Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Unequal probability inverse sampling

SURVEY
METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Statistics Canada Statistique Canada

Canada

by Yves Tillé

Release date: December 20, 2016

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-877-287-4369

Depository Services Program

• Inquiries line	1-800-635-7943
• Fax line	1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^{s}	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, December 2016
Vol. 42, No. 2, pp. 283-295
Statistics Canada, Catalogue No. 12-001-X

Unequal probability inverse sampling

Yves Tillé[1]

Abstract

In an economic survey of a sample of enterprises, occupations are randomly selected from a list until a number r of occupations in a local unit has been identified. This is an inverse sampling problem for which we are proposing a few solutions. Simple designs with and without replacement are processed using negative binomial distributions and negative hypergeometric distributions. We also propose estimators for when the units are selected with unequal probabilities, with or without replacement.

Key Words: Location; Horvitz-Thompson estimator; Negative binomial; Negative hypergeometric; Inverse design; Inclusion probability; Wage.

1 Problem

The problem arose as part of a question on Statistics Canada’s new Job Vacancy and Wage Survey (JVWS). The JVWS comprises a wage component and a job vacancy component. The wage component looks at average wages, minimum wages, maximum wages and starting wages for various occupations.

The objective is to provide wage statistics by economic regions (economic regions are subdivisions of provinces). In the first stage, a sample of 100,000 business locations (also known as local units of enterprises) are selected using a Poisson design stratified by industry and economic region.

For simplicity, the term “enterprise” will be used in the rest of the document instead of “location,” keeping in mind that Statistics Canada defines a location as “a production unit located at a single geographical location at or from which economic activity is conducted and for which a minimum of employment data are available.”

For purposes of managing response burden, it is not possible to identify every occupation in each enterprise. Therefore, proposing a list of occupations and asking whether the listed occupations exist in an enterprise has been considered. Occupations can then be randomly drawn from the list and proposed successively to the head of the enterprise until r occupations have been reached. Since the most common occupations are of specific interest, it is useful to consider cases in which occupations are selected with unequal probabilities from the list in proportion to their prevalence in the total population. Note that this method was not implemented for Statistics Canada’s Job Vacancy and Wage Survey. The survey decided to present a list, of fixed length, of occupations to the surveyed enterprises. Nevertheless, the theoretical properties of the proposed method remain of interest.

“Inverse sampling” refers to a scheme in which units are selected successively until a predetermined number of units with a certain characteristic is obtained. Inverse sampling must not be confused with rejective sampling. In rejective sampling, a sample is selected according to a design, and the sample is rejected if it does not have the desired characteristic (e.g., a specific sample size or an average equal to that of the population). The selection of samples is repeated until a sample with the desired property is obtained.

[1] Yves Tillé, Institute of Statistics, University of Neuchâtel, Avenue de Bellevaux 51, 2000 Neuchâtel, Switzerland. E-mail: yves.tille@unine.ch.

Inverse sampling raises a certain number of theoretical questions. How can such a design be implemented with equal or unequal inclusion probabilities? What is the probability of inclusion of an occupation within each enterprise? How can a variable of interest be estimated using a sample consisting of a few enterprises and a few occupations within them? How can the number of occupations in the enterprise be estimated? More generally, how can this survey be implemented and how can estimation be done?

The key issue is the way in which the occupations are selected. They may be selected using a simple design with or without replacement, or with unequal probabilities. One option would be to select the units with unequal probabilities using the sequential Poisson sampling method proposed by Ohlsson (1998) or the Pareto sampling method proposed by Rosén (1997). The inverse sampling problem has already been discussed by Murthy (1957), Sampford (1962), Pathak (1964), Chikkagoudar (1966, 1969), and Salehi and Seber (2001). However, the parameter to be estimated here is unique, since estimates of average revenue among all enterprises having a specific occupation are desired. We also propose a new unequal-probability inverse design without replacement.

This article is organized as follows: In Section 2, the problem is stated and the notation is defined. The equal probability case with replacement is discussed in Section 3, and the equal probability case without replacement is discussed in Section 4. The unequal probability case with replacement is developed in Section 5. A new selection method for the unequal probability case without replacement is presented in Section 6. Finally, Section 7 contains a short discussion.

