AN EVALUATION OF ROBUST ESTIMATION TECHNIQUES FOR IMPROVING ESTIMATES OF TOTAL HOGS

Susan Hicks and Matt Fetter, USDA/NASS
Susan Hicks, Research Division, 3251 Old Lee Hwy., Room 305, Fairfax, VA 22030

I. Abstract

Outliers are a recurring problem in agricultural surveys. While the best approach is to attack outliers in the design stage, eradicating sources of outliers if possible, large scale surveys are often designed to meet multiple, conflicting needs. Thus the survey practitioner is often faced with outliers in the estimation stage. Winsorization at an order statistic and Winsorization at a cutoff are two procedures for dealing with outliers. The purpose of this paper is to evaluate the efficiency, in terms of true MSE, of Winsorization for improving estimates of total hogs at the state level and to evaluate the efficiency of a data-driven technique for determining the optimal cutoff.

KEY WORDS: Outlier, Winsorization, Minimum Estimated MSE Trimming

II. Design of the Quarterly Agricultural Surveys

The National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture (USDA) provides quarterly estimates of total hogs at the state and national level through its Quarterly Agricultural Surveys (QAS). The QAS uses both a list and an area frame in a multiple frame (MF) approach to provide estimates for a variety of commodities in addition to hogs. Known farm operators are included on the list.

The area frame sampling is based on land use stratification. All land in the contiguous 48 states has a positive probability of selection in the area frame. Thus, the area frame is a complete frame and can be used to measure undercoverage in the list frame. Tract operations found in the area sample are matched to the list frame. Operators not on the list comprise the Non Overlap sample, or NOL.

III. Why are Estimates of Total Hogs so Variable?

Providing reliable estimates of total hogs through the multiple frame approach has always been difficult. The sampling variability in hog estimates is closely associated with the sampling variability in the NOL. While the NOL typically contributes only 25% to the total estimate, its contribution to the total variance is around 75%.

Outliers in the NOL can severely distort the estimates. Rumburg (1992) studied the causes and characteristics of NOL outlier records in five states. He cited three major contributors to outliers in the NOL:

- increased weights due to subsampling,
- the transitory and variable nature of hog production, and
- the location of hog operations on land with little agriculture.

The area frame stratification, based on land use strata, is more efficient for field crops than for livestock items, which tend to be less correlated with land use. Basically the variability in the NOL domain, which is a subset of the area frame, can be attributed to two factors:

1) the population within each strata is highly skewed to the right, and
2) the sample size is small.

IV. What is a Hog Outlier?

Most of the literature on truncation estimators for survey sampling describes its application to the problem of variability in weights. For household surveys, where we're frequently estimating Bernoulli characteristics, outliers are indeed caused by extreme sampling weights. For agricultural surveys extreme observations are caused by a combination of moderate to large weights and moderate to large values.

Lee (1991) addressed this problem by differentiating between outliers defined by classical statistics and influential observations. In classical statistics, outliers are unweighted values situated far away from the bulk of the data. Influential observations are valid reported values that may have a large influence on the estimate. Influential observations may involve outliers, but more frequently are a combination of relatively large

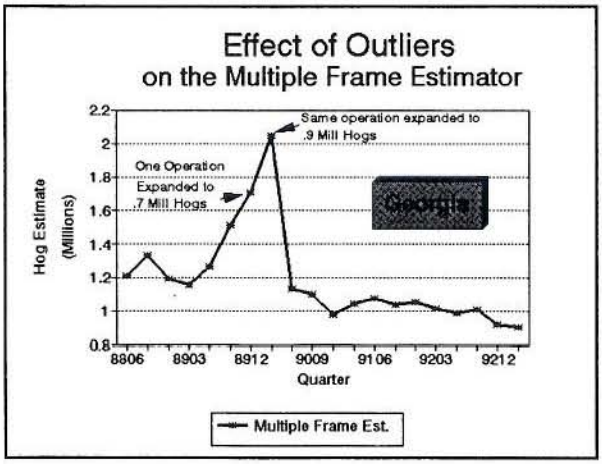
sampling weights and relatively large data values. For our purposes the term outlier will refer to influential weighted survey values, not unweighted values.

V. Current Procedures for Handling Outliers

Currently each state reviews potential outliers during the editing stage. Typically the 20 largest weighted values are listed through the Potential Outlier Prints System (POPS). The state commodity statisticians review these outputs and questionable data are verified. If the weighted data are correct, no changes are made.

However, the preliminary state hog recommendations are adjusted for outliers. Figure 1 shows the effect of extreme outliers in Georgia. In December of 1989, the operation that expanded to .7 million hogs comprised approximately 42.4% of the total multiple frame estimate. This is exactly the kind of situation we want to correct for.

Figure 1



Currently, outliers are adjusted in a somewhat ad hoc fashion at the state level. Treatment of outliers could include truncating the weight to 1.0, truncating the weight to some other value, or not truncating the weight at all. Although the effect of outliers is compensated for in the state recommendation, the records are never changed. This avoids the potential for biasing the national indication. Outliers at the state level are rarely outliers at the national level.

VI. Description of Two Winsorization Estimators

We evaluated two types of robust estimators for improving state level hog indications: Winsorization at a cutoff, t, and Winsorization at r order statistics. The form of the estimator which adjusts to a fixed cutoff is:

$$\hat{Y}_t = \sum_{j=1}^{n} adj\ w_j^t\ y_j \qquad (1)$$

where:

t = *truncation level*

y_j = *reported hogs for* j^{th} *unit*

$$adj = \frac{\sum_{n} w_j}{\sum_{n} w_j^t}$$

$$w_j^t = \begin{cases} w_j & \text{if } w_j y_j \le t \\ t/y_j & \text{if } w_j y_j > t \end{cases}$$

w_j = *design weight for* j^{th} *unit*

In this version of the standard truncation estimator, we truncate the weights of those observations whose weighted value expands larger than t so that the expanded value now equals t. The truncated portions are then "smoothed" over all observations.

We also evaluated estimators which adjust for the r largest values. The form of the estimator is:

$$\hat{Y}_r = \sum_{j=1}^{n} adj\ w_j^r\ y_j \qquad (2)$$

where:

$$w_j^r = \begin{cases} w_j & \text{for } j=1,\ldots,n-r \\ \dfrac{w_{n-r} y_{n-r}}{y_j} & \text{for } j=n-r+1,\ldots,n \end{cases}$$

$$adj = \frac{\sum_{n} w_j}{\sum_{n} w_j^r}$$

To evaluate the efficiency of each estimator for improving estimates of total hogs at the state level we developed a monte carlo simulation.

VII. Description of the Monte Carlo Simulation

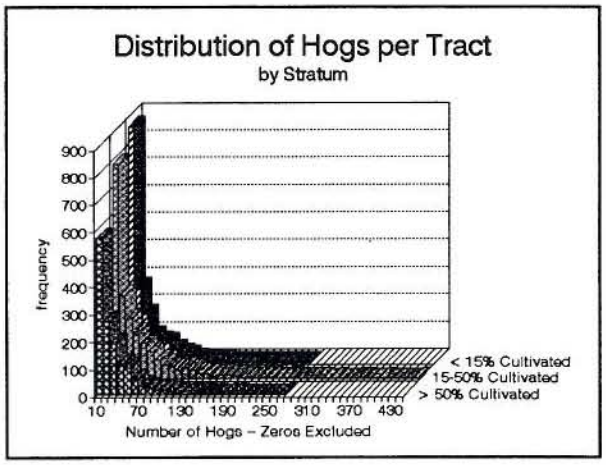
We built our simulation around one state, Georgia. Because of the complexities of the multiple frame design and because the major source of outliers and sampling variability is from the NOL, we restricted the simulation to the NOL domain of the area frame.

A positive NOL segment is defined as any sampled segment that contains at least one NOL hog operation. Separate farm operations are divided into tracts. The

number of sample segments in the area frame is fixed while the number of sample segments in the NOL domain is variable. For the NOL domain the sample unit is the segment, but the reporting unit is the tract. To simplify the simulation we modelled the positive NOL tracts and assumed a fixed NOL sample size.

The tract level weight is a product of the stratum sampling weight and a tract adjustment factor. The adjustment factor prorates an operation's reported value back to the tract level for operations that only partially reside within the sample segment. Based on historical June data from '91 and '92 for Georgia, we developed parametric models of the weighted tract level hog data for positive NOL tracts.

Figure 2



The three models -- one for each strata -- were all gamma density functions. Due to the sparseness of the data it was difficult to validate the model. However, our main interest was in developing a reasonable, highly skewed distribution rather than developing highly accurate models for Georgia NOL tracts.

We created a fixed sample universe for each stratum based on the estimated gamma distribution and the estimated number of positive NOL segments. We also estimated the proportion of positive NOL segments to total NOL segments. With the zero segments included, the result is a highly skewed population with a large spike at zero and a long right tail. See Figure 2.

From the fixed universe, we drew 1000 stratified simple random samples with replacement. Table 1 was created using a SAS program based on 1000 samples. Some of the other graphs were created from a Fortran program using the same data based on 10,000 samples. To compare the performance of the estimators for different sample sizes we chose a sample of size 360 to mimic the June sample and a sample of size 216 to mimic a follow-on sample.

The efficiency of the estimators was estimated as the ratio of the MSE of the unbiased estimator to the MSE of the new estimator. See Table 1 in next section.

VIII. Evaluation of Winsorization at a Cutoff and Winsorization at an Order Statistic

Ernst (1980) compared seven estimators of the sample mean which adjust for large observations. Four of the estimators were modifications of Winsorization at a cutoff, t. The other three estimators were modifications of Winsorization at an order statistic. Ernst showed that for the optimal t, the estimator which substitutes t for the sample values greater than t has minimum mean squared error. Earlier work by Searls (1966) showed that gains are achieved for wide choices of t when the data originate from an exponential distribution. The results from our monte carlo study are consistent with those studies.

Table 1

Truncation Level	MSE Ratio	Number Truncated	MSE Ratio
June			
14000	1.243	1	1.018
12000	1.299	2	.980
10000	**1.321**	3	.918
8000	1.236		
Follow-on			
18000	1.396	1	1.034
15000	**1.440**	2	.993
12000	1.402	3	.908
9000	1.175		
6000	.739		

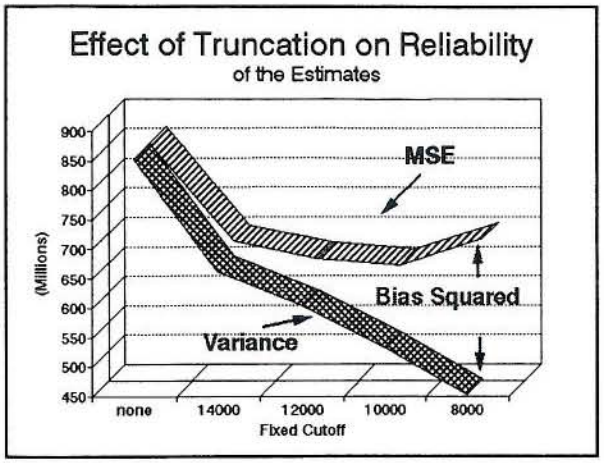
For both the June and follow-on samples, Winsorization at a cutoff is more efficient than Winsorization at an order statistic. Further for smaller cutoffs, more samples are truncated and more observations are truncated per sample. Thus, the bias in the estimator increases. Figure 3 shows the decomposition of variance and bias at each level of trimming evaluated for the June sample size.

In choosing a cutoff for truncation we'd like to minimize the number of samples truncated. In general, we don't want a cutoff that truncates every sample, but rather a cutoff that corrects for rare extreme observation like that depicted in Figure 1.

IX. Minimum Estimated MSE Trimming

In practice, we do not know the underlying distribution of the data and likewise the optimal value for t. The survey practitioner has to weigh the benefits of trimming -- decrease in variance -- with the

Figure 3



costs -- increase in bias.

The obvious criterion for evaluating the effect of different trimming levels is the estimated MSE. Potter (1988) documented a procedure called Minimum Estimated MSE Trimming. Because we do not know the true parameter, Y, we are limited to evaluating the bias based on the unbiased estimator $\hat{Y}$. The estimate of MSE ($\hat{Y}_t$) is derived from the relation:

$$E(\hat{Y}_t-\hat{Y})^2 = Var(\hat{Y}_t)+Var(\hat{Y})-2Cov(\hat{Y}_t,\hat{Y}) +[E(\hat{Y}_t)-E(\hat{Y})]^2 = MSE(\hat{Y}_t)+Var(\hat{Y})-2Cov(\hat{Y}_t,\hat{Y})$$

where:

$\hat{Y}_t$ = *the trimmed estimate*

$\hat{Y}$ = *the unbiased estimate*

Thus, an unbiased estimate of MSE($\hat{Y}_t$) is:

$$\hat{MSE}(\hat{Y}_t)=(\hat{Y}_t-\hat{Y})^2-\hat{V}(\hat{Y})+2\hat{COV}(\hat{Y}_t,\hat{Y})$$

If the correlation between the truncated estimate and the untruncated estimate is approximately 1.0, then this reduces to:

$$\hat{MSE}(\hat{Y}_t) = (\hat{Y}_t-\hat{Y})^2 - \hat{V}(\hat{Y}) + 2\,[\hat{V}(\hat{Y}_t)\hat{V}(\hat{Y})]^{1/2} \quad (3)$$

In this procedure, the estimated MSE is computed for various trimming levels and the trimming level with the minimum MSE is selected for implementation. This procedure can be used to suggest optimal trimming levels in (1) or number of observations to trim in (2). While the minimum MSE technique should identify the optimal trimming level over many samples, for any particular sample it could identify a trimming level far from the optimum. This occurs because our estimate of MSE is conditional on the sample we have drawn.

The efficiency of this estimator, in estimating the true MSE, depends on the efficiency of the variance estimators and the validity of the correlation assumption. In general, for a simple design and ignoring the effects of editing and nonresponse adjustments, we have unbiased estimators for $V(\hat{Y})$. However, obtaining an unbiased estimator for the variance of a truncation estimator is less straightforward. One approximation that is often made is to estimate the sampling variance of $\hat{Y}_t$ by treating the trimmed weights as if they represented the untrimmed weights in the usual variance formulae. We have used this approximation in (3).

