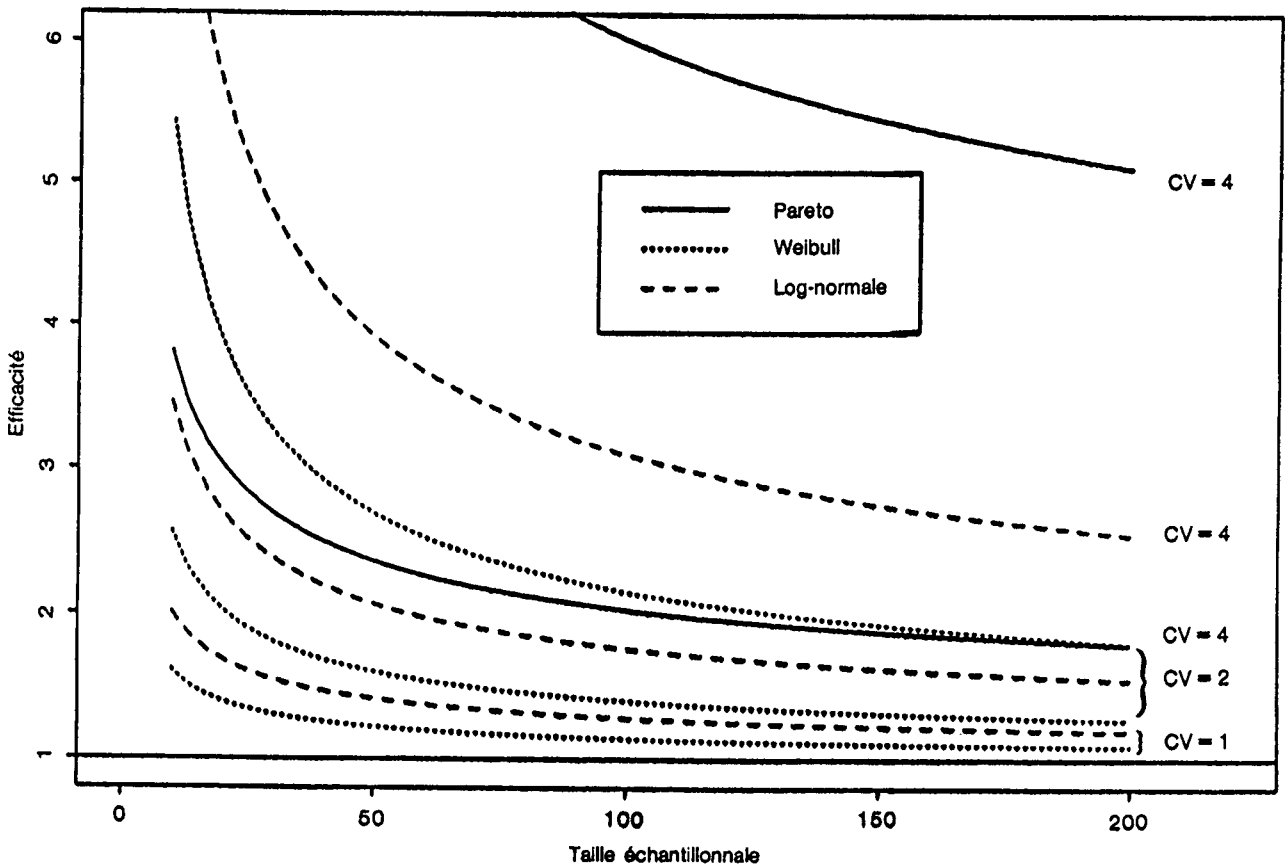


Figure 2. Efficacité de la moyenne winsorisée de Searls.

La figure 1 montre que le nombre prévu d'observations winsorisés $m(F,n)$ diminue avec l'asymétrie croissante de la distribution. Cette observation peut être traduite en un résultat mathématique rigoureux. À cette fin, on suppose que la variable aléatoire Y est plus asymétrique que la variable aléatoire X, si Y a la même distribution que $\psi(X)$, où $\psi(x)$ est une fonction convexe de x. En vertu de cette définition, X^2 est, comme prévu, plus asymétrique que X. Cette notion de l'asymétrie correspond à l'ordonnancement partiel convexe de van Zwet (Barlow et Proschan 1981). Cette définition de l'asymétrie débouche sur la proposition suivante dont nous fournissons la preuve en annexe, en même temps que celles des propositions 2 et 3.

Proposition 1 Si Y est plus asymétrique que X, alors $m(F_X,n) > m(F_Y,n)$, où F_X et F_Y désignent les distributions de X et de Y respectivement.

Les résultats de la présente section s'appliquent également à l'échantillonnage aléatoire simple sans remise. Pour ce plan d'expérience, l'erreur quadratique moyenne de $\bar{X}_R$ est donnée par la formule (2.3) en remplaçant n par $n/(1-f)$, où f désigne la fraction d'échantillonnage. L'algorithme (2.5), avec n divisé par $(1-f)$, peut servir au calcul des valeurs limites optimales pour l'échantillonnage aléatoire simple sans remise.

3. WINSORISATION EN ÉCHANTILLONNAGE STRATIFIÉ

Il existe plusieurs façons d'appliquer la méthode de winsorisation de Searls à l'échantillonnage stratifié. Dans la présente section, chaque strate possède sa propre valeur limite. Désignons par R_h la valeur limite de la strate h. Les valeurs optimales $R_1, R_2, \ldots, R_L$, où L désigne le nombre de strates, sont celles qui minimisent l'erreur quadratique moyenne de $\bar{X}_R = \sum W_h \bar{X}_{Rh}$, où $\bar{X}_{Rh} = \sum \min(X_{hi},R_h)/n_h$, $W_h = N_h/N$ et N_h désigne la taille de la strate h et $N = \sum N_h$. Nous proposons dans la présente section un algorithme pour la détermination de ces valeurs limites optimales.

Désignons par $F_h(x)$, pour $h = 1, \ldots, L$, la distribution de X dans la strate h, et par μ_h et σ_h^2 la moyenne et la variance de F_h respectivement. La détermination de l'erreur quadratique moyenne de $\bar{X}_R$, dans un échantillonnage aléatoire stratifié avec remise, se fait de la façon décrite à la section 2. On obtient:

avec $R_0 = 2\mu$ comme valeur de départ, tend doucement vers la solution de (2.4). Pour les distributions discrètes, les calculs sont faits aisément en se rappelant que

$$\int_R^\infty [1 - F(x)]dx = E[\max(X - R, 0)].$$

Les calculs exacts des points limites optimaux $R(F,n)$ ont été effectués pour les familles d'échantillons de Weibull, log-normale et de Pareto, avec des échantillons dont la taille s oscillait entre 5 et 200. Trois distributions, correspondant aux coefficients de variation (CV) 1, 2, et 4, ont été utilisées pour les deux premières familles alors que pour la famille de Pareto, on utilisait uniquement les coefficients de variation 2 et 4. Le coefficient de variation mesure l'asymétrie de la distribution, les valeurs les plus grandes correspondant à l'asymétrie la plus prononcée. Les valeurs correspondantes des paramètres sont présentées dans le tableau 1.

Tableau 1

Valeurs des paramètres des distributions dont on a évalué les valeurs limites optimales $R(F,n)$

CV	Weibull(α)	Log-normale(ν)	Pareto(γ)
1	1	0.83	-
2	1.84	1.27	2.67
4	2.87	1.68	2.13

On a calculé le point limite optimal pour chaque distribution et pour chaque taille d'échantillon en utilisant l'algorithme (2.5). La figure 1 présente le nombre prévu d'observations winsorisées, $m(F,n) = n\{1 - F[R(F,n)]\}$ en tant que fonction de n, et la figure 2 donne les valeurs correspondantes de l'efficacité. L'efficacité de $\bar{X}_R$ est définie par $\mathrm{Var}(\bar{X})/\mathrm{MSE}(\bar{X}_R)$.

Dans la figure 1, le nombre prévu de valeurs winsorisées en vertu du plan optimal s'approche de 1 pour la plupart des distributions asymétriques, même pour les grands échantillons. En faisant l'approximation de ce nombre à l'aide d'une distribution de Poisson avec le paramètre $m(F,n)$, on constate qu'il existe une probabilité non négligeable, avec le plan de winsorisation optimale, qu'aucun des points de données ne soit winsorisé. Cette probabilité augmente avec l'asymétrie de la distribution puisque $m(F,n)$ diminue avec le coefficient de variation CV. Ainsi, pour les échantillons provenant d'une distribution fortement asymétrique, il n'est pas toujours approprié de winsoriser les valeurs les plus grandes. De telles valeurs ne devraient être winsorisées que lorsqu'elles sont très grandes par rapport aux autres. Comme prévu, dans la figure 2, les meilleurs gains en matière d'efficacité sont obtenus lorsque l'asymétrie est prononcée. Ainsi, la stratégie de winsorisation la plus profitable consiste à repérer les deux ou trois valeurs les plus grandes d'un échantillon et d'en réduire l'impact lorsqu'elles sont très grandes par rapport aux autres.

