Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Simulation-based randomized systematic PPS sampling under substitution of units

by Mary E. Thompson and Changbao Wu

June 2008

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 2008
Vol. 34, No. 1, pp. 3-10
Statistics Canada, Catalogue No. 12-001-X

Simulation-based randomized systematic PPS sampling under substitution of units

Mary E. Thompson and Changbao Wu [1]

Abstract

The International Tobacco Control (ITC) Policy Evaluation Survey of China uses a multi-stage unequal probability sampling design with upper level clusters selected by the randomized systematic PPS sampling method. A difficulty arises in the execution of the survey: several selected upper level clusters refuse to participate in the survey and have to be replaced by substitute units, selected from units not included in the initial sample and once again using the randomized systematic PPS sampling method. Under such a scenario the first order inclusion probabilities of the final selected units are very difficult to calculate and the second order inclusion probabilities become virtually intractable. In this paper we develop a simulation-based approach for computing the first and the second order inclusion probabilities when direct calculation is prohibitive or impossible. The efficiency and feasibility of the proposed approach are demonstrated through both theoretical considerations and numerical examples. Several R/S-PLUS functions and codes for the proposed procedure are included. The approach can be extended to handle more complex refusal/substitution scenarios one may encounter in practice.

Key Words: Inclusion probability; Horvitz-Thompson estimator; Rao-Sampford method; Relative bias; Unequal probability sampling without replacement.

1. Introduction

Construction of survey weights is the first critical step in analyzing complex survey data. It starts with the calculation of the first order inclusion probabilities, which is often straightforward if the original sampling design is well executed without any alterations and/or modifications. For instance, if the sample units are selected with inclusion probability (π) proportional to size (PPS or πps), then the inclusion probabilities are readily available from a simple re-scaling of the size variable. Among existing unequal probability without replacement PPS sampling procedures which are applicable for arbitrary fixed sample sizes, the randomized systematic PPS sampling method is the simplest one to implement. The procedure was first described in Goodman and Kish (1950) as a controlled selection method, and was refined by Hartley and Rao (1962) who studied the important and yet difficult problem of how to compute the second order inclusion probabilities. Let $x_i, i = 1, 2, \ldots, N$ be the values of the known size variable, where N is the total number of units in the population. Let $z_i = x_i / X$ where $X = \sum_{i=1}^{N} x_i$ and assume $nz_i < 1$ for all i. The randomized systematic PPS sampling procedure is as follows: Arrange the N population units in a random order and let $A_0 = 0$ and $A_j = \sum_{i=1}^{j} (nz_i)$ be the cumulative totals of nz_i in that order so that $0 = A_0 < A_1 < \ldots < A_N = n.$ Let u be a uniform random number over [0, 1]. The n units to be included in the sample are those with indices j satisfying $A_{j-1} \le u + k < A_j$ for $k = 0, 1, \ldots, n-1.$ Let s be the set of n sampled units and $\pi_i = P(i \in s)$ be the first order inclusion probabilities. The randomized systematic PPS sampling procedure satisfies the condition

$$\pi_i = nz_i, \ i = 1, 2, \ldots, N. \tag{1.1}$$

Several other without replacement sampling procedures which satisfy (1.1) for an arbitrary fixed sample size n were also proposed in the literature, including the well-known Rao-Sampford unequal probability sampling method (Rao 1965; Sampford 1967) and those of Chao (1982), Chen, Dempster and Liu (1994), Tillé (1996) and Deville and Tillé (1998), among others.

The extensive research work on PPS sampling methods was largely stimulated by the Horvitz-Thompson (HT) estimator $\hat{T} = \sum_{i \in s} y_i / \pi_i$ for the population total $T = \sum_{i=1}^{N} y_i$ of a study variable y. The HT estimator is extremely efficient when y is highly correlated with the size variable x and the sampling procedure satisfies (1.1). It is the unique design unbiased estimator among the class of linear estimators $\sum_{i \in s} w_i y_i$ for T if the weights w_i depend only on i.

While a PPS sampling procedure can be desirable from a theoretical point of view, it is often difficult and/or sometimes impossible to execute due to practical constraints and limitations. Certain modifications and compromises will have to be made. The modified design, however, will no longer satisfy condition (1.1). Direct calculation of the final inclusion probabilities often becomes difficult or even impossible. Among common problems arising from survey practice which require alteration of the original sampling

1. Mary E. Thompson, Department of Statistics and Actuarial Science, University of Waterloo. E-mail: methomps@uwaterloo.ca; Changbao Wu, Department of Statistics and Actuarial Science, University of Waterloo. E-mail: cbwu@uwaterloo.ca.

design, units refusal and substitution of units are the most frequently encountered ones. The scenario is well illustrated by the following example.