2 Formalization of the problem

The following notation is used:

- U : a population of N enterprises, i.e., $U = \{1, \ldots, i, \ldots, N\}$ (U may denote the population of enterprises in an economic region),
- L : the list of occupations,
- M : the number of occupations in the list, i.e., the size of L,
- F_i : the list of occupations in enterprise i, with $F_i \subset L$,
- D_i : the list of occupations absent from enterprise i, with $D_i \subset L$, $F_i \cup D_i = L$ and $D_i \cap F_i = \varnothing$,
- Mp_i : the number of occupations in enterprise i, i.e., the size of F_i,
- r : the number of distinct occupations to be obtained in each enterprise,
- X_i : the number of failures before the r occupations in enterprise i are obtained by selecting the occupations using a given design.

The main objective is to estimate the average wage for an occupation in the total population. Let y_{ik} be the average wage for occupation k in enterprise i, and let z_{ik} be the number of employees with occupation k in enterprise i. The objective is to estimate the average wage for occupation k given by

$$\overline{Y}_k = \frac{\sum_{i \in U \mid F_i \ni k} z_{ik} y_{ik}}{\sum_{i \in U \mid F_i \ni k} z_{ik}}.$$

Assume that a sample of enterprises S_1 is selected from U using some given design with inclusion probabilities π_{1i}. In enterprise i, a sample of occupations S_i is selected using one of the designs described above with inclusion probability $\pi_{k|i}$. If the design is with replacement, $\pi_{k|i}$ represents the expected number of times that occupation k is selected in enterprise i.

$\overline{Y}_k$ can be estimated using a "ratio" type estimator (Hájek 1971):

$$\hat{\overline{Y}}_k = \frac{\sum_{i \in S_1 \mid (S_i \cap F_i) \ni k} \frac{z_{ik} y_{ik}}{\pi_{1i} \pi_{k|i}}}{\sum_{i \in S_1 \mid (S_i \cap F_i) \ni k} \frac{z_{ik}}{\pi_{1i} \pi_{k|i}}}.$$

Therefore, the probability that an occupation will be selected in an enterprise must be known. However, with an inverse type design, the probability is unknown and must therefore be estimated in order to estimate $\overline{Y}_k$. Since the inclusion probabilities appear in the denominator, it is preferable to estimate the inverses of $\pi_{k|i}$. In an enterprise, an occupation's probability of being selected decreases as the number of occupations increases. In addition, the probability depends on the inverse sampling design used in each enterprise.

3 Simple random sampling with replacement

Assume that enterprise i has proportion p_i of the occupations in the list in the enterprise. If the sample of occupations is drawn with replacement in enterprise i until r occupations in the enterprise have been identified, then X_i has a negative binomial distribution denoted by $X_i \sim NB(r, p_i)$. In that case,

$$\Pr(X_i = x_i) = \binom{r + x_i - 1}{x_i} p_i^r (1 - p_i)^{x_i},$$

with $x_i \in \mathbb{N} = \{0, 1, 2, 3, \ldots\}, p_i \in [0, 1], r \in \mathbb{N}^* = \{1, 2, 3, \ldots\}$. Furthermore,

$$\mathrm{E}(X_i) = \frac{r(1 - p_i)}{p_i} \quad \text{and} \quad \mathrm{var}(X_i) = \frac{r(1 - p_i)}{p_i^2}.$$

Let $A_{ik}, k \in L,$ be the number of times that unit k is selected in the sample taken from enterprise i. In a simple design with replacement of size n, the values of A_{ik} have a multinomial distribution. Therefore,

$$\Pr(A_{ik} = a_{ik}, k \in L) = \frac{n!}{M^n} \prod_{k \in L} \frac{1}{a_{ik}!},$$

where $A_{ik} = 0, \ldots, n,$ and

$$\sum_{k\in L} a_{ik} = n.$$

If this multinomial vector is conditioned on a fixed size in a given part of the population, then

$$\Pr\left(A_{ik} = a_{ik}, k \in F_i \,\middle|\, \sum_{k\in F_i} A_{ik} = r\right) = \frac{\Pr\left(A_{ik} = a_{ik}, k \in F_i \text{ and } \sum_{k\in F_i} A_{ik} = r\right)}{\Pr\left(\sum_{k\in F_i} A_{ik} = r\right)}$$

$$= \frac{\dfrac{n!(1-p_i)^{(n-r)}}{(n-r)!M^r}\displaystyle\prod_{k\in F_i}\frac{1}{a_{ik}!}}{\dfrac{n!p_i^r(1-p_i)^{n-r}}{r!(n-r)!}}$$

$$= r!\left(\frac{1}{Mp_i}\right)^r \prod_{k\in F_i}\frac{1}{a_{ik}!},$$

with

$$\sum_{k\in F_i} a_{ik} = r.$$

This shows that, if the sum of A_{ik} is conditioned on one part of the population, the distribution remains multinomial and conditionally there is still a simple design with replacement.

