On Variance minimization for unequal probability sampling

A. Čiginas

Vilnius University, Lithuania; Statistics Lithuania, Lithuania
e-mail: andrius.ciginas@maf.vu.lt

Abstract

We consider sampling designs, where inclusion (to sample) probabilities are mixtures of two components. The first component is proportional to the size of a population unit (described by means of an auxiliary information available). The second component is the same for every unit. We look for mixtures that minimize variances of various estimators of the population total and show how auxiliary information could help to find an approximate location of such mixtures.

We report theoretical and simulation results in the case of Poisson samples drawn from populations which are generated by a linear regression model.

1 Introduction

Consider the population $\mathcal{U} = \{u_1, \dots, u_N\}$ and assume that we want to estimate the population parameter $t_y = \sum_{1 \le i \le N} y_i$, where $y_i = y(u_i)$ denotes a measurement of the population unit u_i. For this purpose we draw a sample s from $\mathcal{U}$. Assume that an auxiliary information is available in the form of the vector $x = (x_1, \dots, x_N)$ with positive coordinates. We call x_i the size of the unit u_i. In the case where the variables y and x are highly correlated it is convenient to take into account the relative weights

$$p_i = x_i / t_x, \qquad t_x = \sum_{i=1}^{N} x_i, \tag{1}$$

when choosing the sampling design. For instance, one can define inclusion (to sample s) probabilities $\pi_i = P(u_i \in s)$ proportional to p_i,

$$\pi_i \approx c p_i, \qquad 1 \le i \le N. \tag{2}$$

Kröger, Särndal and Teikari (2003) give examples of skewed populations and sampling designs with inclusion probabilities close to (2) where the variances of several popular estimators $\hat{t}_y$ of the population total t_y are considerably larger then the variances of the same estimators, but with inclusion probabilities $\pi_i \approx c p_i(h)$, where

$$p_i(h) = (1-h) p_i + h/N, \qquad 1 \le i \le N. \tag{3}$$

Here $h \in [0, 1]$. They consider Horvitz-Thompson, regression and generalized regression (GREG) estimators and sampling designs, where sampling is without replacement and with a fixed sample size. The simulation study shows that the variances of these particular examples are minimized for $h \in (0.2; 0.5)$.

Let us call the value h^* of the parameter h optimal if it minimizes the variance. Generally, it is impossible to find h^* without complete knowledge of the population. Much easier question is whether $h^* > 0$ (i.e., whether inclusion probabilities with the uniform component are preferable) for various sampling designs, various populations and estimators $\hat{t}_y$. Another interesting question is how to make a decision about the location of the minimizer h^*, based on the auxiliary information available.

An attempt to answer these questions in some simple situations is made in the present article. Let us outline our approach. Assume that the population point scatter $\{(y_i, x_i) : 1 \le i \le N\}$ looks as if it had been generated according to a probabilistic model, where $y_1, \dots, y_N$ are assumed to be realized values of independent random variables $Y_1, \dots, Y_N$. Given an estimator $\hat{t}_y$ based on the sample s with the inclusion probabilities $\pi_i \approx cp_i(h)$, let $D_h^* = D_h^*(\hat{t}_y)$ denote the conditional variance of $\hat{t}_y$ given $Y_1 = y_1, \dots, Y_N = y_N$. Furthermore, let D_h denote the expected value of this variance, i.e., $D_h = \mathbf{E}D_h^*$. Assume, for the moment, that in the interval $0 \le h \le 1$ the function $h \to D_h$ has the unique minimizer

$$h_0 = argmin(D_h). \tag{4}$$

Then one may expect that, by the law of large numbers, for large N, the number h_0 is close to the minimizer of the function $h \to D_h^*(\hat{t}_y)$. Therefore, h_0 can be considered as an approximation to the unknown random variable h^*. In order to access the quality of the approximation one would like to evaluate the mean square error $\mathbf{E}(h^* - h_0)^2$ and to compare (expected) values of the target function: $D_{h^*}^*$, $D_{h_0}^*$, D_0^* and D_1^*.

In this article we study the simplest case of the Poisson sample drawn from a population which is generated by a linear regression model (see Särndal, Swensson and Wretman (1992), 226 p.). We have chosen the Poisson sample as a modelling example since here (unique) solutions to the corresponding minimization problems are available and the analysis is relatively simple and lucid.

The article is organized as follows. In Section 2 we introduce the population model and derive the inequality $h_0 > 0$ for two commonly used estimators: Horvitz-Thompson and regression estimator. The approximation h_0 to the random variable h^* can be find numerically, but we also propose explicit approximations to h^*. Examples of a simulation study are reported in Section 3. They demonstrate the empirical evidence of the accuracy of the approximation $h^* \approx h_0$.

