SURVEY METHODOLOGY

Catalogue 12-001

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 1995

•

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

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

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

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

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, $1 + (\nu_I - 1)\rho_I$, 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 ν_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 ν_I tends to be much higher and in extreme cases may equal the average case-load. In the current study values of ν_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

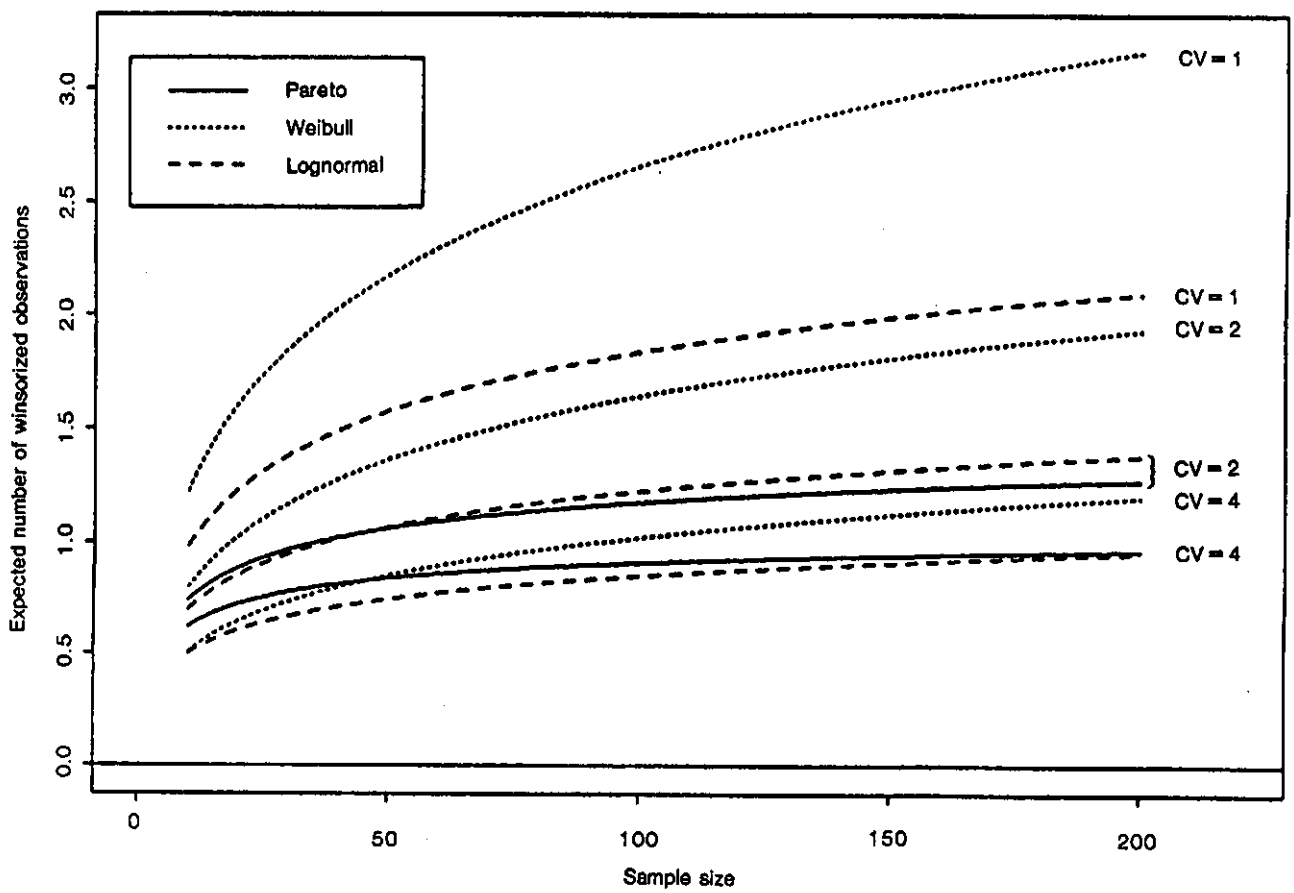


Figure 1. Expected number of winsorized observations for simple and stratified random sampling.