The International Tobacco Control (ITC) Policy Evaluation Survey of China (ITC China Survey) uses a multi-stage unequal probability sampling design for the selection of adult smokers and nonsmokers from seven cities. Each city has a natural hierarchical administrative structure

City → Street District→ Residential Block→ Household → Individual

which was conveniently integrated into the sampling design. At the upper levels, the randomized systematic PPS sampling method is used to select ten street districts from each city, with probability proportional to the population size of the district, and then two residential blocks are selected within each selected district, again using the randomized systematic PPS sampling method, with probability proportional to the population size of the block. Households and individuals within households are further selected, using a modified simple random sampling method. The original plan was to select 40 adult smokers and 10 adult nonsmokers from each of the 20 residential blocks, making the final sample with 800 smokers and 200 nonsmokers for each city.

A difficulty, however, arises in the execution of the survey: several selected upper level clusters (first Street Districts and then Residential Blocks) have refused to participate in the survey, due to time conflict with other activities or unavailability of human resources. These refusing clusters have to be replaced by substitute units, selected from units not included in the initial sample; one possibility is to use once again the randomized systematic PPS sampling method, to achieve the targeted overall sample size.

Under multi-stage sampling designs such as the one used for the ITC China survey, first order inclusion probabilities for individuals selected in the final sample can be calculated by multiplying the inclusion probabilities of units at different stages. When the randomized systematic PPS sampling method is modified due to substitution of units at a certain stage, the condition (1.1) no longer holds for the final sample at that stage. The first order inclusion probabilities under such a scenario are very difficult to calculate and the second order inclusion probabilities become virtually intractable. In Appendix A, we provide a method of direct calculation (5.2) for the π_i when both the initial and the substitute samples are selected using the randomized systematic PPS sampling, assuming random refusal from the initial sample and no refusal from the substitute sample. The expression is valid conditional on the number of refusals and the population order used (after randomization) for the selection of the initial sample. It is apparent that even under such restrictive conditions and assumptions, the expression itself becomes computationally unfriendly with a not-so-large sample size.

In this paper we demonstrate, through both theory and numerical examples, that the first and the second order inclusion probabilities can be accurately estimated through Monte Carlo simulations when complete design information is available. Our numerical examples are motivated by the ITC China survey for which the randomized systematic PPS sampling serves as a baseline method but our theoretical results and the general methodology apply to other unequal probability without replacement sampling procedures as well. Section 2 presents results on the accuracy of simulation based methods. Numerical examples and comparisons are given in Section 3. Several R/S-PLUS functions and codes for the proposed procedure, originally developed for the ITC China survey, are included in Appendix C. Some additional remarks are given in Section 4.

2. Properties of simulation-based methods

When calculation of exact inclusion probabilities is impossible or prohibitive but complete design information is available, Monte Carlo simulation methods can easily be used to obtain estimates of the inclusion probabilities. Denote the completely specified probability sampling design by *p*. The simulation-based method is straightforward: select *K* independent samples, all following the same sampling design *p*; let M_i be the number of samples which include unit *i*. Then the first order inclusion probability $\pi_i = P(i \in s)$ can be estimated by $\pi_i^* = M_i / K$. For a particular *i*, the M_i follows a binomial distribution and the π_i^* satisfies $E(\pi_i^*) = \pi_i$ and $\mathrm{Var}(\pi_i^*) \le (4K)^{-1}$. Suppose for instance that we can afford to take *K* as big as 25×10^6, then $P(|\pi_i^* - \pi_i| < 0.001) \ge 0.99$ for any given π_i.

A more relevant measure of the accuracy of simulation-based methods is the performance of the Horvitz-Thompson estimator using the simulated inclusion probabilities. Let $\hat{T} = \sum_{i \in s} y_i / \pi_i$ and $\tilde{T} = \sum_{i \in s} y_i / \pi_i^*$. For a given sample, the relative bias of using $\tilde{T}$ in place of $\hat{T}$ is defined as $(\hat{T} - \tilde{T}) / \hat{T}$. Without loss of generality, we assume $y_i \ge 0$ for all *i*. It is shown in Appendix B that for any $\varepsilon > 0$ and the given sample *s*,

$$P\left(\frac{|\hat{T} - \tilde{T}|}{\hat{T}} \le \varepsilon\right) \ge 1 - \frac{2(1+\varepsilon^2)}{K\varepsilon^2}\left(\sum_{i \in s} \frac{1}{\pi_i} - n\right). \qquad (2.1)$$

Note that $\sum_{i \in s} (1/\pi_i)$ is the Horvitz-Thompson estimator of the population size *N*, a practical lower bound for $P(|\hat{T} - \tilde{T}| / \hat{T} \le \varepsilon)$ with a small ε is given by

$$\Delta = 1 - \frac{2(N-n)}{K\varepsilon^2}. \qquad (2.2)$$

If one requires that $\varepsilon = 0.01$ and $\Delta = 0.98$, then for $N - n = 100$ the (theoretical) number of independent samples required for the simulation is $K = 10^8$. Since the lower bound given by (2.1) is conservative, and valid for any response variable, one would expect that a smaller K with values around 10^7 or even 10^6 should work well for most practical scenarios where $N - n \le 100$. This is supported by numerical examples presented in Section 3.

