Variance Estimation for Statistics Canada's Small Household Surveys in the Context of the Household Survey Strategy

Sébastien Landry
Statistics Canada, Household Surveys Methods Division, R.H. Coats Building, 16th floor, Ottawa, ON, Canada, K1A 0T6
sebastien.landry@statcan.gc.ca

Abstract
Statistics Canada launched the Household Survey Strategy (HSS) to be able to increase its capacity to conduct on-going and new household surveys in a cost-effective manner. The main methodological component of the HSS is the master sample, which will group the respondents from major household surveys to produce a sample frame which will be used by smaller household surveys to select their samples. Many methodological challenges will confront these multi-phase, multi-stage and multi-frame smaller surveys, including sample coordination, weighting and estimation. This paper will focus on the variance estimation for the smaller surveys. A new variant of the bootstrap method is planned to be used to estimate the variance of these surveys. A simulation study trying to recreate the HSS environment using census data shows that this method should give fairly accurate variance estimates.

Keywords: Variance estimation, Multi-phase surveys, Multi-frame surveys, Bootstrap, Household surveys

1. Introduction

Currently, the demand for data coming from household surveys is ever increasing. However, conducting household surveys has become more expensive and many challenges, including declining response rates, cell phone only households, and others, have arisen. Statistics Canada launched the Household Survey Strategy (HSS) to face these challenges. The HSS will increase Statistics Canada's capacity to conduct household surveys and will do so in a cost-effective manner.

The components of the HSS will be presented in section 2. The methodology of the master sample, which is an important methodological component of the HSS, will be explained in section 3. The weighting, estimation and variance estimation strategies, the latter being the focus of this paper, will be presented in section 4. The results of a simulation study, which has been conducted to validate the variance estimation strategy, will be shown in section 5. An overview of on-going and future HSS-related projects and the conclusion will be presented in section 6.

2. Components of the Household Survey Strategy

The HSS consists of four main components: content harmonization, spreading of the interviewer workload, the master sample and management of response burden.

2.1. Content Harmonization
The HSS will harmonize the concepts and definitions of variables and will provide sets of standardized questions on these harmonized subjects. This will allow all household surveys to ask the same questions when they want to collect data on the same subject.

2.2. Spreading the Interviewer Workload
The objective of this component is to spread the interviewers' workload throughout the year to avoid short periods of time when collection demands would be so high that Statistics Canada would need to hire and train new interviewers for only those short periods of time. This will also prevent having periods of time when collection demand would be too low to fill the current interviewers' workload.

2.3. Master Sample

The idea behind the master sample is a database that will be made up of respondents coming from large household surveys, i.e. surveys that will feed the master sample. It will be used to create a sampling frame for smaller household surveys, i.e. surveys that will use the master sample as their sampling frame. The large surveys collect a portion of their sample at regular intervals (this also addresses the interviewer workload component of the HSS), which means that they will be able to feed the master sample regularly, every one or two months. Each small survey will be able to create its own sampling frame from the master sample, use its own survey design and select its own sample of households or people.

The variables that are stored in the master sample consist of design information from the large surveys, demographic information, both at the household and the person level, paradata, which will be used to measure response burden, and contact information, mainly the telephone number, so that the smaller surveys can reach the household.

2.4. Management of Response Burden

The smaller surveys will select their respective samples from the same pool of households which already responded to a survey. Thus, it will be important to apply a sampling coordination strategy for the smaller surveys in order to keep households from having to respond to too many surveys.

3. Methodology of the Master Sample

A brief overview of the survey design of the two large surveys (there could be more than two large surveys in the future) that will feed the master sample and the template of the survey design for the smaller surveys will be introduced first. The interaction between the large surveys, the master sample and the smaller surveys will then be presented.

3.1. Survey Design of Large Surveys

The two large surveys that will feed the master sample are the Labour Force Survey (LFS) and the Canadian Community Health Survey (CCHS). The LFS uses a clustered area frame covering Canada and selects its household sample through a two-stage design. The clusters from each stratum are grouped in six parts called rotation groups. One cluster per rotation group is selected at the first stage. Then, a systematic sample of households within the selected cluster from the rotation groups is selected at the second stage and that systematic sample remains in the LFS sample for six months. Each month, one sixth of the sample, that is the systematic sample from one rotation group, is replaced by a new systematic sample from the same rotation group. The households leaving the LFS sample become what is called a rotate-out and are fed to the master sample. An LFS person sample is built by selecting every member of the selected households.