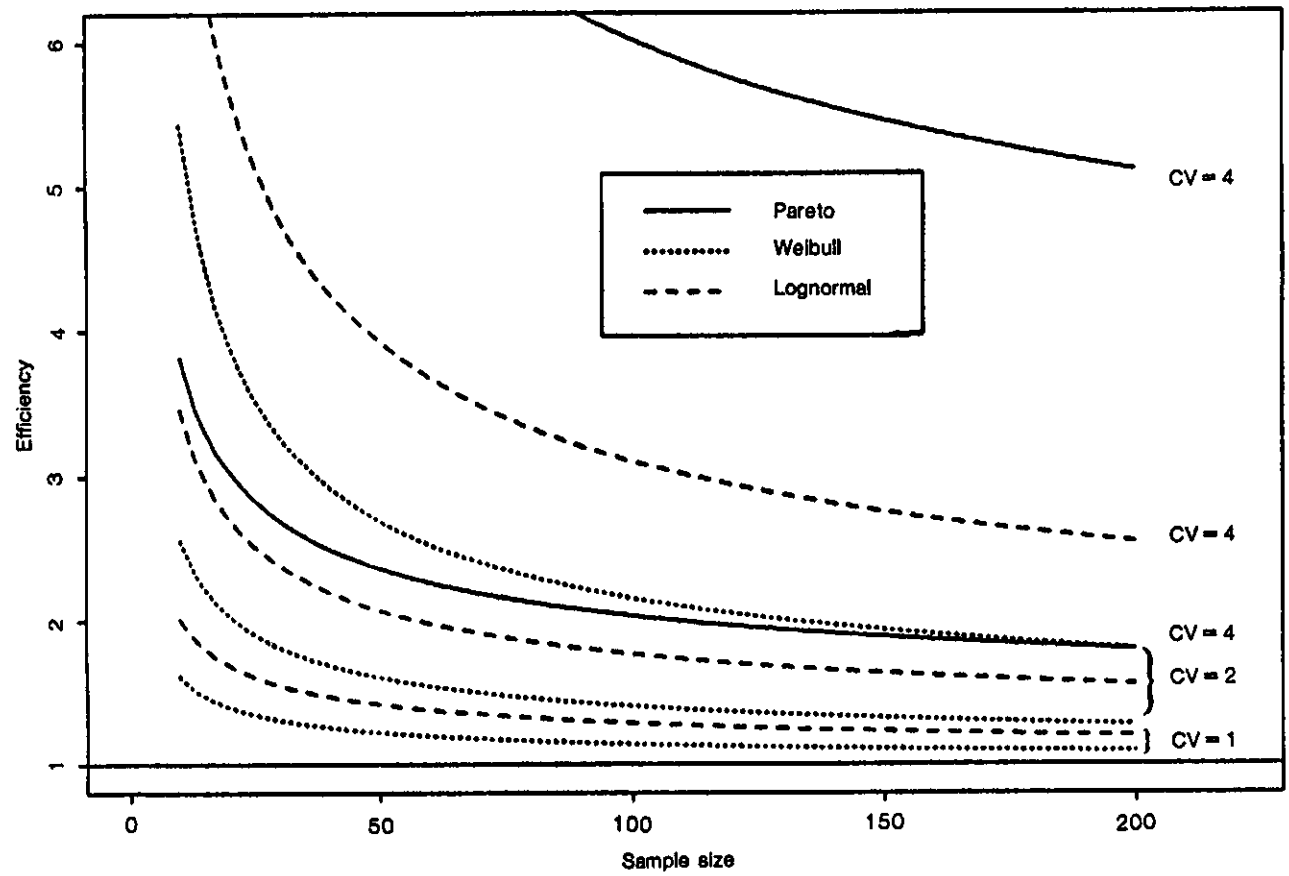


Figure 2. Efficiency of Searls winsorized mean.

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.

