Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2005 : Methodological Challenges for Future Information needs

2005

Statistics Canada   Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2005
Methodological Challenges for Future Information Needs

INDIRECT TWO-PHASE SAMPLING: APPLYING IT TO QUESTIONNAIRE FIELD-TESTING

M.A. Hidiroglou and Pierre Lavallée[1]

ABSTRACT

There are several ways to improve data quality. One of them is to re-design and test questionnaires for ongoing surveys. The benefits of questionnaire re-design and testing include improving the accuracy by ensuring the questions collect the required data, as well as decreased response burden. A number of enterprise surveys questionnaires of the Office for National Statistics are currently being redesigned. In this paper we will focus on the questionnaire redesign of the Annual Survey of Hours and Earnings (ASHE).

KEY WORDS: Two-phase sampling; Questionnaire design; Indirect sampling.

1. INTRODUCTION

The Annual Survey of Hours and Earnings (ASHE) is the main source of data on the distribution of earnings in the United Kingdom. ASHE (formally known as the New Earnings Survey prior to 2004) has been run, broadly in the same form, in every year since 1970. It measures the earnings of employees in employment in the UK across the whole economy in April of each year. ASHE is conducted by the Office for National Statistics (ONS) for Great Britain and the Department of Enterprise, Trade and Investment in Northern Ireland.

The major part of the sample of 1% of employees is drawn by Inland Revenue (the UK's government department responsible for collecting income tax), from the Pay As You Earn (PAYE) system by reference to the last two digits of an employee's National Insurance number. A small proportion of the sample is identified, using the same selection criterion, directly by employers who return their information electronically to the ONS. The sample design is effectively a panel of employees. The sample size in recent years has been about 240,000 employees. For more detail concerning the ASHE, see Pont (2003).

Selected numbers of national insurance are matched with the ONS Interdepartmental Departmental Enterprise Register (IDBR) at the enterprise level. Matched enterprises are required to fill in questionnaires for each one of their employees matching those selected from the Inland Revenue PAYE file. Data requested about the employees include their earnings, hours worked and a description of their occupation, to name a few. In what follows, we assume that an employee is linked to a single enterprise.

A two-page questionnaire was used up to 2004 to collect these data from in-scope enterprises. A new six-page questionnaire was tested in 2004 with the objective to use it in 2005. This new questionnaire included a number of additional questions, and its layout was an improvement over the two-page questionnaire. Response rate is a one of possible measurements that can indicate whether there has been or not an improvement with the newly designed questionnaire. The sample was split into two sub-samples to test whether there was a significant difference in response rates between the two questionnaires. Part of the sample was mailed out 5,000 redesigned questionnaires (six-page version), and the remaining part of the sample was mailed out 235,000 two-page questionnaires. Sampling was constrained to ensure that each enterprise (or cluster) in scope to the survey would either receive the two-page or six-page questionnaire for its matched set of employees. As we will see further, variance estimation is further complicated by this constraint.

[1] M.A. Hidiroglou, Office for National Statistics, Cardiff Road Newport, United-Kingdom, NP10 8XG (mike.hidiroglou@ons.gsi.gov.uk); Pierre Lavallée, Social Survey Methods Division, Statistics Canada, Canada, K1A 0T6 (pierre.lavallee@statcan.ca)

The paper is structured as follows. We describe in more detail the procedure for sampling the two questionnaires, and develop the required notation in section 2. We develop the variance of the estimated response for either questionnaire in section 3. Testing whether the response rates between the two questionnaire types are different implies that a covariance term is required. We provide this development in section 4. Finally, we summarize findings in the concluding section.

2. ASSIGNING THE TWO QUESTIONNAIRES

Let U^A denote the frame of *M* employees (Inland Revenue) with the National Insurance numbers. Let the corresponding universe of *N* matching enterprises on the IDBR be denoted as U^B. Each enterprise on the frame U^B can be regarded as a cluster of employees linked to enterprises on frame U^A. A first-phase sample of *m* employees (elements) s^A is selected from U^A using Bernoulli sampling. In what follows, the estimated variances are derived assuming that s^A is selected with simple random sampling without replacement from U^A. Employees are identified by the index *k*. The sample s^A is split into two sub-samples s_2^A and s_6^A that are respectively sent the two and six-page questionnaires. The corresponding samples sizes of these sub-samples are denoted as m_2^A and m_6^A.

