An example of application of optimal sample allocation in a finite population

Dominik Sieradzki[1], Wojciech Zieliński[2]

Abstract

The problem of estimating a proportion of objects with particular attribute in a finite population is considered. This paper shows an example of the application of estimation fraction using new proposed sample allocation in a population divided into two strata. Variance of estimator of proportion which uses proposed sample allocation is compared to variance of the standard one.

Keywords: survey sampling, sample allocation, stratification, estimation, proportion

JEL Classification: C83, C99

DOI

1. Introduction

In microeconomics, the main subject of interest is human as a managing individual, whereas macroeconomics places the greatest emphasis on households and enterprises (Bartkowiak, 2003). Such objects frequently form multi-million populations. Due to amount of costs it is impossible to subject the population of interest to exhaustive sampling, even for Statistical Office. In economics populations consist of a finite number of units. Survey sampling deals with finite populations. Therefore a sample is drawn from the population. When sampling, two types of errors can be distinguished: sampling error and non-sampling error. Non-sampling error is associated with the non-response problem. Proposal on how to deal with such an issue can be found in Hansen and Hurwitz (1946). This article is focused on sampling error, hence it is assumed that responses were obtained from all of the chosen units in the sample. The sampling error, among others, depends on sampling scheme. In the next part of this paper an example of application of sample allocation proposed in Sieradzki and Zieliński (2018) is presented.

In economics the aim of the research is often to inference about dychotomus occurences, for example farmers' decision about production credit and EU measures (using these funds or not) (Roszkowska-Mądra and Mańkowski, 2010), support for a particular party

[1] Corresponding author: Warsaw University of Life Sciences, Department of Econometrics and Statistics, Nowoursynowska 159, Warsaw, Poland, dominik_sieradzki@sggw.pl.

[2] Warsaw University of Life Sciences, Department of Econometrics and Statistics, Nowoursynowska 159, Warsaw, Poland, wojciech_zielinski@sggw.pl.

or candidate in elections (Szreder, 2010) or business tendency survey (GUS). Consider a problem of support for a particular candidate in the elections. Main issue to consider in the study is a population population $U = \{u_1, \dots, u_N\}$ which contains a finite number of N people who may vote. In this population a number of people who support a particular candidate is observed. All the units in this population could be considered as a vector $\boldsymbol{Y} = (Y_1, \dots, Y_N)^T$, where $Y_k = 1$ if k-th person supports a candidate and $Y_k = 0$ if k-th person doesn't support a candidate, for $k = 1, \dots, N$. Hence $\sum_{k=1}^{N} Y_k$ stands for an unknown number of people in the population who support a candidate. Let us denote this number as M. The aim of the study is to estimate an unknown proportion (fraction) $\theta = \frac{M}{N}$. A sample of size n is drawn using simple random sampling without replacement scheme. In the sample number of people who support a candidate is observed and this number is a random variable. Let ξ denote this random variable. The random variable ξ has hypergeometric distribution (Barnett, 1974; Zieliński, 2010) and its probability distribution function is

$$P_{\theta,N,n}\{\xi = x\} = \frac{\binom{\theta N}{x}\binom{(1-\theta)N}{n-x}}{\binom{N}{n}}, \tag{1}$$

for integer x from set $\{\max\{0, n-(1-\theta)N\}, \dots, \min\{n, \theta N\}\}$. Unbiased estimator with minimal variance of the parameter θ is $\hat{\theta}_c = \frac{\xi}{n}$ (Cochran, 1977; Steczkowski, 1995; Wywiał, 2010). Variance of that estimator equals $D_\theta^2\, \hat{\theta}_c = \frac{1}{n^2} D_\theta^2\, \xi = \frac{\theta(1-\theta)}{n}\frac{N-n}{N-1}$ for all θ. It is obvious the worst variance occurs when θ equals $\frac{1}{2}$.

2. Stratified estimator

In some cases, the population of the study is strongly variable and support for a particular candidate depends on e.g. region or gender of voters. Therefore the sample is drawn due to simple random sampling without replacement scheme, so it can contain only people who support a candidate. To avoid this, stratified random sampling is used. This method of sampling assumes a division of the population among disjoint strata. After such division of the population, random sample is taken in each strata (Cochran, 1977). Let us divide the population U into two disjoint strata U_1 and U_2, $U = U_1 \cup U_2$ of N_1 and N_2 people, respectively. The details of division of the population in stratification can be found in (Horgan, 2006). For example, support in elections can depend on dominant political option at the time. In each strata fraction of people who support a candidate equals θ_1 and θ_2, respectively. The aim of the study is still to estimate the overall proportion θ, not θ_1 and θ_2.