X. Evaluation of Minimum Estimated MSE Trimming as a Data-driven Estimator

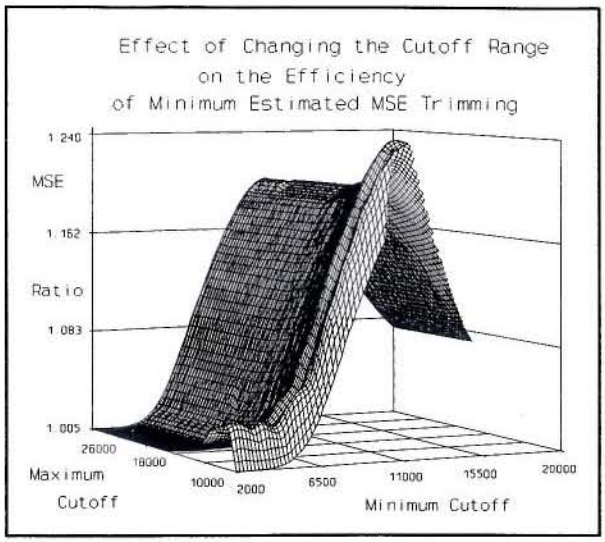
We wanted to evaluate the efficiency of the minimum MSE technique as a data-driven estimator. With this estimator each sample would be truncated at different levels to determine the level that minimized (3). Thus, the cutoff varies from sample to sample. Some preliminary runs showed that the efficiency of this data-driven estimator was highly dependent on the range of trimming levels evaluated. Thus, we evaluated this estimator over the set of possible trimming levels. The minimum trimming levels ranged from 2000 to 20,000 and the maximum trimming levels ranged from 10,000 to 28,000. We evaluated this estimator based on 10,000 monte carlo samples. For each monte carlo sample, the MSE estimator in (3) was used to determine the optimal cutoff for that sample for each range of trimming levels. The minimum and maximum trimming levels were each incremented by 1000 covering all combinations within the ranges specified above.

The level of the estimate, $\hat{Y}_t$, that minimized (3) was retained for each cutoff range and sample. The true MSE of the estimator for each combination of minimum and maximum cutoff was calculated in the usual fashion based on 10,000 $\hat{Y}_t$ estimates. And the MSE ratio is as defined before.

Recall from Table 1 that the optimal cutoff is around 10,000 to 11,000. As Figure 4 shows, this estimator is most efficient when the range of trimming levels evaluated is close to the optimum trimming level. As the minimum cutoff is reduced the efficiency of the estimator drops off rather dramatically. Whereas increasing the maximum cutoff has a minimal effect on the estimator.

For any particular sample the estimated MSE technique could identify a trimming level far from the optimum. This occurs because the estimated MSE is conditional on the sample we have drawn. Thus as Figure 4 shows, the efficiency of this technique as an estimator depends on the range of cutoffs we choose

Figure 4



to evaluate and how close that range is to the true optimum, similar to Winsorization at a cutoff.

Figure 5

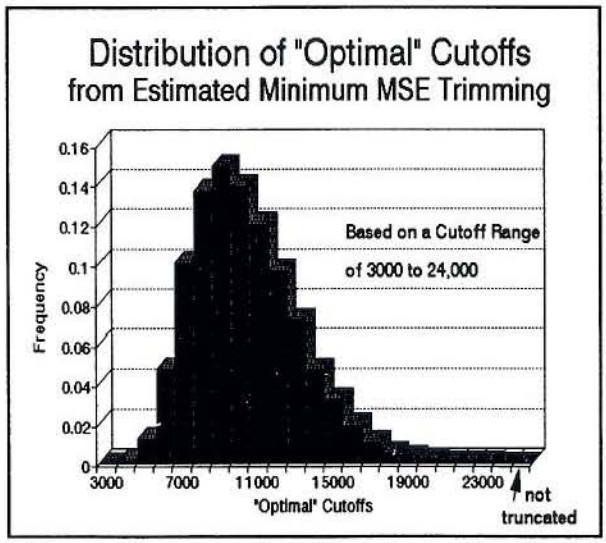


Figure 5 shows the distribution of estimated "optimal" cutoffs in increments of 1000 when this estimator is evaluated over the range 3000 to 24000.

Again, we see that the estimated "optimal" cutoff is data-dependent.

XI. Recommendations

As has been proven theoretically by Ernst (1980), Winsorization at the optimal cutoff is preferable to Winsorization at an order statistic. We believe this estimator holds promise for improving NASS state level hog indications. Further, while the data shows that Winsorization at a cutoff performs well for a wide range of cutoffs, the best efficiencies are obtained for cutoffs greater than or equal to the optimum. However, if we adopt the fixed cutoff estimator, we need to be careful about choosing the cutoff.

Minimum Estimated MSE Trimming provides an alternative to Winsorization at a cutoff when the optimal cutoff is unknown as is frequently the case. However, the efficiency of this data-driven estimator is dependent on how close the range of cutoffs evaluated is to the true optimum. And this technique is computationally intensive.

Minimum Estimated MSE Trimming could be used to suggest optimal cutoffs at the state level, but as Figure 5 shows the estimated cutoff is still highly dependent on the sample data. In the end, determining the optimal cutoff for complex sample designs may remain more art than science.

REFERENCES

Ernst, L. R. (1980), "Comparison of Estimators of the Mean Which Adjust for Large Observations," Sankhyā: The Indian Journal of Statistics, 42, 1-16.

Lee, H. (1991) "Outliers in Survey Sampling," prepared for the Fourteenth Meeting of the Advisory Committee on Statistical Methods.

Potter, F. (1988) "Survey of Procedures to Control Extreme Sampling Weights," Proceedings of the Survey Research Methods Section of the ASA.

Rumburg, S. (1990) "Characteristics of Directly Expanded Hog Data Outliers," NASS Staff Report, No. SRB-92-02, U.S. Department of Agriculture.

Searls, D. T. (1966) "An Estimator for a Population Mean Which Reduces the Effect of Large True Observation," Journal of the American Statistical Association, 61, 1200-1204.

ANALYSIS OF RESPONSE BIAS IN THE JANUARY 1992 CATTLE ON FEED REINTERVIEW PILOT STUDY AND THE JULY 1992 CATTLE ON FEED REINTERVIEW SURVEY

Robert Hood, USDA/NASS
Research Division/3251 Old Lee Hwy., Fairfax, VA 22030

Abstract. To assess the accuracy of reported cattle on feed (COF) inventory, the National Agricultural Statistics Service (NASS) developed a series of reinterview surveys to study response bias and to identify specific reasons for reporting errors in order to improve the survey instruments, training and estimation for COF inventory. A three-phase plan, including a pilot study in January 1992, a semi-operational survey in July 1992 and a fully operational survey in January 1993, was designed to meet these objectives. This paper discusses the results of the January 1992 and July 1992 COF reinterview studies.

For each study, a subsample of respondents reporting for the parent survey was recontacted for face-to-face reinterviews in which a subset of the original questions was re-asked. Differences between the reinterview response and the original parent survey response were reconciled to determine a final "proxy to the true value", which was used to measure response bias.

Although no bias estimates were possible for the January pilot study, useful cognitive information was collected. For July, response bias estimates were generated for several survey items. Although differences were observed between the reinterview responses and the original parent survey responses, the net bias was not significantly different from zero for total COF inventory. The contribution to the bias due to reasons for differences between the responses was also examined to detect any underlying relationships.

I. INTRODUCTION

Over the years, the National Agricultural Statistics Service (NASS) has conducted a variety of reinterview surveys to evaluate the quality of its Agricultural Surveys (AS). The purpose of these reinterview surveys has been to study response bias (as opposed to response variance) and to determine reasons for reporting errors. To assess the accuracy of reported cattle on feed (COF) inventories, a new series of reinterview surveys were developed to study response bias. Specific reasons for reporting errors were obtained to guide efforts to improve the survey instruments, training and estimation for cattle on feed inventory. The main focus of this reinterview program was cattle on feed reporting by smaller farmer-feeder operations, as opposed to larger commercial feedlots. A three-phase plan was designed to implement a reinterview program for COF at NASS. This plan included a one-state pilot study in January 1992, a two-state semi-operational survey in July 1992 and a fully operational five-state survey in January 1993. This paper discusses the setup and results of the first two steps.

In estimating response bias, a "proxy to the true value" must first be obtained. In this study, as in previous reinterview studies at NASS, the reconciled value was considered to be the "true" or final value. Considerable cost and effort was expended to ensure that the value obtained during reconciliation was the best proxy to the true value, as reinterviews were done face-to-face and conducted by supervisory and experienced enumerators. When the original and reinterview responses differed, the enumerators were instructed to determine the "correct" response during the reconciliation process. If there was no difference, i.e. the same response was given during both interviews, this common response was considered the final value. If the respondent could not determine which response was correct, or if a difference was not reconciled by the enumerator, the final value was missing and the observation was not used for that item. If the respondent indicated that either response could be correct, then the average of the two responses was used as the final value. A third response, different from both the original and reinterview responses, was also possible if the reinterview respondent said that neither the original nor the reinterview response was correct.

The formulas used to calculate response bias and variance estimates were based on a stratified random sample design. For the i^{th} observation in stratum h, response bias was measured as: $B_{hi} = O_{hi} - F_{hi}$ for stratum $h = 1,\ldots,L$ and unit $i = 1,\ldots,n_h$ where

O_{hi} = original response

F_{hi} = final or reconciled value.

A negative bias indicates underreporting of a survey item, whereas a positive bias indicates overreporting.

II. REINTERVIEW PROCEDURES

For both January and July 1992, a subsample of respondents reporting for the respective parent Agricultural Survey was recontacted by supervisory and experienced enumerators for face-to-face reinterviews. To get the most accurate data possible, enumerators were instructed to contact the person most knowledgeable about the operation, even if that person was not the same as the parent survey respondent. Reinterviews were to be conducted within ten days of the initial survey in order to minimize recall bias.

Responses to the parent survey were provided to the enumerators in a sealed envelope on a reconciliation form. The reconciliation form contained the questions that appeared on both the parent survey and the reinterview survey; the parent survey responses; and spaces to record the reinterview response, the reconciled "correct" response, and a written explanation in the event that a difference between responses occurred. To maintain the independence between the two responses, the envelopes containing the original parent survey responses were not to be opened until after the reinterview was completed. Having two independent responses and asking the respondent to resolve any discrepancies enabled us to obtain the best possible data.

Immediately after conducting the reinterview, the enumerator would open up the reconciliation form and explain to the respondent that he/she had the information obtained from the initial survey and would like to compare the responses for the few items that appeared on both interviews. Each difference (no matter how small) would then be reconciled to obtain the "correct" response, and a written explanation for why the difference occurred would be recorded on the reconciliation form.

The reinterview questionnaire, used to collect a second independent response for comparison to the original response, was similar to but shorter than the parent survey for both January and July. Reinterview questionnaires for January and July were almost identical. Questions that were common to both the parent survey and reinterview survey included questions pertaining to basic operation description, total cattle on feed inventory and total cattle inventory. Some questions were shortened by dropping "include" and/or "exclude" phrases, while others were reworded in order to ensure that the reinterview/reconciliation process obtained the best "proxy to truth". If a cognitive problem exists with the current operational wording of a particular question, then simply re-asking the question the same way may not uncover an underlying response bias. Since questionnaire wording was to be studied, enumerators were instructed to ask the reinterview questions exactly as worded on the questionnaire. The reinterview questionnaires for both January and July contained additional "cognitive" questions as well as a section on terminology (in which the respondent was asked to give his/her definition of some terms currently being used in our surveys) to be used in evaluating survey definitions and concepts, as well as questionnaire wording. "Probing" questions were also asked to determine if all cattle on feed were being reported and being reported accurately.

II. January 1992 Pilot Study

In January 1992, a reinterview pilot study was conducted in Iowa during the NASS January Agricultural Survey. The objective of this study was to work out the logistics of conducting a reinterview survey for cattle on feed and to field test the reinterview and reconciliation forms. A small non-random subsample of respondents to the January Agricultural Survey who were initially contacted by Computer Assisted Telephone Interviewing (CATI) were selected for face-to-face reinterviews. The subsample was concentrated roughly within a hundred mile radius of the State Statistical Office located in Des Moines. Samples eligible for reinterview were those that reported positive cattle on feed capacity on the initial CATI interview. Of the thirty-two completed reinterviews, twenty-six reported both positive cattle on feed capacity and cattle on feed inventory, while six reported positive capacity but no inventory.

Although no response bias estimates or other statistics were possible for this small non-random sample, the logistics of conducting a reinterview for cattle on feed were worked out and information on the problems of reporting cattle and cattle on feed data were obtained. Some general results from the January pilot study are listed below.

- Cattle were often misclassified. The reference to **heifers** in three of the six breakdowns seemed to confuse the CATI respondent as to which category should be used, often resulting in some animals being counted twice.

- Collecting data by phone can be difficult, especially when a question contains multiple categories, such as the cattle breakdowns consisting of six possible

categories. The respondent cannot see all the possible choices at one time, thus he does not know what his options are and may include animals in one category that should be included in a later category. Several respondents said they would not have had to adjust their numbers as often if they had known all the choices beforehand.

- Placing animals into the correct weight categories was difficult for both CATI and reinterview respondents. There was a lot of guessing as to whether or not cattle were over 500 pounds. Animals less than 500 pounds are considered to be calves by NASS.

- Total cattle inventories were often misreported due to incorrect classification of animals and by the placement of animals into more than one category.

- There was great variability in the definition of a calf among the respondents for this survey. Some respondents used weight as a criterion, while others specified age.

- Reported feedlot capacity for cattle on feed probably indicates the maximum number an operation **could ever** hold, not the maximum number that **would normally** be fed for the slaughter market.