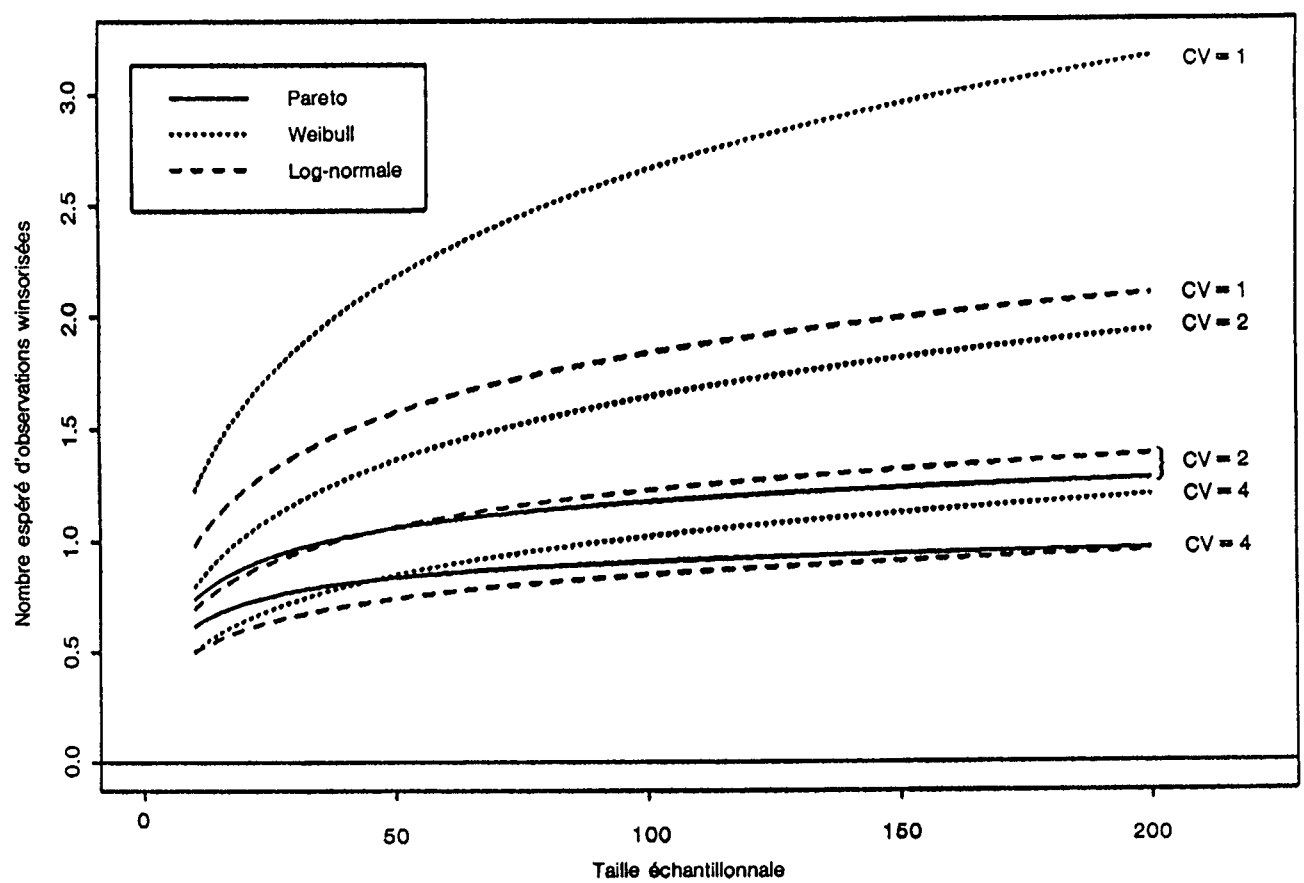


Figure 1. Nombre espéré d'observations winsorisées pour échantillonnage aléatoire simple et stratifié.

à asymétrie positive. On peut en effet choisir entre la famille de Weibull, $F_\alpha(x) = 1 - \exp(-(x/\beta)^{1/\alpha})$ pour $x > 0$; la famille log-normale, $F_\nu(x) = \Phi(\log(x/\beta)/\nu)$ pour $x > 0$; et la famille Pareto, $F_\gamma(x) = 1 - (1 + x/\beta)^{-\gamma}$ pour $x > 0$, où β est un paramètre d'échelle positif et α, ν, et γ sont des paramètres de forme positifs. Les distributions asymétriques discrètes proviennent d'échantillonnages d'enquête. Désignons par $\{y_1, \ldots, y_N\}$ les valeurs de la variable d'intérêt pour les N unités d'une population à échantillonner. Si on tire un échantillon aléatoire simple avec remise, on peut alors désigner par $F(x) = \sum I(y_i \le x)/N$ la distribution sous-jacente où $I(\cdot)$ représente la fonction indicatrice. Pour l'échantillonnage avec ppt, c'est-à-dire l'échantillonnage avec remise assorti de probabilités déterminées par $\{p_i, i = 1, \ldots, N\}$, on utiliserait $F(x) = \sum p_i I(y_i/(Np_i) \le x)$. L'estimateur standard de $\bar{y}$ sous échantillonnage avec ppt,

$$\bar{y}_s = \frac{1}{n} \sum_s \frac{y_i}{Np_i}$$

peut alors être considéré comme la moyenne d'un échantillon aléatoire de taille n tiré d'une distribution F. Fuller (1991) donne des exemples de données d'enquête à distribution asymétrique.

Désignons par $X_1, X_2, \ldots, X_n$ un échantillon tiré de $F(x)$. Pour l'échantillonnage avec ppt, on obtiendra $X_i = y_i/(Np_i)$ où p_i et y_i désignent la probabilité de sélection et la valeur de la variable y pour la i-ième unité sélectionnée dans l'échantillon. La moyenne de la population μ doit être estimée par une moyenne winsorisée,

$$\bar{X}_R = \frac{1}{n} \sum_1^n \min(X_i, R) = \bar{X} - \frac{1}{n} \sum_{i=1}^n \max(X_i - R, 0), \tag{2.1}$$

où $\bar{X}$ est la moyenne des valeurs X_i. L'espérance de $\bar{X}_R$ est égale à

$$E(\bar{X}_R) = \mu - \int_R^\infty (x - R) dF(x) = \mu - \int_R^\infty \int_R^x dy dF(x).$$

En changeant l'ordre d'intégration dans l'intégrale ci-dessus, on prouve que $E(\bar{X}_R) = \mu + B(\bar{X}_R)$, où

$$B(\bar{X}_R) = -\int_R^\infty [1 - F(x)] dx \tag{2.2}$$

représente le biais de la moyenne winsorisée.

Selon la formule (2.1), l'expression de la variance de $\bar{X}_R$ est

$$n \operatorname{Var}(\bar{X}_R) = \sigma^2 - 2\operatorname{cov}[X_1, \max(X_1 - R, 0)] + \operatorname{Var}[\max(X_1 - R, 0)]$$

où X_1 est la première variable aléatoire de l'échantillon et σ^2 est la variance de $F(x)$. Des manipulations semblables à celles qui aboutissent à la formule (2.2) montrent que

$$E[\max(X_1 - R, 0)^2] = 2\int_R^\infty (x - R)[1 - F(x)] dx,$$

et que

$$E[\max(X_1 - R, 0) X_1] = 2\int_R^\infty (x - R)[1 - F(x)] dx - RB(\bar{X}_R).$$

Ainsi,

$$\operatorname{Var}(\bar{X}_R) = \frac{1}{n}\left\{\sigma^2 - 2\int_R^\infty (x - \mu)[1 - F(x)] dx - B^2(\bar{X}_R)\right\},$$

et

$$\operatorname{MSE}(\bar{X}_R) = \frac{\sigma^2}{n} - \frac{2}{n}\int_R^\infty (x - \mu)[1 - F(x)] dx + \frac{n - 1}{n} B^2(\bar{X}_R). \tag{2.3}$$

Searls (1966) a démontré que l'erreur quadratique moyenne de $\bar{X}_R$ comporte un minimum unique qui peut être obtenu en égalant sa dérivée, par rapport à R, à 0. On obtient ainsi l'équation suivante pour la constante optimale de winsorisation $R(F, n)$:

$$\frac{R - \mu}{n - 1} - \int_R^\infty [1 - F(x)] dx = 0. \tag{2.4}$$

Cette équation équivaut à l'équation (14) de Searls (1966). Dans le reste de ce travail, $\bar{X}_R$ désigne la moyenne winsorisée optimale obtenue par la constante de winsorisation $R(F, n)$ qui résout (2.4). À noter que le point limite optimal $R(F, n)$ est équivariant quant à l'emplacement et à l'échelle, c'est-à-dire que si $G(x) = F[(x - b)/a]$, alors $R(G, n) = aR(F, n) + b$.

Il est facile d'établir un algorithme général pour la résolution de (2.4). Notons d'abord qu'en sa qualité de fonction de R, le membre de gauche de l'équation (2.4) est croissant et concave en R puisque sa dérivée, $1/(n - 1) + 1 - F(R)$, est positive et décroissante. En conséquence, l'algorithme de Newton-Raphson (Thisted 1988, 164-167), donné par

$$R_{j+1} = R_j - \frac{(R_j - \mu) - (n - 1)\int_{R_j}^\infty [1 - F(x)] dx}{1 + (n - 1)[1 - F(R_j)]}, \tag{2.5}$$

Techniques d'enquête, décembre 1995
Vol. 21, n° 2, pp. 119–129
Statistique Canada

Moyenne winsorisée de Searls pour populations asymétriques

LOUIS-PAUL RIVEST et DANIEL HURTUBISE[1]

RÉSUMÉ

Nous examinons l'utilité de la moyenne winsorisée comme estimateur de la moyenne d'une distribution à asymétrie positive. On obtient la moyenne winsorisée en remplaçant toutes les observations supérieures à une valeur limite donnée R par cette même valeur R, avant le calcul de la moyenne. La valeur limite optimale, telle qu'elle est définie par Searls (1966), minimise l'erreur quadratique moyenne (mean square error, ou MSE) de l'estimateur winsorisé. Nous proposons des méthodes d'évaluation de cette valeur limite optimale avec divers plans d'échantillonnage, y compris l'échantillonnage aléatoire simple, l'échantillonnage stratifié et l'échantillonnage avec probabilité proportionnelle à la taille (ppt). Pour la plupart des distributions asymétriques, la stratégie de winsorisation optimale aura généralement pour effet de modifier la valeur d'environ une observation dans l'échantillon. On dérive des approximations directes (qui ne fait pas appel à un processus itératif) de l'efficacité de la moyenne winsorisée de Searls en utilisant la théorie des statistiques d'ordre extrême. Une expérience de Monte Carlo nous sert à comparer divers estimateurs réduisant l'impact des valeurs extrêmes.