The allocation of the employees within s^A to the two-sub-samples is such that each enterprise receives a two-page or a six-page questionnaire for its selected employees. The corresponding sample of n^B enterprises (clusters) on the Enterprise Register matched to the elements of the selected National Insurance sample s^A is denoted as s^B, where $s^B = \left\{ i \mid k \in i,\ i \in U^B, k \in s^A \right\}$. The sample s^B is stratified into *H* size strata s_h^B, made up of n_h^B clusters, *h=1,..., H*: $s^B = \bigcup_{h=1}^{H} s_h^B$. A simple random sample s_{6h}^B consisting of n_{6h}^B clusters is selected without replacement from s_h^B. Each element within the selected clusters is mailed a six-page questionnaire. The remaining part of the sample s_{2h}^B consisting of n_{2h}^B clusters is mailed a two-page questionnaire. The corresponding of elements in s^A matched to elements belonging to s_{2h}^B and s_{6h}^B are denoted as s_{2h}^A and s_{6h}^A, respectively, where $s_{2h}^A = \left\{ k \mid k \in i, i \in s_{2h}^B, k \in s^A \right\}$ and $s_{6h}^A = \left\{ k \mid k \in i, i \in s_{6h}^B, k \in s^A \right\}$. Note that $s_h^A = s_{2h}^A \cup s_{6h}^A$, $s_h^B = s_{2h}^B \cup s_{6h}^B$, $m_h = m_{2h} + m_{6h}$ and $n_h^B = n_{2h}^B + n_{6h}^B$, *h*=1,...,*H*.

The sampling scheme can be described as an indirect two-phase sample design. It is indirect, because employees are initially selected from the Inland Revenue frame, matched to enterprises on the Enterprise Register, and data are then requested about these employees from the matched enterprises. It is two-phase design because the two-page or six-page questionnaires are sub-samples of the matched enterprises. The two-phase design is a way to embed experiments in survey sampling.

Once the data have been collected, we test whether there is a significant difference in the response rates between the two-page and six-page questionnaires. The response rate to the two-page questionnaire is estimated as $p_2 = \hat{Y}_2 / \hat{M}_2$ where $\hat{Y}_2$ is the overall weighted response and $\hat{M}_2$ is the corresponding weighted number of two-page questionnaires mailed out. The total number of two-page questionnaires mailed out to the *i*-th cluster within the *h*-th stratum (s_{2hi}^B) is m_{2hi}. The set of elements in s^A that match those in clusters s_{2hi}^B is denoted as s_{2hi}^A: these two subsets will be used interchangeably in what follows. The estimated totals $\hat{Y}_2$ and $\hat{M}_2$ are respectively $\hat{Y}_2 = \dfrac{M}{m} \sum_{h=1}^{H} \dfrac{n_h^B}{n_{2h}^B} \sum_{i=1}^{n_{2h}^B} \sum_{k \in s_{2hi}^A} y_{2k}$ and $\hat{M}_2 = \dfrac{M}{m} \sum_{h=1}^{H} \dfrac{n_h^B}{n_{2h}^B} \sum_{i=1}^{n_{2h}^B} m_{2hi}$. The variable y_{2k} indicates response status: it is equal to one if there has been response to the questionnaire, and zero otherwise. The response rate to the six-page questionnaire, denoted as p_6, is estimated in a similar way.

The difference between the responses rates between the two and six-page questionnaire response rate is tested using the *t*-test given by

$$t = \frac{p_2 - p_6}{\sqrt{\mathrm{v}(p_2) - 2\,\mathrm{cov}(p_2, p_6) + \mathrm{v}(p_6)}} \tag{1}$$

3. ESTIMATING THE VARIANCE OF p_2

In this section, we limit ourselves to the estimation of the variance for the estimated response rate p_2 for the two-page questionnaire. The development for the variance of the estimated proportion for the six-page questionnaire follows along similar lines. We first linearise the proportion $p_2 = \hat{Y}_2 / \hat{M}_2$ (ratio estimator), where $\hat{Y}_2$ and $\hat{M}_2$ are unbiased estimators of $Y_2 = \sum_{k \in U^A} y_k$ and *M*, respectively. Defining the linearised variable as $z_{2,k} = \left(y_{2,k} - p_2 \right) / \hat{M}_2$, the estimated variance for p_2 is given by $\mathbf{v}(p_2) \doteq \mathbf{v}(\widehat{Z}_2)$ where $\widehat{Z}_2 = \frac{M}{m} \sum_{h=1}^{H} \frac{n_h^B}{n_{2h}^B} \sum_{i=1}^{n_{2h}^B} \sum_{k \in s_{2hi}^A} z_{2,k}$, and $\mathrm{v}(\bullet)$ stands for the variance operator that reflects the sample design.