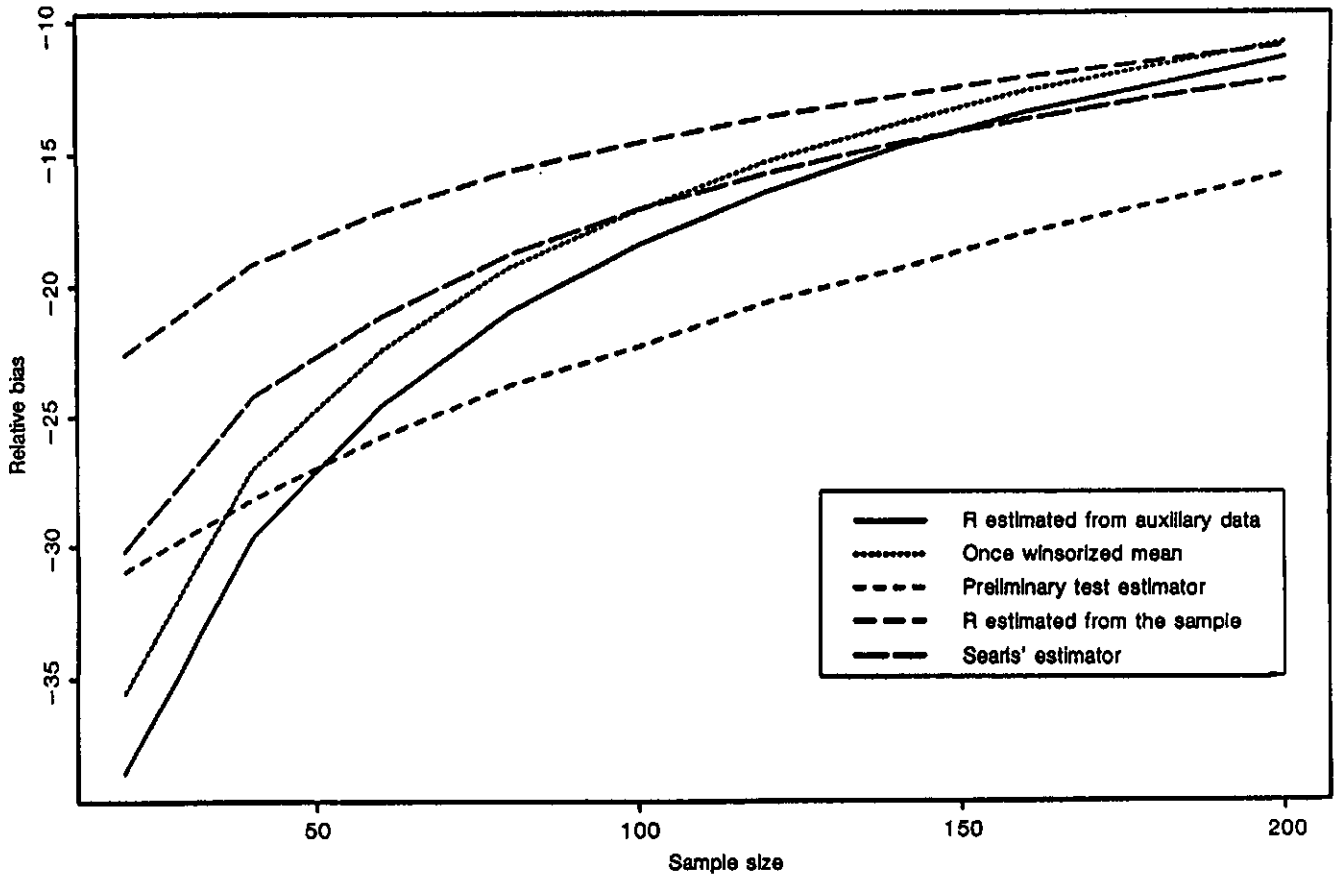


Figure 3. Relative bias of five estimators for the mean of CHICKEN.