MOTS CLÉS: Valeurs aberrantes; domaine d'attraction maximal; erreur quadratique moyenne; échantillonnage aléatoire simple; échantillonnage stratifié.

1. INTRODUCTION

Les échantillons tirés de distributions étalées vers la droite contiennent souvent des valeurs aberrantes beaucoup plus grandes que la plupart des valeurs échantillonnées. On s'efforce habituellement de tenir compte de ces valeurs extrêmes à l'étape de l'élaboration du plan de sondage (Glasser 1962; Hidiroglou 1987). Toutefois, compte tenu des buts multiples de la plupart des sondages, les statisticiens se trouvent souvent aux prises avec des valeurs aberrantes à l'étape de l'estimation. Ces observations nuisent à la stabilité des estimateurs classiques comme la moyenne de l'échantillon. Il paraît donc utile d'étudier des estimateurs de rechange qui réduiront l'incidence des valeurs trop grandes. La winsorisation (Searls 1966) consiste à remplacer les valeurs supérieures à une valeur limite R par cette même valeur R avant le calcul de la moyenne. Searls suggère de choisir un R qui minimise l'erreur quadratique moyenne de la moyenne winsorisée. On peut également choisir un R correspondant à la deuxième valeur en importance dans l'échantillon (Rivest 1994). L'estimateur de Searls s'est avéré le meilleur, parmi toutes les méthodes examinées par Ernst (1980), pour ajuster les valeurs trop grandes. Hicks et Fetter (1993) ont utilisé la méthode de winsorisation de Searls dans le cadre d'un sondage sur l'agriculture. D'autres méthodes ont été proposées pour traiter les valeurs trop grandes présentes dans les échantillons d'enquêtes. Chambers et Kokic (1993) ont étudié les estimateurs dérivés de la théorie des "statistiques robustes" (Huber 1981). Fuller (1991, 1993) a proposé un test préliminaire pour la détection des valeurs extrêmes dans les échantillons; l'incidence de ces valeurs n'est réduite que dans les échantillons pour lesquels ce test est significatif. Lee (1994) présente un bon survol des articles de plus en plus nombreux abordant cette question.

La clé de la mise en oeuvre de la méthode de winsorisation de Searls réside dans la sélection de la valeur limite R. Nous suggérons dans la section 2 un algorithme simple pour le calcul de la valeur limite optimale, pour une population connue soumise à un échantillonnage aléatoire simple ou à un échantillonnage avec ppt. Des calculs répétés de la valeur limite optimale pour plusieurs populations et plusieurs tailles d'échantillons révèlent que dans la plupart des cas, la méthode optimale modifie en moyenne une observation, peu importe la taille de l'échantillon. Dans la section 3 nous reprenons le travail effectué dans la section 2 mais en utilisant cette fois l'échantillonnage stratifié; nous proposons un algorithme simple pour le calcul des valeurs limites dans chaque strate. La règle de la winsorisation d'une observation en moyenne, sans égard à la taille de l'échantillon, s'applique également aux échantillons stratifiés. Dans les sections 4 et 5, nous calculons l'efficacité, en ce qui concerne la moyenne de l'échantillon, de divers estimateurs winsorisés. Dans la section 4, nous dérivons des approximations de l'efficacité de l'estimateur de Searls pour les grands échantillons au moyen de la théorie des statistiques d'ordre extrême. Dans la section 5, nous comparons l'utilité des estimateurs pour la réduction de l'incidence des valeurs extrêmes à l'aide d'une étude de Monte Carlo.

2. PROPRIÉTÉS D'ÉCHANTILLONNAGE DE LA MOYENNE WINSORISÉE

Nous examinons dans la présente section les moyennes winsorisées pour des données tirées d'une distribution discrète ou continue. Plusieurs familles de distributions continues peuvent nous servir de modèles de distributions

[1] Louis-Paul Rivest et Daniel Hurtubise, Département de mathématiques et de statistique, Université Laval, Cité Universitaire (Québec) Canada, G1K 7P4.

Comportement ancien

Paiement des premiers honoraires:	"Vers quel âge environ avez-vous consulté un dentiste pour la première fois et payé, vous ou vos parents, ses honoraires?"
Consultations:	"Avez-vous vu un dentiste au cours des 12 derniers mois?"
Année de la dernière consultation:	"Quelle année avez-vous consulté un dentiste pour la dernière fois?"
Coût l'an dernier:	"Combien avez-vous dépensé pour des soins dentaires au cours des 12 derniers mois?"

BIBLIOGRAPHIE

ANDERSON, D.A. (1985). Variance component models with binary response; interviewer variability. *Journal of the Royal Statistical Society*, Séries B, 47, 203-210.

BIEMER, P.B., et STOKES, S.L. (1985). Optimal design of interviewer variance experiments in complex surveys. *Journal of the American Statistical Association*, 80, 158-166.

BEBBINGTON, A.C., et SMITH, T.M.F. (1977). The effect of survey design on multivariate analysis, *The Analysis of Survey Data*, (C.A. O'Muircheartaigh et C. Payne, Éds.), 175-192. New York: John Wiley.

CHOI, I.C., et COMSTOCK, G.W. (1975). Interviewer effects on responses to a questionnaire relating to mood. *American Journal of Epidemiology*, 101, 84-92.

COLLINS, M., et BUTCHER, B. (1992). Interviewer and clustering effects in an attitude survey. *Journal of the Market Research Society*, 25, 39-58.

CUTRESS, T.W., HUNTER, P.B., DAVIS, P.B., BECK, D.J., et CROXSON, L.J. (1979). *Adult Oral Health and Attitudes to Dentistry in New Zealand*, Medical Research Council, Wellington.

DIJKSTRA, W. (Éd.) (1982). *Response Behaviour in the Survey Interview*. New York: Academic Press.

FEATHER, J. (1973). *A Study of Interviewer Variance*, (WHO/ICS-MCU Saskatchewan Study Area Reports Series 2, No. 3). Department of Social and Preventive Medicine. University of Saskatchewan, Saskatoon.

FELLEGI, I.P. (1974). An improved method of estimating correlated response variance. *Journal of the American Statistical Association*, 69, 496-501.

GROVES, R. (1989). *Survey Errors and Survey Costs*. New York: John Wiley.

GROVES, R., et FULTZ, N.H. (1985). Gender effects among telephone interviewers in a survey of economic attitudes. *Sociological Methods and Research*, 14, 31-52.

HOLT, D., SMITH, T.M.F., et WINTER, P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society*, 143, 474-487.

HOX, J.J. (1994). Hierarchical regression models for interviewer and respondent effects. *Sociological Methods and Research*, 22, 300-318.

KISH, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Society*, 57, 92-115.

KISH, L. (1965). *Survey Sampling*. New York: John Wiley.

KISH, L., et FRANKEL, M.R. (1974). Inferences from complex samples. *Journal of the Royal Statistical Society*, Séries B, 36, 1-37.

KISH, L. (1987). *Statistical Design for Research*. New York: John Wiley.

LESSLER, J.T., et KALSBEEK, W.D. (1992). *Nonsampling Errors in Surveys*. New York: John Wiley.

O'MUIRCHEARTAIGH, C.A. (1976). Response errors in an attitudinal sample survey. *Quality and Quantity*, 10, 97-115.

O'MUIRCHEARTAIGH, C.A., et PAYNE, C. (Éds.) (1977). *The Analysis of Survey Data*, (Volume 2: Model Fitting). New York: John Wiley.

O'MUIRCHEARTAIGH, C.A., et WIGGINS, R.D. (1981). The impact of interviewer variability in an epidemiological survey. *Psychological Medicine*, 11, 817-824.

PANNEKOEK, J. (1988). Interviewer variance in a telephone survey. *Journal of Official Statistics*, 4, 375-384.

PRASAD, N.G., et RAO, J.N.K. (1990). The estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

RAO, J.N.K., et THOMAS, D.R. (1988). The analysis of cross-classified data from complex sample surveys. *Sociological Methodology*, 18, 213-269.

SKINNER, C.J., HOLT, D., et SMITH, T.M.F. (Éds.) (1989). *Analysis of Complex Surveys*. New York: John Wiley.

VERMA, V., SCOTT, C., et O'MUIRCHEARTAIGH, C.A. (1980). Sample designs and sampling errors for the World Fertility Survey. *Journal of the Royal Statistical Society*, Séries A, 143, 431-473.

dans un plan d'échantillonnage à degrés multiples, en fonction des hypothèses de départ du modèle de réponses corrélées (version de base et version élargie).

Le modèle simple sert à déterminer l'incidence relative des effets du groupement de l'échantillon et de l'intervieweur sur l'estimation des moyennes et des proportions. Les résultats confirment plusieurs constatations déjà fort bien documentées: les corrélations intra-catégorie sont habituellement plus faibles pour l'intervieweur que pour le groupement; les corrélations intra-catégorie des groupements varient dans le sens prévu selon le type de question; l'effet global du plan d'échantillonnage se situe entre 2 et 3.5 pour les mêmes types de question; la variabilité de l'intervieweur contribue de manière importante à l'inflation observée et résulte peut-être du facteur d'amplification attribuable à la lourde charge de travail; enfin, l'incidence des effets dus à l'intervieweur varie dans le sens prévu selon le type de question.