III. July 1992 Reinterview Survey

The July 1992 Cattle on Feed Reinterview Survey was designed as a semi-operational survey to facilitate the transition from a research activity to an operational program in January 1993. The primary objectives were to provide real-time response bias estimates for agency use, to expand the domain of samples eligible for reinterview beyond CATI, and to continue collecting cognitive information to improve both the reinterview and operational survey instruments.

Reinterviews were conducted on a subsample of the July Agricultural Survey (AS) respondents originally contacted by CATI in Iowa and Minnesota. A small subsample of non-CATI respondents were also selected for reinterviews in Iowa. Making non-CATI samples eligible for reinterview was an innovation for reinterview studies at NASS. The non-CATI domain was included because it continues to represent a significant amount of our AS data collection, particularly during the January AS for which the reinterview program is designed. A stratified random sample with stratum sampling rates similar to the parent survey stratum rates was allocated for reinterview. There was a total of 440 samples selected for reinterview, with 220 in each state. Of these, only completed parent survey samples, including those coded out-of-business, were eligible for reinterview. Parent survey refusals and inaccessibles were ineligible for reinterview. Out of the 440 units selected for reinterview, 303 units were eligible for reinterview and 266 had both usable reinterview and parent survey data. The reinterview non-response rate (for the 303 eligible units) was only 9.2%.

For July, response bias estimates for total cattle on feed and total cattle and calves were generated at both the state and the two-state combined levels. Response bias estimates were calculated for original response minus final response and for edited data minus final response. Original and edited data produced similar results with respect to statistical significance for the two states. Response bias estimates for edited minus final values are shown in Table 1. No significant response bias was detected for total cattle or cattle on feed at either level. There was wide variability in the response bias estimates in both magnitude and direction (i.e., positive or negative) between the two states for total cattle on feed. Iowa reporting showed negative biases of 2.8% compared to positive biases of 13.4% for Minnesota. Although no significant response bias was detected, differences between the initial and reinterview surveys did occur. Nearly half (48%) of the responses differed between the two surveys for total cattle and about one quarter (24%) of the responses differed for total cattle on feed. The differences simply tended to cancel each other out.

The precision of the bias estimates was very low, as indicated by the large standard errors, relative to the bias estimates. The small sample size was not the only factor influencing the bias estimates and the significance tests. The actual number of non-zero differences played an important role also. Although there were 266 usable observations overall, the actual number of differences was far less for each item. There were 52 non-zero differences for cattle on feed and 112 for total cattle and calves. These few differences were spread over 10 strata in Iowa and 8 strata in Minnesota. With such a structure, the small number of non-zero differences, the large number of zero differences, and the large expansion factors resulted in extreme variances which resulted in low precision for the response bias estimates. This lack of precision of response bias estimates is a problem that continues to plague us with reinterview surveys. Work continues on sample design and estimation improvements to increase our response bias estimation precision.

Table 1. Response Bias Estimates for the July 1992 COF Reinterview Survey.

Item/State	Edited Value - Final Value: Bias	Edited Value - Final Value: % of Edited	Standard Error	95% CI
TOTAL COF				
Iowa	-25,912	-2.8	3.7	(-10.0, 4.4)
Minnesota	60,117	13.4	10.5	(-7.3, 34.0)
Total	34,205	2.5	4.0	(-5.3, 10.3)
TOTAL CATTLE				
Iowa	-74,411	-1.8	2.8	(-7.4, 3.7)
Minnesota	-48,080	-1.9	2.4	(-6.7, 2.9)
Total	-122,491	-1.9	2.0	(-5.8, 2.0)

IV. REASONS

One of the goals of the July reinterview survey was to identify the reasons for discrepancies between the original and reinterview responses in order to evaluate the questionnaires and to determine how much of the bias may be fixable. During the reconciliation process, explanations were recorded by enumerators for each difference that occurred between an original and reinterview response. These reasons were then grouped into three general categories, "estimation or rounding", "definition or interpretation" and "other" (i.e., reasons that could not be attributed to the first two categories). In general, differences due to "definitional" reasons can be viewed as being potentially fixable by changes in the survey instruments, procedures or training. Differences due to "estimation" or "other" reasons probably are not as correctable, if correctable at all.

Since response biases can be positive or negative and therefore cancel each other out, using the net bias could be misleading when analyzing biases. Therefore, the absolute value of each non-zero difference was expanded to obtain the total absolute response error for each reason category. Table 2 shows the frequency of differences by reason category and the percentage of the total absolute response error attributable to each category. While "estimation" reasons accounted for 38.5% and 20.5% of the differences for COF and total cattle, respectively, these reasons contributed the least to the total absolute response error (8.6% for COF and 4.9% for total cattle). "Definitional" reasons were responsible for the majority of the total absolute response error for COF, accounting for 66.4% of the bias, while "other" reasons, responsible for 60.7% of the bias, contributed the most for total cattle. Table 2 shows that there is opportunity for improvement in the survey procedures, instructions and questionnaires. Recall that reasons due to "definitional" problems are considered fixable. "Definitional" reasons accounted for almost two-thirds of the total absolute response error for COF and over one-third for total cattle.

Table 2. Percentage of Total Absolute Response Error by Reason Category for Original Minus Reconciled Values. Frequencies of Response Errors are Shown in Parenthesis.

Item	Reason Category: Estimation	Definition	Other	Total
Total Cattle on Feed	8.6% (20)	66.4% (19)	25.0% (13)	100% (52)
Total Cattle & Calves	4.9% (23)	34.4% (19)	60.7% (70)	100% (112)

Table 3. Frequency Table of Relative Bias by Reason Category (Two States Combined).[1]

	Reason Category					
	Estimation		Definition		Other	
Item/Relative Bias[2]	# of Obs	% of Bias	# of Obs	% of Bias	# of Obs	% of Bias
Total Cattle on Feed						
Bias ≤ 20%	17	(85%)	3	(16%)	5	(38%)
Bias > 20%	3	(15%)	16	(84%)	8	(62%)
Total	20	(100%)	19	(100%)	13	(100%)
Total Cattle & Calves						
Bias ≤ 20%	21	(91%)	9	(47%)	52	(74%)
Bias > 20%	2	(9%)	10	(53%)	18	(26%)
Total	23	(100%)	19	(100%)	70	(100%)

[1] Includes only observations with a bias
[2] Relative Bias = 100 * (Original value - Reconciled value)/Reconciled value

In order to study the relationship between the magnitude of the bias and the reason categories, a relative (percentage) bias was calculated for each observation with a non-zero difference between the original value and reconciled values. Two levels of relative bias were used - less than or equal to 20% in magnitude and greater than 20% in magnitude. Table 3 shows the relationship between the magnitude of the relative bias and the reason categories. The results indicate that there is a significant relationship between the magnitude of the relative bias and the reason categories. "Estimation" reasons tended to be associated with smaller biases for both items. "Definition" reasons were associated with larger biases for cattle on feed but were more evenly distributed for total cattle. "Other" reasons were associated with larger biases for cattle on feed but with smaller biases for total cattle.

A Closer Look at Total Cattle on Feed

The primary focus of this series of reinterview surveys (i.e., the January 1992, July 1992 and January 1993 surveys) was cattle on feed inventory. The reinterview program for cattle on feed grew out of the concern that inventories were being overreported in the farm feeder states. Thus, the results of the July 1992 reinterview study may have been somewhat surprising. No statistically significant bias at the individual or combined state levels was detected. In fact, the results indicated only a slight overreporting of 2.5% at the combined level. Iowa reporting indicated a slight underreporting of 2.8%. Minnesota overreporting was estimated 13.8%, but the variance was large enough for the result to be insignificant. Do these results then indicate that there is no problem? Not necessarily! What must be remembered when looking at the results from July is the sample size was very small. With such a small sample size (recall that there were only 266 usable samples), the results are very volatile and mistakes on just a few reports can have an enormous impact on the final bias estimate.

Table 2 showed the percent of the total absolute response error accounted for by each of the three reason categories. "Definitional" reasons were the major contributor, accounting for 66% of the total absolute response error. "Other" reasons were responsible for 25% and "estimation" reasons for about 9%. The differences attributable to "definitional" reasons are listed below in Table 4. Also shown are their individual percent contribution to the "definitional" absolute error and the number of times each reason was reported.

For each of the five reported "misunderstandings", the reinterview response was determined to be the correct response during reconciliation. The source of the reporting error for these five samples was attributed to either the initial respondent, the initial enumerator or both. The same person responded for two of the five reports. For the five cases of "did not understand question", the reinterview response was also determined

Table 4. Definitional Reasons Reported for Total Cattle on Feed (Two States Combined).

Reason for Difference	% of Definitional Absolute Response Error	Number of Times Reported
Included cattle/calves from another operation	0.5	1
Did not report as of the reference date	4.9	2
Respondent did not figure death loss in total	7.2	2
Respondent did not understand the question	9.6	5
Respondent forgot to include some cattle or calves	13.7	4
Misunderstanding between enumerator & respondent	64.1	5
Total	100.0	19

to be the correct response. The source of error was attributed to the initial respondent in four cases and to both the initial respondent and the initial enumerator in the other case. The same person responded to four of these five cases.

For cattle on feed inventory, there was a total of 52 non-zero differences between the original and reinterview responses (excluding one outlier); 34 in Iowa and 18 in Minnesota. There was variability in the composition of the differences between and within the two states. Iowa had about four times as many negative differences as Minnesota (21 vs. 5). Minnesota had more positive differences than negative (13 vs. 5), while the opposite was true for Iowa (21 negative vs. 13 positive).

Of the 13 negative differences for Iowa, 4 were due to a "misunderstanding between the enumerator and respondent", accounting for 46% of the total negative bias for cattle on feed in Iowa. Two cases in which the "respondent forgot to include some cattle or calves" accounted for almost 21% of the total negative bias. Eight "estimation" reasons accounted for only 12% of the total negative bias. As for the positive differences, the major contributor was one case in which the "respondent had not made a decision on marketings", accounting for almost half of the total positive bias for Iowa.

Whereas the reason "misunderstanding between enumerator and respondent" accounted for 46% of the total negative bias for Iowa, one difference due to this reason was responsible for 70% of the total positive bias in Minnesota. For Minnesota, the five negative differences contributed very little to the overall bias. In all, there were seven "estimation", eight "definitional" and three "other" reasons for Minnesota. To demonstrate just how volatile the bias estimates were, without the one difference due to a "misunderstanding", the percent bias in Minnesota would have dropped from 13.4% to only 3.7%.

In order to reduce response bias and improve data collection, enumerator training should emphasize the reason why a reinterview is being conducted, why it is important to read the questionnaires exactly as worded and the importance of a positive attitude when conducting a reinterview. With the relatively small sample size, data quality is very important. As was seen in the July 1992 reinterview survey, one observation can completely change both the magnitude and direction of the bias estimates for a survey item, so taking the time to collect good data must be stressed.

CONCLUSION

Although the January and July 1992 reinterview studies did not detect any significant overall response bias for cattle on feed and total cattle inventories, useful information on problems associated with reporting cattle on feed, as well as cattle, was obtained. The two studies showed a substantial number of differences between original and reinterview responses which resulted in great variability. However, the differences were nearly offsetting, resulting in non-significant response bias. The results also indicated that there may be room for improvement in the current survey procedures (including questionnaire design and wording) used to collect COF data. "Definition or interpretation" problems were found to account for nearly two-thirds of the total absolute response error for COF. This can be looked upon as being both good and bad. It is bad in the sense that so much "definitional" bias indicates that there may be a problem with the operational survey. However, it is good in the sense that "definitional" problems are considered more "fixable" than "estimation" or "other" problems. In our efforts to reduce response bias and to improve the survey instruments, high priority ought to be given to reducing the errors attributed to "definitional" reasons.

WINSORIZATION OF SURVEY DATA

Louis-Paul Rivest

Département de mathématiques et de statistique, Université Laval, Ste-Foy, Québec, G1K 7P4, Canada

KEY WORDS: Monte Carlo simulations, outliers, Pareto distribution, skew distributions, Weibull distribution

1. Introduction

Survey practitioners are often faced with the problem of estimating the characteristics of skewed populations. These populations contain sampling units that are markedly different from most others. Special estimation techniques have been devised to attenuate the impact of sampled large data values on the survey estimates. These methods involve decreasing the survey weights of large values (Ernst, 1980; Hidiroglou and Srinath, 1981) and using estimation methods that are robust to outlying values (Chambers, 1986; Gwet and Rivest, 1992). Winsorization (Searls, 1966; Fuller, 1991) is a simple method to deal with extreme units when the survey variable only takes positive values. It consists in replacing the largest data values by a predetermined cut-off value. This paper reviews various winsorization schemes for estimating the population means of positive variables using simple and stratified random samples.

There are many probability distributions to model positive skewed data. Following Fuller (1991), we consider a distribution to be skewed if its right tail is heavier than that of the exponential distribution, $F(x)=1-e^{-x}$ for $x>0$. The Weibull distribution, defined as $F_\alpha(x)=1-\exp(-x^{1/\alpha})$, satisfies this condition provided that α is bigger than 1. If X has the exponential distribution then X^α is distributed according to $F_\alpha(x)$. The Pareto distribution and the lognormal distribution defined as $F_\alpha(x)=1-1/(1+x)^\gamma$ and $F_\beta(x)=\Phi\{(\log x)/\beta\}$, for $x>0$, are also skewed probability distributions according to this criterion.

For highly skewed distributions, the distribution of the sample mean retains somes of the skewness of the underlying distribution. Thus, in repeated sampling, the sample mean is in most cases reasonnably close to the population mean. However it may happen that the sample mean is substantially larger than the population mean. Under these circumstances, special techniques are needed to lower the impact of the largest values in order to bring the sample mean closer to the population mean.

2. Winsorization in simple random samples.

Let $x_1<x_2...<x_n$ denote the ordered x-values in a simple random sample of size n drawn from a population U of size N; f represents the sampling fraction. A winsorized mean $\bar{x}_R$ is defined as

$$\bar{x}_R = \frac{1}{n}\sum_{i=1}^{n}\min(x_i,R) \qquad (2.1)$$

where R is the cut-off value. Several methods have been proposed to choose R. R can be chosen to be equal to a convex combination between adjacent extreme order statistics; this is called nonparametric winsorization. Another choice, suggested by Searls (1966), is to take the value of R that minimizes the mean square error of the winsorized mean. This strategy is called optimal winsorization. These two methods for chosing R are presented in this section. Fuller's (1991) preliminary test preliminary test procedure for estimating the mean of a skewed population will also be presented. These methods will then be compared by Monte Carlo simulations in Section 3.

