Published in Journal of Statistical Planning and Inference 141, issue 1, 597-601, 2011
which should be used for any reference to this work

Simple random sampling with over-replacement

Erika Antal *, Yves Tillé

Institute of Statistics, University of Neuchâtel, Pierre à Mazel 7, 2000 Neuchâtel, Switzerland

Keywords:
Survey sampling
Simple random sampling with replacement
Discrete probability distribution
Resampling method

A B S T R A C T

There are several ways to select units with replacement and an equal inclusion expectation. We present a new sampling design called simple random sampling with over-replacement. Its interest lies in the high variance produced for the Horvitz–Thompson estimator. This characteristic could be useful for resampling methods.

0. Introduction

There are several methods for drawing a sample, different goals, and different situations that require different sampling designs. The most basic sampling procedures are simple random sampling with and without replacement. In this paper we show that there are several ways to select units with replacement with an equal inclusion expectation. A new method is proposed where the repetition of the units in the sample is more important than with usual simple random sampling with replacement. This sampling design called simple random sampling with over-replacement provides a larger variance. This property could be interesting for resampling methods. We show how to implement this design and we compare it to simple random sampling with and without replacement.

1. Main concept and notation

A sampling design on a population $U=\{1, \ldots, k, \ldots, N\}$ is a procedure that allows us to randomly select statistical units. Some statistical units can be selected several times in the sample. In survey sampling theory, it is usual to define a sample as a subset of the population U. However, this definition is rather restrictive because it is limited to samples for which the units are selected only once, i.e. when sampling is done without replacement.

A more flexible notation consists in defining a sampling design by a positive, discrete random vector $\mathbf{S}=(S_1,\ldots,S_k,\ldots,S_N)'$, where S_k is the number of times unit k is selected in the sample. The same notation can thus be used to define sampling designs with or without replacement. If the sample is selected without replacement, then S_k can only take the values 0 and 1. If the sample has a fixed sample size n, then $\sum_{k\in U} S_k = n$.

The inclusion expectation of unit k is $\pi_k = \mathrm{E}(S_k)$. Since a unit can be selected several times in the sample, π_k can take any nonnegative value. The joint inclusion expectation of two units k and ℓ is the expectation of the product of S_k and S_ℓ, i.e.

* Corresponding author.
E-mail addresses: erika.antal@unine.ch (E. Antal), yves.tille@unine.ch (Y. Tillé).

$\pi_{k\ell} = \mathrm{E}(S_k S_\ell)$. Moreover, $\Delta_{k\ell} = \mathrm{cov}[S_k, S_\ell] = \pi_{k\ell} - \pi_k \pi_\ell$. If the sample is selected without replacement, then the inclusion expectation is called inclusion probability.

Let $y_1, \ldots, y_N$ denote the values taken on the units of the population by an interest variable y. Suppose now that we want to estimate the total of these values $Y = \sum_{k \in U} y_k$. If all the $\pi_k > 0$, this total can be estimated without bias by $\widehat{Y} = \sum_{k \in U} S_k y_k / \pi_k$. This estimator is called the Horvitz–Thompson estimator if the sample is selected without replacement and the Hansen–Hurwitz estimator if the sample is selected with replacement (see Hansen and Hurwitz, 1949; Horvitz and Thompson, 1952).

The variance of $\widehat{Y}$ is

$$\mathrm{var}(\widehat{Y}) = \sum_{k \in U} \sum_{\ell \in U} \frac{y_k y_\ell}{\pi_k \pi_\ell} \Delta_{k\ell}.$$

If all the $\pi_{k\ell} > 0$, this variance can be estimated without bias by means of the following formula:

$$\widehat{\mathrm{var}}(\widehat{Y}) = \sum_{k \in U} \sum_{\ell \in U} \frac{S_k S_\ell y_k y_\ell}{\pi_k \pi_\ell} \frac{\Delta_{k\ell}}{\pi_{k\ell}}. \tag{1}$$

