On selecting a sample by probability proportional to size with second-order inclusion probabilities and without replacement*

László Mihályffy
Senior statistical adviser (ret.)
Hungarian Central Statistical Office
E-mail: Laszlo.Mihalyffy@ksh.hu

Given appropriate sets of first- and second-order inclusion probabilities, the author provides a method that results in samples including units and pairs of units of the universe with the probabilities specified in advance.

KEYWORDS:
Sampling with probability proportional to size.
Horvitz-Thompson estimator.
Variance estimation.

DOI: 10.20311/stat2016.K20.en083

* The author is indebted to the reviewer for the valuable comments that enabled him to improve the results in the paper.

In the paper the problem of estimating the variance of totals is considered in the case of samples of fixed size selected with probability proportional to size and without replacement. Note that the term "sampling with unequal probabilities" might be used instead of "sampling with probability proportional to size" (abridged πps when sampling is without replacement) throughout the paper; from the aspect of practice, there is no substantial difference between the two notions.

Since the introduction of the Horvitz-Thompson estimator /4/ and the corresponding variance estimator /5/ by *Sen* [1953] and *Yates–Grundy* [1953] (see in the following), a considerable number of publications have been appeared on this topic. The intensive research in this field has been motivated probably by the fact that estimating the variance of an estimated total has proved to be a quite hard job in case of πps sampling in contrast to *pps* sampling, i.e. when sampling is done with replacement. Having extraordinarily ample literature on πps sampling, one should raise the question what is the novelty in this paper.

From the beginning up to our days, the usual way of creating a πps sampling design is as follows:

– Assign a first-order inclusion probability $0 < \pi_i < 1$ to each unit *i* of the universe called also target population $U = \{1, 2, ..., N\}$;

– If n is the sample size, make sure that the equality $\pi_1 + \pi_2 + ... + \pi_N = n$ may hold;

– Define a procedure suitable for selecting samples of size *n* such that the unit *i* is included in the sample with probability π_i;

– On the basis of the sampling procedure derive a rule of determining exact or approximate value of each second-order inclusion probability π_{ij} [1], i.e. the probability of the event that both units *i* and *j* are included in a sample of size $n \left(1 \le i, j \le N, i \ne j\right)$.

Having carried out these operations, samples can be selected and the survey can be conducted; thereafter the Horvitz-Thompson estimator (*Horvitz–Thompson* [1952]) and the Sen-Yates-Grundy estimator (*Sen–Yates–Grundy* [1953]) can be used with the values of the characteristic observed on the units of the sample.

[1] This step is sometimes replaced by providing an approximate formula for the variance estimator.

By contrast, our approach is based on the direct use of the second-order inclusion probabilities π_{ij} in defining the sampling design. The π_{ij}'s should be assessed by means of suitable information obviously other than the design, and the key to solving this problem is given in the following by the relations /1/–/3/ between the first- and second-order inclusion probabilities. Given the set $\{\pi_1, \pi_2, ..., \pi_N\}$, assessing a feasible set of the π_{ij}'s is trivial in certain cases, and then sampling with second-order inclusion probabilities is one of the simplest and fastest method of πps sampling. However, the bulk of this paper is the sampling algorithm on the assumption that the π_{ij}'s are known, and assessing the latter in the general case will be discussed in another paper.

Note that there is a minor looseness of terminology in the paper. A sampling method is obviously a procedure, an algorithm whose result is a sampling design. Nevertheless, in some cases the latter term will refer to the algorithm resulting in the design; this will make the language simpler, hopefully without leading to confusion.

The structure of this paper is as follows. Our sampling algorithm is described in Chapter 1, this is followed by presenting an application in Chapter 2. In Chapter 3 the algorithm is compared with some standard designs of πps sampling from the aspect of the simplicity of usage. It is worth noting here that current research on πps sampling focuses on high entropy of the sampling design – see in the following – rather than on simplicity of computing variance estimates. Hence the goal of this paper is not in the mainstream, but in certain cases simplicity of computing may be more important than high entropy of the design[2]. In the paper the following notations will be used besides those mentioned earlier.

$s = \{i_1, i_2, ..., i_n\}$: sample[3] of size n from U,

$U \setminus \{i\}$: "reduced" universe obtained from U by deleting unit i,

U^*: set of all samples consisting of n units from U,

$C = N!/((N-n)!\,n!)$: total number of samples of size n,

$p(s)$: probability function (abridged pf), positive for all $s \in U^*$,

$\sum_{s \in U} p(s) = 1$,

$s = \{x_1, x_2, ..., x_N\}$: alternative notation for a sample, $x_i = 1$ or $x_i = 0$, if unit $i \in s$ or $i \notin s$, respectively,

[2] The application of the principle of maximum entropy in statistics reduces the chance of receiving unwarranted information (*Jaynes* [1962]).