En deuxième lieu, on s'est servi du modèle élargi pour analyser les effets dus au groupement et à l'intervieweur sur l'estimation de l'écart entre les moyennes et les proportions pour différents domaines, soit deux jeux de comparaisons selon le sexe et la race. L'effet sur l'écart entre les moyennes des domaines est plus faible mais reste significatif pour plusieurs paramètres, notamment les comparaisons selon la race, ce qui donne à penser que l'intervieweur n'a pas la même incidence sur chaque domaine. L'effet est néanmoins suffisamment important pour qu'on s'en inquiète lors de l'élaboration d'études analogues. En général, les effets dus à l'intervieweur ont deux ou trois fois plus d'incidence sur les comparaisons selon la race que sur celles se rapportant au sexe.

Les carences fondamentales de l'analyse signifient que les résultats ont plus une valeur d'indication que de preuve irréfutable. Ils révèlent cependant que le recours à un groupe restreint d'intervieweurs peut avoir de sérieuses conséquences, même lorsqu'on s'intéresse plus aux différences entre domaines qu'aux moyennes ou aux proportions simples. Cette constatation se démarque sans aucun doute des convictions bien ancrées dans certains milieux, comme celui de la santé, et indique que des recherches empiriques considérablement plus poussées s'avéreraient tout à fait justifiées.

Partant des hypothèses du modèle simple de réponses corrélées, on conclut qu'il est possible d'atténuer l'incidence de la variance de l'intervieweur en augmentant le nombre d'intervieweurs, donc en réduisant la charge de travail de chacun d'eux. Bien sûr, il pourrait s'ensuivre une perte au niveau de la qualité des entrevues si on est contraint, pour la même raison, à rogner sur la formation et les méthodes de surveillance. Dans ce cas, on portera une grande attention à la façon dont sont formulées les questions et instructions destinées aux intervieweurs. Pour ce qui est de la version élargie du modèle cependant, une telle stratégie ne suffira sans doute pas. Si l'un des principaux objectifs de l'étude est la comparaison de différents groupes, il importera de veiller à ce que les intervieweurs traitent les groupes concernés de la manière la plus uniforme qui soit. L'enquête devra absolument être conçue pour que chaque intervieweur interroge des participants de chaque groupe. Il s'agit sans doute d'un facteur crucial dans les enquêtes similaires aux études de comparaison avec témoins, en vertu desquelles les issues médicales sont associées à différents degrés d'exposition et le contrôle des variables confusionnelles éventuelles peut avoir une profonde influence sur l'ordre de grandeur de certains paramètres comme le risque relatif.

REMERCIEMENTS

Le projet a bénéficié de l'aide du Medical Research Council of New Zealand. Joanna Broad s'est occupée dans une large mesure des calculs détaillés de l'analyse.

ANNEXE

Questionnaire

Attitudes

Dentistes 1:	"Les dentistes s'intéressent plus à leurs patients qu'à gagner de l'argent."
Dentistes 2:	"Les dentistes recommandent beaucoup plus de choses que le strict minimum."
Dentier:	"Un dentier est aussi bon (meilleur) que les dents naturelles."
Fluoration:	"Que pensez-vous de la fluoration des réserves d'eau publiques?"
Consultations:	"Croyez-vous qu'une personne devrait aller chez son dentiste uniquement lorsqu'elle éprouve un problème ou devrait aussi le consulter quand tout va bien?"
État des dents:	"Si vous aviez rendez-vous chez le dentiste demain, celui-ci trouverait-il quelque chose à redire au sujet de vos dents?"
Gencives:	"Si vous aviez rendez-vous chez le dentiste demain, celui-ci trouverait-il quelque chose à redire au sujet de vos gencives?"

Comportement récent

"Hier,
- avez-vous utilisé un révélateur/un rince-bouche/de la soie dentaire/un cure-dent?
- vous êtes-vous rincé la bouche après avoir mangé?
- vous êtes-vous brossé les dents?"

"Avez-vous acheté des sucreries ou du chocolat la semaine dernière?"

entre les domaines concernés. En supposant que la charge de travail des intervieweurs est répartie de façon uniforme entre les domaines, v_I est égal à zéro quand l'effet dû à l'intervieweur est identique dans les deux cas, car il y a annulation lors de la comparaison. Si les effets dans les deux domaines sont faiblement corrélés, par contre, v_I a tendance à avoir une valeur beaucoup plus élevée, valeur qui, dans les cas extrêmes, pourrait être égale à la charge de travail moyenne. Dans l'enquête qui nous intéresse, les valeurs de v_I se situent entre 0 et 50 pour le sexe et la race. Les effets dus à l'intervieweur spécifiques au domaine peuvent donc avoir une incidence passablement importante sur l'effet du plan d'échantillonnage. Les mêmes remarques s'appliquent à l'incidence du groupement sur la comparaison; si les effets sont les mêmes dans les deux domaines, ils s'annulent dans une large mesure et l'incidence nette est minime, mais cette dernière peut être relativement importante quand l'effet du groupement est spécifique au domaine.

Tableau 2

Effets de l'intervieweur et du groupement sur l'écart entre les domaines

Question	Selon le sexe				Selon la race			
	$\hat{\rho}_I$	$\hat{\rho}_C$	D_2	%Int	$\hat{\rho}_I$	$\hat{\rho}_C$	D_2	%Int
Attitudes:								
Dentistes 2	.004	.043	1.05	0	.010	.027	1.28	46
Consultations	.009	.028	1.08	0	.072	.133	5.19	78
Dents naturelles	.010	.032	1.06	0	.010	.037	1.26	42
Fluoration	.001	.019	1.12	42	.021	.031	2.13	88
Dentier	.007	.018	1.04	0	.011	.035	1.21	33
État des dents	.012	.022	1.52	85	.010	.045	1.40	20
Dentistes 1	.001	.018	1.05	0	.015	.020	1.53	74
Gencives	.006	.022	1.07	0	.003	.104	1.46	9
Moyenne	**.006**	**.025**	**1.12**	**16**	**.019**	**.054**	**1.93**	**49**
Paramètres socio-démographiques:								
Race	.008	.183	1.11	0	–	–	–	
Revenu familial	.004	.095	1.37	24	.004	.099	1.95	40
État civil	.000	.059	1.17	0	.011	.060	1.69	38
Situation de l'emploi	.014	.067	1.42	71	.022	.116	2.09	25
Âge	.007	.052	1.06	0	.006	.093	1.87	24
Sexe	–	–	–		.006	.011	1.09	44
Moyenne	**.007**	**.091**	**1.23**	**19**	**.010**	**.076**	**1.74**	**34**
Comportement récent:								
Brosser les dents	.025	.060	1.62	65	.019	.019	1.68	88
Rincer la bouche	.007	.029	1.28	64	.004	.023	1.20	45
Rince-bouche	.000	.057	1.20	0	.027	.105	2.69	75
Soie dentaire	.003	.021	1.06	0	.015	.036	1.37	32
Cure-dents	.006	.010	1.03	0	.006	.046	1.48	63
Sucreries/chocolat	.012	.009	1.02	0	.013	.022	1.31	48
Dentifrice au fluorure	.010	.007	1.11	100	.007	.000	1.02	100
Moyenne	**.009**	**.028**	**1.19**	**33**	**.013**	**.036**	**1.36**	**64**
Comportement ancien:								
Âge: première consultation payée	.003	.033	1.10	0	.029	.141	2.92	71
Consultation	.005	.035	1.04	0	.020	.018	1.26	50
Année de la dernière consultation	.004	.012	1.20	75	.016	.003	1.83	12
Coût l'an dernier	.007	.021	1.01	0	.076	.117	2.09	42
Moyenne	**.005**	**.025**	**1.09**	**19**	**.035**	**.070**	**2.03**	**44**

Le tableau 2 donne les valeurs de ρ_I et ρ_C pour les comparaisons selon le sexe et selon la race, ainsi que l'effet global D_2 du plan d'échantillonnage et la partie de cet effet attribuable à la variabilité de l'intervieweur. Notons que la question relative à l'usage d'un révélateur a été supprimée au tableau 2, principalement parce que très peu de répondants se servaient d'un tel produit ou en connaissaient l'existence, si bien que la taille de l'échantillon utile dans ce cas est faible et que les résultats ne sont pas significatifs.

L'intervieweur et le groupement ont relativement peu d'incidence sur les comparaisons selon le sexe et les effets du plan d'échantillonnage dépassent à peine l'unité, bien que les valeurs estimées de ρ_I et ρ_C augmentent légèrement, après correction pour cette variable. On ne relève d'effet lié au sexe significatif que pour trois questions, soit l'état des dents et l'usage d'une brosse à dents - pour lesquelles il pourrait exister un biais particulier dû à l'acceptabilité sociale - et la situation de l'emploi - qui présente des connotations fort distinctes pour les hommes et les femmes. Remarquons que l'effet dû à l'intervieweur domine dans chacune des trois comparaisons.