With the procedure in which we draw with replacement until we obtain r occupations in enterprise i, we have

$$\mathrm{E}(A_{ik} \mid X_i) = \begin{cases} \dfrac{r}{Mp_i} & \text{if } k \in F_i \\ \dfrac{X_i}{M - Mp_i} & \text{if } k \in D_i. \end{cases}$$

In fact, conditionally on X_i, in F_i of size Mp_i, r occupations are selected and, in D_i of size $M(1-p_i)$, X_i occupations are selected.

In the case with replacement, what is calculated is not really an inclusion probability, but rather the expected value of A_{ik} which is denoted as $\pi_{k|i}$,

$$\pi_{k|i} = \mathrm{EE}(A_{ik} \mid X_i) = \frac{r}{Mp_i},$$

$k \in L$. The problem is that we know M, r and X_i, but not p_i. We can estimate p_i using the method of moments by solving $\mathrm{E}(X_i) = X_i$, which yields

$$X_i = \frac{r(1-\hat{p}_i)}{\hat{p}_i}$$

and therefore

$$\hat{p}_{i1} = \frac{r}{X_i + r}.$$

The maximum likelihood method provides the same estimator as the method of moments, but this estimator is biased (Mikulski and Smith 1976; Johnson, Kemp and Kotz 2005, page 222). If $r \geq 2$, the unbiased minimum variance estimator of p_i is

$$\hat{p}_{i2} = \frac{r-1}{X_i + r - 1}.$$

However, $1/\hat{p}_{i1}$ is unbiased for $1/p_i$.

Since we are using weights that are inverses of $\pi_{k|i}$, the inverses of $\pi_{k|i}$ are thus estimated as follows:

$$\widehat{1/\pi_{k|i}} = \begin{cases} \dfrac{M\hat{p}_{i2}}{r} & = \dfrac{M(r-1)}{r(X_i + r - 1)} \quad \text{if } k \in F_i \\[2ex] \dfrac{M(1-\hat{p}_{i2})}{X_i} & = \dfrac{M}{X_i + r - 1} \quad \text{if } k \in D_i. \end{cases}$$

However, the case with replacement is not very satisfactory, because selecting r occupations with replacement does not necessarily result in r distinct occupations, since the same occupation may be selected more than once. Furthermore, sampling may be especially long if Mp_i is small. Therefore, sampling without replacement is preferred.

4 Simple random sampling without replacement

For the case without replacement, the notation used is the same as for the draw with replacement. The number of failures X_i therefore has a negative hypergeometric distribution. This probability distribution is little known, to the point that it has been presented as a “forgotten” distribution by Miller and Fridell (2007). This distribution is the counterpart to the negative binomial for the draw without replacement. The general framework is as follows: We consider a population of size M in which there are Mp_i favourable units, namely the occupations in the list that exist in the enterprise. If the draws are equal probability without replacement until r favorable units appear, then the negative hypergeometric variable, $X_i \sim NH(M, r, Mp_i)$, counts the number of failures before r favourable events occur.

The probability distribution is

$$\Pr(X_i = x) = p(x; M, r, Mp_i) = \frac{\begin{pmatrix} x + r - 1 \\ x \end{pmatrix} \begin{pmatrix} M - x - r \\ Mp_i - r \end{pmatrix}}{\begin{pmatrix} M \\ Mp_i \end{pmatrix}},$$

where $x \in \{0,\ldots,M(1-p_i)\}$, $M \in \{1,2,\ldots\}$, $Mp_i \in \{1,2,\ldots,M\}$, and $r \in \{1,2,\ldots,Mp_i\}$.