[3] Samples selected without replacement are only considered.

$p(s) \propto \Phi(s)$: specifying $p(s)$ as a member of some special family of functions,

$p(s) \propto \prod_{i=1}^{N} p_i^{x_i} (1- p_i)^{1- x_i}$: pf of the conditional Poisson design, $0< p_i < 1$,

$H = - \sum_{s \in U^*} p(s) \log p(s)$: entropy of the sampling design,

$p_j = \pi_j / n$: probability of selecting unit j from U, $j = 1, 2, ..., N$.

The following basic relations concerning πps samples will also be referred to in the paper.

$$\pi_1 + \pi_2 + , ..., + \pi_N = n \qquad /1/$$

$$\sum_{j \neq i}^{N} \pi_{ij} = (n- 1)\, \pi_i , \; i = 1, 2, \ldots, N \qquad /2/$$

$$0< \pi_{ij} < \pi_i \pi_j \,^4, \; 1 \leq i , \quad j \leq N , \; i \neq j \qquad /3/$$

$$\hat{Y}_{HT} = \sum_{i \in s} y_i / \pi_i \qquad /4/$$

$$\hat{V}(\hat{Y}_{HT}) = \sum_{i \in s} \sum_{j \in s,\, j > i} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \qquad /5/$$

$$\hat{V}(\hat{Y}_{pps}) = \frac{1}{n(n- 1)} \sum_{j \in s} \left(\frac{y_j}{p_j} - \hat{Y}_{pps} \right)^2 \qquad /6/$$

$\hat{Y}_{HT}$ in /4/ is the sample estimate of the population total $Y = \sum_{k=1}^{N} y_k$ by the Horvitz–Thompson estimator. The variance of $\hat{Y}_{HT}$ is

$$V(\hat{Y}_{HT}) = \sum_{i=1}^{N} \frac{1- \pi_i}{\pi_i} y_i^2 + 2 \sum_{i=1}^{N} \sum_{j>i} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_i y_j ,$$

the sample estimate /5/ of this statistic is by *Sen* [1953] and *Grundy–Yates* [1953]. The order of the sampled units in /5/ should be increasing in terms of their identifi-

[4] In some approximations "≤ " may stand instead of the second "< ".

ers, i.e. of their indices. Estimator /6/ is the counterpart of /5/ in the case of *pps* samples.

1. Sampling by means of second-order inclusion probabilities

Assume we are given the sets of first- and second-order inclusion probabilities satisfying the constraints /1/–/3/. Suppose that a sample of fixed size n should be selected with probability proportional to size from a universe consisting of N units. Using the notations in the introduction, define the following.

Algorithm.

Step 1. Select a unit i from the universe $U = \{1, 2, \ldots, N\}$ with the probability $p_i = \pi_i / n$.

Step 2. Using the probabilities π_{i1}/π_i , π_{i2}/π_i , …, $\pi_{i,\,i-1}/\pi_i$, $\pi_{i,\,i+1}/\pi_i$, …, π_{iN}/π_i , select $n-1$ units from the reduced universe $U \setminus \{i\}$ with probability proportional to size. Denote i_2 , i_3 , … and i_n the selected units. The procedure has finished, resulting in the sample $s = \{i, i_2, i_3, ..., i_n\}$.

Remark. Randomised systematic sampling (*Hartley–Rao* [1962]) is recommended to select the $n-1$ units in the Step 2, since this is the simplest technique between the standard *πps* sampling methods, requiring nearly optimal amount of computing in ingenious applications.[5] For a description of the method see the Appendix.

Theorem. When using this algorithm, each unit i of the universe is included in a sample of n units with probability π_i. In addition, any pair (i, j) of units of the universe $(i \neq j)$ is included in a sample of size n with probability π_{ij}.