L'incidence des mêmes facteurs sur les comparaisons selon la race est beaucoup plus élevée, l'effet global du plan d'échantillonnage s'établissant en moyenne à 1.7. On peut en conclure que la race des répondants entraîne des effets dus à l'intervieweur et au groupement significatifs qui ne s'annulent pas. D'importants effets dus à l'intervieweur sont bien corrélés à la race pour deux questions hypothétiques sur les attitudes (consultations et fluoration), pour une question concernant le comportement récent ainsi que pour l'âge au moment où sont pour la première fois payés les honoraires du dentiste. Ces résultats sont plausibles; tous les intervieweurs étaient européens d'origine, ce qui pourrait avoir donné lieu à une variation systématique de leur interaction avec les répondants d'une autre ethnie. Ici encore, les effets du groupement sont plus marqués pour les paramètres sociodémographiques. Non seulement les effets du plan d'échantillonnage sont-ils en moyenne plus élevés que ceux notés pour les comparaisons selon le sexe, mais les effets dus à l'intervieweur sont aussi deux ou trois plus élevés en général pour l'écart entre les groupes ethniques.

Un examinateur objectif a fort justement souligné qu'étant donné la manière dont les intervieweurs sont déployés (ils travaillaient principalement en équipe dans diverses parties de la Nouvelle-Zélande), il se pourrait fort bien que les effets dus à l'intervieweur soient exagérés parce qu'ils se confondent avec les effets dus à la région. Les différences entre les DL sont si ténues cependant qu'une telle exagération serait minime, mais on ne peut effectivement pas écarter cette possibilité avec un modèle ainsi constitué.

5. DISCUSSION

Le présent article se sert des données empiriques issues d'une enquête typique sur la santé pour évaluer l'incidence de la variabilité de l'intervieweur sur la variance de l'erreur

Tableau 1

Effets du groupement et de l'intervieweur sur les moyennes et les proportions

Question	$\hat{\rho}_I$	$\hat{\rho}_C$	D_0	%Int
Attitudes:				
Dentistes 1	.014	.014	4.61	91
Consultations	.008	.028	3.42	74
Dents naturelles	.008	.027	3.52	76
État des dents	.007	.015	2.97	84
Dentier	.005	.015	2.67	80
Dentistes 2	.004	.033	2.77	57
Gencives	.003	.010	1.96	77
Fluoration	.001	.016	1.66	49
Moyenne	**.006**	**.020**	**2.95**	**73**
Paramètres socio-démographiques:				
Situation de l'emploi	.010	.055	4.20	65
Race	.009	.172	6.87	33
Âge	.004	.042	5.98	52
Revenu familial	.002	.092	3.29	15
État civil	.000	.058	2.34	0
Sexe	.000	.005	1.12	0
Moyenne	**.004**	**.071**	**3.47**	**28**
Comportement récent:				
Brosser les dents	.019	.025	6.16	8
Sucreries/chocolat	.011	.003	3.75	98
Dentifrice au fluorure	.008	.000	3.04	100
Cure-dents	.006	.006	2.66	92
Rincer la bouche	.004	.024	2.43	62
Soie dentaire	.001	.018	1.60	43
Révélateur	.000	.027	1.49	0
Rince-bouche	.000	.018	1.42	0
Moyenne	**.006**	**.012**	**2.82**	**60**
Comportement ancien:				
Âge: première consultation payée	.004	.029	2.34	57
Consultation	.004	.029	2.51	57
Coût l'an dernier	.002	.000	1.19	100
Année de la dernière consultation	.000	.014	1.15	0
Moyenne	**.003**	**.018**	**1.80**	**54**

questions concernant les attitudes produisent des valeurs de ρ_I supérieures à la moyenne, tout comme certains comportements peuvent être sensibles à un biais important résultant de la "désirabilité sociale", par exemple le fait de se brosser les dents ou d'acheter des bonbons et du chocolat. L'origine ethnique et la situation de l'emploi donnent également des valeurs relativement élevées. Les résultats se rapprochent de ceux obtenus lors d'études antérieures, bien qu'ils se trouvent à l'extrémité inférieure de la fourchette des valeurs typiques rapportées ailleurs (Feather 1973; Kish 1962; O'Muircheartaigh 1977; O'Muircheartaigh et Wiggins 1981). On trouvera une étude complète de la question au chapitre 8 de l'ouvrage de Groves (1989). Il se peut que ces résultats viennent en partie de la formation et de la surveillance intensives des intervieweurs, associées au volet de l'enquête effectué sur le terrain. Une "épuration" (correction et vérification) rigoureuse des données après le travail sur le terrain, avant analyse, serait une autre explication. Néanmoins, il se pourrait qu'on doive simplement la situation à l'atténuation qui résulte de l'application du modèle (1) aux proportions mentionnées précédemment.

L'incidence de la variabilité de l'intervieweur sur l'effet global du plan d'échantillonnage, qui combine les effets dus à l'intervieweur et au groupement, a sans doute plus d'importance que les caractéristiques et les valeurs de ρ_I. C'est ce qu'on remarque à la troisième colonne du tableau 1 (D_0), la dernière colonne (int.%) représentant la contribution proportionnelle de la variabilité de l'intervieweur à la valeur globale de D_0. Le plan d'échantillonnage a des effets sensibles puisque la valeur de ces derniers dépasse deux dans tous les cas, à quelques exceptions près. On le doit au groupement et à l'incidence de la lourde charge de travail des intervieweurs, caractéristique à l'étude puisque, selon l'équation (3), la variance augmente d'un facteur de $1 + (\bar{n}_I - 1)\rho_I$, où $\bar{n}_I$ représente la charge de travail moyenne pondérée des intervieweurs. La variabilité de l'intervieweur contribue différemment aux effets du plan d'échantillonnage. Dans le cas des variables socio-démographiques, sa contribution se situe en moyenne un peu au-dessus de la moitié de la contribution attribuable au groupement, alors que pour les questions se rapportant aux attitudes, la contribution de l'intervieweur aux effets du plan d'échantillonnage représente le triple de celle due au groupement. La contribution des deux autres catégories de questions se situe entre ces deux extrêmes.

Les résultats du tableau 1 confirment l'influence de la charge de travail de l'intervieweur sur la variance des estimations de l'échantillon, en raison du facteur d'amplification. En deux mots, un élément associé à l'intervieweur qui présente une très petite corrélation dans la catégorie peut avoir des effets majeurs si la charge de travail de l'intervieweur est suffisamment importante. Dans l'enquête à l'étude, la logistique du déploiement des intervieweurs et les exigences relatives à un contrôle de la qualité soutenu semblent prêcher en faveur de petites équipes, pratique apparemment courante pour bon nombre d'enquêtes sur le terrain dans le secteur de la santé (lire, par exemple, Choi et Comstock 1975). Bref, le nombre d'entrevues se situe en moyenne au-dessus de 250. On se rend tout de suite compte du coût d'une telle stratégie en examinant le tableau 1; de très petites variations entre intervieweurs réduisent considérablement la précision des estimations de l'échantillon.

Passons maintenant à l'objet principal de notre analyse, soit l'incidence de la variabilité de l'intervieweur sur l'écart entre les moyennes ou les proportions des domaines. Nous l'avons évaluée pour deux jeux de comparaisons, le premier selon le sexe (masculin/féminin) et le second selon la race (européenne/non européenne). Ainsi qu'on a pu le constater dans la discussion suivant l'équation (11), la contribution $1 + (v_I - 1)\rho_I$, à D_0 attribuable à la variation entre intervieweurs dépend de la mesure dans laquelle l'effet dû à l'intervieweur est constant dans les deux domaines et de la manière dont la charge de travail individuelle est répartie

intervieweurs, les répondants devraient être attribués au hasard. On le fait rarement avec les enquêtes de grande envergure, car il s'agit d'une méthode onéreuse et compliquée. Il s'ensuit qu'on peut difficilement utiliser les résultats de telles enquêtes pour la recherche méthodologique, car il pourrait y avoir confusion entre les paramètres de l'intervieweur et ceux du répondant. L'analyse à degrés multiples comme celle décrite plus haut, y remédie dans une certaine mesure. Les variables pertinentes des répondants, si elles sont connues, peuvent être intégrées au modèle de régression, ce qui permet d'uniformiser les intervieweurs par des moyens statistiques... Une telle approche a cependant ses limites. En effet, cette dernière s'appuie sur une méthode de contrôle statistique plutôt qu'expérimentale. Elle repose aussi sur l'hypothèse que le modèle intègre constamment toutes les covariables pertinentes. Sans randomisation, nul ne peut conclure que l'influence de toutes les variables confusionnelles a été éliminée" [Traduction]. Dans le cas qui nous intéresse, le déploiement des intervieweurs permet d'identifier formellement tous les éléments de la variance, pourvu que le modèle soit crédible et que l'on accepte que l'attribution du travail aux intervieweurs n'a aucun lien avec les effets du groupement. L'absence de véritable randomisation signifie que les variations dans le genre de réponses obtenues par les intervieweurs pourraient toujours résulter de la répartition de la charge de travail plutôt que du style de l'intervieweur. De toute évidence, les résultats empiriques ne donnent qu'une indication, bref font ressortir des possibilités qu'on devra approfondir au moyen d'études expressément conçues pour cela.

Même en oubliant le manque de randomisation du déploiement des intervieweurs, on se rend compte que le plan d'échantillonnage est beaucoup plus complexe que celui supposé dans la partie théorique qui précède. En effet, l'enquête prévoit trois degrés d'échantillonnage et une stratification régionale des unités à la première étape. Pour l'analyse complète, nous avons ajusté le modèle davantage en y intégrant des effets constants pour la stratification, un modèle hiérarchique à effet aléatoire pour les trois niveaux d'échantillonnage et tous les termes d'interaction du deuxième degré. Malgré cela, les DL correspondant aux unités de la première étapes étaient si flous que l'écart entre les strates et la variance entre les DL sont négligeables pour toutes les variables qui reviennent dans l'analyse subséquente. La variance entre USÉ domine donc et, à toute fin pratique, on peut considérer que l'enquête repose sur un plan d'échantillonnage à deux degrés où les îlots de recensement (regroupés s'il y a lieu) font office d'USÉ. Nous n'avons pas tenu compte des autres éléments dans les résultats examinés plus bas.