$$\mathrm{E}(X_i) = \frac{Mr(1-p_i)}{Mp_i+1}, \mathrm{var}(X_i) = \frac{rM(1-p_i)(M+1)(Mp_i-r+1)}{(Mp_i+1)^2(Mp_i+2)}.$$

Again, A_{ik} denotes the number of times that unit k is selected in the sample. Now, the value of A_{ik} can be only 0 or 1. If n units are selected using a simple design without replacement in $L,$ the sample design is defined as

$$\Pr(A_{ik} = a_{ik}, k \in L) = \binom{M}{n}^{-1},$$

where $a_{ik} \in \{0,1\}$, and

$$\sum_{k \in L} a_{ik} = n.$$

If the vector of A_{ik} is conditioned on a fixed size in one part of the population, we have

$$\begin{aligned}
\Pr\left(A_{ik} = a_{ik}, k \in F_i \,\middle|\, \sum_{k \in F_i} A_{ik} = r\right) &= \frac{\Pr\left(A_{ik} = a_{ik}, k \in F_i \text{ and } \sum_{k \in F_i} A_{ik} = r\right)}{\Pr\left(\sum_{k \in F_i} A_{ik} = r\right)} \\
&= \left[\frac{\binom{Mp_i}{r}\binom{M-Mp_i}{n-r}}{\binom{M}{n}}\right]^{-1} \sum_{\substack{k \in D_i \\ \sum_{k \in F_i} A_{ik} = n-r \\ A_{ik} \in \{0,1\}}} \frac{1}{\binom{M}{n}} \\
&= \left[\frac{\binom{Mp_i}{r}\binom{M-Mp_i}{n-r}}{\binom{M}{n}}\right]^{-1} \frac{\binom{M-Mp_i}{n-r}}{\binom{M}{n}} \\
&= \binom{Mp_i}{r}^{-1},
\end{aligned}$$

with

$$\sum_{k \in F_i} a_{ik} = r.$$

This shows that, if the sum of A_{ik} is conditioned on one part of the population, we still have a simple design without replacement. In the procedure in which we draw without replacement until we obtain r occupations in enterprise $i,$ we therefore have

$$\mathrm{E}\left(A_{ik} \mid X_i\right) = \begin{cases} \dfrac{r}{Mp_i} & \text{if } \; k \in F_i \\[2ex] \dfrac{X_i}{M - Mp_i} & \text{if } \; k \in D_i. \end{cases}$$

The inclusion probability is therefore

$$\pi_{k|i} = \mathrm{EE}\left(A_{ik} \mid X_i\right) = \begin{cases} \dfrac{r}{Mp_i} & \text{if } \; k \in F_i \\[2ex] \dfrac{\mathrm{E}\left(X_i\right)}{M - Mp_i} = \dfrac{r}{Mp_i + 1} & \text{if } \; k \in D_i, \end{cases}$$

for all $k \in L.$ Again, the problem is that we know M, r and $X_i,$ but not $p_i.$ We can estimate p_i using the maximum likelihood method, through a numerical method.

Using the method of moments, an estimate can be obtained by solving for p_i in the equation $X_i = \mathrm{E}\left(X_i\right),$ that is,

$$X_i = \frac{Mr\left(1 - \hat{p}_i\right)}{M\hat{p}_i + 1}.$$

Hence

$$\hat{p}_{i1} = \frac{Mr - X_i}{M\left(r + X_i\right)}.$$

However, in a few lines it is verified that, if $r \geq 2,$

$$\hat{p}_{i2} = \frac{r - 1}{r + X_i - 1}$$

is unbiased for $p_i.$

Again, since we are using weights that are inverses of $\pi_{k|i}.$ The inverses of the inclusion probabilities are thus estimated as follows:

$$\widehat{1/\pi_{k|i}} = \begin{cases} \dfrac{M\hat{p}_{i2}}{r} & = \dfrac{M\left(r - 1\right)}{r\left(X_i + r - 1\right)} & \text{if } \; k \in F_i \\[2ex] \dfrac{M\left(1 - \hat{p}_{i2}\right)}{X_i} & = \dfrac{M}{X_i + r - 1} & \text{if } \; k \in D_i. \end{cases}$$

These weights are also used in the estimator by Murthy (1957), which is unbiased (see also Salehi and Seber 2001). If $Mp_i < r,$ all occupations will be selected in enterprise i and the estimated inclusion probabilities are then equal to 1.