Proof. If i is selected in Step 1, the corresponding selection probability is $p_i = \pi_i / n$. If unit $j \neq i$ is selected in Step 1, the conditional probability $P(i \mid j)$ equals π_{ji} / π_j , that is, the first-order inclusion probability of unit i as a unit selected

[5] This technique starts with arranging the units of the universe in random order, which requires considerable CPU (central processing unit) time in the case of large universes. However, it is not necessary to repeat this ordering whenever a new selection is needed.

from the reduced universe $U \setminus \{i\}$ in a sample of size $n-1$. As for $P(i\,|i)$, the only meaningful interpretation is that it equals 1. Since the events "drawing $\boldsymbol{i}$ given j" in Step 2 constitute a countable partition of "drawing i", by virtue of the law of total probability we have

$$P(i) = \sum_{j=1}^{N} p_j P(i \mid j) = \sum_{j \neq i}^{N} (\pi_j / n)\pi_{ji} / \pi_j + p_i , \qquad /7/$$

which, owing to the relations $\pi_{ji} = \pi_{ij}$, $p_i = \pi_i / n$ and equation /2/, can be re-written as follows:

$$P(i) = \sum_{j \neq i}^{N} \frac{\pi_{ij}}{n} + \frac{\pi_i}{n} = \frac{(n-1)\pi_i}{n} + \frac{\pi_i}{n} = \pi_i .$$

This proves the first part of the Theorem. The proof of the second part is based on the fact that selecting a unit i in the Step 2 of the algorithm – provided that unit j has been selected in Step 1 – is tantamount to selecting the pair of units j and i, ($j \neq i$). The term $\left(\pi_j / n\right) \pi_{ji} / \pi_j = \pi_{ji} / n$ in /7/ is a portion of the first-order inclusion probability π_i, and at the same time it is also a portion of the second-order inclusion probability of the pair of units (j, i). Consider now a sample $s = \{i_1, i_2, ..., i_n\}$ selected from U by means of our algorithm. In the course of the algorithm, this sample occurs on n occasions depending on which of its units is selected in Step 1. Whenever this sample s is selected, all of the $n! / (2!(n-2)!) = n(n-1)/2$ pairs of units contained in it are obviously selected, too. On each occasion, when s is selected, the pairs (j, i) belonging to it will be selected with the same probability. As we have seen above, the case where e.g. i_1 is selected in Step 1 and i_2 in the second contributes the portion $\pi_{i_1 i_2} / n$ to the inclusion probability of the pair (i_1, i_2), thus we conclude that the full inclusion probability of this pair is $n \times \pi_{i_1 i_2} / n$. The proof is thereby complete.

Corollary. For a given set of first-order inclusion probabilities π_i satisfying /1/, the values π_{ij} with $1 \leq i, j \leq N$, $i \neq j$ constitute a set of second-order inclusion probabilities for some πps design if and only if the relations /2/ and /3/ hold.

2. An example of application

As was mentioned in the introduction, the application of our sampling method – called henceforth p_ij method – is especially advantageous in cases where, besides the first-order inclusion probabilities, the second-order ones are also available, or at least there is a simple method to assess them. Such a case and such a "simple method" will be considered in the example below.

Suppose we are given a set of first-order inclusion probabilities $\pi_1, \pi_2, ..., \pi_N$ satisfying constraint /1/. Let

$$p_i = \pi_i / n \quad \text{for} \quad i = 1, 2, ..., N, \qquad /8/$$

$$\tau = \sum_{i=1}^{N} \frac{p_i}{1 - 2p_i}, \qquad /9/$$

$$u_i = \frac{n-1}{n(1+\tau)} \frac{1}{1-2p_i} \quad \text{for} \quad i = 1, 2, ..., N, \qquad /10/$$

$$x_{ij} = u_i + u_j \quad \text{for} \quad i, j = 1, 2, ..., N, \; i \neq j, \quad x_{11} = x_{22} =, ..., = x_{NN} = 0, \qquad /11/$$

and finally

$$\pi_{ij} = x_{ij}\pi_i\pi_j, \; i, j = 1, 2, ..., N, \; i \neq j. \qquad /12/$$

Second-order inclusion probabilities defined by /8/–/12/ can be found often in the literature on πps sampling. They satisfy the basic relations /2/ between the first- and the second-order inclusion probabilities and are positive if each $\pi_i > 0$. In addition, in case $n = 2$, they also satisfy the inequalities /3/ whereby all conditions on second-order inclusion probabilities are fulfilled; these probabilities π_{ij} were derived in the works by *Brewer* [1963], *Rao* [1965] and *Durbin* [1967]. If the relations /2/ and /3/ held in general for *n* greater than 2, the situation would be optimal for our p_ij method, but unfortunately, this is not the case. However, the set of the individual bounds $np_i < 1/2$ for $i = 1, 2, ..., N$ is a sufficient condition on the inequalities /3/, and the latter ensure that the Sen-Yates-Grundy estimate /5/ of the variance may be always non-negative.