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

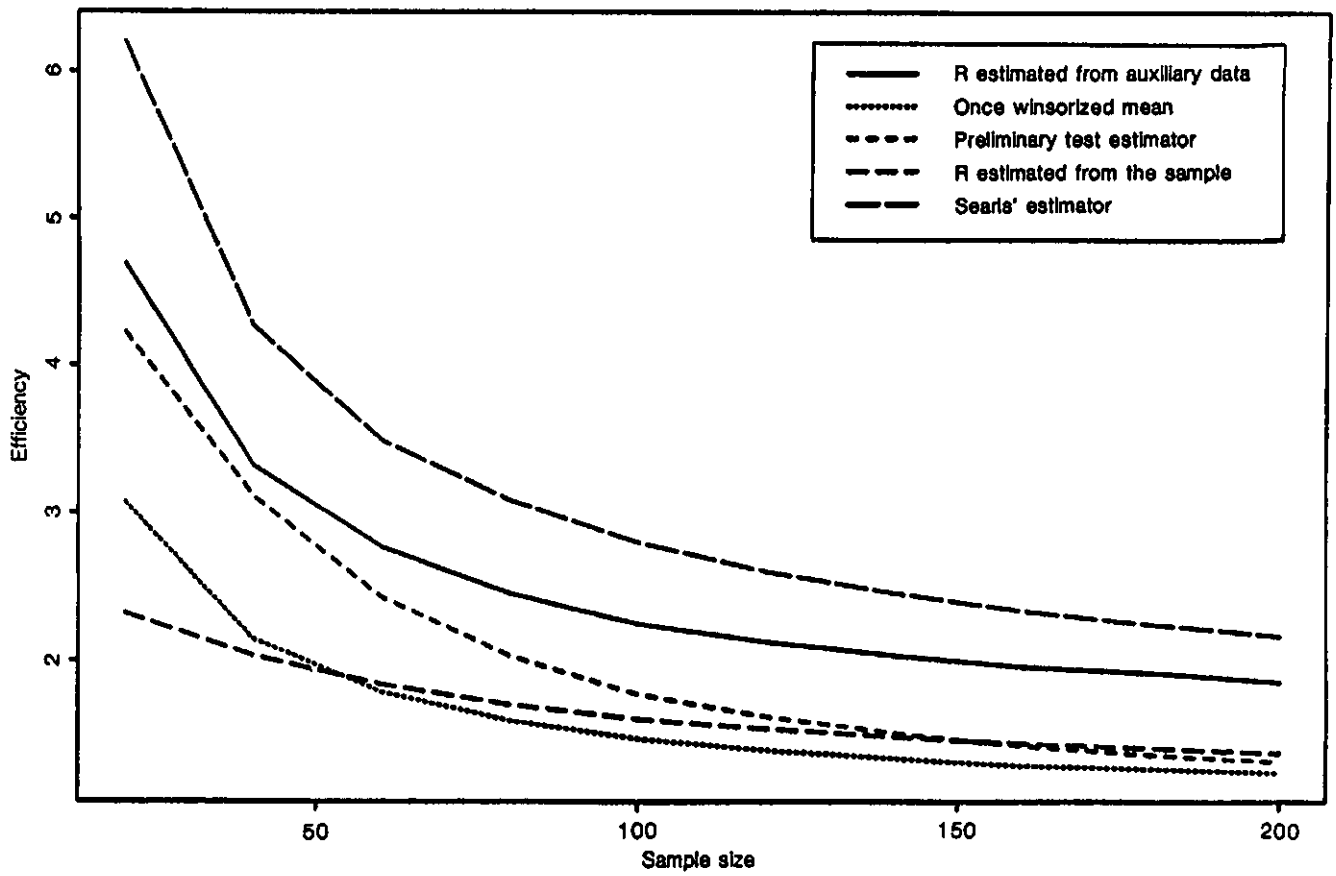


Figure 4. Efficiency of five estimators for the mean of CHICKEN.

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

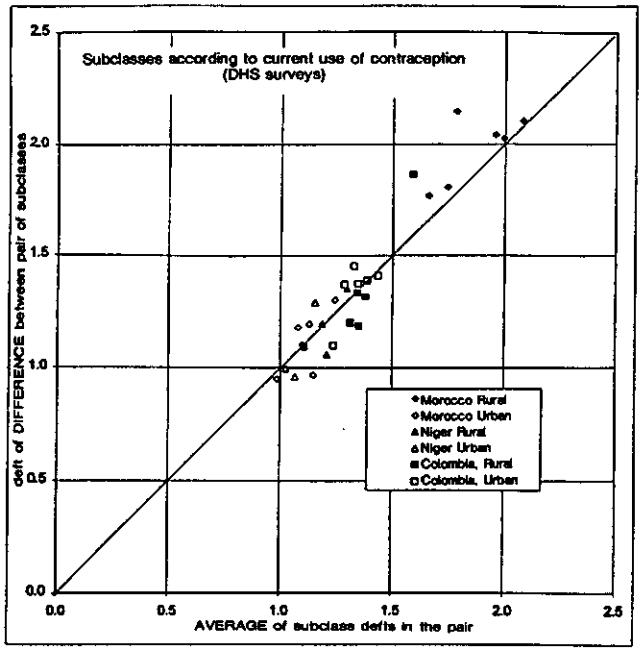


Figure 1. Comparison of deft $(p_i - p_j)$ to the average of deft(p_i),deft(p_j) for categories by current use of contraception*. Illustration of six populations from Demographic and Health Surveys.