Estimation of the second order inclusion probabilities $\pi_{ij} = P(i, j \in s)$ imposes no additional difficulty except that the total number of simulated samples, K, required to achieve the same level of relative accuracy as for the first order case is bigger. Let M_{ij} be the number of simulated samples among the K independent samples which include both i and j. Let $\pi_{ij}^* = M_{ij}/K$ be the estimate for π_{ij}. Suppose the goal is to estimate a quadratic population quantity

$$Q = \sum_{i=1}^{N} \sum_{j=1}^{N} q(y_i, y_j).$$

The Horvitz-Thompson type estimators of Q using π_{ij} or π_{ij}^* are respectively given by

$$\hat{Q} = \sum_{i \in s} \sum_{j \in s} \frac{q(y_i, y_j)}{\pi_{ij}} \text{ and } \tilde{Q} = \sum_{i \in s} \sum_{j \in s} \frac{q(y_i, y_j)}{\pi_{ij}^*}.$$

Following the same argument as that which leads to (2.1), we can show that

$$P\left(\frac{|\hat{Q} - \tilde{Q}|}{\hat{Q}} \le \varepsilon \right) \ge 1 - \frac{2(1+\varepsilon^2)}{K\varepsilon^2} \left(\sum_{i \in s} \sum_{j \in s} \frac{1}{\pi_{ij}} - n^2 \right). \qquad (2.3)$$

Note that $\sum_{i \in s} \sum_{j \in s} (1/\pi_{ij})$ is a design-unbiased estimator of N^2, a practical lower bound for $P(|\hat{Q} - \tilde{Q}|/\hat{Q} \le \varepsilon)$ is given by $1 - 2(N+n)(N-n)/(K\varepsilon^2)$. Comparing this with Δ given by (2.2), it is apparent that we need a much bigger K to achieve the same lower bound, although in both cases the lower bounds are conservative, and the actual K required can be smaller. On the other hand, second order inclusion probabilities are used for the estimation of second order parameters such as the population variance or the variance of a linear estimator. The desired estimation accuracy is less critical than that for first order parameters such as the population total or mean, and therefore a number in between 10^6 and 10^7 for K should be acceptable for many practical situations.

The most critical issue for simulation-based methods is obviously the feasibility of computational implementation. Among other things, it depends largely on the chosen value of K, the complexity of the sampling design, and the computational power available. If $K = 10^6$ and one would like to have the simulation-based results within ten hours, then it is necessary to take 28 simulated samples for every single second. The randomized systematic PPS sampling is the most efficient unequal probability without replacement sampling procedure in terms of computational implementation. It only involves a simple random ordering and selecting a random starting point. Most other competing procedures involve either rejective methods or complicated sequential selections. It takes much longer to select simulated samples with these methods. A comparison of CPU times for computing the simulated π_i between the randomized systematic PPS sampling and the Rao-Sampford unequal probability sampling design is given in Section 4.

3. Numerical examples

The design information used in this section is adapted from the ITC China survey. The number of Street Districts (top level clusters) in each of the seven cities involved in the survey ranges from $N = 20$ to $N = 120$. Within each city $n = 10$ districts are selected using the randomized systematic PPS sampling method. In the case of refusals, substitute districts are selected from the ones not included in the initial sample, again using the randomized systematic PPS sampling method. For the purpose of illustration we use the design information from the smallest city (*i.e.*, $N = 20$). Additional comments on cases where N is large are given in Section 4.

3.1 First order inclusion probabilities

We first demonstrate the accuracy of the simulated π_i when the exact values of π_i are known. We then investigate the impact of substitution of units on the final π_i and the performance of the Horvitz-Thompson estimator for a population total using the simulated π_i. The simulated inclusion probabilities under substitution of units are compared to those assuming the modified design is still PPS sampling.

Example 1. Simulation-based π_i^* when there is no refusal. In this case the exact values of π_i are given by $\pi_i = nz_i$.

(i) Exact values of π_i:

```
0.5840 0.5547 0.6702 0.5331 0.3085 0.2652 0.3930 0.4180 0.6952 0.3471
0.5993 0.5393 0.8240 0.6868 0.4469 0.2191 0.4237 0.4180 0.7567 0.3163
```

(ii) Simulated π_i^*, $K = 10^5$:

```
0.5828 0.5545 0.6656 0.5339 0.3071 0.2656 0.3929 0.4205 0.6969 0.3474
0.6009 0.5429 0.8227 0.6865 0.4446 0.2186 0.4215 0.4179 0.7569 0.3194
```

(iii) Simulated π_i^*, $K = 10^6$:

```
0.5836 0.5558 0.6701 0.5336 0.3081 0.2654 0.3931 0.4180 0.6950 0.3469
0.5994 0.5394 0.8242 0.6864 0.4469 0.2186 0.4237 0.4172 0.7569 0.3166
```

The simulated π_i^* matches π_i to the second decimal point for $K = 10^5$ and to the third for $K = 10^6$ for most cases.