Consider now a universe consisting of $N = 7$ units and assume that the first-order inclusion probabilities pertaining to the latter are the following:

$$0.48, \; 0.29, \; 0.49, \; 0.48, \; 0.41, \; 0.37, \; 0.48. \qquad /13/$$

These add up to $n = 3$, indicating that samples of size 3 should be selected. Denote $\boldsymbol{\pi}$ the vector whose components are the probabilities /13/. Making use of the formulae /8/–/12/, the following results are obtained for the matrices $\mathbf{X} = \left(x_{ij}\right)_{N \times N}$ and $\prod = \left(\pi_{ij}\right)_{N \times N}$:

$$\mathbf{X} = \begin{vmatrix} 0 & 0.7466 & 0.8142 & 0.8102 & 0.7842 & 0.7708 & 0.8102 \\ 0.7466 & 0 & 0.7506 & 0.7466 & 0.7206 & 0.7072 & 0.7466 \\ 0.8142 & 0.7506 & 0 & 0.8142 & 0.7882 & 0.7748 & 0.8142 \\ 0.8102 & 0.7466 & 0.8142 & 0 & 0.7842 & 0.7708 & 0.8102 \\ 0.7842 & 0.7206 & 0.7882 & 0.7842 & 0 & 0.7448 & 0.7842 \\ 0.7708 & 0.7072 & 0.7748 & 0.7708 & 0.7448 & 0 & 0.7708 \\ 0.8102 & 0.7466 & 0.8142 & 0.8102 & 0.7842 & 0.7708 & 0 \end{vmatrix},$$

$$\mathbf{\Pi} = \begin{vmatrix} 0 & 0.1039 & 0.1915 & 0.1867 & 0.1543 & 0.1369 & 0.1867 \\ 0.1039 & 0 & 0.1067 & 0.1039 & 0.0857 & 0.0759 & 0.1039 \\ 0.1915 & 0.1067 & 0 & 0.1915 & 0.1584 & 0.1405 & 0.1915 \\ 0.1867 & 0.1039 & 0.1915 & 0 & 0.1543 & 0.1369 & 0.1867 \\ 0.1543 & 0.0857 & 0.1584 & 0.1543 & 0 & 0.1130 & 0.1543 \\ 0.1369 & 0.0759 & 0.1405 & 0.1369 & 0.1130 & 0 & 0.1369 \\ 0.1867 & 0.1039 & 0.1915 & 0.1867 & 0.1543 & 0.1369 & 0 \end{vmatrix}. \quad /14/$$

It is easy to check that vector $\boldsymbol{\pi}$ and matrix $\mathbf{\Pi}$ given by /14/ satisfy the conditions /1/–/3/ in case $n = 3$. In what follows, a sample of size 3 will be selected with the *p_ij* method described in the previous section, i.e. by means of the first-order inclusion probabilities /13/ and the second-order inclusion probabilities, i.e. the entries of matrix $\mathbf{\Pi}$.

In order to use the *p_ij* method, the order of the units of the universe should be random. Assume that the order of the probabilities π_i in /13/ complies with this requirement. In Step 1 of the algorithm a unit i should be selected from the universe with probability $p_i = \pi_i / n$. Scaling the entries of $\boldsymbol{\pi}$ by $1/n = 1/3$, we get the probabilities

$$0.16,\ 0.29/3,\ 0.49/3,\ 0.16,\ 0.41/3,\ 0.37/3,\ 0.16.$$

From these probabilities the following cumulated totals are obtained for selecting a single unit of the universe: 0.16, 0.257, 0.42, 0.58, 0.717, 0.84, 1.0 (the values are

rounded). The random number generator has selected the value $r = 0.1443637$ from the uniform distribution on the interval (0, 1). Since $r < 0.16$, the first element in the above sequence, we have $i = 1$. This means that further units of the sample should be selected in Step 2 of the algorithm by means of the first row of matrix $\mathbf{\Pi}$, which is

$$\{\pi_{12},\ \pi_{13},\ \pi_{14},\ \pi_{15},\ \pi_{16},\ \pi_{17}\} = \\ = \{0.1039,\ 0.1615,\ 0.1867,\ 0.1543,\ 0.1369,\ 0.1867\}$$

(the vanishing diagonal entry has been omitted). Dividing these probabilities by $\pi_i = \pi_1 = 0.48$, the first-order inclusion probabilities are obtained for selecting samples of size $n - 1$ from the reduced universe consisting of the units 2, 3, 4, 5, 6 and 7. In the special case considered, $n = 3$ and condition /2/ reads as follows:

$$\sum_{j=1,\, j\neq 1}^{7} \pi_{1j} / \pi_1 = 3 - 1 = 2 .$$

The sample of size 2 will be selected from the reduced universe with randomised systematic sampling (see the Appendix). Since the size of the units is measured by the first-order inclusion probabilities π_{1j} / π_1, these are the building blocks of the cumulated totals the last of which is equal to the sample size $n - 1 = 2$. The cumulated totals pertaining to the units of the reduced universe are the following.