4. RÉSULTATS

Nous débuterons par les effets de l'intervieweur et du groupement sur diverses moyennes et proportions. Nous nous sommes servis du modèle (9) pour les deux types de variables. On sait très bien qu'une telle méthode entraîne la sous-estimation des éléments de la variance pour les données binaires (lire Anderson et Aitken 1985 et Pannekock 1988 notamment). Les effets du plan d'échantillonnage sur les proportions que nous avons estimés devraient donc être considérés comme la limite inférieure de la fourchette. Les modèles sont ajustés au moyen du module PROC GLM du SAS. Les effets du groupement sont bien expliqués dans la documentation existante (Kish 1965; Kish et Frankel 1970; 1974). En général, leur ampleur dépend du genre et du nombre d'unités sélectionnées et varie sans doute avec tel ou tel paramètre sociodémographique. Dans l'enquête qui nous intéresse, on prévoyait des effets relativement élevés puisque les îlots de recensement servant d'unités d'échantillonnage devaient se caractériser par une assez bonne homogénéité interne. Face à une telle concentration des caractéristiques de la population, on a supposé que les paramètres démographiques et les éléments connexes donneraient les valeurs les plus importantes de ρ_C. On s'attendait à ce que ρ_I ait une valeur plus faible en raison de l'importante formation des intervieweurs. La documentation existante suggère que ces effets varieront aussi vraisemblablement avec le genre de questions, par exemple celles sur le comportement, les questions d'approfondissement, celles à choix obligatoire ainsi que les questions mal formulées ou ambiguës, particulièrement sensibles à la variabilité de l'intervieweur (Feather 1973, Groves 1989).

Les deux premières colonnes du tableau 1 présentent la mesure estimée des coefficients de corrélation intra-intervieweur et intra-groupe pour diverses questions réparties entre quatre catégories (paramètres sociodémographiques, attitudes, comportement récent et comportement ancien). Une catégorisation de ce genre crée des groupes naturels susceptibles de faire ressortir bon nombre des effets dus à l'intervieweur. Dans chaque groupe, les éléments sont énumérés par ordre de grandeur de la corrélation intra-intervieweur. L'annexe décrit les questions posées (outre celles sur les paramètres sociodémographiques, qui s'expliquent d'elles-mêmes).

Comme on pouvait s'y attendre, les variables sociodémographiques (sauf le sexe) présentent les valeurs les plus élevées pour la corrélation intra-groupe. ρ_C a une valeur moyenne de 0.07 (0.08 lorsqu'on exclut le sexe). La valeur moyenne de ρ_C pour les trois autres catégories ne dépasse pas 0.02. Il se peut que certains aspects que la logique associerait étroitement à la situation sociale – par exemple la fréquence des consultations, le paiement des consultations, le brossage des dents et quelques déclarations relatives aux attitudes – présentent une valeur de ρ_C supérieure à la moyenne. En général cependant, la valeur obtenue reste dans la fourchette signalée par d'autres auteurs (lire, par exemple, Kish 1965, p. 581 pour une série d'enquêtes auprès des consommateurs; Bebbington et Smith 1977 et Verma et coll. 1984 pour les études nationales de l'Enquête mondiale sur la fécondité).

Les valeurs estimées de ρ_I correspondantes apparaissent à la première colonne du tableau 1. Dans l'ensemble, les valeurs obtenues sont nettement plus faibles que celles relevées pour les effets du groupement (habituellement moins de la moitié, et dans certains cas le dixième de ρ_C pour les éléments correspondants). Comme prévu, certaines

l'intervieweur ne s'occupe que d'un seul domaine, $\bar{m}_I$ et $\bar{n}_I$ ont des valeurs analogues et l'effet de l'intervieweur sur l'écart est comparable à celui d'une moyenne simple. En règle générale, les intervieweurs interrogent des personnes des deux domaines et $\bar{m}_I$ est assez faible, ce qui semble corroborer dans une certaine mesure l'hypothèse que la variabilité de l'intervieweur n'exerce qu'une faible influence sur l'écart estimatif entre deux domaines. Les mêmes remarques s'appliquent aux effets du groupement.

L'analyse qui précède repose sur l'hypothèse selon laquelle les effets de l'intervieweur et du groupement, a_i et b_p, sont identiques pour les deux domaines. On peut facilement imaginer des situations où pareille supposition ne tiendrait pas. Certains intervieweurs, par exemple, réagiront très différemment avec un homme ou une femme, ou les membres de différents groupes ethniques. Le modèle qui suit tient compte de telles interactions éventuelles:

$$Y_{ipr}^{(d)} = \mu^{(d)} + a_i^{(d)} + b_p^{(d)} + e_{ipr}^{(d)}, \qquad (9)$$

où $a_i^{(a)}$ et $a_i^{(b)}$ (respectivement $b_p^{(a)}$ et $b_p^{(b)}$) sont désormais censés être des variables aléatoires corrélées selon $r_I(r_C)$. Le modèle simpliste (4) illustre le cas particulier où les variances des effets sont égales et où la valeur de r_I et de r_C est un. D'un autre côté, si les effets de l'intervieweur (du groupement) varient sensiblement d'un domaine à l'autre, $r_I(r_C)$ sera faible (voire aura une valeur négative dans les cas extrêmes). Dans le reste de notre article, nous supposerons pour plus de simplicité que les variances de $a_i^{(a)}$ et $a_i^{(b)}$ (soit $b_p^{(a)}$ et $b_p^{(b)}$ respectivement) sont identiques. Pareille supposition peut ou non être raisonnable dans la pratique, mais la simplification qu'elle permet nous aidera à nous concentrer sur les aspects essentiels. Le modèle de base se rapproche du cas général, mais ses éléments sont légèrement plus complexes. Selon l'équation (9), la variance prévue de $\bar{Y}^{(a)} - \bar{Y}^{(b)}$ est:

$$V(\bar{Y}^{(a)} - \bar{Y}^{(b)}) = (\bar{v}_I\sigma_I^2 + \bar{v}_c\sigma_C^2 + \sigma^2)\left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}}\right) \qquad (10)$$

où

$$\bar{v}_I = \sum_i \left(p_i^{(a)2} - 2r_I p_i^{(a)}p_i^{(b)} + p_i^{(b)2}\right) \Big/ \left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}}\right)$$

et où $\bar{v}_C$ a la même définition dans $q_p^{(a)}$ et $q_p^{(b)}$.

Le facteur d'inflation de la variance en vertu de ce nouveau modèle est:

$$D_2 = 1 + (\bar{v}_I - 1)\rho_I + (\bar{v}_C - 1)\rho_C. \qquad (11)$$

Il s'agit d'une fonction décroissante de r_I, c'est-à-dire de la corrélation entre les effets de l'intervieweur pour les deux domaines; plus faible est la corrélation, plus importante sera la hausse de la variance. Quand $r_I = 1$, $\bar{v}_I$ est égal à $\bar{m}_I$ et l'effet de l'intervieweur est négligeable, pourvu que tous les intervieweurs interrogent un nombre raisonnablement équilibré de personnes dans les deux domaines. Si r_I est faible cependant (signe qu'il existe une forte interaction entre les intervieweurs et les domaines), $\bar{v}_I$ a le même ordre de grandeur que $\bar{n}_I$ et la variabilité de l'intervieweur peut influer sensiblement sur la variance de l'écart entre les domaines.

Dans la pratique, les résultats se situeront entre les deux extrêmes et leur incidence probable devra être établie de manière empirique. Dans la prochaine partie, nous commencerons donc à recueillir des données pratiques en nous servant des réponses à diverses questions posées lors d'une simple enquête sur la santé caractéristique au type de recherches pour lesquelles les comparaisons de domaines présentent une grande importance (même si l'enquête en question n'avait pas été spécifiquement conçue pour l'usage auquel nous la destinons!).

3. EXEMPLE

Notre exemple repose sur les données issues d'une enquête concernant l'hygiène buccale, les attitudes et les habitudes des Néo-zélandais adultes. On trouvera une description détaillée de l'enquête ailleurs (Cutress et coll. 1979). Les aspects de l'étude qui nous intéressent le plus dans le cadre de la recherche actuelle sont le plan d'échantillonnage et le déploiement des intervieweurs.

L'échantillon a été constitué au moyen d'un plan d'échantillonnage stratifié à degrés multiples. Le pays a été divisé en 256 districts locaux (DL) et on a tiré un échantillon géographiquement stratifié de 68 DL des 256 précédents, en fonction de probabilités de sélection proportionnelles à la taille (PPT) de la population à la première étape, la taille de la population correspondant au nombre estimé de personnes de 15 ans et plus. Chaque DL échantillonné a été divisé en unités secondaires d'échantillonnage (USÉ) constituées des îlots de recensement existants, regroupés afin de donner un échantillon minimal de cinquante personnes le cas échéant. Deux USÉ ont ensuite été sélectionnées par PPT pour chaque DL échantillonné à la deuxième étape. Enfin, on a systématiquement retenu 28 adultes dans chaque USÉ. De cette façon, on a uniformisé la probabilité de sélection finale pour tous les adultes, si bien que l'échantillon est (pratiquement) autopondéré.

L'aspect principal de l'enquête concerne le déploiement des intervieweurs. On a recouru aux services de treize intervieweurs, soit au moins trois pour chaque USÉ. Au moins 10% des personnes que devaient interroger les intervieweurs venaient de la même région (Auckland). Idéalement, l'attribution du travail aux intervieweurs devrait faire partie du plan d'échantillonnage, comme le recommandent Fellegi (1974) ou Biemer et Stokes (1985). Malheureusement, l'enquête n'avait pas pour but d'estimer la variance due à l'intervieweur et les répondants ont été répartis de façon anarchique plutôt qu'au moyen d'une véritable méthode de randomisation.