The LFS area frame is also used by the CCHS, but a two-stage, two-phase sampling design is used to select households for that survey. To simplify, the CCHS first uses the LFS sampling design to select available systematic samples of households. Then, a sub-sample of the first-phase sample is selected using the CCHS' own stratification criteria. The sample, which covers one year, is divided into two-month collection periods, so the CCHS will feed the master sample after a collection period has been completed. The CCHS also uses a telephone frame, which covers around 80% of the Canadian population. A stratified simple random sampling without replacement (SRSWOR) strategy is used to select phone numbers from that frame for a two-month collection period. After gathering demographic information about the household members, one person from each selected household is then selected for the CCHS person sample.

Both surveys have the same area frame first-phase design. However, while they share selected clusters, the households selected in the clusters are ultimately assigned to only one survey. A result of this is that the area frame portion of the master sample will be considered to have been selected from a two-stage, two-phase design and that the telephone frame portion of the master sample will be considered to have been selected from a stratified SRSWOR design.

More information about the LFS methodology can be obtained in Statistics Canada (2008a), while the methodology of the CCHS is presented in Statistics Canada (2008b).

3.2. Survey Design of Smaller Surveys

As stated earlier, a smaller survey that wishes to use the master sample as a sampling frame will be able to select its household sample using its own sampling design. However, in order to facilitate the management of the response burden of the master sample households, the smaller survey will need to use a stratified SRSWOR design. Also, for weighting purposes, the selection of households from the area frame and the telephone frame will have to be independent. So the area frame smaller survey sample will be selected from a three-phase design, while the telephone frame smaller survey sample will be selected from a two-phase design.

If the smaller survey needs to sample people, an additional stage of sampling is performed and one (or more) person from every selected household is sampled.

3.3. Interaction Between Large Surveys, Master Sample and Smaller Surveys

Figure 1 provides a visual representation of the master sample process, from its creation to its use by smaller surveys. The area and telephone frames are kept in separate steps until the smaller survey collection.