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

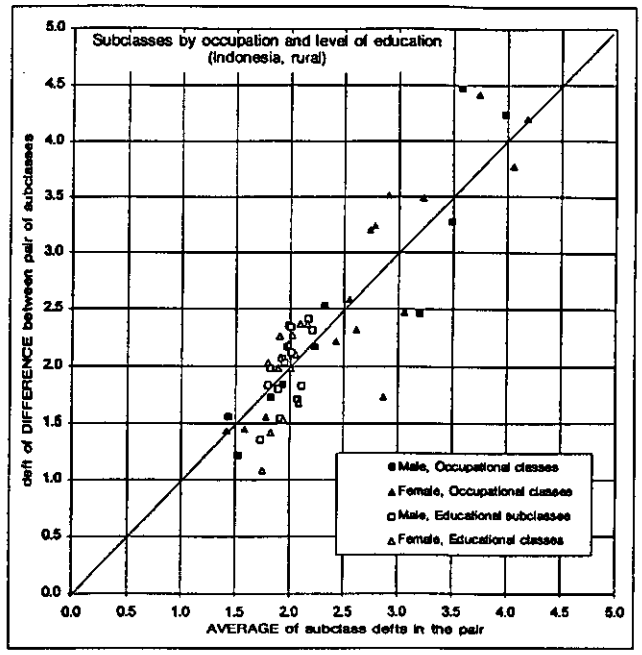


Figure 2. Comparison of deft $(p_i - p_j)$ to the average of deft(p_i),deft(p_j) for categories by occupation and level of education by sex. Illustration from a population census.

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		**D. Problems in Country**			
3-5	2.32	1.18	1.75	2.04					
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.

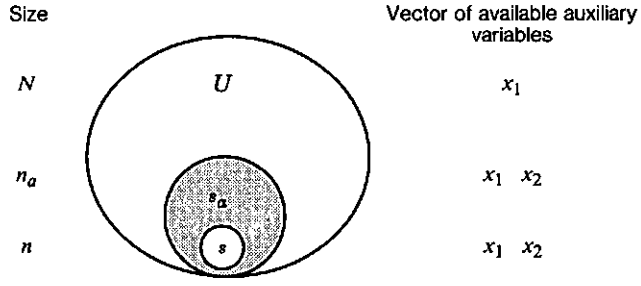
In a situation in which we wish to adjust a survey, and in which we wish simultaneously to correct the biases that would result from the use of gross weightings and to reduce the variance, the findings must be adapted: the changes introduced in the weightings to correct the biases are greater than the 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

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

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) $y_i = x_{i1}' b_1 + u_{i1}$,

and

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

$\hat{Y}_5 = \hat{Y}_2$ 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 (1) $y_i = x_{i1}' b_1 + u_{i1}$ 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

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

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.

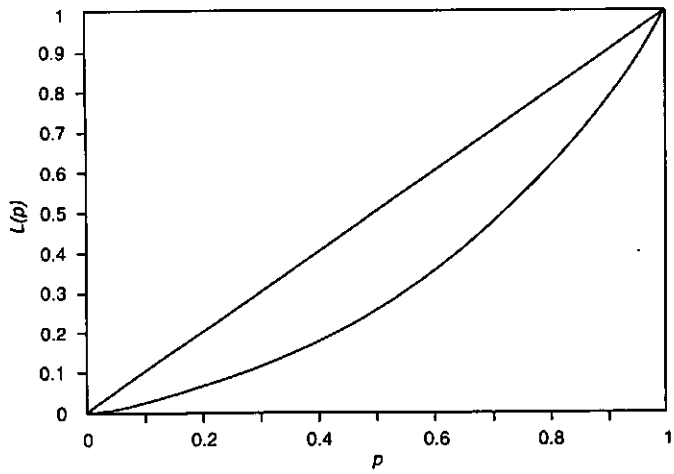


Figure 1. Lorenz Curve for the Weibull Distribution with Shape Parameter $\alpha = 1.6$.

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

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

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

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

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,

$$\int [u(y,\hat{\Theta}) - u(y,\Theta_o)] dF(y)$$

$$= (\hat{\Theta} - \Theta_o) \left. \frac{\partial E\{u(y,\Theta)\}}{\partial \Theta} \right|_{\Theta=\Theta_o} + o_p(|\hat{\Theta} - \Theta_o|).$$

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

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

$$\times \int u(y,\Theta_o) d\hat{F}(y) = \int u^*(y) d\hat{F}(y), \tag{2.4}$$

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:

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

Letting

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

and

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

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 $u^*_{hc} = \sum_i w_{hci}[I\{y_{hci} \le y\} - \hat{F}(y)]$.