Il s'agit d'une situation assez caractéristique aux études importantes. La citation de Hox (1994) que voici la résume bien: "Idéalement, dans les enquêtes où interviennent des

intervieweurs. Voici un modèle de réponses corrélées simple qui convient aux observations effectuées dans le cadre d'une telle enquête:

$$Y_{ipr} = \mu + a_i + b_p + e_{ipr}, \quad (1)$$

où i désigne l'intervieweur, p l'unité primaire d'échantillonnage (UPÉ) et r, le répondant. La moyenne μ est constante et on suppose que les autres éléments (a_i, b_p et e_{ipr}) sont des variables aléatoires indépendantes dont la variance respective est σ_I^2, σ_C^2 et σ^2. Les modèles de ce genre sont abondamment utilisés dans les recherches théoriques sur la variance des réponses (lire Prasad et Rao 1990 pour un exemple récent; pour les travaux antérieurs, se reporter à l'analyse exhaustive qu'on retrouve à la section 11.3 de Lessler et Kalsbeek 1992).

Puisqu'il y a pondération automatique, la moyenne de l'échantillon, $\bar{Y}$, correspond à l'estimateur naturel de la moyenne de la population. Sa variance selon le modèle de réponses corrélées (1) est donc:

$$V(\bar{Y}) = (\bar{n}_I\sigma_I^2 + \bar{n}_C\sigma_C^2 + \sigma^2)/n$$

équation dans laquelle $\bar{n}_I = \sum_i n_i^2/n$, où n_i représente le nombre de répondants interrogés par le i-ième intervieweur et $n = \sum_i n_i$, la taille de l'échantillon, et dans laquelle $\bar{n}_C = \sum_p m_p^2/n$, où m_p indique le nombre de répondants de la p-ième UPÉ. Soulignons que $\bar{n}_I$ dépasse toujours la moyenne arithmétique simple des facteurs n_i et que sa valeur peut être considérablement plus élevée si n_i varie beaucoup.

Examinons maintenant la variance prévue correspondante, $V_0(\bar{Y})$ par exemple, qu'on obtiendrait si les n observations étaient indépendantes (à savoir, si on avait prélevé un échantillon au hasard dans une très vaste population d'UPÉ au moyen d'un grand nombre d'intervieweurs). De (1), on déduit que

$$V_0 = \sigma_{I0t}^2/n \quad (2)$$

où

$$\sigma_{I0t}^2 = \sigma_I^2 + \sigma_C^2 + \sigma^2.$$

Le relèvement de la variance prévue attribuable à l'effet combiné de la variabilité de l'intervieweur et de la corrélation dans le groupe est calculé au moyen du ratio

$$D_0 = V(\bar{Y})/V_0$$

$$= 1 + (\bar{n}_I - 1)\rho_I + (\bar{n}_C - 1)\rho_C \quad (3)$$

où $\rho_I = \sigma_I^2/\sigma_{I0t}^2$ et $\rho_C = \sigma_C^2/\sigma_{I0t}^2$. Nous appellerons ce ratio "effet du plan d'échantillonnage", même s'il diffère légèrement de la définition courante, laquelle se rapporte à la variance réelle plutôt qu'à la variance prévue. De l'équation (3) il ressort que la variabilité de l'intervieweur peut passablement influer sur la variance de la moyenne d'un échantillon quand le nombre moyen de cas confiés à l'intervieweur $\bar{n}_I$ est assez important et cela, même si la corrélation intra-intervieweur ρ_I demeure assez faible.

Supposons maintenant que nous nous intéressions à l'écart entre les moyennes de deux domaines plutôt qu'à une moyenne simple, par exemple aux différences liées au sexe ou à la race. L'extension la plus simple du modèle de réponses corrélées (1) pourrait donner le modèle suivant:

$$Y_{ipr}^{(d)} = \mu^{(d)} + a_i + b_p + e_{ipr}^{(d)} \quad (4)$$

pour les observations du d-ième domaine. Dans ce cas, la moyenne $\mu^{(d)}$ de chaque domaine pourrait différer, mais on suppose que les effets attribuables à l'intervieweur et au groupement sont identiques.

Soit $p_i^{(d)} = n_i^{(d)}/n^{(d)}$, où $n_i^{(d)}$ représente le nombre de répondants du domaine d interrogés par le i-ième intervieweur, et $n^{(d)}$ correspond au nombre total de répondants du domaine d. Parallèlement, soit $q_p^{(d)} = m_p^{(d)}/n^{(d)}$, où $m_p^{(d)}$ indique le nombre de répondants du domaine d pour la p-ième UPÉ. En vertu de l'équation (4), la variance prévue de $\bar{Y}^{(a)} - \bar{Y}^{(b)}$, c'est-à-dire l'écart de la moyenne de l'échantillon pour les deux domaines serait:

$$V(\bar{Y}^{(a)} - \bar{Y}^{(b)}) =$$

$$(\bar{m}_I\sigma_I^2 + \bar{m}_C\sigma_C^2 + \sigma^2)\left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}}\right), \quad (5)$$

où

$$\bar{m}_I = \sum_i (p_i^{(a)} - p_i^{(b)})^2 \Big/ \left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}}\right) \quad (6)$$

et

$$\bar{m}_C = \sum_p (q_p^{(a)} - q_p^{(b)})^2 \Big/ \left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}}\right). \quad (7)$$

Si les observations étaient indépendantes, la variance prévue correspondante serait:

$$V_1 = \sigma_{I0t}^2\left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}}\right)$$

si bien que l'inflation attribuable à la variabilité de l'intervieweur et à la corrélation dans le groupe s'établit désormais à:

$$D_1 = \mathrm{Var}(\bar{Y}^{(a)} - \bar{Y}^{(b)})/V_1$$

$$= 1 + (\bar{m}_I - 1)\rho_I + (\bar{m}_C - 1)\rho_C. \quad (8)$$

L'ampleur de l'effet dépend de la façon dont la charge de travail des intervieweurs et les UPÉ coupent les domaines. À une extrémité, lorsque chaque intervieweur interroge le même nombre de répondants dans chaque domaine (à savoir, quand $p_i^{(a)} = p_i^{(b)}$), $\bar{m}_I$ est égal à zéro et les effets de l'intervieweur s'annulent. À l'autre cependant, lorsque

Techniques d'enquête, décembre 1995
Vol. 21, n° 2, pp. 111–118
Statistique Canada

La variance de l'intervieweur et ses effets sur les comparaisons de domaines

PETER DAVIS et ALASTAIR SCOTT[1]

RÉSUMÉ

Le présent article étudie les effets de la variabilité de l'intervieweur sur la précision des écarts estimés entre les moyennes de domaines. La première partie est consacrée à l'élaboration d'un modèle des éléments corrélés de la variance, permettant d'identifier les facteurs qui déterminent l'ampleur des effets, aspect qui influe sur le plan d'échantillonnage et la formation de l'intervieweur. Dans la deuxième partie, on trouvera une analyse empirique des données provenant d'une vaste enquête à plusieurs degrés sur l'hygiène dentaire. On s'est servi du sexe et de la race du répondant pour créer les deux jeux de domaines servant à la comparaison. Les effets globaux dus à l'intervieweur et au groupement de l'échantillon n'ont guère d'incidence sur la variance des comparaisons entre sexes masculin et féminin, mais la variance augmente pour certains écarts entre les deux groupes ethniques retenus. De fait, les effets de l'intervieweur sont deux ou trois fois plus élevés pour la comparaison des groupes ethniques que pour celle des deux sexes. Ces constatations présentent une certaine utilité pour les enquêtes dans le secteur de la santé où il est courant de recourir à un petit nombre d'intervieweurs ayant subi une formation poussée.

MOTS CLÉS: Variance due à l'intervieweur; comparaisons de domaines; effet du plan d'échantillonnage.

1. INTRODUCTION

Les enquêtes qui exigent une formation spéciale et poussée des intervieweurs, comme c'est souvent le cas dans le secteur de la santé, doivent fréquemment se contenter d'un petit nombre d'intervieweurs. Évaluer l'incidence de la variabilité de l'intervieweur sur des statistiques simples comme la moyenne et les proportions a suscité passablement de recherches, et on sait pertinemment que le recours à un nombre restreint d'intervieweurs, à qui on attribue une lourde charge de travail, peut contribuer de manière assez importante à l'erreur totale. Groves (1989, ch. 8) et Lessler et Kalsbeek (1992, n° 11.3) font une très bonne synthèse de la documentation existante à ce sujet. Néanmoins, la plupart des enquêtes médicales et sociales portent essentiellement sur des aspects plus complexes, par exemple la comparaison de sous-groupes ou l'estimation des effets d'un paramètre sur l'issue d'une maladie, si bien qu'on a tendance à croire que la variabilité de l'intervieweur est beaucoup moins importante et qu'il n'y a guère de danger à utiliser un nombre relativement restreint d'intervieweurs. Après les travaux de défrichage de Kish et Frankel (1974), les effets du groupement sur l'ajustement des modèles de régression multiple ou des modèles logarithmiques-linéaires de données catégoriques ont fait l'objet de maintes recherches théoriques et empiriques. Skinner et ses collaborateurs (1989) et Rao et Thomas (1988) ont bien dépouillé la documentation pertinente. Le problème de la comparaison des moyennes des sous-groupes, plus simple du point de vue conceptuel mais tout aussi important sur le plan pratique, a également donné lieu à certaines recherches empiriques (lire Kish 1987 et Skinner 1989, par exemple), mais on s'est relativement peu intéressé au côté théorique.

