On influence of clustering population on estimation accuracy of population total

Janusz L. Wywiał, Grzegorz Sitek

Department of Statistics, Econometrics and Mathematics
University of Economics in Katowice, Poland,
janusz.wywial@ue.katowice.pl, grzegors12@wp.pl

Survey Sampling in Economic and Social Research,
Katowice,5-6 June 2017

This paper is a result of the grant supported by *National Science Centre, Poland*, no. 2016/21/B/HS4/00666.

Contents:

Estimation based on cluster sample

Basic notation

- U is population of size N,
- $\boldsymbol{y}_k = [y_{k,1} ... y_{k,m}]$ where $k \in U$,
- $\mathcal{D} = \{U_1, ..., U_h, ..., U_G\}$ is partition of U into clusters,

$$\bar{y} = \sum_{k \in U} y_k / N, \quad y_U = N\bar{y} = \sum_{k \in U} y_k,$$

$$\bar{y}_{U_h} = \sum_{k \in U_h} y_k / N_h, \quad v_{y,U_h} = \sum_{k \in U_h} (y_k - \bar{y}_{U_h})^2 / (N_h - 1),$$

$$v_{*,y} = \frac{1}{N-G} \sum_{h=1}^{G} \sum_{k \in U_h} (y_k - \bar{y}_{U_h})^2 = \frac{1}{G} \sum_{h=1}^{G} w_h v_{y,U_h}, \quad w_h = \frac{N_h - 1}{N - G}.$$

$$\bar{y}_{\mathcal{U}} = \sum_{h=1}^{G} y_{U_h} / G = y_U / G, \quad v_{y,\mathcal{U}} = \sum_{h=1}^{G} (y_{U_h} - \bar{y}_{\mathcal{U}})^2 / (G - 1).$$

Estimation based on cluster sample

Simple cluster sample estimator

$$\tilde{y}_S = \frac{G}{g} \sum_{h \in s} \sum_{k \in U_h} y_k = \frac{G}{g} \sum_{h \in s} y_{U_h},$$

$$V(\tilde{y}_S, P_1(s)) = \frac{G(G-g)}{g} v_{y,\mathcal{U}}, \quad P_1(s) = \binom{G}{g}^{-1}, \qquad (1)$$

$$V(\tilde{y}_S, P_1(s)) = \frac{G(G-g)}{g} \bar{N} v_y \left(1 + \frac{N-G}{G-1} \delta_y\right) + \frac{G(G-g)}{g} a_y,$$

homogeneity coefficient: $\delta_y = 1 - \frac{v_{*,y}}{v_y}, \quad \delta_y \in \left[-\frac{G-1}{N-G}; 1\right],$

$$a_y = \frac{1}{G-1} \sum_{h=1}^{G} (N_h - \bar{N}) N_h \bar{y}_{U_h} \bar{y}_{U_h}.$$

Estimation based on cluster sample

Simple cluster sample estimator

$$y_S = \frac{N}{n}\sum_{k\in S} y_k, \quad V(y_S, P_0(s)) = \frac{N(N-n)}{n} v_y, \quad P_0(s) = \binom{N}{n}^{-1}.$$

Relative efficiency:

$$deff(\tilde{y}_S) = 1 + \frac{N-G}{G-1}\delta_y + \frac{a_y}{\bar{N} v_y}.$$

If $N_h = const$ for all $h = 1, ..., G$:

$$0 \leq deff(\tilde{y}_S) = 1 + \frac{N-G}{G-1}\delta_y \leq \frac{N-1}{G-1}.$$

Estimation based on cluster sample
Sampford sampling scheme

Sampford's (1967) sampling design:

$$P_2(s) = c_1 \left(\sum_{h=1, h \notin s} \pi_h \right) \prod_{h \in s} \frac{\pi_h}{1 - \pi_h}, \qquad s \in \mathbf{S}_{\mathcal{U}}$$

where c_1 is such that $\sum_{s \in \mathbf{S}_{\mathcal{U}}} P_2(s) = 1$.

The sampling scheme:

- The first unit is drawn with probab. π_h/g, $h = 1, ..., G$,
- next $g - 1$ units are drawn with probab. $\propto \frac{\pi_h}{1-\pi_h}$,
- $\pi_h = \frac{g x_{U_h}}{\sum_{i \in \mathcal{U}} x_{U_i}} = \frac{g x_{U_h}}{x_U}$.