5 Unequal probability sampling with replacement

Unequal probability sampling is not really more difficult to process when the draw is with replacement. Now let p_{ik} denote the probability of an occupation being drawn in each draw with

$$\sum_{k \in L} p_{ik} = 1.$$

Let P_i be the sum of p_{ik} limited to the occupations in enterprise i:

$$P_i = \sum_{k \in F_i} p_{ik}.$$

In this case, X_i has a negative binomial distribution with parameters r and P_i. Therefore,

$$\mathrm{E}(X_i) = \frac{r(1-P_i)}{P_i} \quad \text{and} \quad \mathrm{var}(X_i) = \frac{r(1-P_i)}{P_i}.$$

Let $A_{ik}, k \in L$ be the number of times that unit k is selected in the sample. In an unequal probability design with replacement of size $n,$ the values of A_{ik} have a multinomial distribution. Therefore,

$$\Pr(A_{ik} = a_{ik}, k \in L) = n! \prod_{k \in L} \frac{p_{ik}^{a_{ik}}}{a_{ik}!},$$

where $A_{ik} = 0, \ldots, n,$ and

$$\sum_{k \in L} a_{ik} = n.$$

If this multinomial vector is conditioned on a fixed size in one part of the population, then

$$\begin{aligned}
\Pr\left(A_{ik} = a_{ik}, k \in F_i \,\middle|\, \sum_{k \in F_i} A_{ik} = r\right) &= \frac{\Pr\left(A_{ik} = a_{ik}, k \in F_i \text{ and } \sum_{k \in F_i} A_{ik} = r\right)}{\Pr\left(\sum_{k \in F_i} A_{ik} = r\right)} \\
&= \frac{\dfrac{n!(1-P_i)^{(n-r)}}{(n-r)!} \displaystyle\prod_{k \in F_i} \frac{p_{ik}^{a_{ik}}}{a_{ik}!}}{\dfrac{n! P_i^r (1-P_i)^{n-r}}{r!(n-r)!}} \\
&= r! \prod_{k \in F_i} \left(\frac{p_{ik}}{P_i}\right)^{a_{ik}} \frac{1}{a_{ik}!},
\end{aligned}$$

with

$$\sum_{k \in F_i} a_{ik} = r.$$

This shows that, if the sum of A_{ik} is conditioned on one part of the population, the distribution remains multinomial and conditionally there is still an unequal probability design with replacement.

With the procedure in which we draw with replacement until we obtain r occupations in enterprise i, we have

$$\mathrm{E}\left(A_{ik} \mid X_i\right) = \begin{cases} \dfrac{rp_{ik}}{P_i} & \text{if } \ k \in F_i \\ \dfrac{X_i p_{ik}}{1-P_i} & \text{if } \ k \in D_i. \end{cases}$$

The expected value of A_{ik} is

$$\pi_{k \mid i} = \mathrm{EE}\left(A_{ik} \mid X_i\right) = \frac{rp_{ik}}{P_i},$$

$k \in L.$ The problem is that we know p_{ik}, r and $X_i,$ but not $P_i.$ We can estimate P_i using the method of moments by solving $\mathrm{E}\left(X_i\right) = X_i,$ which gives

$$X_i = \frac{r\left(1-\hat{P}_i\right)}{\hat{P}_i}$$

and therefore

$$\hat{P}_{i1} = \frac{r}{X_i + r}.$$

The maximum likelihood method provides the same estimator as the method of moments, but this estimator is biased (Mikulski and Smith 1976; Johnson et al. 2005, page 222). In fact, the unbiased minimum variance estimator is

$$\hat{P}_{i2} = \frac{r-1}{X_i + r - 1}.$$

However, $1/\hat{P}_{i1}$ is unbiased for $P_i.$

Again, since we are using weights that are inverses of $\pi_{k \mid i}.$ The inverses of $\pi_{k \mid i}$ are thus estimated as follows:

$$\widehat{1/\pi_{k \mid i}} = \begin{cases} \dfrac{\hat{P}_{i2}}{rp_{ik}} & = \dfrac{r-1}{\left(X_i + r - 1\right) rp_{ik}} \quad \text{if } \ k \in F_i \\ \dfrac{1-\hat{P}_{i2}}{X_i p_{ik}} & = \dfrac{1}{\left(X_i + r - 1\right) p_{ik}} \quad \text{if } \ k \in D_i. \end{cases} \tag{5.1}$$