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

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

which is the solution of the estimating equation

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

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

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

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

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

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

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

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

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

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

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

The resulting estimate is

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

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)

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

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 $\emptyset, \{1\},\{2\},\{3\},\{1,3\},\{2,3\}$; 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 $f(k)$, $k = 1, \ldots, n$, 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 $\pi_{f(k)}/p^{(k)}_{f(k)}$ 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

$$\pi_{f(k),g_k(\ell)}/p^{(\ell)}_{f(k),g_k(\ell)} = \max\{\pi_{f(k),j}/p^{(\ell)}_{f(k),j} : j \in T_{k\ell}\},$$

where

$$T_{k1} = \{1, \ldots, n\} \sim \{f(1), \ldots, f(k)\},$$

$$T_{k\ell} = T_{k(\ell-1)} \sim \{g_k(\ell-1)\}, \quad \ell = 2, \ldots, n-k,$$

$$T^*_{k\ell} = T_{k\ell} \cup \{f(k)\}, \quad \ell = 1, \ldots, n-k,$$

$$p^{(\ell)}_{f(k),j} = P(f(k), \quad j \in I \text{ and } I \subset T^*_{k\ell}), \quad \ell = 1, \ldots, n-k, \quad j \in T_{k\ell}. \tag{3.13}$$

Note that $p^{(\ell)}_{f(k),j}$ 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, $p^{(\ell)}_{f(k),g_k(\ell)}$ is the probability that $I^* = \{f(k), g_k(\ell)\}$. Furthermore, if $I_i = \{f(k), g_k(\ell)\}$ then $p_i^* = p^{(\ell)}_{f(k),g_k(\ell)}$, 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 $p^{(\ell)}_{f(k),j}$, it is established in Ernst and Ikeda (1994) that if $f(k) \in F_\alpha$, $j \in F_\beta$, then

$$p^{(\ell)}_{f(k),j} = p_{f(k),j} \prod_{\substack{t=1 \\ t \neq \alpha}}^{r} p'_t(T^*_{k\ell}) \text{ if } \alpha = \beta,$$

$$= p''_{f(k),\alpha}(T^*_{k\ell}) p''_{j\beta}(T^*_{k\ell}) \prod_{\substack{t=1 \\ t \neq \alpha,\beta}}^{r} p'_t(T^*_{k\ell}) \text{ if } \alpha \neq \beta. \tag{3.14}$$

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 $\pi_{2j}/p_j^{\{1\}}$. 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, $p_i^* = p^{(\ell)}_{f(k),g_k(\ell)}$. 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.

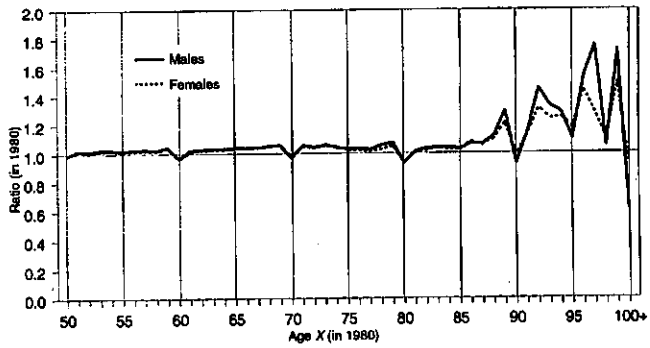


Figure 1A. Intercensal Ratios of Observed to Expected Population: Whites, 1970-1980.

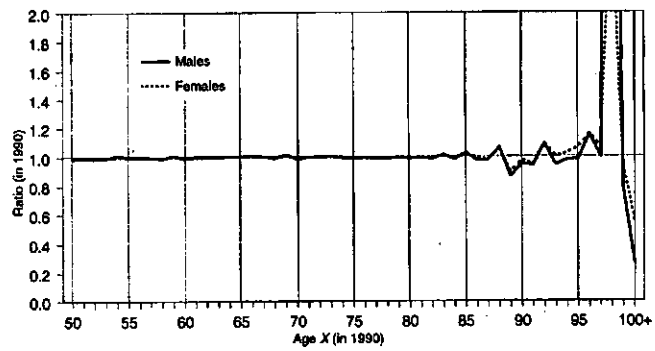


Figure 1B. Intercensal Ratios of Observed to Expected Population: Whites, 1980-1990.