Let π_{1k} and $\pi_{1k\ell}$ denote the first-order and second-order inclusion probabilities for the first-phase sample s^A. Similarly, let $\pi_{2i|s^A}$ and $\pi_{2ij|s^A}$ be the conditional first-order and second-order inclusion probabilities for the second phase sample $s_2^A = \bigcup_{h=1}^{H} s_{2h}^A$, given the first-phase sample s^A. We use the Sen-Yates-Grundy (SYG) variance estimator for two-phase sampling given in Hidiroglou and Rao (2003) to obtain the estimated variance for $\hat{Z}_2$. This variance estimator is less prone to yield negative variance estimates than the Horvitz-Thompson (HT) form for non-fixed sampling designs. The resulting variance is always non-negative for our design. The two-phase SYG variance estimator is given by

$$\mathrm{v}_{\mathrm{SYG}}\left(\widehat{Z}_2\right) = \mathrm{v}_{2,SYG}\left(\widehat{Z}_1\right) + \mathrm{v}_{\mathrm{SYG}}\left(\widehat{Z}_2 \middle| s^A\right) \tag{2}$$

where $\widehat{Z}_1 = E\left(\widehat{Z}_2 \middle| s^A\right) = \sum_{s^A} z_k / \pi_{1k}$, $\mathrm{v}_{2,SYG}\left(\widehat{Z}_1\right) = \sum\sum_{k<\ell \in s_2^A} \frac{\pi_{1k}\,\pi_{1\ell} - \pi_{1k\ell}}{\pi_{1k\ell}\pi_{2k\ell|s^A}} \left(\frac{z_{2,k}}{\pi_{1k}} - \frac{z_{2,\ell}}{\pi_{1\ell}} \right)^2$, and

$$\mathrm{v}_{\mathrm{SYG}}\left(\widehat{Z}_2 \mid s^A\right) = \sum\sum_{k<\ell \in s_2^A} \frac{\pi_{2k|s^A}\pi_{2\ell|s^A} - \pi_{2k\ell|s^A}}{\pi_{1k\ell}\pi_{2k\ell|s^A}} \left(\frac{z_{2,k}}{\pi_{1k}\pi_{2k|s^A}} - \frac{z_{2,\ell}}{\pi_{1\ell}\pi_{2\ell|s^A}} \right)^2 .$$

The first-order and second-order inclusion probabilities in phase 1 are respectively $\pi_{1k} = m/M$ and $\pi_{1k\ell} = \left\{ m\left(m-1\right) \right\} / \left\{ M\left(M-1\right) \right\}$. The first-order inclusion probabilities in phase 2 are $\pi_{2k|s^A} = n_{2h}^B / n_h^B$ if $k \in s_{2hi}^A$ and $i \in U_h^B$. The second-order inclusion probabilities depend as to which stratum and cluster the elements *k* and ℓ belong to. The different cases are as follows:

i. Same stratum *h* and cluster $i \in U_h^B$: $\pi_{2k\ell|s^A} = n_{2h}^B / n_h^B$ if $k, \ell \in s_{2hi}^A$.

ii. Same stratum *h*, but different clusters: $\pi_{2k\ell|s^A} = \left[n_{2h}^B \left(n_{2h}^B - 1 \right) \right] / \left[n_h^B \left(n_h^B - 1 \right) \right]$ if *i* and $i' \in U_h^B$, $i \neq i'$, $k \in s_{2hi}^A$ and $\ell \in s_{2hi'}^A$.

iii. Different strata $h \neq h'$, and different clusters across the strata: $\pi_{2k\ell|s^A} = \left(n_{2h}^B n_{2h'}^B \right) / \left(n_h^B n_{h'}^B \right)$ if $i \in U_h^B$, $i' \in U_{h'}^B$, $k \in s_{2hi}^A$ and $\ell \in s_{2h'i'}^A$.