2 Results

1. Population. We shall assume that $y_1, \dots y_N$ are realized values of independent random variables $Y_1, \dots, Y_N$ such that for every k,

$$\mathbf{E}(Y_k) = \beta_1 + \beta_2 x_k, \qquad \mathbf{V}(Y_k) = \sigma_k^2. \tag{5}$$

Here $\sigma_1, \dots \sigma_N$ and $x_1, \dots, x_N$ are non-random numbers and $x_k > 0$ for every k. We assume in what follows that $\beta_2 \ne 0$. Later we will assume that $\sigma_k^2 = \sigma^2 x_k^\gamma$, $1 \le k \le N$, $\gamma \in [0, 2]$.

2. Poisson sample includes the unit u_k in the sample s with probability π_k so that the inclusion events for different units are independent. In particular, the random variables $\mathbb{I}_k := \mathbb{I}_{\{u_k \in s\}}$ are independent. Given $n < N$ and $h \in [0,1]$ we choose probabilities

$$\pi_k = \pi_k(h) = np_k(h), \quad 1 \le k \le N. \tag{6}$$

Then the expected sample size

$$E(\mathbb{I}_1 + \cdots + \mathbb{I}_N) = \pi_1(h) + \cdots + \pi_N(h) = n.$$

For simplicity of notation we shall assume in what follows that

$$\pi_k(0) < 1, \qquad \text{for every} \qquad k = 1, \ldots, N. \tag{7}$$

Then $\pi_k(h) < 1$ for every $h \in [0,1]$ and $k = 1, \ldots, N$.
We shall show that in the case of the Poisson sample the functions $h \to D_h^*$ and $h \to D_h$ are convex for Horvitz-Thompson and regression estimator. Therefore, the numbers $h_0 = argminD_h$ and $h^* = argminD_h^*$ are well defined.

3. Horvitz-Thompson estimator (HT estimator for short)

$$\hat{t}_{yHT} = \sum_{i=1}^{N} \mathbb{I}_i y_i \pi_i^{-1}$$

is unbiased and its variance

$$D_h^* = \sum_{i=1}^{N} y_i^2 (1 - \pi_i) \pi_i^{-1}. \tag{8}$$

Proposition 1. *The functions $h \to D_h$ and $h \to D_h^*$ are convex. These functions are constants whenever $p_i = N^{-1}$ for every $i = 1, \ldots, N$.*

The next Proposition 2 shows that very often we have $h_0 > 0$. Therefore, the inclusion probabilities (6) with equal probability sampling component of size $h_0 > 0$ lead to a lower variance of HT estimator than the traditional choice of inclusion probabilities (2).

Proposition 2. *Assume that $\sigma_i^2 = \sigma^2 x_i^\gamma$, $\gamma \in [0,2]$. Assume that at least two of probabilities $\{p_i\}$ are distinct.*
(i) Assume that $\gamma \in [0,2)$. If $\beta_1\beta_2 > 0$ then $0 < h_0 < 1$. If $\beta_1 = 0, \beta_2 \neq 0$ then we have $0 < h_0 < 1$ for $\sigma^2 > 0$ and $h_0 = 0$ for $\sigma^2 = 0$.
(ii) Assume that $\gamma = 2$. If $\beta_1\beta_2 > 0$ then $0 < h_0 < 1$. If $\beta_1 = 0, \beta_2 \neq 0$ then $h_0 = 0$.

In our presentation at the conference we shall refer results of a simulation study where the values of variances D_h^* are compared for $h = h^*$, $h = h_0$, $h = 0$ and $h = 1$.

4. Regression estimator. It is convenient to treat the cases $\beta_1 = 0$ and $\beta_1 \neq 0$ separately.
4.1. Assume that $\beta_1 = 0$. In this case the regression estimator can be written in the form (see Särndal, Svensson, Wretman (1992))

$$\hat{t}_{yr} = \hat{t}_{yHT} + \hat{B}(t_x - \hat{t}_{xHT}),$$

where

$$\hat{t}_{xHT} = \sum_{k=1}^{N} \mathbb{I}_k x_k \pi_k^{-1}, \qquad \hat{B} = \Big(\sum_{k=1}^{N} \mathbb{I}_k \frac{x_k^2}{\sigma_k^2 \pi_k}\Big)^{-1} \sum_{k=1}^{N} \mathbb{I}_k \frac{x_k y_k}{\sigma_k^2 \pi_k}.$$

The variance formula is rather complex and, therefore, it is convenient to deal with the approximate variance (see ibidem),

$$D_h^* = \sum_{k=1}^{N} (y_k - Bx_k)^2 (1-\pi_k)\pi_k^{-1}, \quad \text{where} \quad B = D^{-1} \sum_{k=1}^{N} \frac{x_k y_k}{\sigma_k^2}$$

and $D = \sum_{k=1}^{N} \sigma_k^{-2} x_k^2$. A simple calculation shows that the expected value $D_h = \mathbf{E} D_h^*$ can be written in the form

$$D_h = \sum_{k=1}^{N} \Big(\frac{1}{n} \frac{1}{p_k(h)} - 1\Big)(\sigma_k^2 - D^{-1} x_k^2). \tag{9}$$

The same argument as above shows that the functions $h \to D_h^*$ and $h \to D_h$ are convex.