Nous examinerons surtout ici les comparaisons de sous-groupes (ou domaines). Nous débuterons par la partie théorique en utilisant un modèle simple des éléments de la variance. Théoriquement, l'impact de la variabilité de l'intervieweur dépend de deux choses: la répartition de la charge de travail de l'intervieweur entre les domaines et l'interaction entre intervieweur et domaine. Nous appliquerons ensuite la théorie aux données issues d'une enquête assez typique dans le secteur de la santé, en établissant deux jeux de domaines selon le sexe et la race du répondant. Malheureusement, l'enquête en question n'avait pas été conçue pour estimer les effets de l'intervieweur. Plus précisément, les intervieweurs n'avaient pas été déployés de façon aléatoire, de sorte que nos résultats sont plus indicatifs que probants. Quoi qu'il en soit, ils sont assez troublants pour que le problème justifie une étude plus poussée. Les résultats venant de la comparaison des groupes ethniques, en particulier, donnent à penser que l'utilisation d'un petit nombre d'intervieweurs pourrait susciter des difficultés dans certains cas, même lorsque l'analyse porte sur des comparaisons plutôt que des moyennes ou de simples proportions.

2. THÉORIE

Pour plus de simplicité, nous débuterons avec un plan d'échantillonnage autopondéré à deux degrés. Ce dernier est assez complexe pour mettre en relief les grandes idées, mais reste assez simple pour qu'on ne s'enlise pas dans une foule de détails. Respectant la suggestion de Collins et Butcher (1982), nous nous attaquerons simultanément au problème de la variance et à celui du groupement des

[1] Peter Davis, Department of Community Health, University of Auckland, Private Bag 92019, Auckland, New Zealand; et Professor Alastair Scott, Department of Mathematics and Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand.

Ernst et Ikeda présentent un algorithme à taille réduite pour maximiser la rétention de certaines unités primaires d'échantillonnage (UPÉ) quand un nouvel échantillon (c.-à-d. présentant une nouvelle stratification et une nouvelle répartition) est tiré pour une enquête répétée. Pour commencer, les auteurs décrivent la méthode de transport élaborée par Causey, Cox et Ernst (1985). Cette dernière permet d'optimiser la rétention des UPÉ, mais le problème de transport pourrait être trop grand pour se prêter à une solution pratique. Puis, les auteurs exposent leur algorithme, qui est une approximation de la méthode précédente, mais qui présente l'avantage d'être de taille plus petite, donc, utilisable dans de nombreuses situations pratiques. Enfin, ils présentent une application de l'algorithme à la Survey of Income and Program Participation.

Shrestha et Preston évaluent la cohérence des données sur les aînés tirées du Recensement et des registres de l'état civil aux États-Unis. Pour commencer, ils décrivent les données sur lesquelles porte l'étude et leurs sources. Puis, ils présentent la méthodologie utilisée pour évaluer la qualité des données sur les aînés et expliquent comment il convient d'interpréter les résultats de l'application de cette méthodologie. Enfin, ils présentent les résultats de l'application de la méthodologie aux données recueillies de 1970 à 1990.

Dillman, Clark et Sinclair comparent l'incidence de diverses stratégies d'envoi postal et de suivi sur le taux de réponse obtenu lors du Recensement des États- Unis. La comparaison des stratégies utilise un plan factoriel et un échantillon de 50,000 unités de logement. Les auteurs analysent les résultats en s'appuyant sur des comparaisons multiples, par paire, des moyennes de traitement et sur la régression logistique.

Forster et Snow évaluent l'utilisation d'ordinateurs portatifs pour la réalisation d'enquêtes démographiques dans les pays en développement. Une étude de collecte de données a été effectuée en vue de comparer l'utilisation de questionnaires imprimés et de questionnaires informatisés dans le cadre de l'enquête sur la mortalité des adultes menée auprès des habitants de la côte kenyane. Les résultats indiquent que l'utilisation d'ordinateurs portatifs permet de réduire le temps de traitement des données, d'améliorer la qualité de ces dernières et de diminuer les coûts d'enquête à long terme.

Le rédacteur en chef

Techniques d'enquête, décembre 1995
Vol. 21, n° 2, pp. 109–110
Statistique Canada

Dans ce numéro

Ce numéro de *Techniques d'enquête* contient des articles traitant de divers sujets. Dans le premier, Davis et Scott examinent l'incidence éventuelle des effets dus à l'intervieweur sur les comparaisons entre moyennes de domaines. Grâce à un modèle des éléments de la variance, ils montrent de façon théorique que l'incidence dépend de la répartition de la charge de travail de l'intervieweur entre les domaines, d'une part, et de l'interaction entre intervieweurs et domaines, d'autre part. Ils appliquent le modèle aux données d'une enquête sur la santé afin d'estimer la grandeur des effets dus à linterviewer aux fins de comparaisons entre sexes et entre groupes ethniques. Les auteurs ont noté que, dans certains cas, les effets dus à l'intervieweur qui sont particuliers à un domaine ont une incidence considérable sur l'exactitude des comparaisons entre domaines.

Rivest et Hurtubise examinent l'utilité de la moyenne winsorizée comme estimateur de la moyenne d'une population ayant une distribution asymétrique vers la droite. Une moyenne winsorizée est obtenue en remplaçant toutes les observations supérieures à une valeur limite donnée R par cette même valeur R, avant le calcul de la moyenne. Les auteurs suggèrent un algorithme simple pour le calcul de R qui minimise l'erreur quadratique de l'estimateur. Les auteurs appliquent cette méthode à plusieurs tailles d'échantillon et à divers plan de sondage incluant l'échantillonnage stratifié et avec probabilités proportionnelles à la taille. Ils dérivent des approximations directes de l'efficacité de la moyenne winsorizée. Ils terminent leur article par une simulation de Monte Carlo pour comparer divers estimateurs qui réduisent l'impact des valeurs extrêmes.

Kish, Frankel et Verma examinent l'incidence possible et l'importance des effets du plan de sondage (eps) sur un ensemble de statistiques apparentées. En se fondant sur 14 enquêtes menées dans six pays, les auteurs présentent une approche empirique mettant en rapport l'effet du plan de sondage de statistiques analytiques, $\text{eps}(p_i - p_j)$, aux effets du plan de sondage des statistiques séparées, $\text{eps}(p_i)$ et $\text{eps}(p_j)$, pour deux des nombreuses catégories de la même variable. L'approximation proposée doit faire l'objet de vérifications constantes. Pourtant, elle semble s'appliquer largement aux données étudiées et est nettement préférable aux hypothèses énoncées jusqu'à maintenant sur l'$\text{eps}(p_i - p_j)$.

Dupont discute de l'estimation d'un total à partir d'un échantillon à deux phases et en présence d'information auxiliaire. Dans un premier temps, on présente trois estimateurs par régression chacun employant l'information auxiliaire d'une façon différente. Ensuite, quatre estimateurs par calage sont proposés chacun correspondant à une stratégie particulière d'utilisation de l'information auxiliaire. Par la suite, Dupont montre que les stratégies de calage peuvent être associées à une modélisation par régression. Dans cet article, on discute aussi de l'estimation de la variance pour les sept estimateurs présentés, du choix de l'estimateur en présence de non-réponse ainsi que de l'utilisation a priori ou a posteriori de l'information auxiliaire.

Spisak discute du recours au contrôle statistique du processus pour assurer la qualité d'une base de sondage construite selon un processus en continu et utilisée pour une enquête qui se répète périodiquement. Les tailles de la base de sondage constituent une série chronologique pour laquelle il convient de déterminer le modèle qui permettra d'estimer la variance du processus nécessaire pour l'établissement des graphiques de contrôle. L'auteur se sert des données du United States Unemployment Insurance Benefits Quality Control Program pour illustrer la méthode.

Binder et Kovacevic montrent comment on peut se servir de la méthode des équations d'estimation pour établir des procédures appropriées d'estimation de la variance dans le cas de données tirées d'une enquête dont le plan d'échantillonnage est complexe. La méthode est surtout utile lorsque la grandeur à estimer correspond à une fonction non linéaire complexe des valeurs de la population observée, comme dans le cas de bon nombre de mesures courantes de l'inégalité du revenu. Les auteurs mettent au point les détails de la méthode proposée pour plusieurs statistiques complexes de la répartition du revenu, y compris le coefficient de Gini, l'ordonnée de la courbe de Lorenz, la part du quantile et la mesure du faible revenu. Ils donnent un exemple numérique fondé sur des données de l'Enquête sur les finances des consommateurs au Canada.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada
Volume 21, numéro 2, décembre 1995

TABLE DES MATIÈRES

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à faire parvenir le texte rédigé en anglais ou en français au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes-ménages, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de Techniques d'enquête (n° 12-001 au catalogue) est de 45 $ par année au Canada, 50 $ (É.-U.) aux États-Unis, et de 55 $ (É.-U.) par année à l'étranger. Prière de faire parvenir votre demande d'abonnement à Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête et la Société Statistique du Canada.

TECHNIQUES D' ENQUÊTE

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

DÉCEMBRE 1995 • VOLUME 21 • NUMÉRO 2

Publication autorisée par le ministre
responsable de Statistique Canada

Décembre 1995

Prix : Canada : 45 $
États-Unis : 50 $ US
Autres pays : 55 $ US

N° 12-001 au catalogue

ISSN 0714-0045

Ottawa

Statistique Canada Statistics Canada

Canada

TECHNIQUES D'ENQUÊTE

Catalogue 12-001

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

NUMÉRO 2

Statistique Canada Statistics Canada

Canada