Let w_1 denote a contribution of the first strata, i.e. $w_1 = {}^{N_1}/_{N}$. Obviously the overall proportion equals

$$\theta = w_1\theta_1 + w_2\theta_2\,, \tag{2}$$

where $w_2 = 1 - w_2$ is a contribution of the second strata. It seems intuitively obvious to take as our estimate of θ,

$$\hat{\theta}_w \ = w_1\frac{\xi_1}{n_1} + w_2\frac{\xi_2}{n_2}, \tag{3}$$

where n_1 and n_2 denote sample sizes from the first and the second strata, respectively. Now, there are two random variables describing the number of units with a particular attribute in samples drawn from each strata:

$$\xi_1 \sim H(N_1, \theta_1 N_1, n_1), \xi_2 \sim H(N_2, \theta_2 N_2, n_2) \tag{4}$$

Let us consider costs of sampling. Suppose that cost of sampling from the first strata equals c_1 and from the second one c_2. Funds for the poll are limited. Cost function is of the form:

$$C = c_1 n_1 + c_2 n_2. \tag{5}$$

The main goal is to estimate the overall fraction θ, not fraction in each strata. The parameter θ_1 will be considered as a nuisance one. This parameter will be eliminated by appropriate averaging. Note that for a given $\theta \in [0, 1]$, parameter θ_1 is a fraction M_1/N_1 (it is treated as the number, not as the random variable) from the set

$$A = \left\{a_\theta, a_\theta + \frac{1}{N_1}, \dots, b_\theta\right\}, \tag{6}$$

where

$$a_\theta = \max\{0, \frac{\theta - w_2}{w_1}\} \text{ and } b_\theta = \min\{1, \frac{\theta}{w_1}\} \tag{7}$$

and let L_θ be cardinality of A.

It is facile to prove that estimator $\hat{\theta}_w$ is an unbiased estimator of fraction θ (Sieradzki and Zieliński, 2017). Hence it is necessary to compare variances of both estimators. Averaged variance of estimator $\hat{\theta}_w$ having regard to cost, could be as follows:

$$D_\theta^2 \hat{\theta}_w = \frac{1}{L_\theta} \sum_{\theta_1 \in A} \left(\frac{w_1^2}{n_1} \theta_1 (1-\theta_1) \frac{N_1 - n_1}{N_1 - 1} + \frac{w_2^2}{(C - c_1 n_1)/c_2} \frac{\theta - w_1\theta_1}{w_2} \left(1 - \frac{\theta - w_1\theta_1}{w_2}\right) \frac{N_2 - (C - c_1 n_1)/c_2}{N_2 - 1} \right) \tag{8}$$

Detailed analysis of variance $D_\theta^2 \hat{\theta}_w$ can be found in (Sieradzki and Zieliński, 2017; Sieradzki and Zieliński, 2018). In further steps: firstly, finding 'the worst' situation, i.e. such value of proportion for which variance $D_\theta^2 \hat{\theta}_w$ takes on its maximal value is needed. Then it is necessary to find such (n_1^{opt}, n_2^{opt}), that minimizes this maximal variance. The optimal allocation of the sample is $(n_2^{opt} = (C - c_1 n_1^{opt})/c_2)$:

$$n_1^{opt} = \begin{cases} \dfrac{C\sqrt{(N_2-1)}w_1}{c_1\sqrt{(N_2-1)}w_1 - \sqrt{c_1 c_2 w_2 (N(w_1^2 - 3w_1 + 1.5) - w_1)}} & \text{for } w_1 \le w_1^* \\ \text{numerical solution available} & \text{for } w_1 > w_1^*, \end{cases} \tag{9}$$

where w_1 equals about 0.46 (Sieradzki and Zieliński, 2018).

In order to compare effectiveness of both estimators, it is necessary to determine sample size for the classical estimator $\hat{\theta}_c$. Let n_c denote a sample size for estimator $\hat{\theta}_c$. Sample size could be described as follows (Sieradzki and Zieliński, 2018):

$$n_c = \frac{C}{w_1 c_1 + w_2 c_2}. \tag{10}$$

Example of application of this sample allocation will be considered in the next section.

3. Example

Suppose that the aim of the research is to estimate support for a candidate (it will be referred to as a candidate "A") in second round of presidential elections in Poland. In Poland there are more than 30 milion people who are entitled to vote (due to official statistics, in 2015 there were N=30709281 voters). The standard way of estimation θ is to take a sample of size n_c due to the scheme of simple sampling without replacement. In the sample the number of answers "yes, I will vote for candidate A" is counted. Let us denote this number as ξ. Obviously the standard estimator of the support is $\frac{\xi}{n_c}$.

In 2015 some party which is linked with candidate 'A' won in 7 of 16 voivodeships. In those voivodeships there were 14526524 people who may vote. In the remaining ones there were 16182757 voters. Hence let us divide the population of electorate into two strata: the first one of the weight w_1=14526524/30709281=0.47 and the second one of the weight w_2=16182757/30709281=0.53. Suppose that costs of sampling from the first and the second strata equal c_1=3 and c_2=1, respectively. Funds for the sampling for this poll equal e.g. C=1200. Sample size n_c equals 618. The optimal division (n_1^{opt}, n_2^{opt}) of the sample for this numerical case could be calculated. After some calculations (which can be done in e.g. Mathematica) optimal sample allocation is obtained: n_1=242, n_2=474.