6 Unequal probability sampling without replacement

6.1 Sequential sampling without replacement

For the draw without replacement, the first problem is determining the design. One option is to use the method by Ohlsson (1995) called sequential Poisson sampling. This method involves generating M uniform random variables in the interval $[0,1]$, denoted u_{ik}. Next, we select the n units corresponding to the smallest values of $u_{ik} / \pi_{k|i}$. This method has the advantage of being usable for any sample size and providing a sequence of samples that are included in each other. Unfortunately, it only satisfies approximately the fixed inclusion probabilities. However, the approximations are very accurate according to the simulations given in Ohlsson (1995).

Methods have also been proposed by Sampford (1962) and Pathak (1964). We propose an exact solution to the problem in the sense that the inclusion probabilities are exactly satisfied. We begin by calculating the inclusion probabilities for a design of fixed size n with inclusion probabilities proportional to a strictly positive auxiliary variable $b_k, k \in L.$ The probabilities are determined by

$$\pi_{k|i}(n) = \min\left(1, C_n \frac{b_k}{\sum_{\ell \in L} b_\ell}\right),$$

where C_n is determined such that

$$\sum_{k \in L} \pi_{k|i}(n) = \sum_{k \in L} \min\left(1, C_n \frac{b_k}{\sum_{\ell \in L} b_\ell}\right) = n.$$

A simple algorithm for calculating these probabilities is described in Tillé (2006, page 19), among others. The probabilities can be calculated simply using the function `inclusionprobabilities` in the R sampling package.

A sequential selection method must therefore select a sample of size n with inclusion probabilities $\pi_{k|i}(n)$. It must then make it possible to go from size n to size $n+1$ by simply selecting an additional unit such that the completed sample has an inclusion probability of $\pi_{k|i}(n+1)$. It appears that the only method that allows that to be achieved is the elimination method (Tillé 1996). This method starts with the entire population (the list of occupations) and eliminates one unit in each step. In step $j = 1, \ldots, N,$ the unit is eliminated from among the remaining units with the probability

$$1 - \frac{\pi_{k|i}(N-j)}{\pi_{k|i}(N-j+1)}.$$

This method can thus be used to create a sequence of samples included in each other that verify the inclusion probabilities in relation to their size.

Therefore, we can simply apply the elimination method for sample size $n = 1$ so that the algorithm successively eliminates all the units. Taking them in the reverse order of elimination, we obtain a sequence of units. The first n units of the sequence are selected with inclusion probability $\pi_{k|i}(n)$. The appendix contains a function written in R that can be used to generate this sequence. The code is executed in a simulation that shows that the probabilities obtained through simulations by applying this function are equal to the fixed inclusion probabilities for all sample sizes.

6.2 Inverse or negative design with unequal probabilities

Now that the design is defined, the inverse design can be defined. The units in the list of occupations are taken using the elimination method until r occupations in the enterprise are selected. In this case, the probability distribution of the number of failures X_i seems impossible to calculate. Calculating the conditional inclusion probability $\mathrm{E}\left(A_{ik} \mid X_i\right)$ is also problematic.

However, we can proceed by analogy and estimate the inclusion probabilities on the basis of expression (5.1) developed for the case with replacement, where p_{ik} can simply be replaced by

$$\frac{\pi_{k|i}\left(r + X_i\right)}{r + X_i}.$$

Therefore, we obtain

$$\widehat{1/\pi_{k|i}} = \begin{cases} \dfrac{(r-1)\left(r + X_i\right)}{r\left(X_i + r - 1\right)\pi_{k|i}\left(r + X_i\right)} & \text{if} \quad k \in F_i \\[2ex] \dfrac{r + X_i}{\left(X_i + r - 1\right)\pi_{k|i}\left(r + X_i\right)} & \text{if} \quad k \in D_i. \end{cases}$$

7 Discussion

The selection problem can therefore be resolved for all cases, with or without replacement and with equal or unequal probabilities. The proposed solution based on the elimination method respects the inclusion probabilities exactly, which is not true for Ohlsson's sequential sampling. The implementation is especially simple, since the program provides an ordered sequence of occupations to propose until the objective has been met.