Example 2. To assess the performance of the Horvitz-Thompson (HT) estimator for a population total using the true π_i and the simulated π_i^* from Example 1, we generated the response variable from the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \ldots, N,$ where x_i is the size variable and ε_i are independent and identically normally distributed with mean 0 and variance σ^2. We considered three populations (three values of σ^2) where the population correlation coefficients between *x* and *y* are respectively 0.3, 0.5 and 0.8. For each of the three populations, $B = 2{,}000$ repeated samples of size $n = 10$ were selected using the randomized systematic PPS sampling, and for each sample three HT estimators were computed using the true π_i, the simulated π_i^* with $K = 10^5$ and the π_i^* with $K = 10^6$, respectively. The results, not reported here to save space, showed that all three HT estimators have relative bias less than 0.04% and almost identical mean squared errors.

Example 3. When there are refusals in the initial PPS sample and substitute units are selected from units not included in the initial sample using the same PPS sampling procedure, there are two questions of interest: (1) how to compute the inclusion probabilities π_i for the final sample; and (2) to what extent the substitution procedure has altered the original PPS sampling design. We can compute the simulated π_i^* and compare them with $\tilde{\pi}_i$ obtained by assuming a PPS sampling after the refusing units are removed from the sampling frame. In simulating the π_i^*, we assume for simplicity that there is no possible refusal from any unit outside the initial sample, and hence there is no refusal among the substitute units. The number of replications *K* is chosen as 10^6 for the simulation. We consider two scenarios where there are three refusing units in the population, and all are among the initial sample of size $n = 10$.

(i) Three large units refuse: Simulated π_i^* (first two rows) versus $\tilde{\pi}_i$ (last two rows) assuming PPS.

```
0.7231 0.6981 0.7947 0.6773 0.4354 0.3811 0.5339 0.5619 0.0000 0.4815
0.7363 0.6826 0.0000 0.8070 0.5919 0.3210 0.5678 0.5615 0.0000 0.4441

0.7560 0.7182 0.8677 0.6901 0.3994 0.3434 0.5088 0.5412 0.0000 0.4494
0.7759 0.6983 0.0000 0.8892 0.5786 0.2837 0.5486 0.5412 0.0000 0.4096
```

(ii) Three small units refuse: Simulated π_i^* (first two rows) versus $\tilde{\pi}_i$ (last two rows) assuming PPS.

```
0.6326 0.6049 0.7167 0.5829 0.0000 0.0000 0.4415 0.4668 0.7406 0.3937
0.6482 0.5901 0.8558 0.7330 0.4965 0.0000 0.4728 0.4664 0.7976 0.3590

0.6343 0.6025 0.7280 0.5790 0.0000 0.0000 0.4268 0.4540 0.7550 0.3770
0.6510 0.5858 0.8949 0.7459 0.4854 0.0000 0.4602 0.4540 0.8218 0.3436
```

It is apparent that the sizes of the refusing units have dramatic impact on the distribution of the final inclusion probabilities. If one ignores the alteration of the sampling design due to substitution of units and treats the design as if it is still a PPS sampling, then the inclusion probabilities for large units are inflated and the role of small units is down-played. This trend is more pronounced when there are large units among the refusals, *i.e.*, case (i) where $\pi_{14}^* = 0.8070$ compared to $\tilde{\pi}_{14} = 0.8897$ and $\pi_{16}^* = 0.3210$ to $\tilde{\pi}_{16} = 0.2837$.

3.2 Second order inclusion probabilities

There have been considerable research activities on the randomized systematic PPS sampling, mainly for obtaining second order inclusion probabilities π_{ij} and variance estimators. Hartley and Rao (1962) derived exact formulas for the π_{ij} when $n = 2$ and $N = 3$ or $N = 4$; Connor (1966) extended the results and derived the exact formula for general *n* and *N*, and the related computational procedure was later implemented in the Fortran language by Hidiroglou and Gray (1980). The procedure is quite heavy as evidenced by the 165 lines of Fortran code.

The most intriguing result is probably the asymptotic approximation to π_{ij} derived by Hartley and Rao (1962). In a recent paper Kott (2005) showed that the variance estimator of a Horvitz-Thompson estimator based on the Hartley-Rao approximation not only performs well under the design-based framework but also has good model-based properties. The Hartley-Rao approximation was initially derived based on the assumption that *n* is fixed and *N* is large and is correct to the order of $O(N^{-4})$ (Hartley and Rao 1962: Equation (5.15) on page 369). In a private conversation with J.N.K. Rao during the 23rd International Methodological Symposium of Statistics Canada, he pointed out that the approximation is still valid even if *n* is large, as long as n/N is small. For cases where *N* is not large and/or n/N is not small, such as the ITC China survey example considered here, the goodness of the Hartley-Rao approximation has not been documented.