Probability	Index of the unit					
	2	3	4	5	6	7
Cumulated total	0.2165	0.6155	1.0044	1.3259	1.6111	2.0000

The starting value in the randomised systematic sampling is a positive random number not exceeding the distance $d = 1$; the value obtained with the random number generator was $k_1 = 0.4915$. The next (and in this case also the last) auxiliary variable will be $k_2 = k_1 + d = 1.4915$. Since $0.2165 < k_1 \leq 0.6155$ and $1.3259 < k_2 \leq 1.6111$, the second and the third unit of the sample to be selected are $i_2 = 3$ and $i_3 = 6$, respectively. The sample of size 3 from the universe with 7 units consists of the units 1, 3 and 6, respectively.

3. Comparison of the *p_ij* method with some standard πps designs

The introduction of the *p_ij* method was motivated by the aim to find a πps design facilitating a very simple way of computing variance estimates. The comparison of the method with some standard designs of πps sampling should report on the results of this endeavour. For the purpose of the comparison the following sampling methods have been chosen:

– Sunter's sequential method (*Sunter* [1986]),
– conditional Poisson sampling (*Hájek* [1964], [1981]; *Chen–Dempster–Liu* [1994]), and
– Sampford sampling (*Sampford* [1967]).

Owing to their fine theoretical properties, these methods lead the field in terms of number of references; as for practical applications, they are dominated by the randomised systematic and the ordinary Poisson sampling. Our p_ij method can be regarded as a variant of randomised systematic sampling, since the first unit is determined by some selection probability π_i / n, and the remaining $n-1$ units are selected with the randomised systematic method.

The criteria of comparison will be run time needed to select a sample on the one hand and the complexity of computing or estimating the first- and second-order inclusion probabilities on the other. Consider first run time, which will be estimated by the number of operations needed to perform sampling.

The *randomised systematic sampling* stipulates random order of the units of the universe. Fortunately, the sorting need not be repeated whenever a new selection is required. Using the properly ordered universe, each unit should be scanned to find neighbouring units i and $i+1$ such that $i \le t_o + kd < i+1$ where $t_o + kd$ is the member of an arithmetic sequence of length n (see the Appendix). To sum up, the total number of operations needed with this method can be estimated as

$$O(N \log N) + O(N) \qquad /15/$$

where the first term stands for the operations of sorting and the second for scanning the individual units. According to the remark above, this estimate applies also to the *p_ij* method.

Sunter's sequential method (earlier version) stipulates ordering the units by decreasing first-order inclusion probabilities and scans each unit in this order. A unit *i* is selected if $\pi_i < \pi_i^*$ where π_i^* is the current value from the random number generator, and if this is the case, i is deleted from the universe, and the first-order inclu-

sion probabilities belonging to the remaining units are recalculated properly. With this method, units are included in the samples with the given (i.e. original) π_i's but the sample size n is a random number. This undesirable property of the method has been eliminated in the current version; however, at the cost of growing complexity of the method. The estimate /15/ of the number of operations is valid for the earlier as well as the current version of the sequential method.

Conditional Poisson sampling (CP) is derived from the ordinary Poisson sampling (*Hájek* [1964]). Its probability function (pf) belongs to the exponential family (see the notations in the introduction). Selecting a sample with the CP is performed with the rejection-acceptance method: ordinary Poisson sampling (see the Appendix) is repeated with the parameters $p_1, p_2, ..., p_N$ until a sample of size n is obtained. Samples of size less or greater than n are rejected. If the parameters are known, the first-order inclusion probabilities can be computed by means of a closed form expression requiring $O\left(n^2N\right)$ operations (*Chen–Dempster–Liu* [1994]). In practice, the inverse problem when $\pi_1, \pi_2, ..., \pi_N$ are given and the corresponding parameters p_i are unknown is of key importance; this is solved by an iterative method using $O\left(n^2N\right)$ operations per iteration (*Chen–Dempster–Liu* [1994]). Thus the total number of operations needed to select a sample with the CP if the first-order inclusion probabilities are known amounts to $kO\left(n^2N\right) + L \times O(N)$ where k is the number of iterations needed to achieve proper convergence, and the ordinary Poisson sampling has to be repeated L times to obtain a sample of size n. Note that there is such an alternative algorithm for CP sampling that the term $L \times O(N)$ is replaced by $O(nN)$ (*Chen–Dempster–Liu* [1994]) .