Nevertheless, this variance estimator is often very unstable. It can take negative values (see, for instance Tillé, 2006, pp. 26–29). When the sampling design has a fixed sample size, the variance can be written as

$$\mathrm{var}(\widehat{Y}) = \frac{-1}{2} \sum_{k \in U} \sum_{\ell \in U} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 \Delta_{k\ell},$$

and can be estimated by

$$\mathrm{var}(\widehat{Y}) = \frac{-1}{2} \sum_{k \in U} \sum_{\ell \in U} S_k S_\ell \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 \frac{\Delta_{k\ell}}{\pi_{k\ell}},$$

which can also be written under the quadratic form

$$\widehat{\mathrm{var}}_D(\widehat{Y}) = \sum_{k \in U} \sum_{\ell \in U} \frac{S_k S_\ell y_k y_\ell}{\pi_k \pi_\ell} D_{k\ell}, \tag{2}$$

with

$$D_{k\ell} = \begin{cases} \sum\limits_{\substack{j \in U \\ j \neq k}} S_j \dfrac{\Delta_{kj}}{\pi_{kj}} & \text{if } k = \ell, \\ \dfrac{\Delta_{k\ell}}{\pi_{k\ell}} & \text{if } k \neq \ell. \end{cases}$$

2. Simple random sampling without replacement

A sampling design is said to be simple and without replacement if $\Pr(\mathbf{S} = \mathbf{s}) = n!(N-n)!/N!$, for all $\mathbf{s} \in \mathcal{S}_n^N$, where $\mathcal{S}_n^N = \{\mathbf{s} \in \{0,1\}^N \mid \sum_{k=1}^N s_k = n\}$. In simple random sampling without replacement, $\Delta_{k\ell} = -n(N-n)/\{N^2(N-1)\}$, if $k \neq \ell \in U$ and $\Delta_{kk} = n(N-n)/N^2, k \in U$, which gives the variance of the estimator of the total $\mathrm{var}(\widehat{Y}) = N^2(N-n)\sigma^2/\{(N-1)n\}$, where $\sigma^2 = N^{-1}\sum_{k \in U}(y_k - \overline{Y})^2$, and $\overline{Y} = N^{-1}\sum_{k \in U} y_k$. Moreover, we have $\Delta_{k\ell}/\pi_{k\ell} = -(N-n)/\{N(n-1)\}$ when $k \neq \ell \in U$ and $\Delta_{kk}/\pi_{kk} = (N-n)/N, k \in U$, which gives $\widehat{\mathrm{var}}(\widehat{Y}) = N^2(N-n)\widehat{\sigma}^2/(Nn)$, where $\widehat{\sigma}^2 = (n-1)^{-1}\sum_{k \in U} S_k (y_k - \widehat{\overline{Y}})^2$, and $\widehat{\overline{Y}} = n^{-1}\sum_{k \in U} S_k y_k$.

3. Simple random sampling with replacement

A sampling design is said to be simple and with replacement if

$$\Pr(\mathbf{S} = \mathbf{s}) = \frac{1}{N^n} \binom{n}{s_1 \cdots s_k \cdots s_N}^{-1} \quad \text{for all } \mathbf{s} \in \mathcal{R}_n^N,$$

where $\mathcal{R}_n^N = \{\mathbf{s} \in \mathbb{N} \mid \sum_{k=1}^N s_k = n\}$. Vector $\mathbf{S}$ therefore has a multinomial distribution. A well-known result is that a multinomial distribution can be derived from a sequence of Poisson independent random variables that are conditioned on their sum. More formally, consider N random Poisson variable $X_1, \ldots, X_N$ with the same parameter λ, i.e.

$\Pr(X_k = x_k) = e^{-\lambda}\lambda^{x_k}/x_k!, x_k = 0,1,2,3,\ldots$. Then, one can prove that

$$\Pr\left(X_1 = x_1, \ldots, X_N = x_n \middle| \sum_{i=1}^{N} X_i\right) = \frac{1}{N^n}\binom{n}{s_1 \cdots s_k \cdots s_N}^{-1},$$

for all $(x_1, \ldots, x_N) \in \mathcal{R}_n^N$. The conditional distribution no longer depends on λ anymore (see Bol'shev, 1965; Johnson et al., 1997, p. 65).