The first part of $\mathrm{v}_{\mathrm{SYG}}\left(\widehat{Z}_2\right)$ of the variance given by $\mathrm{v}_{2,\mathrm{SYG}}\left(\widehat{Z}_1\right) = F \sum\sum_{k<\ell \in s_2^A} \frac{1}{\pi_{2k\ell|s^A}} \left(z_{2,k} - z_{2,\ell}\right)^2$, where $F = \frac{M^2\left(1-f^A\right)}{m^2\left(m-1\right)}$ and $f^A = m/M$. Using the appropriate inclusion probabilities given above, the estimated variance expression for $\mathrm{v}_{2,SYG}\left(\widehat{Z}_1\right)$ can be simplified to

$$\mathrm{v}_{2,\mathrm{SYG}}\left(\widehat{Z}_1\right) = F\left\{\begin{array}{l} \sum_{h=1}^{H} \frac{n_h^B}{n_{2h}^B} \frac{\left(n_h^B - n_{2h}^B\right)}{\left(n_{2h}^B - 1\right)} \sum_{i=1}^{n_{2h}^B} \sum_{k \in s_{2hi}^A} m_{hi} \left(z_{2,k} - \bar{z}_{2hi}\right)^2 \\ \qquad + \sum_{h=1}^{H} \frac{n_h^B}{n_{2h}^B} \frac{\left(n_h^B - n_{2h}^B\right)}{\left(n_{2h}^B - 1\right) n_{2h}^B} m_{2h} \sum_{i=1}^{n_{2h}^B} \sum_{k \in s_{2hi}^A} \left(z_{2,k} - \bar{z}_{2h}\right)^2 \\ \qquad + \tilde{m}\left[\sum_{h=1}^{H} \frac{n_h^B}{n_{2h}^B} \sum_{i=1}^{n_{2h}^B} \sum_{k \in s_{2hi}^A} z_{2,k}^2 - \frac{1}{\tilde{m}}\left(\sum_{h=1}^{H} \frac{n_h^B}{n_{2h}^B} \sum_{i=1}^{n_{2h}^B} \sum_{k \in s_{2hi}^A} z_{2,k}\right)^2\right] \end{array}\right\} \tag{3}$$

where $\bar{z}_{2hi} = \sum_{k \in s_{2hi}^A} z_{2,k} / m_{2hi}$ and $\bar{z}_{2h} = \sum_{i=1}^{n_{2h}^B} \sum_{k \in s_{2hi}^A} z_{2,k} / \sum_{i=1}^{n_{2h}^B} m_{2hi}$ are weighted means for the elements, and $\tilde{m} = \left(\sum_{h=1}^{H} \frac{n_h^B}{n_{2h}^B} m_{2h}\right)$. Note that this estimator is always non-negative.

The second part of the variance estimator $\mathrm{v}_{\mathrm{SYG}}\left(\widehat{Z}_2\right)$ is given by

$$\mathrm{v}_{\mathrm{SYG}}\left(\widehat{Z}_2 \middle| s^A\right) = \frac{M^2}{m^2} \sum_{h=1}^{H} \frac{\left(n_h^B\right)^2}{n_{2h}^B} \left(1 - \frac{n_{2h}^B}{n_h^B}\right) \frac{1}{n_{2h}^B - 1} \sum_{i=1}^{n_{2h}^B} \left(z_{2hi} - \frac{\sum_{i=1}^{n_{2h}^B} z_{2hi}}{n_{2h}^B}\right)^2 \tag{4}$$

where $z_{2hi} = \sum_{k \in s_{2hi}^B} z_{2,k}$.