Estimation based on cluster sample

Horvitz-Thompson estimator from Sampford's sample

$$y_{HT} = \sum_{h \in \mathcal{U}} \frac{y_{U_h}}{\pi_h},$$

When the effective sample size is fixed, then:

$$V\left(y_{HT}, P_2(s)\right) = \sum_{h=1}^{G} \sum_{i=1, i \neq h}^{G} \left(\frac{y_{U_h}}{\pi_h} - \frac{y_{U_i}}{\pi_i} \right)^2 (\pi_{h,i} - \pi_h \pi_i).$$

When $y_{U_h} = \beta x_{U_h} + e_{U_h}$, $k \in \mathcal{U}$, then:

$$\left(\frac{y_{U_h}}{\pi_h} - \frac{y_{U_i}}{\pi_i} \right)^2 = \beta^2 \left(\frac{e_{U_h}}{x_{U_h}} - \frac{e_{U_i}}{x_{U_i}} \right)^2 \to 0, \quad \text{for all} \quad h \in \mathcal{U}$$

and $V\left(y_{HT}, P_2(s)\right) \to 0$.

Clustering algorithms
First systematic algorithm $\mathcal{U}_1$

- let us assume that $x_k < x_{k+1}$ for $k = 1, ..., N-1$,
- h-th cluster is identified by such $k \in U_h$ so that $k = (i-1)G + h$, for $i = 1, ..., M$ and $h = 1, ..., G$,

The set $\mathcal{U}_1$ is the simple systematics sample space.

- When M is even and $N = MG$, then

$$U_h = \{(h-1)\frac{M}{2} + i; N - (h-1)\frac{M}{2} - i + 1\}$$

for $h = 1, ..., G$ and $i = 1, ...M/2$.

- Particularly, when $M = 2$ and $N = MG$:

$$U_h = \{h; N - h + 1\}.$$

- In this case:

$$\pi_h = \frac{g(x_{N-h+1} + x_h)}{x_U}$$

Clustering algorithms
Permutation algorithm $\mathcal{U}_3$

- Let $\mathcal{U}^{(0)} = \{U_1^{(0)}, ..., U_G^{(0)}\}$ be any start partition of population into clusters of the same sizes,
- in the t-h (t=0,1,...) iteration partition $\mathcal{U}^{(t)} = \{U_1^{(t)}, ..., U_G^{(t)}\}$ is generated through permutation pop. elements of U,
- the intra-cluster variance of variable x is evaluated for $\mathcal{U}^{(t)}$:

$$v^{(t)} = v_{*,x,\mathcal{U}^{(t)}}$$

- $\mathcal{D}_3$ is treated as optimal when

$$\mathcal{D}_3 = arg(max_{\{t=1,..,T\}}(v^{(t)})) = arg(min_{\{t=1,..,T\}}(\delta^{(t)}))$$

where $\delta^{(t)} = 1 - \frac{v^{(t)}}{v_x}$.

Clustering algorithms
Algorithm $\mathcal{U}_4$

- clusters sizes of $\mathcal{U}^{(t)}$ are not necessary the same;
- let $f : U \to \mathcal{U}^{(t)}$, $f_t(k) = h$, if and only if $k \in U_h^{(t)}$;
- at stage $t+1$ $k \in U_h^{(t)}$ is moved $U_z^{(t)}$, $z \neq h$, $z = 1, ..., G$,

$$(\underline{k}, \underline{z}) = arg \left(min_{\{z=1,...,G; k \in U\}} \left(v^{(t)}(k, z) \right) \right);$$

- $v^{(t)}(k, z)$ is the intra-cluster variance of x evaluated for $\mathcal{U}^{(t)}$ where $U_z^{(t)}$ and $U_h^{(t)}$ are replaced by $\{U_z^{(t)} \cup \{k\}\}$ and $\{U_h^{(t)} - \{k\}\}$, respectively;
- If $v^{(t)}(\underline{k}, \underline{z}) < v^{(t)}$, then $v^{(t+1)} = v^{(t)}(\underline{k}, \underline{z})$ and $\mathcal{U}^{(t+1)}$ is replaced with $\mathcal{U}^{(t)}$ where $U_{\underline{z}}^{(t)}$ and $U_{\underline{h}}^{(t)}$,$\underline{h} = f_t(\underline{k})$, are replaced with $U_{\underline{z}}^{(t+1)} = \{U_{\underline{z}}^{(t)} \cup \{\underline{k}\}\}$ and $U_{\underline{h}}^{(t+1)} = \{U_{\underline{h}}^{(t)} - \{\underline{k}\}\}$, respectively;
- the iteration process is stopped when $v^{(t)}(\underline{k}, \underline{z}) \geq v^{(t)}$.