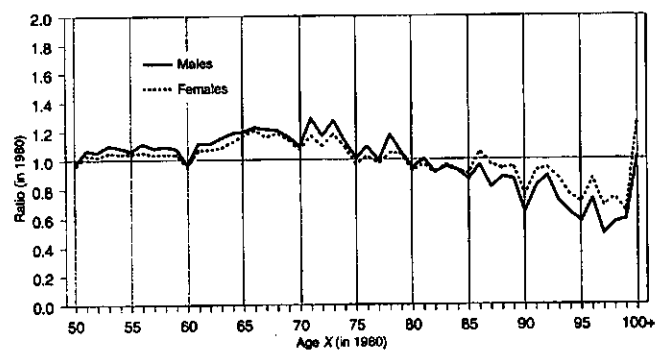


Figure 2A. Intercensal Ratios of Observed to Expected Population: Blacks, 1970-1980.

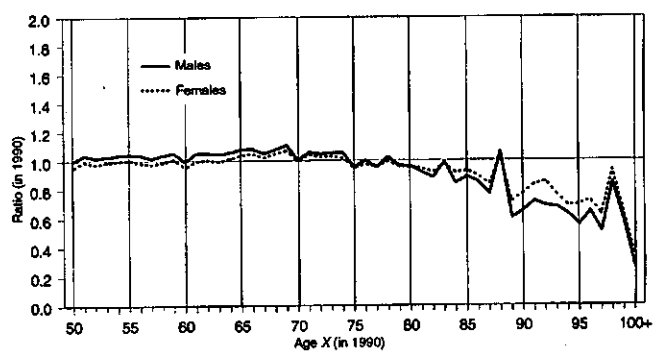


Figure 2B. Intercensal Ratios of Observed to Expected Population: Blacks, 1980-1990.

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.

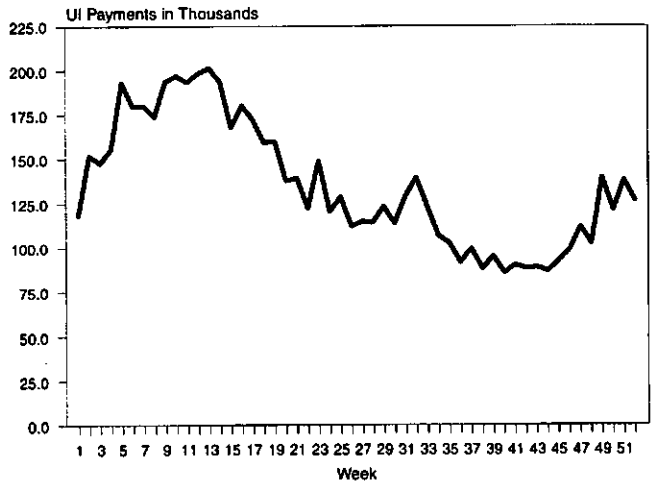


Figure 1. Number of UI payments per week.

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.

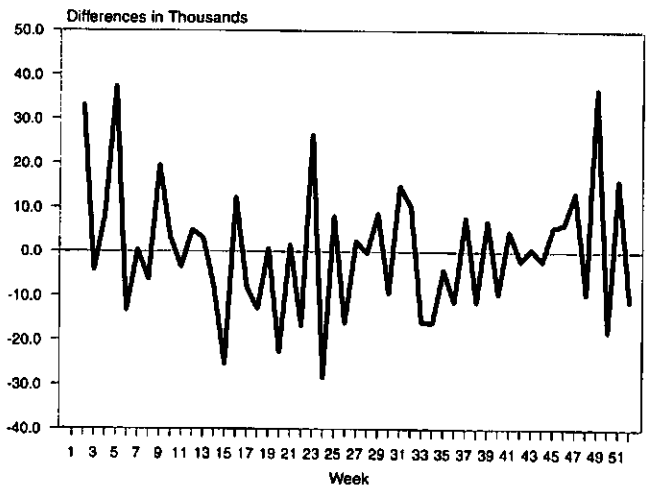
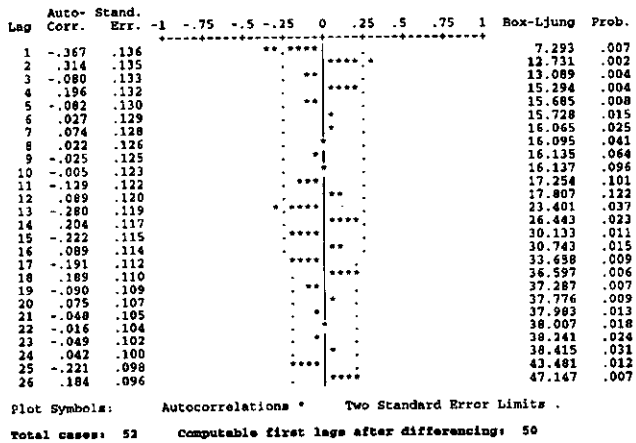


Figure 3. First differences of UI payments.