The HT version of the variance is

$$\mathrm{v}_{\mathrm{HT}}\left(\widehat{Z}_2\right) = \mathrm{v}_{2,HT}\left(\widehat{Z}_1\right) + \mathrm{v}_{\mathrm{HT}}\left(\widehat{Z}_2 \middle| s^A\right) \tag{5}$$

The first part of $\mathrm{v}_{\mathrm{HT}}\left(\widehat{Z}_2\right)$ is given by

$$\mathrm{v}_{2,HT}\left(\widehat{Z}_1\right) = F\left\{\begin{array}{l} \sum_{h=1}^{H} \frac{n_h^B}{n_{2h}^B} \frac{\left(n_h^B - n_{2h}^B\right)}{\left(n_{2h}^B - 1\right)} \sum_{i=1}^{n_{2h}^B} \left(z_{2hi} - \frac{1}{n_{2h}^B} \sum_{i=1}^{n_{2h}^B} z_{2hi}\right)^2 \\ \qquad + m\left[\sum_{h=1}^{H} \frac{n_h^B}{n_{2h}^B} \sum_{i=1}^{n_{2h}^B} \sum_{k \in s_{2hi}^A} z_{2,k}^2 - \frac{1}{m}\left(\sum_{h=1}^{H} \frac{n_h^B}{n_{2h}^B} \sum_{i=1}^{n_{2h}^B} \sum_{k \in s_{2hi}^A} z_{2,k}\right)^2\right] \end{array}\right\} \tag{6}$$

The variance estimators $\mathrm{v}_{2,HT}\left(\widehat{Z}_1\right)$ and $\mathrm{v}_{2,SYG}\left(\widehat{Z}_1\right)$ are similar. However, $\mathrm{v}_{2,HT}\left(\widehat{Z}_1\right)$ may be negative, and this never the case for $\mathrm{v}_{2,SYG}\left(\widehat{Z}_1\right)$. They are exactly equal if $\tilde{m} = m$ and the number of elements m_{2hi} in each cluster is the same. The second component of (5), given by $\mathrm{v}_{\mathrm{HT}}\left(\widehat{Z}_2 \middle| s^A\right)$, is the same as the one for $\mathrm{v}_{\mathrm{SYG}}\left(\widehat{Z}_2 \middle| s^A\right)$ in (2).

4. ESTIMATING THE COVARIANCE BETWEEN p_2 AND p_6

The population difference $P_d = P_2 - P_6$ where $P_2 = Y_2 / M$ and $P_6 = Y_6 / M$ is estimated by $p_d = p_2 - p_6$. The variance of the difference is obtained using the linearised versions of p_2 and p_6 given by $z^*_{2,k} = \left(y_{2,k} - P_2\right)/M$ and $z^*_{6,k} = \left(y_{6,k} - P_6\right)/M$, respectively. Conditioning on s^A and using Tam (1984)'s results on covariance of overlapping samples, it can be shown that the variance of p_d is given by

$$V(p_d) = M^2\left(1 - \frac{m}{M}\right)\frac{S_d^2}{m} + E\left[\frac{M^2}{m^2}\sum_{h=1}^{H} n_h^B\left(\frac{n_{6h}^B}{n_{2h}^B}\widetilde{S}_{2h}^2 + \frac{n_{2h}^B}{n_{6h}^B}\widetilde{S}_{6h}^2 - 2\widetilde{S}_{26h}\right)\right] \tag{7}$$

where $S_d^2 = \frac{1}{M-1}\sum_{k=1}^{M}(d_k^* - \overline{D}^*)^2$, $\overline{D}^* = \frac{1}{M}\sum_{k=1}^{M} d_k^*$, $d_k^* = z^*_{2,k} - z^*_{6,k}$, and

$$\widetilde{S}_{26h} = \frac{1}{n_h^B - 1}\sum_{i \in s_h^B}\left(z^*_{2hi} - \sum_{i \in s_h^B} z^*_{2hi} / n_h^B\right)\left(z^*_{6hi} - \sum_{i \in s_h^B} z^*_{6hi} / n_h^B\right) \tag{8}$$

with $z^*_{2hi} = \sum_{i \in s^B_{2hi}} z^*_{2k}$ and $z^*_{6hi} = \sum_{i \in s^B_{6hi}} z^*_{6k}$.

Both the estimation of S_d^2 and $\widetilde{S}_{26h}$ require that both y_{2k} and y_{6k} be known for the same unit. This is not possible as each enterprise fills up a single version of the questionnaire for each one of its designated employees. An imputed value of y_{2k} can be used to match to against the y_{6k} response. Such procedures have been proposed by Dumais and Lavallée (1990), and more recently by Van Brakel and Binder (2001). In our context, since ASHE is a panel survey, the value is generated by assuming that the response pattern is the same as the previous year. Denoting this value as y_{2k}^{IMP} , where $k \in s^B_{6hi}$, one can generate the required variables to estimate y_{2k} and y_{6k} . Hence $V(p_d)$ can now either be estimated using the SYG given by (2) or the HT approach given in Section 3.