When the randomized systematic PPS sampling procedure is altered due to substitution of units, it is virtually impossible to derive the second order inclusion probabilities or some sort of approximations. With the simulation-based approach, however, it remains straightforward to obtain very reliable estimates of the π_{ij} through a large number of simulated samples, given that the altered sampling procedure is completely specified. In what follows we examine the performance of variance estimators using the simulated π_{ij}^* when there is no alteration to the randomized systematic PPS sampling procedure. In this case $\pi_i = nz_i$ and the Hartley-Rao approximation $\tilde{\pi}_{ij}$ to π_{ij} can also take part in the comparison.

Example 4. We first compare π_{ij}^* to $\tilde{\pi}_{ij}$ for each of the individual entries. To save space, we only present the results for $i = 1, \ldots, 5$ and $j = 1, \ldots, 10,$ which are sufficient to

show the general picture. The Hartley-Rao approximation $\tilde{\pi}_{ij}$ is very close to the simulated π^*_{ij}, matching to the second decimal point for the majority of the entries. This is clearly an interesting observation given that $N = 20$ and $n = 10$.

(i) Simulated π^*_{ij}, $K = 10^6$:

```
0.0000 0.3121 0.3821 0.2975 0.1669 0.1442 0.2116 0.2249 0.3975 0.1873
0.3121 0.0000 0.3623 0.2816 0.1590 0.1372 0.2025 0.2141 0.3766 0.1784
0.3821 0.3623 0.0000 0.3469 0.1899 0.1640 0.2483 0.2659 0.4586 0.2153
0.2975 0.2816 0.3469 0.0000 0.1523 0.1312 0.1938 0.2061 0.3606 0.1717
0.1669 0.1590 0.1899 0.1523 0.0000 0.0742 0.1124 0.1197 0.1968 0.0988
```

(ii) Hartley-Rao approximation $\tilde{\pi}_{ij}$:

```
0.0000 0.3079 0.3769 0.2952 0.1668 0.1427 0.2143 0.2286 0.3921 0.1884
0.3079 0.0000 0.3569 0.2795 0.1579 0.1351 0.2029 0.2164 0.3712 0.1784
0.3769 0.3569 0.0000 0.3421 0.1932 0.1654 0.2484 0.2649 0.4544 0.2183
0.2952 0.2795 0.3421 0.0000 0.1514 0.1296 0.1946 0.2075 0.3559 0.1710
0.1668 0.1579 0.1932 0.1514 0.0000 0.0732 0.1099 0.1172 0.2010 0.0966
```

Example 5. For second order inclusion probabilities the main focus is on variance estimation. With fixed sample size, an unbiased variance estimator for the Horvitz-Thompson estimator $\hat{Y}_{\mathrm{HT}} = \sum_{i\in s} y_i/\pi_i$ is given by the well-known Yates-Grundy format,

$$v(\hat{Y}_{\mathrm{HT}}) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2. \qquad (3.1)$$

We consider the three synthetic populations described in *Example 2*. The true variance $V = \mathrm{Var}(\hat{Y}_{\mathrm{HT}})$ is obtained through simulation using $B = 10^5$ simulated samples and is computed as $B^{-1}\sum_{b=1}^{B}(\hat{Y}_b - Y)^2$, where Y is the true population total and $\hat{Y}_b$ is the Horvitz-Thompson estimator of Y computed from the b^{th} simulated sample. Three variance estimators in the form of (3.1), denoted respectively by v_1, v_2 and v_3, are examined, with the π_{ij} in (3.1) being respectively replaced by the Hartley-Rao approximation $\tilde{\pi}_{ij}$, the simulated π^*_{ij} for $K = 10^5$ and the π^*_{ij} for $K = 10^6$. The performance of these estimators is measured through the simulated relative bias $\mathrm{RB} = B^{-1}\sum_{b=1}^{B}(v^{(b)} - V)/V$ and the simulated instability $\mathrm{INST} = \{B^{-1}\sum_{b=1}^{B}(v^{(b)} - V)^2\}^{1/2}/V$, where $v^{(b)}$ is the variance estimate computed from the b^{th} sample, using another set of $B = 10^5$ independent samples. The results are summarized in Table 1 below. The three populations are indicated by the correlation coefficient ρ between y and x.

Table 1 Relative bias and instability of variance estimators

	RB(%)			INST		
Population	v_1	v_2	v_3	v_1	v_2	v_3
$\rho = 0.30$	6.1%	1.4%	-0.3%	0.66	0.65	0.65
$\rho = 0.50$	4.3%	2.5%	-1.1%	0.42	0.44	0.42
$\rho = 0.80$	2.6%	1.2%	-0.2%	0.61	0.60	0.60

In terms of relative bias, all three variance estimators are acceptable, with the one (v_1) based on the Hartley-Rao approximation $\tilde{\pi}_{ij}$ having the largest bias. For variance estimators using the simulated π^*_{ij}, increasing the value of K from 10^5 (*i.e.*, v_2) to 10^6 (*i.e.*, v_3) makes the bias to be negligible, although the one with $K = 10^5$ is clearly acceptable in practice. All three versions of the variance estimator have similar measures in terms of instability.