2.1 Searls' optimal winsorization.

Let F represent the population distribution of the X_i's,

$$F(x) = \frac{1}{N}\sum_{i=1}^{N}1\{X_i\le x\}$$

The bias of $\bar{x}_R$ as an estimator of $\bar{X}$, is given by $-B(\bar{x}_R)$, where

$$B(\bar{x}_R) = \int_R^\infty (1-F(x))dx.$$

The mean square error of estimator $\bar{x}_R$, $MSE(\bar{x}_R)$, can be expressed in terms of F as (Searls,1966; Rivest and Hurtubise, 1993)

$$MSE(\bar{x}_R) = \frac{1-f}{n}\left(S_x^2 - 2\int_R^\infty (x-\bar{X})(1-F(x))dx - B^2(\bar{x}_R) + B^2(\bar{x}_R)\right) \qquad (2.2)$$

where $\bar{X}$ and S_x^2 are the population mean and the population variance of X.

Differentiating with respect to R yields the following equation for the value of R minimizing (2.2),

$$\frac{R-\bar{X}}{n/(1-f)-1} = B(\bar{x}_R) = \frac{1}{N}\sum_{i=1}^{N}\max(X_i-R,0) \qquad (2.3)$$

A simple algorithm to solve this equation (see Rivest and Hurtubise, 1993 for a detailed investigation) is to update the current value of the cut-off value R^t with R^{t+1},

$$R^{t+1} = R^t - \frac{\frac{R^t-\bar{X}}{n/(1-f)-1} - \frac{1}{N}\sum_{i=1}^{N}\max(X_i-R^t,0)}{\frac{1-f}{n-1+f}+\frac{1-F(R^t)}{N}}. \qquad (2.4)$$

Setting the starting value R^0 to $\bar{x}$, this algorithm will converge rapidly to the value of R minimizing (2.2). Let R_{opt} denote the value of R for which (2.2) is minimum and $\bar{x}_{opt}$ the corresponding winsorized mean. From (2.3), the bias of $\bar{x}_{opt}$ is equal to $(R_{opt}-\bar{X})/(n/(1-f)-1)$. Note that a linear tranformation in the data, $Y_i=aX_i+b$, produces the same linear transformation in R_{opt} and in $\bar{x}_{opt}$. This means that $R_{opt}(Y)=aR_{opt}(X)+b$ and $\bar{y}_{opt}=a\bar{x}_{opt}+b$.

Rivest and Hurtubise (1993) established an intriguing property of R_{opt}: the expected number of winsorized observations, $m_n(F)=n(1-F(R_{opt}))$, increases as the skewness of the distribution decreases. This is illustrated in Table 1 for the Weibull family. For instance, in samples of size 200 drawn from the exponential distribution, the optimal scheme winsorizes an average of 3.16 observations; this leads a meager gain in efficiency of less than 10%. On the other hand, for a Weibull with a CV of 4, winsorizing on average 1.2 observations is optimal in samples of size 200. Winsorization can then bring a substantial gain in efficiency.

Supposing that k samples, $\{x_{1i}\}$, $\{x_{2i}\},\ldots,\{x_{ki}\}$ were drawn from population U and that they were available to estimate the optimal winsorization parameter R. One can standardize the x-values to accomodate for a possible change, over time, in the distribution of X. For sample j, substracting the median $\hat{m}_j$ and divide by the interquartile range $\hat{iq}_j$, defined as the third quartile minus the first one. The standardized samples $\{(x_{ji}-\hat{m}_j)/\hat{iq}_j\}$, for j=1,...,k, can then be pooled together and an estimate $\hat{R}$ of the optimal winsorization parameter for a standardized sample of size n can be calculated from the pooled data using algorithm (2.3). The optimal winsorization constant for the current sample is then given by $\hat{R}_{opt}=\hat{m}+\hat{iq}\,\hat{R}$ where $\hat{m}$ and $\hat{iq}$ are the median and the interquartile range of the current sample. For most skewed distributions $m_n(F)$ is approximately equal to 1. Therefore an alternative strategy for choosing the winsorization parameter is to set $R=F^{-1}(1-1/n)$ and to estimate R using the $M(1-1/n)^{th}$ order statistics of the pooled sample where M is the size of the pooled sample. The corresponding estimator for $\bar{X}$ is considered in the simulations given in Section 3 under the label $\bar{x}_{1/n}$.

One can use Searls' winsorization technique to construct an estimator for $\bar{X}$ even if no auxiliary information is available to estimate R_{opt}. The winsorization parameter is estimated by the value of R that minimizes an estimator of (2.2). Replacing, in (2.2), the population characteristics F(x), S_x^2, and $\bar{X}$ by their sample values s^2, $\bar{x}$, and $\hat{F}(x)$, where

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n}1\{x_i\le x\},$$

yields a simple estimator of $MSE(\bar{x}_R)$. Substituting in (2.4) the population characteristics by their sample counterparts provides an algorithm for calculating the optimal value of R. Let $\bar{x}_e$ denote the corresponding winsorized estimator. An appealing property of this approach is that it generalizes easily to complex sample designs. For instance minimizing an estimator of the mean square error was investigated by Potter (1990) as a method to control large sampling weights .

2.2 Nonparametric winsorization

This section assumes that the population under study is infinite. F(x) represents the cumulative distribution function of X. Let x_{n-2}, x_{n-1}, and x_n are the three largest data values in the sample. The cut-off values R_λ under study are equal to $(1-\lambda)x_n+\lambda x_{n-1}$, for λ in (0,1], and to $(2-\lambda)x_{n-1}+(\lambda-1)x_{n-2}$ for λ in (1,2]. Let $\bar{x}_\lambda$ denote the winsorized mean obtained with cut-off value R_λ. Estimators $\bar{x}_\lambda$ are investigated in Rivest (1993). The major result in this paper is that winsorizing one observation or less, i.e. taking $\lambda=1$, is nearly optimal for all distributions F having a finite variance.

Nonparametric winsorization provides interesting reductions in mean square error at the cost of some bias. One drawback of this approach is that nonparametric winsorization lowers the estimate of the population mean for all samples even if most samples do not contain outliers. For outlier free samples nonparametric winsorization increases the error of estimation. The winsorization schemes proposed in Section 2.1, based on auxiliary information, and the ones considered in the next section do not have these deficiencies. They leave some of the samples unchanged.

Rivest (1993) shows that, for the purpose of estimating the mean square error of $\bar{x}_1$, treating the

winsorized sample $\{x_1,x_2,...,x_{n-1},x_{n-1}\}$ as if it were a random sample yields an estimator with a severe negative bias. The relative biases ranged between -40% and -60% for several skewed distributions. A mean square error estimator having a small bias is given by:

$$\frac{s^2}{n} - \frac{1}{n^2}(x_n+x_{n-1}-2\bar{x}_1)(x_n-3x_{n-1}+2x_{n-2})$$

where s^2 is the x-sample variance including all n units.

2.3 Fuller's preliminary test estimator.

Fuller's estimator winsorizes the sample only if its upper tail is heavier than what would be expected from an exponential sample. The key to the construction of $\bar{x}_{pt}$, Fuller's preliminary test estimator, is the following formula for the sample mean,

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{T-j} x_i + jx_{T-j} + \sum_{i=1}^{j} i(x_{n+1-i}-x_{n-i})\right) \quad (2.5)$$

where j is a positive integer. In other words, $\bar{x}$ is the j^{th} winsorized mean plus a sum of normalized spacings involving extreme order statistics. Estimator $\bar{x}_{pt}$ winsorizes this sum only when it is large. If $\{x_i\}$ is an ordered exponential sample, then the normalized spacings defined, for i=1,...,n-1, as $i(x_{n-i+1}-x_{n-i})$ are independent and distributed according to the exponential distribution. Thus, for any T>j, the ratio F_{Tj} defined as

$$F_{Tj} = \frac{\frac{1}{j}\sum_{i=1}^{j} i(x_{n+1-i}-x_{n-i})}{\frac{1}{T-j}\sum_{i=j+1}^{T} i(x_{n+1-i}-x_{n-i})}$$

should be close to 1 if the sample is exponential. When the upper tail of the underlying distribution is heavier than an exponential upper tail, F_{Tj} can be quite large (see Fuller, 1991). If, for some constant K_j, F_{Tj} is smaller than K_j, one will consider that the right tail of the sample is not heavy. $\bar{x}$ is then appropriate as an estimator of the population mean. If F_{Tj} is larger than K_j, then the sum of the normalized spacings involving the largest observations appearing in (2.5) is winsorized. The winsorized estimator is obtained by replacing this sum by $jK_j\,\bar{d}_{Tj}$ where $\bar{d}_{Tj}$ is the numerator of F_{Tj},

$$\bar{d}_{Tj} = \frac{1}{T-j}\sum_{i=j+1}^{T} i(x_{n+1-i}-x_{n-i}).$$

Thus Fuller's estimator is equal to $\bar{x}$ if $F_{Tj}<K_j$, and to

$$\frac{1}{n}\left(\sum_{i=1}^{T-j} x_i + j(x_{T-j}+K_j\,\bar{d}_{Tj})\right)$$

otherwise. Fuller (1991) discusses the choice of j, T, and K_j. For the simulations presented in Section 3, j was chosen equal to 3, T equal to $[4n^{1/2}-10]$ and K_3 equal to 3.5.

3. A Monte Carlo experiment

The estimators in Section 2 were compared by using exact and Monte Carlo calculations. The estimators under study are presented in Table 2. For estimator $\bar{x}_{1/n}$ the cut-off value R was estimated by the second largest order statistic from an auxiliary sample of size 2n. Each Monte Carlo sample had its own auxiliary sample. Samples of size 20, 40, 60, 100, and 200 were considered in the study. The populations under study were the Weibull and the Pareto distributions with a CV of 4 as well as Acre introduced by Fuller (1991), whose CV is equal to 5. For Acre the sample design was with replacement random sampling. The relative biases and the efficiencies with respect to the sample mean are presented for the 5 estimators under study in Tables 3, 4 and 5. In all cases R_{opt} was evaluated using the algorithm (2.4) and the efficiencies were calculated exactly using Splus. Thus the biases and the efficiencies of $\bar{x}_{opt}$ are exact. The biases and the efficiency of $\bar{x}_1$ in Pareto samples were also calculated exactly (see Rivest, 1993).

For the Weibull and the Pareto distributions 100,000 Monte Carlo samples were generated for each sample size. For the Pareto distribution, efficiencies with respect to $\bar{x}_1$ were calculated by simulation. To get the efficiencies reported in Table 4, the simulated efficiencies were multiplied by the exact efficiency of $\bar{x}_1$ with respect to $\bar{x}$. Acre results are based on 200,000 Monte Carlo replicates.

For the three distributions, the estimators have similar rankings: $\bar{x}_{opt} > \bar{x}_{1/n} > \bar{x}_{pt} > \bar{x}_e$ and $\bar{x}_1$. $\bar{x}_{opt}$ has been included in the simulations as a standard. The expected number of winsorized observations $m_n(F)$ under the optimal scheme are given in Table 1 for the Weibull distribution. For the Pareto and for Acre, the corresponding expectations are (.72,.81,.86,.91,.96) and (.32,.38,.42,.50,.50) respectively. The good performance of $\bar{x}_{1/n}$ shows that it is possibile to do almost as well as $\bar{x}_{opt}$ with limited auxiliary information. For the Pareto and Acre distribution, the optimal scheme winsorizes on average less than one observation. Therefore the relatively large bias of $\bar{x}_{1/n}$ can be reduced and its efficiency

increased by using a larger auxiliary sample. Additional simulations not reported here showed that taking the second largest observation in an auxiliary sample of size 3n as an estimator of the winsorization parameter improves on $\bar{x}_{1/n}$ in Tables 4 and 5.

Estimator $\bar{x}_{pt}$ is the best among the estimators which do not use auxiliary information. Its relatively weak performance at n=20 can be attributed to the value of T, T=7, which is low. A value of T=10 would have yielded better results for samples of size 20. Estimator $\bar{x}_{pt}$ is substantially better than $\bar{x}_1$ for the Pareto and the Acre distributions, when the skewness is large. When the skewness is moderate, $\bar{x}_{pt}$ is comparable to $\bar{x}_e$ and $\bar{x}_1$. One explanation for the superiority of $\bar{x}_{pt}$ is that it leaves the sample mean unchanged when there are no outliers in the sample. The performance of $\bar{x}_e$ is poor. It is worst than the simple $\bar{x}_1$ for the Weibull and the Pareto distributions. Many reasons can be put forward to explain this phenomenon. First among all estimators in the study, $\bar{x}_e$ is the only one that is not robust to outliers. If one lets the largest sample value x_n go to infinity, then all the estimators of Table 2 remain bounded except for $\bar{x}_e$ which goes to infinity. This might explain the poor showing of $\bar{x}_e$ and the erratic behavior of its efficiency in Pareto samples. These samples sometimes contain wild values that have a large impact on $\bar{x}_e$. Also, the intriguing property of Searl's winsorization scheme noted in Section 2.1 implies that $\bar{x}_e$ winsorizes more in normal samples than in skewed sample producing unnecessary bias.