Sampford sampling is a rejective method: the first unit i_1 is selected with the probability $p_{i_1} = \pi_{i_1} / n$, and $n-1$ other units are selected with the probabilities $\lambda_j \Big/ \sum_{k=1}^{N} \lambda_k$, $j = i_2, i_3, ..., i_n$ where $\lambda_k = p_k\left(1 - np_k\right)$, $k = 1, 2, ..., N$. The latter units are selected with replacement, and the sample consisting of the units $i_1, i_2, i_3, ..., i_n$ is accepted only if the units are all different, otherwise it is rejected. The probability function pertaining to the method is of the following form: $p(s) = nK_n\lambda_{i_1}\lambda_{i_2} ... \lambda_{i_n}\left(1 - \sum_{h=1}^{n} p_{i_h}\right)$ where K_n is a constant. It is pointed out that with these definitions π_i is the first-order inclusion probability of unit i. There is an exact closed form representation for K_n, this needs $O\left(N^2\right)$ operations to be com-

puted. The probability of obtaining an acceptable sample is $P_n = (n-1)!\Big/\left(K_n\left(\sum_{i=1}^{N} \lambda_i \right)^{n-1} \right)$, and $1/P_n$ is the expected number of samples that must be drawn to obtain an acceptable sample. Therefore, the expected number of operations needed to obtain a sample with this method is

$$O\left(N^2\right) + \frac{O(n)}{P_n} \ . \qquad /16/$$

Comparing the sampling methods considered above from the aspect of run time, we see that both the p_ij method and Sunter's sequential method use $O(N \log N) + O(N)$ operations to select a sample of size *n*. Nevertheless, the p_ij method is simpler and therefore also somewhat faster than its sequential counterpart, since the latter cannot guarantee the fixed size of the sample without a specific routine if correction is needed. CP sampling is a frequently used method with favourable properties such as high entropy and analytic form of the probability function. As was mentioned above, sampling with CP requires $kO\left(n^2 N\right) + L \times O(N)$ operations provided that k iterations are needed to adjust the parameters p_i of the pf to the given first-order inclusion probabilities π_i, and ordinary Poisson sampling should be repeated L times to obtain a sample consisting of *n* different units. $O\left(n^2 N\right)$ and $O(N)$ are estimated numbers of operations used per iteration and performing ordinary Poisson sampling once, respectively. Due to expert judgment, *k* is of moderate size, occasionally quite small. In any case, rejection-acceptance methods are usually slower than sequential methods. This holds for Sampford sampling, too, though owing to some improvement that method has become more efficient, i.e. faster (see *Bondesson–Traat–Lundqvist* [2006]). Our conclusion is that from the aspect of run time both *p_ij* and Sunter's method are faster than CP and Sampford sampling; this is reflected also in the bounds /15/ and /16/ of the numbers of operations needed by the methods in question.

Each of the sampling methods considered above is suitable to provide second-order inclusion probabilities. However, in the case of the current version of Sunter's method, the π_{ij}'s are exact only for $i < j \le N - n$, otherwise they have approximate values; $O\left(N^2\right)$ operations are needed to compute them. In case of conditional Poisson sampling, exact values of the π_{ij}'s can be computed by an explicit formula

requiring $O\left(n^2N^2\right)$ operations for the $N(N-1)/2$ probabilities (see *Chen–Dempster–Liu* [1994]). Sampford sampling provides also an explicit expression for computing the π_{ij}'s by means of the probability function. Provided that K_n has been computed and the π_{ij}'s are needed for the sampled units only, the computational load amounts to $O\left(n^2N\right)$; if all second-order inclusion probabilities are needed, $O\left(N^3\right)$ operations should be carried out.