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.

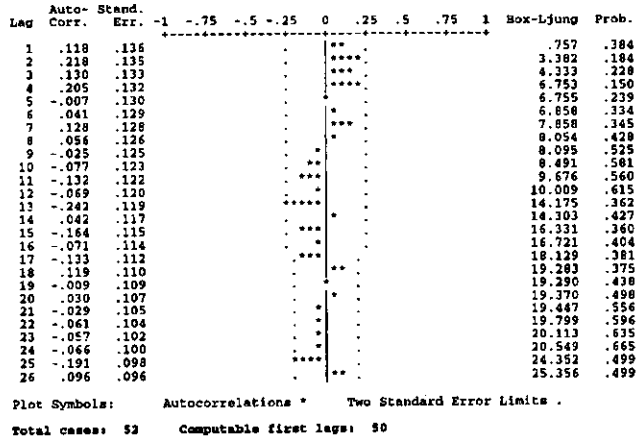


Lag	Auto-Corr.	Stand. Err.	Box-Ljung	Prob.
1	-.367	.136	7.293	.007
2	.314	.135	12.731	.002
3	-.080	.133	13.089	.004
4	.196	.132	15.294	.004
5	-.082	.130	15.685	.008
6	.027	.129	15.728	.015
7	.074	.128	16.065	.025
8	.022	.126	16.095	.041
9	-.025	.125	16.135	.064
10	-.005	.123	16.137	.096
11	-.129	.122	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	33.638	.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.	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 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.

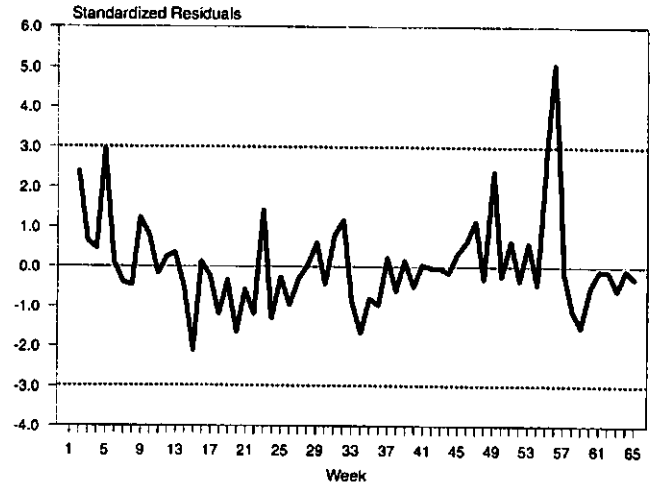


Figure 6. Control chart for model residuals (baseline data + next quarter).

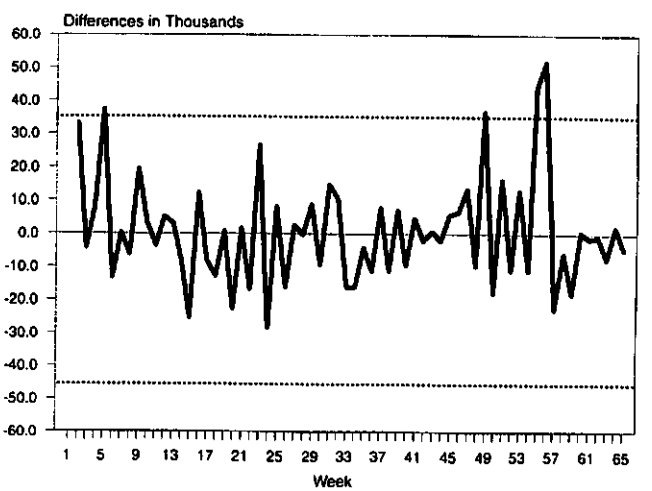


Figure 7. Control chart for UI payments (first differences − baseline + next quarter).

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.

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.