The estimation issue is slightly more difficult. For the unequal probability sampling without replacement, we must make do with a heuristic solution. As well, it can be seen that, in the second stage, there tends to be lower inclusion probabilities in enterprises that have many occupations. This should lead us to select with greater probabilities the enterprises that may have a larger number of occupations, to avoid selecting occupations with probabilities that are too unequal.

Acknowledgements

The author wishes to thank Pierre Lavallée for submitting this interesting problem and providing thoughtful comments on an earlier version of this article. The author also thanks Audrey-Anne Vallée for her meticulous proofreading, and a referee and writer of *Survey Methodology* for their pertinent remarks, which made it possible to improve this article.

Appendix

```
#
# Load sampling package, which contains the function inclusionprobabilities().
#
library(sampling)
#
# The function returns a vector with the sequence numbers of the eliminations.
# The last (resp. first) unit eliminated is the first (resp. last)
# component of the vector.
# The function therefore provides the numbers of the units to be presented
# successively for the inverse selection.
# The argument x is the vector of values of the auxiliary variable used to calculate
# the inclusion probabilities.
#
elimination<-function(x)
       {
       pikb=x/sum(x)
       M = length(pikb)
       n = sum(pikb)
       sb = rep(1, M)
       b = rep(1, M)
       res=rep(0, M)
       for (i in 1:(M)) {
              a = inclusionprobabilities(pikb, M - i)
              v = 1 - a/b
              b = a
              p = v * sb
              p = cumsum(p)
              u = runif(1)
              for (j in 1:length(p)) if (u < p[j])
                     break
              sb[j] = 0
              res[i]=j
              }
       res[M:1]
       }
#
# 500,000 simulations with a size in a list of size M=20.
# By taking the first m components of vector v, we obtain a sample
# of size m.
#
M=20
x=runif(M)
Pik=array(0,c(M,M))
#
# Calculate the inclusion probabilities for all sample sizes from 1 to 20.
#
for(i in 1:M) Pik[i,]=inclusionprobabilities(x, i)
rowSums(Pik)

SIM=50000
SS=array(0,c(M,M))
for(i in 1:SIM)
{
S=array(0,c(M,M))
v=elimination(x)
for(i in 1:M) S[i,v[1:i]]=1
SS=SS+S
}
SS=SS/SIM
#
# Compare actual and empirical inclusion probabilities.
#
Pik
SS
SS-Pik
```

References

Chikkagoudar, M.S. (1966). A note on inverse sampling with equal probabilities. *Sankhyā,* A28, 93-96.

Chikkagoudar, M.S. (1969). Inverse sampling without replacement. *Australian Journal of Satistic*, 11, 155-165.

Hájek, J. (1971). Discussion of an essay on the logical foundations of survey sampling, part on by D. Basu. In *Foundations of Statistical Inference*, (Eds., V.P. Godambe and D.A. Sprott), page 326, Toronto, Canada. Holt, Rinehart, Winston.

Johnson, N.L., Kemp, A.W. and Kotz, S. (2005). *Univariate Discrete Distributions*. New York: John Wiley & Sons, Inc.

Mikulski, P.W., and Smith, P.J. (1976). A variance bound for unbiased estimation in inverse sampling. *Biometrika*, 63(1), 216-217.

Miller, G.K., and Fridell, S.L. (2007). A forgotten discrete distribution? Reviving the negative hypergeometric model. *The American Statistician*, 61(4), 347-350.

Murthy, M.N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhyā*, 18, 379-390.

Ohlsson, E. (1995). Sequential Poisson sampling. Research report 182, Stockholm University, Sweden.

Ohlsson, E. (1998). Sequential Poisson sampling. *Journal of Official Statistics*, 14, 149-162.

Pathak, P.K. (1964). On inverse sampling with unequal probabilities. *Biometrika*, 51, 185-193.

Rosén, B. (1997). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, 62, 159-191.

Salehi, M.M., and Seber, G.A.F. (2001). A new proof of Murthy’s estimator which applies to sequential sampling. *The Australian and New Zealand Journal of Statistics*, 43, 281-286.

Sampford, M.R. (1962). Methods of cluster sampling with and without replacement for clusters of unequal sizes. *Biometrika*, 49(1/2), 27-40.

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

Tillé, Y. (2006). *Sampling Algorithms*. New York: Springer.