The biases of winsorized estimators are large. In repeated surveys a systematic bias of more than 5% on individual population estimates is not acceptable. In this case, minimizing the mean square error of one estimate, $MSE(\bar{x}_R)$ might not be a good criterion. One should possibly attempt to minimize the mean square error of a sum a successive estimates, $MSE(\bar{x}_{R1}+\bar{x}_{R2}+...+\bar{x}_{Rk})$. When the $\bar{x}_{Rj}$'s are independent, i.e. the successive samples are not overlapping, this can be done relatively easily. This amounts to finding the R that minimizes $V(\bar{x}_R)+kB^2(\bar{x}_R)$ where $B(\bar{x}_R)$ is the bias. Reasoning as in Section 2.1, it is easily shown that the optimal R is the solution of the following equation,

$$\frac{R-\bar{X}}{nk/(1-f)-1} = B(\bar{x}_R). \qquad (3.1)$$

Thus, neglecting the sampling fraction, the R that minimizes $V(\bar{x}_R)+kB^2(\bar{x}_R)$, in a sample of size n, is the value that minimizes $MSE(\bar{x}_R)$ in a sample of size kn. Section 2.1 suggests that for many distributions $F^{-1}(1-1/(kn))$ would be a good approximation to the winsorization constant. To apply this estimation scheme successfully, one needs approximations for the extreme quantiles of F. Large auxiliary samples are needed.

Optimal winsorization for the sum of k=3 successive estimates was investigated by Monte Carlo simulations for Acre. All the estimators of Table 2 were modified in order to reduce their biases. For nonparametric winsorization $R_{1/3}=2x_n/3+x_{n-1}/3$ was selected as cut-off value. The preliminary test estimator of Table 7 uses j=1 and $K_1=5.8$. For $\bar{x}_e(k=3)$ and $\bar{x}_{opt}(k=3)$ the winsorization parameter was estimated by solving equation (3.1) with the population characteristics and their sample values respectively. The winsorization parameter of $\bar{x}_{1/(3n)}$ was set equal to the second largest value of an auxiliary sample of size 6n.

In Table 6, the efficiency of $\bar{x}_{1/(3n)}$ is large, much larger than that of $\bar{x}_{pt}(j=1)$ for similar biases. However none of the good estimators of Table 7 has a small bias. It is worth noting that very little winsorization produces large biases. This might be caused by the extreme skewness of population Acre.

4. Winsorization in stratifed samples

Let L denote the number of strata, F_h, for h=1,...,L the distribution of X in stratum h, and N_h the size of stratum h. In this section we consider a winsorization scheme were each stratum has its own winsorization parameter R_h, h=1,...,L. The winsorized estimator of $\bar{X}$ is given by $\bar{x}_R=\sum W_h \bar{x}_{Rh}$ where $W_h=N_h/\Sigma N_h$ and $\bar{x}_{Rh}=\sum \min(x_{hi},R_h)/n_h$. One is looking for the values of R_h that minimize $MSE(\bar{x}_R)$. Neglecting the sampling fractions, one has

$$MSE(\bar{x}_R)=\sum_{h=1}^{L}\frac{W_h^2}{n_h}\left(S_h^2-2\int_{R_h}^{\infty}(x-\bar{X}_h)(1-F_h(x))dx-B^2(\bar{x}_{Rh})\right)+\left(\sum_{h=1}^{L}W_hB(\bar{x}_{Rh})\right)^2$$

where F_h represents the distribution of X in stratum h and $B(\bar{x}_{Rh})$ is the negative bias of $\bar{x}_{Rh}$ as an estimator of $\bar{X}_h$

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

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

$$\frac{W_h}{n_h}\{R_h - \bar{X}_h + B(\bar{X}_{Rh})\} = \sum_{h=1}^{l} W_h B(\bar{X}_{Rh}) \quad (4.1)$$

for h=1,...,L.

These equations have explicit solutions in one interesting case. Suppose that the distributions of X within the strata are equal up to a change in location and scale, i. e. $F_h = F((x - \bar{X}_h)/S_h)$ for some distribution F, and that the sample sizes are determined according to Neymann optimal allocation, that is $n_h \sim W_h S_h$. In this case, the optimal solution is easily seen to be equal to $R_h = S_h R_{opt} + \bar{X}_h$ where R_{opt} is the optimal winsorization parameter for a simple random sample of size $n = \Sigma n_h$ drawn from F. R_{opt} can be calculated by solving equation (2.3). The expectation of the total number of winsorized observations is then $m_n(F) = n(1-F(R_{opt}))$. Values for $m_n(F)$ are reported in Table 1 for Weibull distributions.

If the assumptions that the distributions F_h's are equal, up to a change in location and scale, is tenable, then a generalization to stratified designs of estimator $\bar{X}_{1/n}$ is easily constructed. In each auxiliary stratified sample, the observations in each stratum are standardized, using order statistics as proposed in Section 2.1. The strata of all auxiliary samples are then pooled together and the second largest observation of the pooled sample, say S, is noted and put aside. Winsorization parameters for the current stratified sample can be calculated as $\hat{m}_h + \hat{iq}_h S$, for h=1,...,L, where $\hat{m}_h$ and $\hat{iq}_h$ are the median and the inter-quartile range in stratum h. Further investigations are needed to evaluate the performance of this winsorization scheme for stratified designs.

5. Conclusions

Simple winsorized estimators can yield substantial gains in efficiency when sampling a skewed distribution. If the survey is repeated over time, an appealing procedure for estimating the winsorization parameter is to use the data from past surveys. Section 3 shows that estimating R using the second largest data values from the combined samples of the last two or three surveys yields large gains in efficiency. The size of the auxiliary sample for estimating R depends on the skewness of the distribution, distributions with heavy tails demand large auxiliary samples, and on the bias that one can tolerate. This paper has focussed on simple and stratified random sampling; extensions to unequal probabilities of selection sampling designs with several levels of sampling need further investigation.

Acknowledgements

I am grateful to Daniel Hurtubise for some of the computer programs that were used in the Monte Carlo simulations and to Mike Hidiroglou for his careful reading of the manuscript. This research was supported by an operating grant from N.S.E.R.C..

Bibliography

Chambers R. L. (1986), "Outlier robust finite population estimation," *Journal of the American Statistical Association,* **81**, 1063-1069.

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

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

Gwet J.P., and L.P Rivest (1992) "Outlier resistant alternatives to the ratio estimator," *Journal of the American Statistical Association,* **87**, 1174-1182.

Hidiroglou M. A. and K.P. Srinath (1981) "Some estimators of a population total from a simple random sample containing large units," *Journal of the American Statistical Association,* **76**, 690-695.

Potter F. J. (1990) "A study of procedure to identify and trim extreme sampling weights," *Proceedings of the Section on Survey Research Methods,* ASA, 225-230.

Rivest L.P. (1993) "Statistical properties of winsorized means for skewed distributions," *Biometrika.* To appear

Rivest L.P. and D. Hurtubise (1993) "On Searls' winsorization method for estimating the mean of a skewed population," Submitted for publication

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

n	CV=1 α =1			CV=2 α =1.84			CV=4 α =2.87		
	$m_n(F)$	Eff.	bias	$m_n(F)$	Eff.	bias	$m_n(F)$	Eff.	bias
20	1.60	1.39	-8	1.02	2.02	-16	0.64	3.87	-29
40	2.03	1.25	-5	1.28	1.68	-12	0.80	2.91	-22
60	2.30	1.19	-4	1.43	1.54	-9	0.89	2.52	-18
100	2.65	1.14	-3	1.64	1.40	-7	1.02	2.15	-14
200	3.16	1.08	-2	1.93	1.27	-4	1.20	1.79	-10

Table 1. Expected numbers of winsorized observations ($m_n(F)$), relative biases(in percentage) and efficiencies of the optimal winsorized mean for three Weibull distributions and 5 sample sizes.

Estimator	Description
$\bar{x}_1$	Once winsorized mean defined in Section 2.2
$\bar{x}_{pt}$	Preliminary test estimator of Section 2.3
$\bar{x}_e$	Optimal winsorized mean with R estimated from the sample
$\bar{x}_{1/n}$	Winsorized mean with $R=F^{-1}(1-1/n)$ estimated with auxiliary samples.
$\bar{x}_{opt}$	Optimal winsorized mean with R_{opt} known

Table 2. Estimators in the Monte Carlo investigation.

n	$\bar{x}_1$		$\bar{x}_{pt}$		$\bar{x}_e$		$\bar{x}_{1/n}$		$\bar{x}_{opt}$	
20	2.02	-34	2.01	-26	1.86	-25	3.09	-36	3.87	-29
40	1.65	-23	1.76	-19	1.63	-20	2.46	-24	2.91	-22
60	1.50	-18	1.62	-15	1.50	-17	2.19	-19	2.52	-18
100	1.36	-13	1.47	-11	1.38	-14	1.92	-14	2.15	-14
200	1.23	-8	1.32	-7	1.24	-10	1.64	-8	1.79	-10

Table 3. Efficiencies, with respect to the sample mean, and relative biases (in percentages) of 5 estimators for samples drawn from the Weibull distribution with a coefficient of variation of 4 (α =2.87).

n	$\bar{x}_1$		$\bar{x}_{pt}$		$\bar{x}_e$		$\bar{x}_{1/n}$		$\bar{x}_{opt}$	
20	5.92	-18	5.66	-9	2.84	-15	8.62	-19	9.80	-16
40	5.04	-13	5.42	-7	3.40	-11	7.00	-13	7.78	-12
60	4.62	-10	5.09	-6	3.42	-10	6.35	-11	6.90	-10
100	4.19	-8	4.70	-5	2.66	-8	5.55	-8	6.01	-8
200	3.73	-5	4.23	-4	2.65	-6	4.80	-6	5.10	-6

Table 4. Efficiencies, with respect to the sample mean, and relative biases (in percentages) of 5 estimators for samples drawn from the Pareto distribution with a coefficient of variation of 4 (γ =2.1333).

n	$\bar{x}_1$		$\bar{x}_{pt}$		$\bar{x}_e$		$\bar{x}_{1/n}$		$\bar{x}_{opt}$	
20	4.87	-34	6.66	-28	2.64	-21	7.30	-36	9.21	-28
40	3.19	-28	5.00	-26	2.40	-19	4.97	-30	6.26	-24
60	2.51	-20	3.95	-25	2.20	-17	4.03	-26	5.05	-21
100	1.90	-22	2.93	-23	1.95	-15	3.07	-22	3.90	-18
200	1.37	-14	1.99	-19	1.64	-12	2.20	-16	2.83	-14

Table 5. Efficiencies, with respect to the sample mean, and relative biases (in percentages) of 5 estimators for samples drawn from population ACRE .

n	$\bar{x}_{1/3}$		$\bar{x}_{pt(j=1)}$		$\bar{x}_{e(k=3)}$		$\bar{x}_{1/(3n)}$		$\bar{x}_{opt(k=3)}$	
20	1.83	-11	3.91	-21	1.58	-11	7.40	-26	7.71	-21
40	1.67	-9	2.96	-19	1.53	-9	4.70	-20	5.23	-17
60	1.55	-8	2.35	-17	1.48	-8	3.57	-17	4.23	-15
100	1.42	-7	1.85	-14	1.42	-7	2.60	-13	3.30	-12
200	1.24	-5	1.34	-11	1.31	-6	1.69	-7	2.37	-9

Table 6. Efficiencies, with respect to the sample mean, and relative biases (in percentages) of 5 estimators for samples drawn from population ACRE .

GENERALIZED REGRESSION ESTIMATION FOR A TWO-PHASE SAMPLE OF TAX RECORDS

John Armstrong and Hélène St-Jean, Statistics Canada
John Armstrong, Statistics Canada, 11-N R.H. Coats Bldg., Tunney's Pasture, Ottawa, Canada, K1A 0T6

KEY WORDS: Calibration, domain

1. Introduction

The two-phase tax sample is part of a general strategy for production of annual estimates of Canadian economic activity at Statistics Canada. Annual economic data for large businesses are collected through mail-out sample surveys. Data for small businesses are obtained from the tax sample. Tax data rather than survey data are used to obtain small business estimates in order to reduce costs and response burden.

Administrative files containing information on business taxfilers are provided to Statistics Canada by Revenue Canada, the Canadian government department responsible for tax collection. There are two reasons why sampling of tax records is used rather than simple tabulation from these administrative files. First, although taxfilers are classified by Revenue Canada using the Standard Industrial Classification (SIC) code system (Statistics Canada 1980), only the first two digits of SIC (SIC2) can be determined with sufficient accuracy using business activity information reported on tax returns. Estimates are required for domains defined using all four digits of SIC. The cost of improving the accuracy of four-digit SIC (SIC4) codes for all tax records would be substantial. Second, estimates are required for many variables that are not available in machine-readable form and must be obtained from source documents. The cost of transcription of this information for all records would be prohibitive.

A two-phase approach to sampling of tax records was adopted to provide better control of sample sizes in SIC4 domains. The estimation methodology that has been used in production since 1989 involves use of population counts but does not employ all available auxiliary information. The work on the use of additional auxiliary information reported here was motivated by the potential to reduce sample sizes required to obtain specified levels of precision. The framework of generalized regression estimation facilitates extensions of the current estimator to employ additional auxiliary data.

The two-phase sample design is briefly described in Section 2. Section 3 includes a description of the estimation methods currently used in production. A derivation of the generalized regression estimator is presented in Section 4. Section 5 includes the results of an empirical study.

2. Sample Design

The target (in-scope) population for tax sampling is the population of businesses with gross income over $25,000, excluding large businesses covered by mail-out sample surveys. There are two types of taxfilers - T1s and T2s. A T1 taxfiler is an individual, who may own all or part of one or more unincorporated businesses, while a T2 taxfiler is an incorporated business. Information concerning numbers of businesses owned by T1 taxfilers and ownership shares is not available from Revenue Canada. Geographical information, as well as gross business income and net profit are provided for both T1 and T2 taxfilers.

Estimates are required for about 35 financial variables that are not provided as frame data. Data for these variables are captured from copies of tax returns for taxfilers in the second-phase sample.

Information about the population of taxfilers for a particular tax year is provided by Revenue Canada over a period of two calendar years as tax returns are received and processed. If sample selection for a particular tax year did not begin until a complete frame was available, data capture operations would lead to considerable additional delays before estimates could be produced. Bernoulli sampling is used to select both first- and second-phase samples to reduce delays and provide a relatively uniform workload to operations staff.

The first-phase sample is a sample of taxfilers selected from a frame created using Revenue Canada information. Strata are defined by SIC2, province and size (gross business income). The first-phase sample is a longitudinal sample. All taxfilers that are included in the first-phase sample in TY(T) (tax year T) and are still in-scope for tax sampling in TY(T+1) are included in the first-phase sample for TY(T+1). Taxfilers are also added to the first-phase sample annually to replace taxfilers sampled in previous years that are no longer in-scope.