Two ways of implementing simple random sampling with replacement are given in Tillé (2006, pp. 60–61). In simple random sampling with replacement, $\pi_k = n/N$ for all $k \in U$, and $\Delta_{k\ell} = -n(N-1)/\{N^2(N-1)\}$, when $k \neq \ell \in U$ and $\Delta_{k\ell} = n(N-1)/N^2, k \in U$, which gives the variance of the Hansen–Hurwitz estimator of the total $\operatorname{var}(\widehat{Y}) = N^2\sigma^2/n$. Moreover, we have

$$\frac{\Delta_{k\ell}}{\pi_{k\ell}} = \begin{cases} \dfrac{N-1}{N-1+n} & \text{if } k = \ell, \\ -\dfrac{1}{n-1} & \text{if } k \neq \ell. \end{cases}$$

Although it is possible to construct an unbiased estimator of the variance by using expression (1), the result obtained is very strange and should not be used (see Tillé, 2006, p. 58). It is nevertheless possible to construct an unbiased estimator by using the quadratic form based on the $D_{k\ell}$ given in expression (2)

$$D_{k\ell} = \begin{cases} 1 & \text{if } k = \ell, \\ -\dfrac{1}{n-1} & \text{if } k \neq \ell, \end{cases}$$

which gives $\widehat{\operatorname{var}}(\widehat{Y}) = N^2\widehat{\sigma}^2/n$.

4. Simple random sampling with over-replacement

Simple random sampling with replacement can be viewed as a conditional distribution of independent Poisson variables. What happens if instead of using the Poisson distribution, we use another discrete distribution? If we use a sequence of geometric random variables conditioned on their size, we obtain another sampling design with replacement with a fixed sample size. We have called this design simple random sampling with over-replacement because the repetitions of the units are more frequent than in a usual simple random sampling with replacement.

First, consider a sequence of N independent geometric random variables X_k: $\Pr(X_k = x_k) = (1-p)p^{x_k}, x_k = 0,1,2,3,\ldots$ with parameters $\pi_k \in (0,1)$. The sample size $n_s = \sum_{k=1}^{N} X_k$ is random. Let us now calculate the conditional geometric sample design. If S_k denotes the random variable that gives the number of times unit k is selected in the sample, we have

$$\begin{aligned} \Pr(S_1 = x_1, \ldots S_N = x_N) &= \Pr\left(X_1 = x_1, \ldots X_N = x_N \middle| \sum_{k=1}^{N} X_k = n\right) \\ &= \frac{\prod_{k=1}^{N}(1-p)p^{x_k}}{\sum_{\mathcal{R}_n^N}\prod_{k=1}^{N}(1-p)p^{x_k}} = \frac{q^N p^n}{\sum_{\mathcal{R}_n^N} q^N p^n} = \frac{1}{\operatorname{card}\mathcal{R}_n^N} = \frac{1}{\binom{N+n-1}{n}}. \end{aligned}$$

All the samples have exactly the same probability of being selected. By noting that

$$\#\mathcal{R}_n^N = \binom{N+n-1}{n} \quad \text{and} \quad \#\mathcal{R}_{n-j}^{N-1} = \binom{N-1+n-j-1}{n-j},$$

we can derive the marginal distribution of S_k:

$$\Pr(S_k = j) = \frac{\binom{N-1+n-j-1}{n-j}}{\binom{N+n-1}{n}}, \quad j = 0, \ldots, n,$$

which is an inverse (or negative) hypergeometric distribution (see Johnson et al., 1993, pp. 239, 264). We thus have $\mathrm{E}(S_k) = n/N$ and

$$\operatorname{var}(S_k) = \frac{n(N-1)(N+n)}{N^2(N+1)}.$$

This sampling design has a fixed sample size, which implies that $\sum_{k\in S}\operatorname{cov}(S_k, S_\ell) = \operatorname{cov}(n, S_\ell) = 0$. Moreover, since all the units are treated symmetrically, $\operatorname{cov}(S_k, S_\ell) = -\operatorname{var}(S_k)/(N-1)$. The matrix of $\Delta_{k\ell}$ is thus given by

$$\Delta_{k\ell} = \frac{(N-1)(N+n)n}{N^2(N+1)} \times \begin{cases} 1 & \text{if } k = \ell, \\ -\dfrac{1}{N-1} & \text{if } k \neq \ell, \end{cases}$$

Table 1
Comparison of the variance of the three simple designs.