Clustering algorithms
Algorithms $\mathcal{U}_5$ and $\mathcal{U}_6$

Algorithm 5: $h = 1, ...G$, $k = 1, .., N-1$,

$$N_h = N_1 + \Delta(h-1), \quad \Delta \leq \frac{1}{3}N - N_1, \quad \Delta = 1, 2, ...$$

$$N = \sum_{h-1}^{G} N_h = G(N_1 - \Delta) + \Delta G(G+1)/2, \quad \bar{N} = N_1 - \Delta + \Delta(G+1)/2;$$

if $x_k \geq x_{k+1}$, then $U_h = \{x_{\underline{N}_{h-1}+1}, x_{\underline{N}_{h-1}+2}, ..., x_{\underline{N}_{h-1}+N_h-1}, x_{\underline{N}_h}\}$

$$\underline{N}_h = \sum_{i=1}^{h} N_i = \underline{N}_{h-1} + N_h, \quad \underline{N}_0 = 0 \quad \underline{N}_1 = N_1 \quad \underline{N}_G = N.$$

If $N_1 = 2$, $\Delta = 1$, then $U_1 = \{x_1, x_2\}$, $U_2 = \{x_3, x_4, x_5\}$,
$U_3 = \{x_6, x_7, x_8\}$, $U_4 = \{x_9, x_{10}, x_{11}\}$,...

Algorithm 6: The same as Algorithm 5 but for $x_k \leq x_{k+1}$.

Sampling from stratified population
Estimation under several localisation of samples in strata

- Sampling design: $P_3(s) = \prod_{h=1}^{H} \binom{N_h}{n_h}^{-1}, \quad n = \sum_{h=1}^{H} n_h;$
- Estimator: $\hat{y}_s = \sum_{h=1}^{H} N_h \bar{y}_{s_h};$
- if $n_h = n w_h, \quad deff\,(\hat{y}_s) = \sum_{h=1}^{H} w_h v_{y,U_h} / v_y,\ w_h = N_h/N;$
- if $n_h = n/H, \quad deff\,(\hat{y}_s) = \sum_{h=1}^{H} w_h v_{y,U_h} / (H v_y);$
- if $n_h \propto N_h \sqrt{v_{U_h}}, \quad deff\,(\hat{y}_s) = \frac{\left(\sum_{h=1}^{H} N_h \sqrt{v_{U_h}}\right)^2 - n \sum_{h=1}^{H} N_h v_{U_h}}{N(N-n)v}.$

Sampling from stratified population

Stratification lgorithms $\mathcal{U}_7$ and $\mathcal{U}_8$

- Algorithm 7: The interval with values of an auxiliary variable is divided into sub-intervals (strata) of the same length treated as strata U_h of sizes N_h, $h = 1, ..., H$;
- Algorithm 8: The interval with values of an auxiliary variable is partitioned into sub-intervals in such a way that amounts of data in the strata are the same.

Sampling from stratified population

Stratification algorithm $\mathcal{U}_9$ by Dalenius et al. (1957, 1959)

- Evaluation of such intermediate boundaries $\underline{x}_h$ of the intervals that $\frac{1}{n}\left(\sum_{h=1}^{H} N_h\sqrt{v_{U_h}}\right)^2 - \sum_{h=1}^{H} N_h v_{U_h} = minim$;
- the range of x is partitioned into intervals $I_i = (\check{x}_i; \check{x}_i + q]$, $i = 1, ..., B-1$ and N_i is the number of x_k observed in I_i.
- the intermediate boundaries are determined as follows:

$$\underline{x}_h = \check{x}_i : \quad (h-1)\bar{z} < z_i \leq h\bar{z}, \quad h = 1, ..., H-1, \quad i = 1, ..., B.$$

where

$$z_i = \sum_{j=1}^{i} \sqrt{N_j}, \quad i = 1, ..., B, \qquad \bar{z} = \frac{z_B}{H}. \tag{2}$$

Sampling from stratified population

Stratification algorithm $\mathcal{U}_{10}$

- This algorithm is equivalent to $\mathcal{U}_9$ except for expression (2) which now takes form:

$$z_i = \sum_{j=1}^{i} \sqrt{\hat{f}(x_j)}, \quad i = 1, ..., N, \qquad \bar{z} = \frac{z_B}{H} \tag{3}$$

$\hat{f}(x_j)$ is the density estimator. The kernel estimator is:

$$\hat{f}(x) = \frac{1}{n\omega} \sum_{i=1}^{N} K\left(\frac{x - x_j}{\omega}\right)$$

where ω is the the bandwidth. Usually in practice:

$$K(z) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}z^2\right).$$

Analysis of accuracy using simulation
Data

- Data concerning Swedish municipalities published in the monograph by Särndal, Swensson and Wretman (1992) are used;
- variable under study y: the real estate values (according to a 1984 assessment, in millions of kronor);
- auxiliary variable x: number of municipal employees (in millions of kronor);
- the population size (without outliers) is $N = 280$;
- parameters: $\rho_{xy} = 0.9924$. Moreover, $\bar{x}_U = 51945.99$, $\bar{y}_U = 378859$, $v_x = 35954.39$, $v_y = 2008981$.