The *p_ij* method was introduced on the assumption that for a *πps* design feasible sets of first- and second-order inclusion probabilities are given. However, if the method should be compared with the above standard designs from the aspect of convenience when it comes to variance estimation, one needs a tool, that is, some procedure to provide a feasible set of the π_{ij}'s if the first-order inclusion probabilities π_i satisfying /1/ are given. For the time being, there is no better option than the second-order inclusion probabilities defined by the relations /8/–/12/ in Chapter 2. They are actually very simple, all in all, $N^2 + 2$ additive and $3N^2 + 3$ multiplicative operations are needed to compute them. Unfortunately, they also have a drawback, namely, there is only a sufficient condition on their feasibility: $\pi_i < 1/2$ for $i = 1, 2, ..., N$. Research is underway to find an algorithm for computing π_{ij}'s not subject to this restriction. Summarising the conclusions of the comparisons above, the following can be stated: the *p_ij* method is faster than the designs using the rejection-acceptance method such as conditional Poisson sampling and Sampford sampling. It is at least as fast as Sunter's sequential method and, in contrast with that method, always yields exact results. From the aspect of variance estimation with the Sen-Yates-Grundy formula, the *p_ij* method combined with the formulae /8/–/12/ is more efficient than Sunter's method, the conditional Poisson sampling as well as the Sampford sampling provided that $\pi_i < 1/2$ is satisfied for each first-order inclusion probability.

Besides the above comparisons, there is a by-product of the *p_ij* method and the Theorem that may deserve some attention. There are several publications on *πps* designs under titles similar to that of the present paper, e.g. "Sampling with prescribed second-order inclusion probabilities" (see *Bondesson* [2012], *Gabler–Schweigkoffer* [1990], *Herzel* [1986], *Sinha* [1973], *Lundqvist–Bondesson* [2009], etc.). The goal of their authors is similar: given the sets of appropriate second-order inclusion probabilities, define a sampling design so that the units of the universe and pairs of them may be included in a sample of fixed size with the given probabilities. The aim of using prescribed second-order inclusion probabilities is to control the size of the variance of some specific estimates on the one hand and to achieve high entro-

py of the design on the other. The difference between this approach and that of the present paper was stressed in the introduction. Up to now, the usual approach to treat the problem has been the following: choose a design with known probability function and adjust the parameters of the pf so that the units of the universe may have the inclusion probabilities specified in advance.

The following important result of this trend of research was achieved by *Bondesson* [2012]: for a set of π_{ij}'s satisfying the necessary and sufficient conditions on second-order inclusion probabilities, there is a set of the parameters a_{ij} of the probability function of the conditional Poisson design of order 2 yielding the prescribed second-order inclusion probabilities. In addition, the entropy of this design is maximal among the designs having the same second-order inclusion probabilities. Our Corollary is simpler than the necessary and sufficient conditions used in Bondesson's paper on second-order inclusion probabilities, and might replace them. The conditional Poisson design of order 2 is a modified version of CP with probability function $p(s) \propto \exp\left(\sum_{i,j} a_{ij} x_i x_j\right)$, a_{ij} is symmetric, $i, j = 1, 2, ..., N$, $i \neq j$; its application uses considerably long run time.

Appendix

1. Randomised systematic sampling

Arrange the N units of the universe in random order, and compute cumulated totals of the quantities representing their size in the following way: $t_1 = a_1$, $t_2 = t_1 + a_2$, $t_3 = t_2 + a_3$, ..., $T = t_N = t_{N-1} + a_N$ Introduce the pace $d = T/n$ where n denotes sample size. Choose a positive real number $k_1 < d$ and define the sequence k_1, $k_2 = k_1 + d$, $k_3 = k_2 + d$, $k_4 = k_3 + d$, … The unit v will be selected in the sample if there is such an element k_l in the sequence that $t_{v-1} < k_l \leq t_v$ (the case $t_0 = 0$ is not excluded). The unit v is included in the sample with a probability proportional to $a_v = t_v - t_{v-1}$. The quantities a_i representing the size of the units of the universe may be identical with the first-order inclusion probabilities.