Sampling design	Variance of the estimator of the total
Simple without replacement	$\frac{(N-n)N^2\sigma^2}{(N-1)n}$
Simple with replacement	$\frac{N^2\sigma^2}{n}$
Simple with over-replacement	$\frac{(N+n)N^2\sigma^2}{(N+1)n}$

which allows us to compute the variance of the Hansen–Hurwitz estimator:

$$\mathrm{var}(\widehat{Y}) = \frac{(N+n)N^2\sigma^2}{(N+1)n}.$$

This variance is much larger than the variance obtained under simple random sampling with replacement.

Simple random sampling with over-replacement can be implemented by a rejective procedure that consists in selecting geometric samples until a sample size n is obtained. Tillé (2006, p. 34) also proposed a general sequential algorithm in order to quickly generate multivariate random variables. This algorithm is based on the computation at each step of the conditional distribution probabilities of the S_k, that is

$$\Pr(S_k = j | S_{k-1}, \ldots, S_1) = \frac{\binom{N-k-1+n_k-j}{n_k-j}}{\binom{N-k+n_k}{n_k}}, \quad j = 0,1,2,3,\ldots,n_k,$$

where $n_1 = n$ and

$$n_k = n - \sum_{j=1}^{k-1} S_j, \quad k = 2,\ldots,N.$$

Algorithm 1is the application of the general algorithm presented in Tillé (2006, p. 34) to sampling with over-replacement. It provides an efficient implementation of sampling with over-replacement.

Algorithm 1. Algorithm for simple random sampling with over-replacement

- For k=1,…,N unit k is selected S_k times, where
$$\Pr(S_k = j) = \frac{\binom{N-k-1+n_k-j}{n_k-j}}{\binom{N-k+n_k}{n_k}}, \quad j = 0,1,2,3,\ldots,n_k.$$

5. Discussion

Table 1 shows the three variances of simple designs. Compared to simple random sampling with replacement, we find that simple random sampling without replacement and simple random sampling with over-replacement have a symmetric position. Indeed, for random sampling without replacement, the finite population correction factor is $(N-n)/(N-1)$ and for simple random sampling with over-replacement, the over-replacement correction factor is $(N+ n)/(N+1)$.

Simple random sampling with over-replacement is interesting because it shows that there are several methods of sampling with replacement that have an equal inclusion expectation in the sample. It is also possible to define a large range of simple random sampling by combining several simple random sampling designs. For instance, one can select a subset of observations by simple random sampling with replacement and a second subset by simple random sampling with over-replacement. So, a large range of sampling designs with replacement can be defined with different variances of the estimator of the total. Antal and Tillé (2010) have used simple random sampling with over-replacement to construct new bootstrap methods for complex sampling designs. The main idea consists of mixing simple random sampling with over-replacement with other sampling designs in order to construct ad hoc resampling designs for reproducing the correct estimator of variance in a complex sampling design. Sampling with over-replacement is thus not only a simple mathematical curiosity but can be used in practical applications.

References

Antal, E., Tillé, Y., 2010. A direct bootstrap method for complex sampling designs from a finite population, submitted for publication. Technical Report, University of Neuchâtel.

Bol'shev, L.N., 1965. On a characterization of the Poisson distribution. Teoriya Veroyatnostei i ee Primeneniya 10, 64–71.

Hansen, M.H., Hurwitz, W.N., 1949. On the determination of the optimum probabilities in sampling. Annals of Mathematical Statistics 20, 426–432.
Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47, 663–685.
Johnson, N., Kotz, S., Kemp, A., 1993. Univariate Discrete Distributions. Wiley, New York.
Johnson, N.L., Kotz, S., Balakrishnan, N., 1997. Discrete Multivariate Distributions. Wiley, New York.
Tillé, Y., 2006. Sampling Algorithms. Springer, New York.