Accuracy analysis

Population partitioned into clusters of the same sizes.

Table 1. Relative efficiencies

n	(M,g)	$\mathcal{U}_1$		$\mathcal{U}_2$		$\mathcal{U}_3$	
		$\tilde{y}_S$	y_{HT}	$\tilde{y}_S$	y_{HT}	$\tilde{y}_S$	y_{HT}
1	2	3	4	5	6	7	8
16	(2,8)	0.73	0.014	0.52	0.015	0.72	0.014
16	(4,4)	0.53	0.017	1.03	0.014	0.44	0.021
16	(8,2)	0.28	0.017	2.03	0.014	0.27	0.018
28	(2,14)	0.73	0.014	0.52	0.015	0.72	0.011
28	(4,7)	0.53	0.018	1.03	0.014	0.44	0.021
28	(14,2)	0.19	0.017	3.51	0.020	0.16	0.019
48	(2,24)	0.73	0.014	0.52	0.015	0.72	0.012
48	(4,12)	0.53	0.018	1.03	0.014	0.44	0.020
48	(8,6)	0.28	0,018	2.03	0.014	0.27	0.018

Source: Own calculations.

Accuracy analysis

Population partitioned into clusters.

Table 2. Relative efficiency coefficients.

			$\mathcal{U}_4$		$\mathcal{U}_5$		$\mathcal{U}_6$	
G	g	$g\bar{N}$	$\tilde{y}_S$	$y_{HT,S}$	$\tilde{y}_S$	$y_{HT,S}$	$\tilde{y}_S$	$y_{HT,S}$
1	2	3	4	5	6	7	8	9
23	2	24.348	0.511	0.0151	2.048	0.0171	23.73	0.0066
23	4	48.696	0.511	0.0150	2.048	0.0168	23.73	0.0042
23	8	97.391	0.511	0.0151	2.048	0.0155	23.73	0.0020
17	2	32.941	0.664	0.0153	3.084	0.0162	29.96	0.0064
17	4	65.882	0.664	0.0146	3.084	0.0154	29.96	0.0024
14	2	40	0.772	0.0236	3.7166	0.0166	34.08	0.0075
14	4	80	0.772	0.0242	3.7166	0.0154	34.08	0.0071

Source: Own calculations.

Accuracy analysis

Population partitioned into strata.

Table 3. Relative efficiencies.

n	H	$\mathcal{U}_7$	$\mathcal{U}_8$	$\mathcal{U}_9$	$\mathcal{U}_{10}$
1	2	3	4	5	6
16	2	0.372	0.627	0.243	0.219
16	4	0.136	0.353	0.061	0.058
16	8	0.052	0.137	0.022	0.021
32	2	0.372	0.627	0.231	0.213
32	4	0.136	0.353	0.058	0.055
32	8	0.052	0.137	0.021	0.019
48	2	0.372	0.627	0.218	0.204
48	4	0.136	0.353	0.054	0.052
48	8	0.052	0.137	0.019	0.018

Source: Own calculations.

Conclusions

- The accuracy of the Sampford cluster sampling strategy is comparable to the accuracy of Dalenius-Hadges stratified sampling strategy and its modification.
- The ordinary estimator of the total, which uses the simple cluster sampling is evidently less accurate than estimating the total using Dalenius-Hadges stratified sampling strategy and its modification.
- The Sampford sampling design from a population clustered according algorithm $\mathcal{U}_6$ as well as estimation based on stratified samples drawn from a population partitioned according to $\mathcal{U}_{10}$ are more accurate than other considered strategies in this paper.

Reference

- Cochran W. G. (1977). *Sampling Techniques*. John Wiley & Sons, New York-Chichester-Brisbane-Toronto-Singapore.
- Dalenius T., Hodges J. L. Jr. (1957). The choice of stratification points. *Skandinavisk Aktuarietidskrift* Vol. 3-4, pp. 198-203.
- Dalenius T., Hodges J.L. Jr. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, Vol. 54, No. 285, pp. 88-101.
- Horvitz D.G., Thompson D.J. (1952). A generalization of sampling without replacement from finite universe. *Journal of the American Statistical Association* vol. 47, pp. 663-685.
- Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika* 54, pp. 499-513.
- Särndal C.E., Swensson B., Wretman J. (1992). *Model Assisted Survey Sampling* Springer-Verlag, New York, Berlin, Heidelberg.
- Yates F., Grundy P.M. (1953). Selection without replacement from within strata with probability proporcional to size. *Journal of the Royal Statistical Society* vol. B15, pp. 235-261.

Thank you very much