Let $I=\{i\}$ denote the population of taxfilers. Similarly, let $J=\{j\}$ denote the population of businesses that is the target population for tax sampling. Each T1 tax return includes income and

expense data for each business wholly or partially owned by the taxfiler, as well as ownership share information. A statistical entity, denoted by *(i,j)*, is created for every taxfiler-business combination in the first-phase sample. (The correspondence between businesses and T2 taxfilers is assumed to be one to one.) Statistical entities are assigned SIC4 codes by Statistics Canada. These codes are determined using business activity descriptions reported on tax returns as well as supplementary data sources and are more accurate in digits three and four than codes assigned by Revenue Canada.

Conceptually, the second-phase sample is a sample of businesses. Operationally, it is a sample of taxfilers selected using statistical entities. Statistical entities are stratified using SIC4 codes assigned by Statistics Canada, as well as province and size. The total revenue of business j is used as the size variable for statistical entity *(i,j)*. If one statistical entity corresponding to a T1 taxfiler is selected for the second-phase sample, then all statistical entities corresponding to the taxfiler are selected. Consequently, the second-phase selection probability for statistical entity *(i,j)* depends only on i. For more details concerning the sample design, refer to Armstrong, Block and Srinath (1993).

3. Estimation

3.1 Horvitz-Thompson Estimator

If business j is a partnership, it will be included in the second-phase sample if any of the corresponding taxfilers are selected. The Horvitz-Thompson estimator must be adjusted for partnerships. Let δ_{ij} denote the proportion of business j owned by taxfiler i and suppose that statistical entity *(i,j)* is selected for the second-phase sample. The data for business j is adjusted by multiplying it by δ_{ij} so that only the component of income and expense items corresponding to taxfiler i is included in estimates. Rao (1968a) describes a similar adjustment in a slightly different context.

Let y_j denote the value of the variable y for business j. The Horvitz-Thompson estimate of the total of y over domain d, incorporating adjustment for partnerships, is given by

$$\hat{Y}_{H\text{-}T}(d)=\sum_{i\in s2}\sum_{j\in J_i}\delta_{ij}\, y_j(d)/(p_{1i}\, p_{2i})$$

where $s2$ is the second-phase sample, J_i is a set containing the indices of the businesses wholly or partially owned by taxfiler i, p_{1i} is the first-phase selection probability for taxfiler i, p_{2i} is the second-phase selection probability for statistical entity *(i,j)* and $y_j(d)=y_j$ if business j falls in domain d and is otherwise zero.

Noting that selection probabilities depend only on the taxfiler index i, $\hat{Y}_{H\text{-}T}(d)$ can be written as

$$\hat{Y}_{H\text{-}T}(d)=\sum_{i\in s2} y_i(d)/(p_{1i}\, p_{2i})$$

where

$$y_i(d)=\sum_{j\in J_i}\delta_{ij}y_j(d)\ .$$

The variance of $\hat{Y}_{H\text{-}T}(d)$ is

$$V(\hat{Y}_{H\text{-}T})=\sum_i [(1-p_{1i}\, p_{2i})/(p_{1i}\, p_{2i})]y_i(d)^2$$

and this variance is estimated by

$$\hat{V}(\hat{Y}_{H\text{-}T}(d)) = \sum_{i\in s2}\frac{(1-p_{1i}\, p_{2i})}{(p_{1i}\, p_{2i})^2}y_i(d)^2\ .$$

3.2 Poststratified Horvitz-Thompson Estimator

Sunter (1986) shows that the estimator analogous to $\hat{Y}_{H\text{-}T}(d)$ has a large variance for a one-phase design using Bernoulli sampling. He considers a ratio form of the estimator, adjusted for differences between actual and expected sample sizes as suggested by Brewer, Early and Joyce (1972). He notes that the ratio form has a small bias and a variance that is considerably smaller than the unadjusted version. The methodology used to produce tax estimates incorporates ratio adjustments to account for differences between actual and expected sample sizes.

Ratio adjustments are applied within poststrata during weighting of both the first- and second-phase samples. Choudhry, Lavallée and Hidiroglou (1989) provide a general discussion of weighting using a poststratified ratio adjustment. Following their notation, let $U=\{u\}$ denote a set of first-phase poststrata and suppose that poststratum u contains N_u taxfilers. An estimate of the number of taxfilers in the population that fall in first-phase poststratum u, based on the first-phase sample, is

$$\check{N}_u=\sum_{i\in s1\cap u}(1/p_{1i})\ .$$

The poststratified first-phase weight for taxfiler i, $i\in u$ is

$$W_{1i}=(1/p_{1i})(N_u/\check{N}_u)\ .$$

Similarly, let $V=\{v\}$ define a set of second-phase poststrata. An estimate of the number of taxfilers in second-phase poststratum v, based on the first-phase sample, is

$$\tilde{N}_v=\sum_{i\in s1\cap v}W_{1i}\ .$$

An alternative estimate, using only units in the second-phase sample, is

$$\dot{N}_v = \sum_{i \in s2 \cap v} W_{1i}/p_{2i} \ .$$

The poststratified second-phase weight for a statistical entity associated with taxfiler i is

$$W_{2i} = (1/p_{2i})(\tilde{N}_v/\dot{N}_v)$$

and the final weight is

$$W_i = W_{1i} W_{2i} \ .$$

The poststratified estimate of the total of y over domain d is given by

$$\hat{Y}(d) = \sum_{i \in s2} \sum_{j \in J_i} \delta_{ij} W_i y_j(d) \ .$$

Choudhry, Lavallée and Hidiroglou (1989) provide an expression for the approximate variance of $\hat{Y}(d)$ and propose the estimator

$$\hat{V}(\hat{Y}(d)) = \sum_u \sum_v \left(\frac{N_u \tilde{N}_v}{\check{N}_u \dot{N}_v}\right)^2 \sum_{i \in s2 \cap u \cap v} \frac{(1-p_{1i})}{p_{1i}^2 \, p_{2i}} \left(y_i(d) - \frac{\hat{Y}_u(d)}{\hat{N}_u}\right)^2$$
$$+ \sum_u \sum_v \left(\frac{N_u \tilde{N}_v}{\check{N}_u \dot{N}_v}\right)^2 \sum_{i \in s2 \cap u \cap v} \frac{(1-p_{2i})}{(p_{1i} \, p_{2i})^2} \left(y_i(d) - \frac{\hat{Y}_v(d)}{\hat{N}_v}\right)^2$$

where $\hat{N}_u$ and $\hat{N}_v$ are calculated using final weights. The inclusion of the factor $(N_u \tilde{N}_v)^2/(\check{N}_u \dot{N}_v)^2$ can be motivated by an improvement in the conditional properties of the estimator (Royall and Eberhardt 1975).

4. Generalized Regression Estimation

A generalized regression estimator is described by Deming and Stephan (1940). Recent applications of generalized regression estimation at Statistics Canada include the work of Lemaître and Dufour (1987) and Bankier, Rathwell and Majkowski (1992). Hidiroglou, Särndal and Binder (1993) discuss generalized regression estimation for business surveys using a number of examples.

Deville and Särndal (1992) derive the generalized regression estimator using calibration. Discussion of generalized regression estimation for the two-phase tax sample is facilitated by the use of calibration ideas. During generalized regression weighting of the first-phase sample, the design weights $1/p_{1i}$ are adjusted to yield weights $W_{1i} = g_{1i}/p_{1i}$ that respect the calibration equations

$$\sum_{i \in s1 \cap u} W_{1i} \, x_i = X_u$$

for each first-phase poststratum u, where x_i is an $L_1 \times 1$ vector of auxiliary variables known for all units in the population and X_u is the vector of auxiliary variable totals for poststratum u. The adjusted weights minimize the distance measure $\sum_{i \in s1} (g_{1i}-1)^2/p_{1i}$.

Weighting of the second-phase sample involves calibration conditional on the results of first-phase weighting. The initial weights, W_{1i}/p_{2i}, are adjusted to give final weights, $W_i = g_{2i} W_{1i}/p_{2i}$, that satisfy the calibration equations

$$\sum_{i \in s2 \cap v} W_i \, z_i = \tilde{Z}_v$$

for each second-phase poststratum v, where z_i is an $L_2 \times 1$ vector of auxiliary variables known for all units in the first-phase sample and $\tilde{Z}_v = \sum_{i \in s1 \cap v} W_{1i} \, z_i$. The final weights minimize the distance measure $\sum_{i \in s2} W_{1i} \, (g_{2i}-1)^2/p_{2i}$.

Using first- and second-phase "g-weights", the generalized regression estimator can be written as

$$\hat{Y}_{GREG}(d) = \sum_{i \in s2} y_i(d) g_{1i} \, g_{2i} \, /(p_{1i} \, p_{2i}) \ .$$

The first-phase g-weight for taxfiler i, $i \in u$, is $g_{1i} = 1 + \boldsymbol{\lambda}_u' \boldsymbol{x}_i$, using the definitions $\boldsymbol{\lambda}_u' = (\boldsymbol{X}_u - \check{\boldsymbol{X}}_u)' \boldsymbol{M}_u^{-1}$, $\check{\boldsymbol{X}}_u = \sum_{i \in s1 \cap u} \boldsymbol{x}_i \, /p_{1i}$ and $\boldsymbol{M}_u^{-1} = (\sum_{i \in s1 \cap u} \boldsymbol{x}_i \, \boldsymbol{x}_i'/p_{1i})^{-1}$. The second-phase g-weight for taxfiler i, $i \in v$, is $g_{2i} = 1 + \boldsymbol{\lambda}_v' \boldsymbol{z}_i$, using the defintions $\boldsymbol{\lambda}_v' = (\tilde{\boldsymbol{Z}}_v - \dot{\boldsymbol{Z}}_v)' \boldsymbol{M}_v^{-1}$, $\dot{\boldsymbol{Z}}_v = \sum_{i \in s2 \cap v} W_{1i} \, \boldsymbol{z}_i \, /p_{2i}$ and $\boldsymbol{M}_v^{-1} = (\sum_{i \in s2 \cap v} W_{1i} \, \boldsymbol{z}_i \, \boldsymbol{z}_i'/p_{2i})^{-1}$.

Ignoring variability due to the estimation of regression coefficients, the variance of $\hat{Y}_{GREG}(d)$ can be approximated as

$$V(\hat{Y}_{GREG}(d)) \approx \sum_i \frac{1-p_{1i}}{p_{1i}} (E_{1i}(d))^2$$
$$+ E_1[\sum_{i \in s2} \frac{1-p_{2i}}{p_{2i}} (W_{1i} E_{2i}(d))^2] \ ,$$

where E_1 denotes expectation with respect to the first phase of sampling, $E_{1i}(d) = y_i(d) - \boldsymbol{x}_i' \boldsymbol{B}_u(d)$ for each taxfiler in first-phase poststratum u and $\boldsymbol{B}_u(d)$, the vector of estimated coefficients from the regression of $y(d)$ on $\boldsymbol{x}$ that would be obtained if $y(d)$ was available for all taxfilers in first-phase poststratum u, is given by

$$\boldsymbol{B}_u(d) = (\sum_{i \in u} \boldsymbol{x}_i \, \boldsymbol{x}_i' \,)^{-1} (\sum_{i \in u} \boldsymbol{x}_i \, y_i(d)) \ .$$

Similarly, $E_{2i}(d) = y_i(d) - \boldsymbol{z}_i' \boldsymbol{B}_v(d)$ for each taxfiler in second-phase poststratum v and

$$\boldsymbol{B}_v(d) = (\sum_{i \in s1 \cap v} W_{1i} \, \boldsymbol{z}_i \, \boldsymbol{z}_i' \,)^{-1} (\sum_{i \in s1 \cap v} W_{1i} \, \boldsymbol{z}_i \, y_i(d)) \ .$$

An estimator of the approximate variance of $\hat{Y}_{GREG}(d)$ is

$$\hat{V}(\hat{Y}_{GREG}(d)) = \sum_i \frac{1-p_{1i}}{p_{1i}^2\, p_{2i}}(g_{1i}\, e_{1i}(d))^2 + \sum_i \frac{1-p_{2i}}{(p_{1i}\, p_{2i})^2}(g_{1i}\, g_{2i}\, e_{2i}(d))^2 .$$

Since $y(d)$ is available only for units in $s2$, the best available estimate of $B_u(d)$ is

$$\hat{B}_u(d)=(\sum_{i\in s2\cap u} W_i\, x_i\, x_i')^{-1}(\sum_{i\in s2\cap u} W_i\, x_i\, y_i(d)) .$$

Similarly, the best available estimate of $B_v(d)$ is

$$\hat{B}_v(d)=(\sum_{i\in s2\cap v} W_i\, z_i\, z_i')^{-1}(\sum_{i\in s2\cap v} W_i\, z_i\, y_i(d)) .$$

The sample residuals needed to compute the variance estimator are $e_{1i}(d)=y_i(d)-x_i'\hat{B}_u(d)$ and $e_{2i}(d)=y_i(d)-z_i'\hat{B}_v(d)$.

If a single auxiliary variable with value one for all taxfilers is employed during both first- and second-phase weighting, $g_{1i}=N_u/\check{N}_u$ for all taxfilers in first-phase poststratum u, $g_{2i}=\check{N}_v/\dot{N}_v$ for all taxfilers in second-phase poststratum v and $\hat{Y}_{GREG}(d)$ is equivalent to $\hat{Y}(d)$.

If y is strongly correlated with x and z, the variance of the generalized regression estimator of the population total of y will be relatively small. However, it is important to note that strong correlations between y and x and z do not imply that the variance of $\hat{Y}_{GREG}(d)$ will be small, since $y(d)$ may be poorly correlated with x and z within poststrata that include at least one sampled unit falling in domain d.