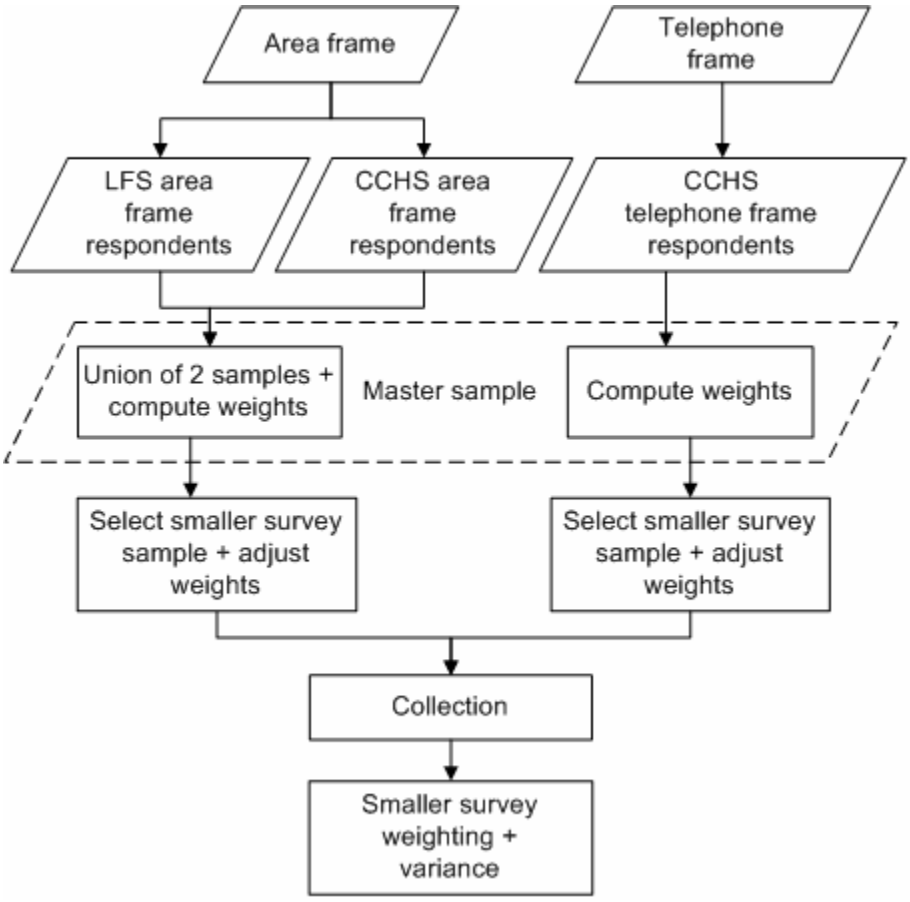

Figure 1: The Master Sample Process

The area frame is used to select both the LFS sample and a portion of the CCHS sample. Respondents from these samples are then merged and the weights are adjusted (greater detail is given in section 4) to form the area frame portion of the master sample. This portion is then used by the smaller survey to select a part of its sample. The telephone frame is only used by the CCHS to select a portion of its sample. Once weighting is performed, the respondents of that sample are put in the telephone frame portion of the master sample. The telephone frame portion of the master sample is then used by the smaller survey to select the other part

of its sample. The two parts of the smaller survey sample are then merged for collection, final weighting, estimation and variance estimation.

4. Weighting, Estimation and Variance Estimation Under the HSS

As stated in section 3.3, many sampling weight adjustments have to be performed to transform the weights coming from large surveys into weights that can be used by the smaller surveys to produce estimates. The weighting steps are presented in section 4.1. The estimation process is then explained in section 4.2. Finally, the variance estimation strategy is presented in section 4.3.

4.1. Weighting Process

The survey frames are kept in separate steps in most of the master sample process. Therefore, their respective weighting processes are also kept separate and will be presented accordingly. Then, the weight adjustment necessary to combine the survey frames will be shown.

4.1.1. Master Sample Area Frame Weighting

Since the first-phase sampling design is the same for the LFS and the CCHS, the master sample weighting strategy will assume that the LFS and CCHS first phase samples were drawn during the same step, i.e. the first-phase sampling selected the union of both samples. Once the first phase has been selected, the households are assigned to the LFS or the CCHS deterministically, which means that this step will not need to be taken in account when adjusting weights. So, let w_{hgij}^{A1} be the first-phase weight of household *j* from cluster *i* which is in rotation group *g* and in stratum *h* for a large survey. This weight is calculated assuming that the usual number of systematic samples (generally one, sometimes two or more) is selected in the selected cluster of every rotation group of the stratum for an LFS collection period. When the weights at the master sample level need to be derived, the weight adjustment will depend on the number of systematic samples selected during the first phase of each large survey. For the LFS, this is determined by the number of rotate-outs in the master sample, while for the CCHS, the whole first-phase sample is used. The weights also need to take in account the fact that, sometimes, not all six rotation groups from the stratum were used. The master sample first-phase weight $w_{hgij}^{A,MS1}$ is calculated as follows:

$$w_{hgij}^{A,MS1} = \frac{w_{hgij}^{A1}}{S_{hg}^{LFS} + S_{hg}^{CCHS}} * \frac{6}{G_h}$$

where S_{hg}^{LFS} is the number of LFS systematic samples coming from rotation group *g* of stratum *h* in the master sample, S_{hg}^{CCHS} is the number of CCHS first-phase systematic samples coming from rotation group *g* of stratum *h* and G_h is the number of rotation groups used in stratum *h*.

While the LFS has only one phase of sampling, the CCHS selects its final sample by performing a second phase of sampling, its own sampling design being used. A non-response adjustment is also performed, since only the respondents will be kept in the master sample. So the master sample design weight $w_j^{A,MS}$ is calculated as follows:

$$w_j^{A,MS} = w_{hgij}^{A,MS1} * (I_{hgij}^{LFS} + I_{hgij}^{CCHS} * \frac{S_r}{s_r}) * r_j^A$$

where I_{hgij}^{LFS} is an indicator variable showing whether the household *j* from cluster *i* in rotation group *g* of stratum *h* is in the LFS first-phase sample or not, I_{hgij}^{CCHS} is an indicator variable showing whether the household *j* from cluster *i* in rotation group *g* of stratum *h* is in the CCHS first-phase sample or not, S_r is the number of households from CCHS second-phase stratum *r* in the CCHS first-phase sample, s_r is the number of households from CCHS second-phase stratum *r* in the CCHS second-phase sample and r_j^A is the large survey area frame non-response adjustment factor for household *j*.