4.2. Assume that $\beta_1 \neq 0$. In this case the population size N can be considered as an auxiliary information and we have the regression estimator (see Särndal, Svensson, Wretman (1992))

$$\hat{t}_{yr} = \hat{t}_{yHT} + \hat{B}_1 (N - \hat{t}_{1HT}) + \hat{B}_2 (t_x - \hat{t}_{xHT}).$$

Here $\hat{t}_{1HT} = \sum_{i=1}^{N} \mathbb{I}_i \pi_i^{-1}$. The coefficients

$$\begin{pmatrix} \hat{B}_1 \\ \hat{B}_2 \end{pmatrix} = \Big(\sum_{i=1}^{N} \mathbb{I}_i X_i X_i' / \sigma_i^2 \pi_i\Big)^{-1} \sum_{i=1}^{N} \mathbb{I}_i X_i y_i / \sigma_i^2 \pi_i,$$

where $X_i = \binom{1}{x_i}$. The variance formula of this estimator is rather complex and we shall consider the approximate variance instead (see ibidem)

$$D_h^* = \sum_{k=1}^{N} (\pi_k^{-1} - 1)(y_k - B_1 - x_k B_2)^2, \tag{10}$$

where

$$\begin{pmatrix} B_1 \\ B_2 \end{pmatrix} = \Big(\sum_{i=1}^{N} X_i X_i' / \sigma_i^2\Big)^{-1} \sum_{i=1}^{N} X_i y_i / \sigma_i^2.$$

It is convenient to write the function $D_h = \mathbf{E} D_h^*$ in the form

$$D_h = \sum_{k=1}^{N} \Big(\frac{1}{n} \frac{1}{p_k(h)} - 1\Big) \Big(\sigma_k^2 - \frac{1}{W}(D - 2Gx_k + Hx_k^2)\Big), \tag{11}$$

where we denote

$$D = \sum_{k=1}^{N} \frac{x_k^2}{\sigma_k^2}, \qquad G = \sum_{k=1}^{N} \frac{x_k}{\sigma_k^2}, \qquad H = \sum_{k=1}^{N} \frac{1}{\sigma_k^2}, \qquad W = DH - G^2.$$

The same argument as above shows that functions $h \to D_h^*$ and $h \to D_h$ are convex.

In both cases expressions of the functions (9) and (11) are complicated for further theoretical analysis (the minimization problem of the functions (9) and (11) can be easily solved numerically), so we shall consider the approximation (see Särndal, Svensson, Wretman (1992))

$$D_h \simeq \sum_{k=1}^{N} \Big(\frac{1}{n}\frac{1}{p_k(h)} - 1\Big)\sigma_k^2. \tag{12}$$

This approximation is convex function too.

Proposition 3. *Assume that $\sigma_i^2 = \sigma^2 x_i^\gamma$, $\gamma \in [0,2]$. Assume that at least two of probabilities $\{p_i\}$ are distinct. Assume that the functions (9) and (11) are changed by approximation (12). Let $\sigma^2 > 0$.*
(i) Assume that $\gamma = 0$. Then $h_0 = 1$.
(ii) Assume that $\gamma \in (0,2)$. Then $0 < h_0 < 1$.
(iii) Assume that $\gamma = 2$. Then $h_0 = 0$.

5. Explicit approximations to h^*. Assume that $\sigma_k^2 = \sigma^2 x_k^\gamma$, $1 \le k \le N$, $\gamma \in [0,2]$. For HT estimator (after some analytical and statistical assumptions) we have

$$h^* \approx h_{HT} = \frac{\beta_1 + \frac{cv(y)}{2}(1-\frac{\gamma}{2})\sigma}{\beta_1 + \beta_2\mu_x + \frac{cv(y)}{2}\mu_\sigma}, \tag{13}$$

where $\mu_x = t_x/N$, $\mu_\sigma = \frac{1}{N}\sum_{i=1}^N \sigma_i$ and $cv(y)$ is the coefficient of variation of y in the population $\mathcal{U}$.
For regression estimator can be similarly derived

$$h^* \approx h_R = \frac{(1-\frac{\gamma}{2})\sigma}{\mu_\sigma}. \tag{14}$$

3 Simulation examples

We fix population size $N = 1000$, expected sample size $n = 100$. Consider auxiliary information vector $\tilde{x}_E$ with coordinates