4. Some additional remarks

In theory, the simulation-based method for computing inclusion probabilities is applicable to any sampling design, as long as the complete design information is available. It is an effective approach to handling more complex substitution scenarios or other types of modifications to the original design. In the ITC China survey, one of the refusing units has to be substituted by a unit from a particular region of the city due to workload constraints and field work restrictions. In a Canadian national survey of youth, there were second and third round refusing units (schools) and hence substitute units before achieving the targeted sample size. As pointed out by an Associate Editor, a similar situation was also reported in the 57th Round of the National Sample Survey Organization, Government of India (www.mospi.gov.in) where a modification was made to the circular systematic sampling with probability proportional to size in order to select two distinct sub-samples. Gray (1973) described a method on increasing the sample size (number of psu's) when the initial sample was selected by the randomized systematic PPS method. Calculation of second order inclusion probabilities under the proposed procedure is difficult even for a very small sample size. In all these cases analytic solutions to the inclusion probabilities are either difficult to use or not available but the simulation-based approach can be applied without any extra difficulty.

The recent paper by Fattorini (2006) discussed the use of the simulation-based method for spatial sampling where the units are selected sequentially. When a PPS sampling design is altered due to one or more rounds of substitution of units, the modified design can also be viewed as sequential. Our theoretical results on the accuracy of simulation-based methods, however, are different from those of Fattorini. We have used a conditional argument and proposed to assess the performance of the estimator using the simulated inclusion probabilities for a given sample, which is of interest for practical applications.

The central issue related to simulation-based methods is the feasibility of computational implementation. The randomized systematic PPS sampling has a major advantage in computational efficiency. The Rao-Sampford unequal probability sampling method (Rao 1965; Sampford 1967), for instance, is another popular PPS sampling procedure. It has several desirable features such as closed form expressions for the second order inclusion probabilities and

is more efficient than the randomized systematic PPS sampling (Asok and Sukhatme 1976). The following is a comparison of CPU times between the randomized systematic PPS sampling and the Rao-Sampford PPS sampling for simulating the first order inclusion probabilities. The sample size is fixed at $n = 10$ and the number of simulated samples is $K = 10^6$. The results are obtained using R on a dual-processor unix machine.

N	Systematic PPS	Rao-Sampford PPS
200	4.7 hours	7.5 hours
100	2.5 hours	5.0 hours
50	1.6 hours	4.4 hours
20	1.2 hours	8.9 hours

It is interesting to note that, although in general the Rao-Sampford procedure takes longer time to obtain the results, it takes much longer for the case of $N = 20$. This is because the Rao-Sampford method uses a rejective procedure and it usually takes many rejections to arrive at a final sample when the sampling fraction n/N is large. The randomized systematic PPS sampling, on the other hand, is not affected by this and the simulation-based method can provide results with desired accuracy in a timely fashion for $N = 400$ or even bigger. Several R/S-PLUS functions and major codes for the proposed approach are included in Appendix C and are applicable to other substitution scenarios after minor modifications.

One of the reasons for the use of the randomized systematic PPS sampling in selecting upper level clusters in the ITC China survey is that the final design is self-weighting. An interesting question arises when there are refusals: how to select the substitute units such that the final altered sampling design is still (approximately) self-weighting? In some other circumstances such as rotating samples, this is achievable; see, for instance, Fellegi (1963). How to accomplish this goal with the ITC China survey design is currently under investigation.

Acknowledgements

This research is partially supported by grants from the Natural Sciences and Engineering Research Council of Canada. The authors also thank the International Tobacco Control (ITC) Policy Evaluation Project and the ITC China Survey Project for assistance and support. The ITC project is supported in part by grants from the National Cancer Institute of the United States (P50 CA11236) Roswell Park Transdisciplinary Tobacco Use Research Center and the Canadian Institutes of Health Research (57897). Funding for the ITC China project is provided by the Ministry of Health and the Ministry of Finance of China.

Appendix A

A direct calculation under random refusal

Under the randomized systematic PPS sampling design and assuming random refusal, it is possible in principle to calculate the inclusion probabilities under a substitution rule directly. The starting point is to enumerate all possible initial samples and their probabilities based on the particular population order used to select the initial sample.

Recall that $A_0 = 0, A_j = \sum_{i=1}^{j} (nz_i)$ and $A_N = n$. For a chosen uniform starting value $u \in [0, 1]$, unit j is to be selected if

$$A_{j-1} \le u + k < A_j \tag{5.1}$$

for some $k = 0, 1, \ldots, n-1$. Let k_j be the largest integer less than A_j, and let the remainder e_j be given by $e_j = A_j - k_j$. Let $0 < e_{(1)} \le e_{(2)} \le \ldots \le e_{(N)}$ be the order statistics of the remainders, and let $k_{(1)}, \ldots, k_{(N)}$ be the corresponding k_j's. Note that $e_{(N)} = 1$. We could then generate N possible samples $s_1, \ldots, s_N$ with respective probabilities