4.1.2. Master Sample Telephone Frame Weighting

The weight of a household (represented by its telephone number) coming from the telephone frame for a collection period *c* is $w_{rcj}^{T} = \frac{1}{\pi_{rcj}} = \frac{M_{rc}}{m_{rc}}$, where $\pi_{rcj} = \frac{m_{rc}}{M_{rc}}$ is the probability of selection of household *j* in stratum *r* for collection period *c*, M_{rc} is the number of telephone numbers in the population of stratum *r* for collection period *c* and m_{rc} is the number of telephone numbers from stratum *r* in the sample for collection period *c*. Independent samples are drawn for each two-month collection period and the telephone frame is updated every six months. Since π_{rcj} is fairly small, the probability that a household gets selected in more than one telephone frame sample is negligible. So, after applying a telephone frame adjustment factor r_j^T, which includes non-response adjustment and adjustment for households with multiple telephone numbers, the master sample design weight of a household coming from the telephone frame $w_j^{T,MS}$ is calculated as follows:

$$w_j^{T,MS} = \frac{r_j^T}{\sum_{c=1}^{C} \pi_{rcj}}$$

where *C* is the number of telephone frame collection periods included in the master sample.

4.1.3. Smaller Survey Weighting

Since the selection of the smaller survey sample is done separately for the two survey frames, these frames have to be considered as separate strata. The objective being to calculate a double-expansion estimator, the design weight for a household selected by the smaller survey is $w_{kj}^{D,SS} = w_j^{D,MS} * \frac{N_k^D}{n_k^D}$, where *D* is either *A* or *T*, that is the survey frame household *j* comes from, N_k^D is the number of households from frame *D* in smaller survey stratum *k* and n_k^D is the number of households from frame *D* and stratum *k* in the smaller survey sample.

To obtain an estimate from a sample that was selected using two frames, a dual-frame weight adjustment needs to be performed. The adjustment will be based on the one proposed by Hartley (1962). An assumption is made that all households that are in the telephone frame can be found on the area frame, but not all households that are in the area frame can be found on the telephone frame. Furthermore, the telephone frame membership status is unknown for some area frame households. The telephone frame membership status will be modeled through a logistic regression, using area frame households for which the telephone frame membership status is known. A probability of belonging to the telephone frame will then be imputed for the area frame households for which the telephone frame membership status is unknown.

After applying the dual-frame weight adjustment, the smaller survey household weight w_j^{SS} will become:

$$w_j^{SS} = w_{kj}^{A,SS} * (I_j^{A\setminus T} + \alpha * I_j^{A\cap T} + ((1-p_j) + p_j * \alpha) * I_j^{A,U}) + w_{kj}^{T,SS} * (1-\alpha) * I_j^{T}$$

where $I_j^{A\setminus T}$ is a binary variable indicating if household *j* is in the area frame sample, but it is known that it is not in the telephone frame, $I_j^{A\cap T}$ is a binary variable indicating if household *j* is in the area frame sample and it is known that it is in the telephone frame, $I_j^{A,U}$ is a binary variable indicating if the household *j* is in the area frame sample, but the telephone frame membership status is unknown, I_j^{T} is a binary variable indicating if the household *j* is in the telephone frame sample, p_j is the probability that area frame household *j* is also in the telephone frame and α is the dual-frame adjustment factor.

When the smaller survey is at the person level, the weight for the selected person *q* from household *j* will be $w_{jq}^{SS} = \frac{w_j^{SS}}{z_{jq}}$, where z_{jq} is the probability of selection of person *q* in household *j*, which may have been calculated by the CCHS or the smaller survey.

After the dual-frame adjustment, a non-response adjustment will be performed within non-response groups containing households with similar probability of response. The household and person weights will become $w_j^{SS,NR} = w_j^{SS} * r_j^{SS}$ and $w_{jq}^{SS,NR} = w_{jq}^{SS} * r_{jq}^{SS}$ respectively, where r_j^{SS} and r_{jq}^{SS} are the household and person non-response adjustment factors for household *j*.
The final weighting step is calibration where the weights are adjusted to match population totals and/or estimates at the master sample level. The final household and person weights will be $w_j^{SS,F} = w_j^{SS,NR} * c_j^{SS}$ and $w_{jq}^{SS,F} = w_{jq}^{SS,NR} * c_{jq}^{SS}$ respectively, where c_j^{SS} and c_{jq}^{SS} are the household and person calibration factors for household *j*.