$$x_i = \Big|\log(1 - \frac{i-0.5}{N})\Big|, \qquad 1 \le i \le N.$$

Note that this auxiliary information vector satisfy the condition (7).
Given an auxiliary information vector $\tilde{x}_E$ consider the population models $y_i = 2 + x_i + \sigma_i\eta_i$, where $\sigma_i^2 = \sigma^2 x_i^\gamma$, $\gamma \in \{0; 0.5; 1; 1.5; 2\}$, $1 \le i \le N$. Here $\eta_1, \eta_2, \ldots$ denotes the sequence of independent standard normal random variables. For every γ we choose the value of σ so that the expectation of the coefficient of correlation between $\tilde{x}_E$ and y is near 0.9.
The first table report the simulation study of the HT estimator variance (8) and the second table report the simulation study of the regression estimator variance (10). Columns

Table 1 HT estimator

γ	h_0	$\frac{D_{h_0}}{D_0}$	$\frac{D_{h_0}}{D_1}$	E_1	E_2	E_3	E_4
0.0	0.675	0.2106	0.8940	0.2112	0.8937	0.9999	0.9999
0.5	0.668	0.2178	0.8905	0.2176	0.8895	0.9999	0.9999
1.0	0.663	0.2185	0.8877	0.2189	0.8889	0.9999	0.9999
1.5	0.660	0.2183	0.8856	0.2182	0.8865	0.9999	0.9999
2.0	0.658	0.2182	0.8836	0.2184	0.8836	0.9999	0.9999
	$\mathbf{E}(h^* - h_0)^2$	$cv(y)$	h_{HT}	V_1	V_2	V_3	V_4
0.0	3.37E-05	0.372	0.676	4.29E-04	8.33E-06	1.42E-09	1.60E-09
0.5	2.93E-05	0.361	0.671	2.73E-05	1.31E-05	1.27E-09	2.75E-09
1.0	3.07E-05	0.363	0.664	3.49E-06	2.02E-05	1.19E-09	1.10E-09
1.5	4.96E-05	0.356	0.658	2.91E-06	3.03E-05	3.56E-09	4.63E-09
2.0	6.22E-05	0.390	0.652	2.70E-06	4.76E-05	8.00E-09	1.61E-08

Table 2 Regression estimator

γ	h_0	$\frac{D_{h_0}}{D_0}$	$\frac{D_{h_0}}{D_1}$	E_1	E_2	E_3	E_4
0.0	1.000	0.1099	1.0000	0.1227	1.0000	0.9998	0.9998
0.5	0.716	0.4496	0.9362	0.4460	0.9345	0.9994	0.9903
1.0	0.419	0.7831	0.7832	0.7814	0.7857	0.9995	0.9866
1.5	0.161	0.9587	0.6045	0.9582	0.6048	0.9994	0.9895
2.0	0.000	1.0000	0.4458	1.0000	0.4434	0.9999	0.9999
	$\mathbf{E}(h^* - h_0)^2$	$cv(y)$	h_R	V_1	V_2	V_3	V_4
0.0	1.98E-04	0.372	1.000	1.18E-03	4.09E-31	2.21E-07	2.21E-07
0.5	8.33E-04	0.361	0.827	1.06E-03	2.05E-04	3.96E-07	2.36E-05
1.0	8.33E-04	0.363	0.564	4.47E-04	4.86E-04	5.44E-07	2.73E-05
1.5	6.41E-04	0.356	0.272	8.58E-05	1.01E-03	6.45E-07	2.47E-05
2.0	1.65E-05	0.390	0.000	2.09E-31	6.04E-04	2.65E-08	2.65E-08

E_1-E_4 shows the means of the ratios $\frac{D^*_{h_0}}{D^*_0}$, $\frac{D^*_{h_0}}{D^*_1}$, $\frac{D^*_{h^*}}{D^*_{h_0}}$, $\frac{D^*_{h^*}}{D^*_{h_{HT}}}$ (or $\frac{D^*_{h^*}}{D^*_{h_R}}$ for regression estimator) respectively and V_1-V_4 - their variances. Expected values given in the columns E_1-E_4, V_1-V_4 and the mean square error $\mathbf{E}(h^* - h_0)^2$ are evaluated using a tiny Monte Carlo study. We generate 50 independent copies of a given population and evaluate empirical mean values of the parameters of interest. Quantity $cv(y)$ is evaluated using first copy of a given population.

References

Särndal, C.E., Swensson, B., Wretman, J. (1992) *Model assisted survey sampling.* (Springer series in Statistics) Springer-Verlag Berlin, Heidelberg, New York.

Kröger, H., Särndal, C.E., Teikari, I. (2003). Poisson mixture sampling combined with order sampling, *Journal of Official Statistics*, 19, 59–70.

M. Bloznelis and A. Čiginas, On Variance minimization for unequal probability sampling, *Vilnius Univ. Preprint 05-11.*