The correlation between $y(d)$ and x and z within a poststratum that includes at least one sampled unit falling in domain d will be low if domain d includes a small proportion of all the sampled units in the poststratum. This situation may arise for two reasons. First, poststrata may be defined to include many domains. If each first-phase poststratum is formed by combining one or more first-phase sampling strata, most first-phase poststrata will include more than one SIC4 domain. Second, domains may be divided between a number of poststrata if the SIC codes used for stratification contain errors.

5. Empirical Study

In order to compare the performance of $\hat{Y}_{H-T}(d)$, $\hat{Y}(d)$ and $\hat{Y}_{GREG}(d)$, an empirical study was conducted using data from the province of Québec for tax year 1989. Since the estimator $\hat{Y}(d)$ is a special case of $\hat{Y}_{GREG}(d)$, it will be called $\hat{Y}_{GREG-TPH}(d)$ in subsequent discussion. (TPH is an abbreviation for two-phase Hájek.) Two other generalized regression estimators were considered. In both cases, x and z contained a variable with value one for all taxfilers. One generalized regression estimator (GREG-R2) involved calibration on taxfiler revenue during second-phase weighting. The second estimator (GREG-R1R2) involved calibration on taxfiler revenue at both phases of weighting.

The universe used for the study included approximately 140,000 T2 taxfilers. The first- and second-phase selection probabilities employed during sampling for production for tax year 1989 were used. The first-phase sample included approximately 31,000 taxfilers and there were about 23,000 businesses in the second-phase sample. Estimates were produced for total expenses. The correlation between taxfiler revenue and total expenses was 0.960.

Large proportions of units in the first- and second-phase samples were selected with certainty. All units with first-phase selection probability one were excluded from first-phase weighting and the corresponding g-weights were set to one. Units with second-phase selection probability one were treated analogously during second-phase weighting. There were 12812 units in the first-phase sample with first-phase selection probabilities different from one and 910 units in the second-phase sample with second-phase selection probabilities different from one.

Each first-phase poststratum consisted of one or more of the first-phase sampling strata used during sampling for 1989 production. All the sampling strata included in any particular first-phase poststratum corresponded to the same revenue class. (There were five revenue classes.) Each first-phase poststratum contained a minimum of 20 sampled units. The use of a minimum sample size was motivated by concerns about the bias in the approximate variance estimator for $\hat{Y}_{GREG}(d)$ when the sample size is very small (Rao 1968b). If a first-phase sampling stratum included fewer than 20 sampled units, it was combined with sampling strata for similar SIC2 codes and the same revenue class until a poststratum containing at least 20 sampled units was obtained. Application of this procedure led to 166 first-phase poststrata. Second-phase poststrata were formed analogously. There were 36 second-phase poststrata.

First- and second-phase g-weights for the GREG estimators were calculated using a modified version of the SAS macro CALMAR (Sautory 1991). The set of first-phase sampling weights calculated for GREG-R1R2 included 18 negative

weights. There were no negative second-phase weights calculated for either GREG-R2 or GREG-R1R2. (Negative weights are not possible for the GREG-TPH estimator.) Estimates of total expenses were produced for 77 SIC2 domains, 256 SIC3 domains and 587 SIC4 domains using the three GREG estimators, as well as $\hat{Y}_{H\text{-}T}(d)$. Since GREG-R1R2 did not produce any negative estimates, the negative weights associated with the estimator were used without modification.

Results of comparison of the GREG-TPH and H-T estimators are presented in Table 1. The GREG-TPH estimator performs better than the H-T estimator for the majority of domains. The gains obtained using GREG-TPH are particularly large for SIC2 domains. At the SIC4 level, the estimated coefficient of variation (CV) for the GREG-TPH estimate of total expenses is lower than the estimated CV for the H-T estimate for 60.5% of domains. In cases in which the estimated CV for GREG-TPH is lower it is only 5.5% smaller, on average, than the estimated CV for H-T. When the estimated CV for GREG-TPH is higher it is 7.9% larger than the estimated CV for H-T, on average. Examination of the relative magnitudes of estimates calculated using GREG-TPH and H-T provides more compelling evidence to prefer GREG-TPH over H-T. The GREG-TPH estimate of total expenses was larger then the H-T estimate for over 93% of the SIC4 domains. Actual two-phase tax sample sizes are typically lower than expected sample sizes for various operational reasons. Use of the GREG-TPH estimator provides an automatic non-response adjustment.

A large proportion of units in the second-phase tax sample have second-phase selection probability one and both GREG-R2 and GREG-TPH use the same auxiliary variables during first-phase weighting. Although GREG-R2 performed somewhat better than GREG-TPH, differences between these two estimators were marginal. The GREG-R1R2 estimator is compared to GREG-TPH in Table 2. Estimated CVs for GREG-R1R2 are generally smaller than estimated CVs for GREG-TPH and the relative performance of GREG-R1R2 improves as domain size increases. Nevertheless, GREG-R1R2 is superior to GREG-TPH for only 64% of SIC4 domains, and the average increase in estimated CVs for those domains in which GREG-R1R2 did worse than GREG-TPH is larger than the average decrease in estimated CVs for domains in which GREG-R1R2 performed better.

The results in Tables 2 indicate that, although the GREG-R1R2 estimator shows some promise, it would be inappropriate to completely replace the GREG-TPH estimator currently used in production by GREG-R1R2.

The results reported in Table 3 were obtained after SIC codes assigned to taxfilers by Revenue Canada and SIC codes used for stratification of the second-phase sample were changed for sampled units, where necessary, to eliminate inconsistencies between these codes and those used to determine domain membership. A comparison of Tables 2 and 3 indicates that the relative performance of GREG-R1R2 for large domains is considerably better when there are no classification errors. GREG-R1R2 reduces estimated CVs by over 22% (on average) for over 85% of SIC2 domains.

6. Conclusions

Generalized regression estimation provides a convenient framework for the use of auxiliary information. It can be applied to Statistics Canada's two-phase tax sample. Bernoulli sampling is used in both phases of tax sampling because it has considerable operational advantages. The estimation method currently used in production incorporates poststratified ratio adjustments to compensate for differences between actual and expected sample sizes. It can be derived as a generalized regression estimator.

In an empirical study, the generalized regression estimator currently used in production (GREG-TPH) performed much better then the Horvitz-Thompson estimator. Two generalized regression estimators using additional auxiliary information were compared to GREG-TPH. The alternative estimators produced improvements for large domains. However, their performance for the smaller domains of particular interest to users of tax estimates was less convincing. The feasibility of using one of the alternative estimators on a limited scale is under study.

Acknowledgements

The authors would like to thank René Boyer for modifying the SAS macro CALMAR for use in the empirical study, as well as K.P. Srinath and Michael Hidiroglou for helpful discussions. Thanks are also due to Michael Bankier and Jean Leduc for comments on an earlier version of this paper.

References

ARMSTRONG, J., BLOCK, C. and SRINATH, K.P. (1993). Two-phase sampling of tax records for business surveys. *Journal of Business and Economic Statistics*, (in press).

BANKIER, M., RATHWELL, S. and MAJKOWSKI, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census of Population. *Statistics Sweden, Workshop on the Uses of Auxiliary Information in Surveys.*

BREWER, K.R.W., EARLY, L.J., and JOYCE, S.F. (1972). Selecting several samples from a single population. *Australian Journal of Statistics*, 14, 231-239.

CHOUDHRY, G.H., LAVALLÉE, P. and HIDIROGLOU, M. (1989). Two-phase sample design for tax data. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 646-651.

DEMING, W.E. and STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 34, 911-934.

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

LEMAÎTRE, G. and DUFOUR, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-207.

HIDIROGLOU, M.A., SÄRNDAL, C.-E. and BINDER, D.A. (1993). Weighting and estimation in establishment surveys. Paper presented at the International Conference on Establishment Surveys, Buffalo, New York.

RAO, J.N.K. (1968a). Some nonresponse sampling theory when the frame contains an unknown amount of duplication. *Journal of the American Statistical Association*, 63, 87-90.

RAO, J.N.K. (1968b). Some small sample results in ratio and regression estimation. *Journal of the Indian Statistical Association*, 6, 160-168.

ROYALL, R.M. and EBERHARDT, K.R. (1975). Variance estimates for the ratio estimator. *Sankhyā*, Ser. C, 37, 43-52.

SAUTORY, O. (1991). La macro SAS: CALMAR. Unpublished manuscript. Institut national de la statistique et des études économiques.

STATISTICS CANADA (1980). Standard Industrial Classification, Catalogue 12-501E, Statistics Canada.

SUNTER, A.B. (1986). Implicit longitudinal sampling from administrative files: a useful technique. *Journal of Official Statistics*, 2, 161-168.

Table 1. Comparison of GREG-TPH and H-T estimators for total expenses, estimated coefficients of variation

Domain Type	Gains using GREG-TPH		Losses using GREG-TPH	
	No.	Mean	No.	Mean
SIC2	57	0.773	20	1.100
SIC3	175	0.910	81	1.082
SIC4	355	0.945	232	1.079

Table 2. Comparison of GREG-R1R2 and GREG-TPH estimators for total expenses, estimated coefficients of variation

Domain Type	Gains using GREG-R1R2		Losses using GREG-R1R2	
	No.	Mean	No.	Mean
SIC2	51	0.867	26	1.170
SIC3	160	0.934	96	1.093
SIC4	377	0.954	210	1.074

Table 3. Comparison of GREG-R1R2 and GREG-TPH estimators for total expenses, estimated coefficients of variation, no misclassification

Domain Type	Gains using GREG-R1R2		Losses using GREG-R1R2	
	No.	Mean	No.	Mean
SIC2	66	0.778	11	1.057
SIC3	184	0.916	72	1.047
SIC4	402	0.944	185	1.034

ESTIMATING SAMPLING VARIANCES AND OTHER ERRORS IN THE SWEDISH CONSUMER PRICE INDEX

Jörgen Dalén, Statistics Sweden, Price Statistics, S - 11581 Stockholm, Sweden

Key words: errors, allocation, two-dimensional.

1. Introduction

This is an abridged version of two reports to appear in Statistics Sweden's series of R&D Reports, on variance estimation (Ohlsson and Dalén 1993) and on the measurement of other KPI (Konsumentprisindex, the Swedish CPI) errors (Dalén 1993). Here we deal mainly with variance estimation. Reference lists and other background material are excluded.

Large parts of the KPI are based on price quotations from two-dimensional samples, each of which is the cross-classification of a sample of *outlets* (shops, restaurants, etc.) and a sample of *products* (items, commodities). In the sequel, such a procedure for sampling from a two-dimensional population will be called *Cross-Classified Sampling*. In this report we will use general results from Ohlsson (1992) to derive estimators of the sampling variance of the KPI. Numerical results will also be given.

We have also developed a model for dealing with other types of KPI errors: We include two types of **errors between surveys** in this model.

1) **Errors in consumption weights**, i.e. the weights used in the aggregation process from item group indices to the all-item KPI.

2) **Errors due to non-coverage of item groups.** Certain products, mainly services, are not included in the KPI, notably (1992) financial services, public child care and care of the elderly and certain international transport services.

We also deal with two types of **errors within surveys.**

3) **Sampling errors in the price surveys.**

4) **Non-sampling errors in the price surveys.** Here we refer to errors due to quality adjustment, formula errors, selection bias due to purposive sampling, low level weighting errors, errors in the recorded price as well as the familiar non-response and coverage errors. These errors generally give rise to biases or bias risks.

We use a mean square error model with both variance and bias components in two stages. This model is not, however, dealt with further in this summary.

2. The KPI sampling structure

A distinctive feature of the Swedish KPI is that it is composed of many (50-60) independent price surveys for different product groups. Some of these are large in terms of the total weight covered but several are very small.

The all item KPI could be written as a sum of weighted one-survey indices:

$$I = \sum_k w_k I_k \quad , \qquad (2.1)$$

where I_k is the one-survey index and w_k the aggregate weight of all items in that survey. Based on sample data the I_k are estimated independently by $\hat{I}_k$ giving rise to the estimated KPI, $\hat{I}$, and the aggregate variance can be computed simply as

$$V(\hat{I}) = \sum_k w_k^2 \; V(\hat{I}_k) \qquad (2.2)$$

where the $V(\hat{I}_k)$ are estimated for each survey separately.

The real problem is how to estimate variances for each individual survey. In the search for such estimates we have given priority to large surveys. For surveys covering 50% of the weight, thorough variance estimates have been done. For surveys covering an additional 30% of the weight we have made crude estimates of the order of size of the variances. Based on general knowledge of the price data, the many independent, small surveys making up for the remaining 20% of the weight are not believed to influence a measure of total KPI sampling error significantly although thorough variance estimates for these surveys have not yet been done.

There are two major sampling dimensions in a CPI for measuring price change - of outlets and of products. In many countries purposive sampling is used in one or both of these dimensions. In fact, the only country attempting an all out probability design for its CPI is the United States. In Sweden, outlet sampling is mainly

done by probability. Product sampling is done by probability for product groups covering about 18% of the KPI weight while as purposive selection is used for about 40%. For other product groups there is either no product sampling at all (all products are covered) or the distinction between outlets and products is not clear. In the latter cases various mixtures of probability-based and purposive procedures are used.

There is also a third sampling dimension in a CPI - for the estimation of weights. But in the KPI weights are mainly taken from the National Accounts that in its turn use retail trade surveys together with other sources. The weights are thus not estimated according to probability based sampling theory. Hence the weight dimension is not further discussed here but in Dalén (1993).

The approach in this paper is basically design-based. From this point of view there is no such thing as a variance from a purposive sample. But there is an obvious need to compute some measure of the contribution to the KPI error, due to purposive sampling, at least for the purpose of obtaining a rational allocation with respect to the size of the outlet sample vs. the size of the product sample. Our approach on this issue is to calculate variances from purposive samples, with a design similar to the probability samples in some other parts of the KPI. This requires setting up a design-based model for the purposive selection reflecting as closely as possible how this selection was actually done. Below (in Section 4.2) we describe the construction of a postulated design in the largest KPI price survey.

3 Variance structure

For the exact variance formulas and estimators the reader is referred to the full report. Here we only give a brief hint as to the variance structure.