4.2. Estimation Process

As stated in section 4.1.3, estimates at the master sample level may be required to perform smaller survey calibration. In this case, the dual-frame adjustment will need to be applied to the master sample weights and calibration will need to be performed so that the weights match to population totals. So, the master sample weights used in estimation will be:

$$w_j^{MS,F} = \left[w_j^{A,MS} * (I_j^{A\setminus T} + \alpha * I_j^{A\cap T} + ((1-p_j) + p_j * \alpha) * I_j^{A,U}) + w_j^{T,MS} * (1-\alpha) * I_j^{T}\right] * c_j^{MS}$$

where p_j is calculated using the smaller sample telephone frame membership status model, and the calibration factor c_j^{MS} is derived using the weights after the dual-frame adjustment.

The master sample level and smaller survey level estimates for the total of a variable of interest *y* will be calculated using the following formulae:

$$\hat{Y}_{MS} = \sum_{j\in MS} w_j^{MS,F} * y_j \text{ or } \hat{Y}_{MS} = \sum_{j\in MS}\sum_{q\in j} w_{jq}^{MS,F} * y_{jq}, \text{ and } \hat{Y}_{SS} = \sum_{j\in SS} w_j^{SS,F} * y_j \text{ or } \hat{Y}_{SS} = \sum_{j\in SS}\sum_{q\in j} w_{jq}^{SS,F} * y_{jq},$$

whether *y* is at the household or person level.

4.3. Variance Estimation Strategy

Since the smaller surveys end up using a multi-stage, multi-phase and multi-frame design, finding an exact formula to calculate the variance of their estimates is nearly impossible, so a replication method will need to be used. The bootstrap method, as presented by Rao and Wu (1988), is the preferred method in this case, mainly because of its fixed number of replication samples. This number is fixed to 500, which has proven effective for Statistics Canada surveys. For the area frame, the bootstrap sampling will be performed at the first-stage level, which means that clusters from the first-phase master sample will be resampled. For the telephone frame, telephone numbers from the master sample will be resampled during the bootstrap sampling.

Because the HSS has a multi-phase design and plans to use a double-expansion estimator, the usual bootstrap variance estimate can lead to undesirable results. An alternative estimate proposed by Kim, Navarro and Fuller (2006) (later referred to as KNF) will be used instead. For a two-phase design, the authors propose a weight adjustment estimating the second-phase weight in order to obtain the second-phase bootstrap weight.

Assuming that $w_{hgij}^{A,MS1(b)}$ is the usual bootstrap weight from bootstrap sample *b* for household *j* which is in the first-phase area frame sample, KNF states that the master sample second-phase (b) bootstrap weight for area frame household *j* will become:

$$w_j^{MS(b)} = \sum_{r=1}^{R} w_{hgij}^{A,MS1(b)} \left(\frac{\sum_{j \in MS1} w_{hgij}^{A,MS1(b)} (w_{hgij}^{A,MS1})^{-1} x_{jr}}{\sum_{j \in MS} w_{hgij}^{A,MS1(b)} (w_{hgij}^{A,MS1})^{-1} x_{jr}} \right)$$

where x_{jr} is a second-phase indicator for household j in second-phase stratum r, *MS1* is the master sample first-phase sample, and R is the number of second-phase strata.
For a household coming from the telephone frame, since there was only one phase of sampling performed, the usual bootstrap weight will be used. Hence, $w_j^{MS(b)} = w_j^{T,MS(b)}$.

Since the smaller survey introduces an additional sampling phase, another KNF weight adjustment will need to be performed in order to obtain proper bootstrap weights for the smaller survey. So, regardless of where the frame household j is coming from, its (b) bootstrap weight for the smaller survey will be:

$$w_j^{SS(b)} = \sum_{k=1}^{K} w_j^{MS(b)} \left(\frac{\sum_{j \in MS} w_j^{MS(b)} (w_j^{MS})^{-1} I_{jk}}{\sum_{j \in SS} w_j^{MS(b)} (w_j^{MS})^{-1} I_{jk}} \right)$$

where w_j^{MS} is the master sample design weight for household j (either $w_j^{A,MS}$ or $w_j^{T,MS}$, depending on which frame the household comes from), I_{jk} is an indicator variable for household j in smaller survey stratum k and K is the number of smaller survey strata. The (b) bootstrap weight at the person level will be $w_{jq}^{SS(b)} = \frac{w_j^{SS(b)}}{z_{jq}}$.