$$e_{(1)}, e_{(2)} - e_{(1)}, \ldots, e_{(N)} - e_{(N-1)},$$

some of which may be 0. We begin by generating s_1. From each $j = 1, \ldots, N$, put j in s_1 if $A_{j-1} \le k < A_j$ for some $k = 0, 1, \ldots, n-1$, *i.e.*, s_1 is selected using $u = 0$ in (5.1). As we move u from 0 to 1, different possible samples can be identified sequentially. Now given $s_1, \ldots, s_m$, let s_{m+1} be the same as s_m except that the $(k_{(m)} + 1)^{\text{th}}$ element is advanced by 1. For example, suppose $n = 4$ and $s_m = \{1, 3, 6, 9\}$, and suppose $k_{(m)} = 0$, then $s_{m+1} = \{2, 3, 6, 9\}$. On the other hand, if $k_{(m)} = 2$, then $s_{m+1} = \{1, 3, 7, 9\}$. The sample s_{m+1} will have probability $e_{(m+1)} - e_{(m)}$.

By construction, $\pi_i = nz_i$ for $i = 1, \ldots, N$. If only first and second order inclusion probabilities are desired, a similar but simpler algorithm can be used to calculate the second order inclusion probabilities directly, conditional on the initial order. However, for applications where the probabilities of all samples are needed, the sample generation algorithm can be carried out. For example, for small populations, it is then also possible to calculate the first order inclusion probabilities when there is refusal and substitution. Suppose we first select a sample of size n with randomized systematic PPS sampling. Suppose n_1 of these agree to respond and an additional $n_2 = n - n_1$ are selected, again using randomized systematic PPS sampling, from those units not sampled the first time. Assume for simplicity that refusal in the first sample occurs at random, and that there is no refusal in the second substitute sample. Note that this is a different assumption from the one used in Example 3, where the set of refusals is considered to be non-random. The

inclusion probability for unit *i*, conditional on the assumed initial population order, is

$$nz_i \times \frac{n_1}{n} + \sum_{m:i \notin s_m} p_1(s_m) \frac{n_2 z_i}{\sum_{j:j \notin s_m} z_j}. \qquad (5.2)$$

The outer sum is taken over all samples s_m of size n, generated according to the procedure described above but without having unit i, with probabilities $p_1(s_m) = e_{(m)} - e_{(m-1)}$. The inner sum involved in the denominator is taken over all j not included in s_m from the outer sum. The unconditional inclusion probability can be obtained by appropriate averaging over all population orders which give distinct values. Clearly this is feasible only when the population is small, or when z takes a small number of values.

Appendix B

Derivation of (2.1)

In this appendix we show that for any $\varepsilon > 0$ and a given sample s,

$$P\left(\frac{|\hat{T} - \tilde{T}|}{\hat{T}} \le \varepsilon\right) \ge 1 - \frac{2(1+\varepsilon^2)}{K\varepsilon^2}\left(\sum_{i \in s} \frac{1}{\pi_i} - n\right),$$

where $\hat{T} = \sum_{i \in s} y_i / \pi_i, \tilde{T} = \sum_{i \in s} y_i / \pi_i^*,$ and π_i^* are the simulated first order inclusion probabilities based on K independent samples. Noting that $E(\pi_i^*) = \pi_i$ and $\mathrm{Var}(\pi_i^*) = \pi_i(1-\pi_i)/K,$ by Chebyshev's inequality we have $P(|\pi_i^* - \pi_i| > c) \le \pi_i(1-\pi_i)/(Kc^2)$ for any $c > 0$. It follows that

$$\begin{aligned}
&P\left(\frac{|\pi_i^* - \pi_i|}{\pi_i^*} > \varepsilon\right) \\
&= P(\pi_i^* - \pi_i > \pi_i^* \varepsilon) + P(\pi_i^* - \pi_i < -\pi_i^* \varepsilon) \\
&= P(\pi_i^* - \pi_i > \varepsilon\pi_i/(1-\varepsilon)) + P(\pi_i^* - \pi_i < -\varepsilon\pi_i/(1+\varepsilon)) \\
&\le P(|\pi_i^* - \pi_i| > \varepsilon\pi_i/(1-\varepsilon)) + P(|\pi_i^* - \pi_i| > \varepsilon\pi_i/(1+\varepsilon)) \\
&\le \frac{(1-\varepsilon)^2 \pi_i(1-\pi_i)}{K\varepsilon^2\pi_i^2} + \frac{(1+\varepsilon)^2 \pi_i(1-\pi_i)}{K\varepsilon^2\pi_i^2} \\
&= \frac{2(1+\varepsilon^2)}{K\varepsilon^2}\left(\frac{1}{\pi_i} - 1\right).
\end{aligned}$$

If $y_i \ge 0$ for all i, then

$$|\hat{T} - \tilde{T}| \le \sum_{i \in s} \frac{y_i}{\pi_i} \frac{|\pi_i^* - \pi_i|}{\pi_i^*} \le \max_{i \in s}\left\{\frac{|\pi_i^* - \pi_i|}{\pi_i^*}\right\} \hat{T}.$$