The theory of cross-classified sampling gives us three major variance components, each of which is further subdivided into several strata. These are the variance between products (V_{PRO}), variance between outlets (V_{OUT}) and interaction variance (V_{INT}):

$$V_{PRO} = \sum_g \frac{1}{m_g} \sigma_g^2, \qquad (3.1)$$

$$V_{OUT} = \sum_h \frac{1}{n_h} \sigma_h^2 \text{ and} \qquad (3.2)$$

$$V_{INT} = \sum_g \sum_h \frac{1}{n_h m_g} \sigma_{gh}^2, \qquad (3.3)$$

where m_g is the sample size in product stratum g and n_h is the sample size in outlet stratum h.

The various σ^2 are the underlying population variances which in practice display a complicated structure.

We have $V_{TOT} = V_{PRO} + V_{OUT} + V_{INT}$.

4 Variance estimation

The cross-classified variance estimation procedures were applied to two of the major price measurement systems in the KPI. These are called the List Price System (LIPS) and the Local Price System (LOPS).

4.1 The List Price System (LIPS)

In the LIPS probability sampling is used for products as well as outlets. Prices are taken from wholesalers' price lists. LIPS accounts for about 18% of the total CPI weight.

LIPS covers most food products (not fresh food such as fruits and vegetables, bread and pastries or fish) and other daily commodities such as those typically found in a supermarket. Sampling of products is based on historic sales data from the three major wholesalers in Sweden using systematic PPS sampling of products, stratified into about 60 strata.

Sampling of outlets is done by pps sampling, viz. sequential Poisson sampling introduced in Ohlsson (1990). The size measure is number of employees plus 1, used as a rough measure of turnover.

There were about 1200 products in the sample in 1991-1992 . The number of outlets was 58 in 1991 and 59 in 1992.

Our postulated design (which differs somewhat from the one actually used) uses an asymmetric stratification structure. We start by stratifying the whole product-outlet population into three wholesaler strata. In each of these **primary strata** cross-classified sampling is done independently. In each primary stratum an unstratified PPS sample of outlets is cross-classified with a stratified PPS sample of products. This gives the following variance structure

$$V_{TOT} = \Sigma_h w^2_h V_{TOT\,h} \qquad (4.1)$$

where the w_h are market share weights of the three wholesalers.

4.2 The Local Price System (LOPS)

In the LOPS interviewers collect prices each month. This system accounts for about 21% of the total CPI weight.

Outlets are divided into 25 strata according to the SNI code (Swedish Code of Industrial Classification which closely follows the ISIC code) of the outlet. Within a stratum sequential Poisson sampling is used as in the LIPS.

For products, however, purposive sampling is used in several steps. The products are divided into product groups according to National Accounts and other sources of information . Within a product group one or more products are chosen, often only one. All in all there are some 140 "representative products" in the LOPS. For each of these 140 products a commodity specification is done in the central office. The interviewer is then asked to find the particular variety according to the specification that is the most sold within the surveyed outlet.

The actual sampling procedure in LOPS is thus a combination of random sampling of outlets and purposive sampling of products.

Our postulated design is, however, the cross-classified design described above. There are 20(1991) or 21(1992) outlet strata (some strata are collapsed) and 48 (1991) or 43(1992) product strata.

The forming of the product strata was based on the actual procedures used in the selection of the representative products. There the starting point is often information on the consumption value of a rather narrow product group like "bananas", "dish washers", or "towels". The final sample is then one or several representative products in each group. In our variance estimation procedures we consider these final products as randomly chosen from their respective product groups. For imputing a subjective "inclusion probability" into the variance formulas we basically ask the question. How much of the consumption value *in the average outlet* is accounted for by the product finally selected by the interviewer? For example for the product group *bananas* there is typically only one brand and one price in an outlet and we therefore set the inclusion probability to one for the representative product *one kg of bananas* within the product group *bananas*. On the contrary, a particular brand of the representative product *towel, terry cloth, 100% cotton, hemmed, about 50x70 cm* within the product group *towels* is likely to have a rather small share of an outlet's total sales value of towels and in this and many similar cases we set the "inclusion probability" to 0.1.

We believe that, although there are subjective elements in our procedure, it gives a fairly realistic picture of the random error arising from product sampling.

4.3 Other price surveys

Variance computations were done for other price surveys too. The methods used were cruder and more simplified. The exact procedures are not of general interest so a short summary will be sufficient.

The **apparel survey** (covering about 6% of the total KPI weight) used sequential sampling of outlets and a purposive sample of 24 garments. But in each outlet several (up to 8) varieties of each garments were priced. According to various comparability criteria only a certain portion of the varieties are included in the index. Here we used a simplified one-dimensional variance estimation procedure based on the effective sample size for each garment.

The **rental survey** (directly 10% and with imputation about 13.5% of the weight) is based on a random sample of about 1000 apartments and is thus one-dimensional. The estimator is post-stratified into different size groups crossed with newly built versus old ones. The estimated index is a kind of unit-value index comparing the average rent per m^2 for all apartments at two points of time, with corrections for differences in quality (equipment etc.) Our variance computations are so far simplified in that they do not take account of the changing population.

The **interest survey** (8.5% of the weight) estimates the amount of interest paid or foregone for owner-occupied homes by multiplying their average purchase values with the average interest rate. For estimating average debt a sample of some 800 homes is drawn and for estimating interest rate there is a sample of financial institutes. No direct computations of sampling error have so far been done for this survey but a crude assessment indicates that sampling variances must be rather small compared with other large KPI surveys (and compared with other errors and uncertainties in the index for owner-occupied housing itself).

The **car survey** (3%) was, in 1991, based on 31 brands of cars (60 in 1992). The design could be interpreted as stratified with one purposively selected unit in each stratum. Variances have been computed based on a crude assumption of simple random sampling, which is judged to be reasonable.

The **petrol survey** (4%) covers five types of petrol (almost all existing) at 120 petrol stations, sampled with sequential Poisson sampling so variance estimation was done in a straight-forward way.

The **survey of alcoholic beverages** (2.6%) relies on information from the Swedish state monopoly for its total sales. It thus has zero sampling error (leaving aside taxfree sales at airports etc.).

Other product groups also have zero or very small sampling error. This is the case for, e.g., **gambling and lotteries, TV license fees, hospital care** and **dental care**.

For other surveys weights are so small that they could not influence a measure of total KPI sampling error significantly. Direct variance estimates have not yet been done, however.

5. Numerical results for 1992

Up to now we have computed sampling errors for two years. Here we give results for 1992. In Table 1-3 we present December-to-month variance estimates (within an annual link) for the three most important surveys.

The interpretation of these results is the following with the LIPS as our example. V_{TOT} for June 1992 is 0.0500. If the LIPS part of the KPI for June 1992 with December 1991 as reference period was 102 an appropriate 95% confidence interval for its sampling error would be $102\pm1.96\times\sqrt{0.0500}=102\pm0.44$. If the LIPS had been the only source of sampling error in the KPI and the KPI figure for June 1992 was 103 then a 95% confidence interval for it would be $103\pm1.96\times\sqrt{0.0014}=105\pm0.07$. The contribution to the KPI variance is calculated according to (2.2) remembering that the weight for the whole LIPS system was 0.18.

For the LOPS (table 2) we note that the V_{PRO} component is always larger than V_{OUT}. We have here some 800 outlets and 140 products. Since it costs less to include one more product than one more outlet in the sample, this means that we have had a poor allocation. These results have generated a successive movement towards a more efficient allocation with more products and fewer outlets in this survey which is the most expensive one in the KPI.

The Apparel survey (table 3) shows the relatively largest variances of all price surveys and is now the greatest concern in our KPI sampling design. Unfortunately, we have not so far been able to produce a good decomposition of this variance. A major problem here is that our criteria for comparability are sometimes so restrictive that the result is a very small effective sample size. But more liberal comparability criteria lead to a risk for increasing biases instead. For 1993, measures have been taken to increase the efficiency of the comparability criteria and for 1994 we aim at introducing a hedonic technique that will enable us to use all data more efficiently.

For other price surveys than those discussed in some detail above, some crude estimates are given here only to show that their contributions to KPI variance is of a smaller order of size than for those surveys, where we have used more careful procedures:

TABLE 4 Crude variance estimates for some price surveys

Survey	Variance Estimate	KPI Weight 1992	Contribution to KPI Variance
Rental Survey	0.1	0.1348	0.002
Car Survey	1	0.0298	0.0015
Petrol Survey	0.001	0.0402	0.000002

6. The allocation problem

In order to give a survey an optimal sampling allocation two things are neccessary: a variance function and a cost function. The variance function with the estimates it has provided so far is given above. Now we turn to the cost function.

In the Swedish KPI practice there are two levels of the allocation problem - between surveys and within surveys. So far we have not been able to give direct cost estimates to different surveys. As proxies we could use sample sizes which gives the following picture, with reference to 1991, for the surveys discussed above.

TABLE 5

Survey	Variance estimate, on average	Contribution to KPI variance	Effective sample size	Mode of data collection
List Price System	0.09	0.003	18000	Price Lists (Visits)
Local Price System	0.19	0.009	8500	Visits, Some Telephone
Apparel Survey	2.06	0.009	1000	Visits
Rental Survey	0.1	0.002	1000	Mail
Car Survey	1	0.0015	31	Telephone
Petrol Survey	0.001	0.000002	600	Telephone

Now, of course, costs are not generally proportional to sample size. In the LIPS, for example, a mode of data collection (mainly price lists) was used which makes it much less expensive than the LOPS (where interviewers to a large extent visit the outlets for price measurement), despite its larger sample size.

Some obvious misallocation is readily seen from table 5, however. The sample sizes of the car and the petrol survey should rather be reversed, for example. Also there is an obvious need for increasing the sample sizes for apparel items.

When it comes to allocation within surveys, we take our most expensive survey, LOPS, as our example. Here we use the following cost function:

$$C = C_0 + \Sigma_h\, n_h\{a_h + b_h \Sigma_g m_g(\Sigma_{k\in g} r_{ghk})\}, \qquad (6.1)$$

where

C_0 is a fixed cost for administration etc.,
n_h is the number of outlets in outlet stratum h,
m_g is the number of items in item stratum g,
a_h is the fixed cost per outlet in stratum h, mainly due to travel time,
b_h is the cost of measuring one item in outlets of stratum h and
r_{ghk} is the relative frequency of item $k\in g$ in outlets of stratum h.

In practice a_h depends on the extent to which telephone interviews could be used for price measurement in stratum h. This could be done in most months, if there are few and simply defined items in the outlet.

The b_h are the same for most strata but food items are generally simpler to measure and so b_h is smaller for the daily commodity stores.

Now if we try to combine (6.1) with the variance functions we run into a non-linear optimization problem for which it seems impossible to find an explicit solution. A direction to go would therefore be to try some kind of numerical optimization procedure. So far we have not made attempts in this direction since the necessary work could be expected to be large.

Although a formal optimization has not been done, the variance and cost functions have made sizeable improvements in the LOPS allocation possible by simple inspection of the numerical results combined with calculation of marginal changes. We have increased the number of products (in particular in the highly variable product group *fresh fruit and vegetables*) in the survey while cutting down on the number of outlets. All together this has led to lower costs and smaller variance at the same time!

7. References

Dalén, J. (1993): An Error Model for the Swedish Consumer Price Index. To appear in R&D Reports, Statistics Sweden.

Ohlsson, E. (1990): Sequential Poisson Sampling from a Business Register and its Application to the Swedish Consumer Price Index. Statistics Sweden R&D Report.

Ohlsson, E. (1992: Cross-classified Sampling for the Consumer Price Index. Statistics Sweden R&D Report.

Ohlsson, E. and Dalén, J (1993): Variance Estimation in the Swedish Consumer Price Index. To appear in R&D Reports, Statistics Sweden.

TABLE 1: Variance estimates in the LIPS 1992

Month, stix=short term index ltix=long term index	VPRO	VOUT	VINT	VTOT	Contribution to KPI Variance
January	0.0106	0.0134	0.0016	0.0256	0.0007
February	0.0117	0.0144	0.0018	0.0279	0.0008
March	0.0121	0.0146	0.0019	0.0286	0.0008
April	0.0108	0.0241	0.0020	0.0369	0.0010
May	0.0167	0.0299	0.0022	0.0488	0.0013
June	0.0171	0.0303	0.0026	0.0500	0.0014
July	0.0200	0.0299	0.0026	0.0525	0.0014
August	0.0173	0.0400	0.0027	0.0600	0.0016
September	0.0207	0.0357	0.0027	0.0591	0.0016
October	0.0192	0.0307	0.0028	0.0527	0.0014
November	0.0151	0.0599	0.0030	0.0780	0.0021
December, stix	0.0164	0.0721	0.0031	0.0916	0.0025
December, ltix	0.0177	0.1076	0.0065	0.1318	0.0036

TABLE 2: Variance estimates in the LOPS 1992

Month	VPRO	VOUT	VINT	VTOT	Contribution to KPI variance
January	0.0394	0.0258	0.0125	0.0777	0.0032
February	0.0744	0.0294	0.0172	0.1210	0.0049
March	0.0487	0.0335	0.0178	0.1000	0.0041
April	0.0603	0.0442	0.0215	0.1260	0.0051
May	0.0678	0.0387	0.0244	0.1309	0.0053
June	0.0797	0.0447	0.0276	0.1520	0.0062
July	0.1452	0.0531	0.0317	0.2300	0.0094
August	0.0991	0.0627	0.0334	0.1952	0.0080
September	0.0718	0.0529	0.0338	0.1585	0.0065
October	0.1450	0.0587	0.0484	0.2121	0.0086
November	0.1476	0.0600	0.0533	0.2609	0.0106
December	0.1416	0.0705	0.0561	0.2682	0.0109

TABLE 3 Variance estimates in the Apparel Survey 1992

Month	Total variance	Contribution to KPI variance
January	1.28	0.0041
February	2.10	0.0067
March	3.38	0.0109
April	1.94	0.0088
May	3.01	0.0109
June	7.82	0.0176
July	11.06	0.0356
August	11.19	0.0360
September	3.33	0.0107
October	2.01	0.0065
November	3.62	0.0122
December	3.68	0.0124