The estimators of S_d^2 and $\widetilde{S}_{26h}$ require the linearised versions of p_2 and p_6 defined by $z_{2,k} = \left(y_{2,k} - p_2\right)/\hat{M}_2$ and $z_{6,k} = \left(y_{6,k} - p_6\right)/\hat{M}_6$, respectively. The corresponding linearised difference is $d_k = z_{2,k} - z_{6,k}$. Using the HT version, S_d^2 is estimated by

$$\hat{S}_d^2 = \frac{1}{(m-1)}\left\{\sum_{h=1}^{H}\frac{n_h^B}{n_{6h}^B}\sum_{i=1}^{n_{6h}^B}\sum_{k \in s^B_{6hi}}^{i} d_{hik}^2 - \frac{1}{m}\left(\sum_{h=1}^{H}\frac{n_h^B}{n_{6h}^B}\sum_{i=1}^{n_{6h}^B} d_{hi}\right)^2 \right.$$

$$\left. + \frac{1}{m}\sum_{h=1}^{H}\frac{n_h^B(n_h^B - n_{6h}^B)}{n_{6h}^B(n_{6h}^B - 1)}\left(\sum_{i=1}^{n_{6h}^B} d_{hi}^2 - \frac{\left(\sum_{i=1}^{n_{6h}^B} d_{hi}\right)^2}{n_{6h}^B}\right)\right\} \tag{9}$$

where $d_{2hi} = \sum_{k \in s^B_{2hi}} d_{2,k}$. Similarly, the estimator of $\widetilde{S}_{26h}$ is given by

$$\hat{\widetilde{S}}_{26h} = \frac{1}{n_{6h}^B - 1}\sum_{i=1}^{n_{6h}^B}\left(z_{2,hi}^{IMP} - \frac{\sum_{i=1}^{n_{6h}^B} z_{2hi}^{IMP}}{n_{6h}^B}\right)\left(z_{6,hi} - \frac{\sum_{i=1}^{n_{6h}^B} z_{6hi}}{n_{6h}^B}\right) \tag{10}$$

where $z_{2hi}^{IMP} = \sum_{k \in s_{2hi}^B} z_{2,k}^{IMP}$ and $z_{6hi} = \sum_{k \in s_{2hi}^B} z_{6,k}$.

5. CONCLUSION

Variance computations were complicated given the embedded nature of the survey design, and its associated constraints. Using the *t*-test given by (1), we found that the response rates were slightly better for the six-page questionnaire than for the two- page questionnaire. We opted for a four-page version of the six-page questionnaire in 2005, as it was cheaper to process. All the questions in the six-page questionnaire were transferred to the four-page questionnaire. Recent results obtained using the new four-page questionnaire clearly demonstrated its improvements in terms of data quality.

REFERENCES

Dumais, J., and Lavallée, P. (1990), "Une approche pour l'échantillonnage et l'estimation d'enquêtes trimestrielles: l'Enquête sur le camionnage pour compte d'autrui", *Actes du colloque sur les méthodes et domaines d'application de la statistique,* Bureau de la statistique du Québec, Québec, 1990, pp 51-58.

Hidiroglou, M.A. and Rao, J.N.K. (2003), "Variance Estimation in Two-Phase Sampling", *Proceedings of Symposium 2003*,*Challenges in Survey Taking for the Next Decade*, Statistics Canada, Catalog No. 11-522-XIE, Ottawa.

Pont, M. (2003), "Redesigning the UK's Annual Structural Earning Survey", paper presented at the *Federal Committee on Statistical Methods*, Washington, D.C., November 2003.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Tam, S.M. (1984), "On Covariances From Overlapping Samples", *The American Statistician*, Vol. 38, No. 4, pp. 288-289.

Van den Brakel, J. and Binder, D.A. (2001), "Variance Estimation for Experiments Embedded in Complex Sampling Schemes", *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 805-810.