To take in account the whole weighting process in the variance estimation process, the dual-frame adjustment, non-response adjustment and calibration (using first-phase estimates calculated from the bootstrap sample if applicable) are applied to each bootstrap weight to obtain final bootstrap weights, $w_j^{SS,F(b)}$ and $w_{jq}^{SS,F(b)}$, which will be used in bootstrap variance estimation. The (b) bootstrap estimate at the household and person levels will be $\hat{Y}_{SS}^{(b)} = \sum_{j \in SS} w_j^{SS,F(b)} * y_j$ and $\hat{Y}_{SS}^{(b)} = \sum_{j \in SS} \sum_{q \in j} w_{jq}^{SS,F(b)} * y_{jq}$. The KNF variance estimate will be $\hat{V}(\hat{Y}_{SS}) = \frac{1}{500} \sum_{b=1}^{500} (\hat{Y}_{SS}^{(b)} - \hat{Y}_{SS})^2$.

5. Simulation Study

A simulation study has been conducted to verify if the KNF variance estimate will provide adequate results in the context of the HSS. The objective of the study was to simulate the HSS process 1,000 times to obtain 1,000 estimates which would be used to calculate a Monte Carlo (MC) variance. The MC variance would then be used as a standard to compare with the KNF variance estimates. To achieve this, an area frame has been simulated using data extracted from the long-form questionnaires of the 2001 Canadian Census. Demographic information, which was available for all households from the frame, has been used as calibration variables and variables of interest. A telephone frame has been created by selecting a large sub-sample of the area frame. The frames have then been clustered and/or stratified according to the LFS and the CCHS sampling designs.

For each MC iteration, an LFS and a CCHS sample have been selected according to their respective sampling designs. Both samples were drawn to get a two-month collection period. The households from these samples have been fed to the master sample. A smaller survey sample has then been selected from the master sample. It consisted of about a quarter of the master sample households from which one person has been sampled. The weighting process described in section 4 has been applied to the smaller survey sample. Estimates and KNF variance estimates have then been produced for two variable of interest: income and labour force status, which is either the person was employed, unemployed or not in labour force.

Table 1 presents the results from the simulation study. For each variable of interest, the average estimate over the 1,000 simulations are shown, followed by the average KNF CV and the ratio between the average KNF variance and the MC variance, which is the most interesting result to analyze.

Table 1: Results from the Simulation Study

Variable	Avg. Estimate	Avg. KNF CV	Ratio Avg. KNF Var. vs MC Var.
Income ($)	8,870,677,134	13.1%	1.058
Labour Force Status			
Employed, worked	182,483	14.3%	1.046
Employed, absent	8,094	100.1%	1.030
Unemployed	54,303	36.1%	1.019
Not in Labour Force	180,565	12.9%	1.009

The ratios between the KNF variance and the MC variance are slightly larger than 1, which usually happens when the correction for a "without replacement" design is not used, like in this case. However, they are close enough to 1 to conclude that the KNF variance estimates will properly estimate the real variance in the context of the HSS.

During previous tests of the KNF variance, it was observed that, if the second-phase sample size is small in some second-phase strata, some bootstrap samples could have no selected households in these strata, which could result in undesired variance estimates (since the bootstrap weight adjustments presented in section 4.3 would involve a division by 0).

6. On-Going HHS-Related Projects and Conclusion

A prototype of the master sample was built in early 2008 and a pilot survey using the master sample methodology has been conducted to test this component of the HSS on the field. The results of this pilot survey are currently being analyzed. If the results are satisfactory, the master sample could go under production as early as in 2009.

Although the HSS has other components that are still in development or already in use at Statistics Canada, the master sample, assuming its implementation in production, will provide a methodological framework that will allow smaller household surveys to select their sample from a survey frame and potentially benefit from the household information already collected by large surveys to produce weights and to calculate proper variance estimates.

Acknowledgements

The author would like to thank Statistics Canada's New Household Survey Strategy Methodology Committee and the reviewers for their contribution to this paper.

References

Hartley, H. O. (1962), "Multiple Frame Surveys", 1962 Proceedings of American Statistical Association, the Social Statistics Section, pp. 203–206.

Kim, J.K., Navarro, A. and Fuller, W.A. (2006). "Replication Variance Estimation for Two-Phase Stratified Sampling", Journal of the American Statistical Association, Vol. 101, No. 473, pp. 312-320.

Rao, J.N.K. and Wu, C.F.J. (1988). "Resampling Inference With Complex Survey Data", Journal of the American Statistical Association, Vol. 83, No. 401, pp. 231-241.

Statistics Canada (2008a). "Methodology of the Canadian Labour Force Survey", Catalogue no. 71-526-X.

Statistics Canada (2008b). "Canadian Community Health Survey (CCHS) 2007 Microdata files User guide", Internal publication.