2. Poisson sampling

"Poisson sampling is a sampling process where each element of the population that is sampled is subjected to an independent Bernoulli trial which determines whether the element becomes part of the sample during the drawing of a single sample. Each element of the population may have a different probability of being included in the sample. The probability of being included in a sample during the drawing of

a single sample is denoted as the first-order inclusion probability of that element. If all first-order inclusion probabilities are equal, Poisson sampling becomes equivalent to Bernoulli sampling, which can therefore be considered to be a special case of Poisson sampling. Mathematically, the first-order inclusion probability of the i^{th} element of the population is denoted by the symbol π_i, and the second-order inclusion probability that a pair consisting of the i^{th} and j^{th} element of the population that is sampled is included in a sample during the drawing a single sample is denoted by π_{ij}. The following relation is valid during Poisson sampling: $\pi_{ij} = \pi_i \times \pi_j$." (*Wikipedia* [2008])

References

BONDESSON, L. – TRAAT, I. – LUNDQVIST, A. [2006]: Pareto sampling versus conditional Poisson and Sampford sampling. *Scandinavian Journal of Statistics*. Vol. 33. Issue 4. pp. 699–720. http://dx.doi.org/10.1111/j.1467-9469.2006.00497.x

BONDESSON, L. [2012]: On sampling with prescribed second-order inclusion probabilities. *Scandinavian Journal of Statistics*. Vol. 39. Issue 4. pp. 813–829. http://dx.doi.org/10.1111/j.1467-9469.2012.00808.x

BREWER, K. W. R. [1963]: A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*. Vol. 5. Issue 1. pp. 5–13. http://dx.doi.org/10.1111/j.1467-842X.1963.tb00132.x

CHEN, X. H. – DEMPSTER, A. P. – LIU, J. S. [1994]: Weighted finite population sampling to maximize entropy. *Biometrika* Vol. 81. No. 3. pp. 457–469. http://dx.doi.org/10.1093/biomet/81.3.457

DURBIN, J. [1967]: Design of multi-stage surveys for estimation of sampling error. *Applied Statistics*. *Series C*. Vol. 16. No. 2. pp. 152–164. http://dx.doi.org/10.2307/2985777

GABLER, S. – SCHWEIGKOFFER, R. [1990]: The existence of sampling designs with pre-assigned inclusion probabilities. *Metrika* Vol. 37. Issue 1. pp. 87–96.

HÁJEK, J. [1964]: Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*. Vol. 35. No. 4, pp. 1491–1528. http://dx.doi.org/10.1214/aoms/1177700375

HÁJEK, J. [1981]: *Sampling from a Finite Population*. Marcel Dekker. New York.

HARTLEY, B. G. – RAO, J. N. K. [1962]: Sampling with unequal probabilities and without replacement. *The Annals of Mathematical Statistics*. Vol. 33. No. 2. pp. 350–374. http://dx.doi.org/10.1214/aoms/1177704564

HERZEL, A. [1986]: Sampling without replacement with unequal probabilities: Sample designs with preassigned joint inclusion probabilities of any order. *Metron.* Vol. XLIV. No. 1. pp. 49–68.

HORVITZ, D. G. – THOMPSON, D. J. [1952]: A generalisation of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. Vol. 47. pp. 663–685. http://dx.doi.org/10.1080/01621459.1952.10483446

JAYNES, E. T. [1963]: Information theory and statistical mechanics. In: *Ford, K.* (ed.): *Statistical Physics*. W. A. Benjamin. New York. pp. 181–218.

LUNDQVIST, A. – BONDESSON, L. [2009]: *On sampling with desired inclusion probabilities of first and second order*. Research report in mathematical statistics. Umeå University. Umeå**.** http://snovit.math.umu.se/Forskning/MathStat/reports/Lundqvist05-3.pdf

RAO, J. N. K. [1965]: On two simple schemes of unequal probability sampling without replacement. *Journal of Indian Statistical Association*. Vol. 3. No. n. d. pp. 173–180.

SAMPFORD, M. R. [1967]: On sampling without replacement with unequal probabilities of selection. *Biometrika*. Vol. 54. Nos. 3–4. pp. 499–513. http://dx.doi.org/10.2307/2335041

SEN, A. R. [1953]: On the estimate of variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*. Vol. 5. No. 2. pp. 119–127.

SINHA, B. K. [1973]: On sampling schemes to realize preassigned sets of inclusion probabilities of first two orders. *Calcutta Statistical Association Bulletin*. Vol. 22. Nos. 85–88. pp. 89–110.

SUNTER, A. B. [1977]: List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics*. Vol. 26. No. 3. pp. 261–268. http://dx.doi.org/10.2307/2346966

SUNTER, A. B. [1986]: Solutions to the problem of unequal probability sampling without replacement. *International Statistical Review*. Vol. 54. No. 1. pp. 33–50. http://dx.doi.org/10.2307/1403257

Wikipedia [2008]: Poisson sampling. https://en.wikipedia.org/wiki/Poisson_sampling

YATES, F. – GRUNDY, P. M. [1953]: Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*. *Series B*. Vol. 15. No. 2. pp. 253–261.