For any $\varepsilon > 0$ and the given sample s,

$$\begin{aligned}
P\left(\frac{|\hat{T} - \tilde{T}|}{\hat{T}} \le \varepsilon\right) &\ge P\left(\max_{i \in s}\left\{\frac{|\pi_i^* - \pi_i|}{\pi_i^*}\right\} \le \varepsilon\right) \\
&\ge 1 - \sum_{i \in s} P\left(\frac{|\pi_i^* - \pi_i|}{\pi_i^*} > \varepsilon\right) \\
&\ge 1 - \frac{2(1+\varepsilon^2)}{K\varepsilon^2}\left(\sum_{i \in s} \frac{1}{\pi_i} - n\right).
\end{aligned}$$

Appendix C

R/S-PLUS Implementation

C1. An R function for randomized systematic PPS sampling.

The input variables of the function are `x`: the population vector of size variable and `n`: the sample size. The function `syspps` returns the set of *n* selected units.

```
syspps<-function(x,n){
N<-length(x)
U<-sample(N,N)
xx<-x[U]
z<-rep(0,N)
for(i in 1:N) z[i]<-n*sum(xx[1:i])/sum(x)
r<-runif(1)
s<-numeric()
for(i in 1:N){
if(z[i]>=r){
s<-c(s,U[i])
r<-r+1
 } }
return(s[order(s)])
}
```

C2. An R function for simulating the second order inclusion probabilities.

The input variables of the function are `x`: the population vector of size variable and `s`: the set of labels of units in the sample. The default sampling procedure is the randomized systematic PPS sampling method and the number of repeated samples is $K = 10^6$. The function `piij` returns an $n \times n$ matrix with the $(ij)^{\text{th}}$ entry being the simulated $\pi_{ij}^*, i, j \in s$.

```
piij<-function(x,s){
N<-length(x)
n<-length(s)
p<-matrix(0,n,n)
for(k in 1:1000000){
ss<-syspps(x,n)
for(i in 1:(n-1)){
for(j in (i+1):n){
if(min(abs(ss-s[i]))+min(abs(ss-s[j]))==0)
    p[i,j]<-p[i,j]+1
  } } }
p<-(p+t(p))/1000000
return(p)
   }
```

C3. An R function for PPS sampling under substitution of units.

```
sysppssub<-function(x,n,refus){
s<-syspps(x,n)
sub<-numeric()
for (i in 1:n){
if(min(abs(s[i]-refus))==0) sub<-c(sub,i)
}
m<-length(sub)
if(m>0){
s<-s[-sub]
U1<-(1:length(x))[-c(refus,s)]
x1<-x[-c(refus,s)]
s1<-syspps(x1,m)
s<-c(s,U1[s1])
}
return(s[order(s)])
}
```

The default procedure for the selection of the initial sample and the substitute sample is the randomized systematic PPS sampling. The following R function `sysppssub` is used for simulating the inclusion probabilities under substitution of units. The input variables are `x`: the population vector of size variable, `n`: the sample size, and `refus`: the set of refusing units from the initial sample. The function returns a set of units for the final sample.

C4. R codes for simulating the π_i under substitution of units.

```
pi<-rep(0,N)
for(i in 1:1000000){
s<-sysppssub(x,n,refus)
for(j in 1:N){
if(min(abs(s-j))==0) pi[j]<-pi[j]+1
 } }
pi<-pi/1000000
```

References

Asok, C., and Sukhatme, B.V. (1976). On Sampford's procedure of unequal probability sampling without replacement. *Journal of the American Statistical Association*, 71, 912-918.

Chao, M.T. (1982). A general purpose unequal probability sampling plan. *Biometrika*, 69, 653-656.

Chen, X.H., Dempster, A.P. and Liu, J.S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika*, 81, 457-469.

Connor, W.S. (1966). An exact formula for the probability that two specified sampling units occur in a sample drawn with unequal probability and without replacement. *Journal of the American Statistical Association*, 61, 384-390.

Deville, J.C., and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85, 89-101.

Fattorini, L. (2006). Applying the Horvitz-Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. *Biometrika*, 93, 269-278.

Fellegi, I.P. (1963). Sampling with varying probabilities without replacement: Rotating and non-rotating samples. *Journal of the American Statistical Association*, 58, 183-201.

Goodman, R., and Kish, L. (1950). Controlled selection - A technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-372.

Gray, G.B. (1973). On increasing the sample size (number of psu's). Technical Memorandum, Statistics Canada.

Hartley, H.O., and Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374.

Hidiroglou, M.A., and Gray, G.B. (1980). Construction of joint probability of selection for systematic P.P.S. sampling. *Applied Statistics*, 29, 107-112.

Kott, P.S. (2005). A note on the Hartley-Rao variance estimator. *Journal of Official Statistics*, 21, 433-439.

Rao, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal Indian Statist. Association*, 3, 173-180.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

Tillé, Y. (1996). An elimination procedure for unequal probability sampling without replacement. *Biometrika*, 83, 238-241.