Suppose that in the whole sample 100 ’yes’ answers were obtained. The point estimate of the support with classical estimator equals $\hat{\theta}_c$=100/618 = 16.18% and its estimated variance equals

$$\hat{v}_c(100) = \frac{\hat{\theta}_c(1-\hat{\theta}_c)}{618}\frac{30709281-618}{30709281-1} = 0.00021946, \tag{11}$$

Suppose that in the sample of size n_1 from the first strata there were 10 ’yes’ answers and the number of ’yes’ answers in the sample of size n_2 equals 128. The point estimate of the support would be $\hat{\theta}_w = 16.14\%$. The estimated variance of the estimator $\hat{\theta}_w$ equals

$$\hat{v}(10,128) = (\frac{14\,526\,524}{30\,709\,281})^2 \frac{\frac{10}{242}\left(1-\frac{10}{242}\right)}{616}\frac{14\,526\,524-242}{14\,526\,524-1} + (\frac{16\,182\,757}{30\,709\,281})^2 \frac{\frac{128}{474}\left(1-\frac{128}{474}\right)}{474}\frac{16\,182\,757-474}{16\,182\,757-1} = 0.00001516. \tag{12}$$

The relative reduction of estimated variance equals

$$reduction = \left(1 - \frac{\hat{v}_w(10,128)}{\hat{v}_c(100)}\right) \cdot 100\% = 30.94\%. \tag{13}$$

Table 1 shows other possible results of the poll, assuming that the overall ’yes’ answers equal to 100, total funds equal 1200, costs of sampling from the first and the second stratum equal 3 and 1, respectively.

Table 1. Possible results for ξ=100, $\hat{v}_c(100) = 0.00021946$, c_1=3, c_2=1, C=1200, n_1=242, n_2=474, n_c=618, $\hat{\theta}_w$=16.18%.

ξ_1	ξ_2	**support**	**variance**	**reduction**
10	128	16.14%	0.0001516	30.94%
20	111	16.22%	0.0001750	20.28%
30	93	16.19%	0.0001930	12.08%

40	75	16.16%	0.0002061	6.10%
50	57	16.13%	0.0002143	2.33%
60	40	16.21%	0.0002188	0.31%
70	22	16.18%	0.0002174	0.94%
80	4	16.15%	0.0002112	3.77%

In Tables 2 and 3 there are given possible results assuming, that the overall positive answers is 300 and 400 respectively.

Table 2. Possible results for ξ=200, $\hat{v}_c(200) = 0.000354193$, c_1=3, c_2=1, C=1200, n_1=242, n_2=474, n_c=618, $\hat{\theta}_w$=32.36%.

ξ_1	ξ_2	**support**	**variance**	**reduction**
10	274	32.31%	0.0001788	49.53%
20	257	32.39%	0.000215	39.29%
30	239	32.36%	0.0002466	30.37%
40	221	32.33%	0.0002733	22.83%
50	204	32.41%	0.0002954	16.6%
60	186	32.38%	0.0003125	11.78%
70	168	32.35%	0.0003247	8.32%
80	150	32.32%	0.0003321	6.24%
90	133	32.40%	0.0003352	5.37%
100	115	32.37%	0.0003329	6.01%
110	97	32.33%	0.0003258	8.02%
120	80	32.41%	0.0003146	11.17%
130	62	32.38%	0.0002979	15.89%
140	44	32.35%	0.0002763	21.99%
150	26	32.32%	0.0002498	29.46%
160	9	32.40%	0.0002197	37.97%

Table 3. Possible results for ξ=300, $\hat{v}_c(300) = 0.000404188$, c_1=3, c_2=1, C=1200, n_1=242, n_2=474, n_c=618, $\hat{\theta}_w$=48.54%.

ξ_1	ξ_2	**support**	**variance**	**reduction**
10	421	48.59%	0.0000947	76.57%

20	403	48.56%	0.0001447	64.19%
30	385	48.53%	0.0001899	53.01%
40	367	48.5%	0.0002303	43.03%
50	350	48.58%	0.0002652	34.4%
60	332	48.55%	0.0002959	26.8%
70	314	48.52%	0.0003217	20.41%
80	297	48.6%	0.0003424	15.29%
90	279	48.57%	0.0003586	11.28%
100	261	48.54%	0.0003699	8.47%
110	243	48.51%	0.0003764	6.87%
120	226	48.59%	0.0003781	6.45%
130	208	48.55%	0.000375	7.22%
140	190	48.52%	0.000367	9.2%
150	172	48.49%	0.0003541	12.38%
160	155	48.57%	0.0003368	16.66%

Tables 4-6 contain proper columns, assuming that cost of sampling in the second strata is greater than in the first strata, i.e. c_1=1 and c_2=3.

Table 4. Possible results for ξ=100, $\hat{v}_c(100) = 0.00021946$, c_1=1, c_2=3, C=1200, n_1=405, n_2=265, n_c=582, $\hat{\theta}_w$=17.18%.

ξ_1	ξ_2	support	variance	reduction
10	81	17.22%	0.0002342	4.23%
20	75	17.2%	0.0002372	2.98%
30	69	17.19%	0.0002385	2.45%
40	63	17.17%	0.0002381	2.63%
50	57	17.16%	0.0002359	3.52%
60	51	17.14%	0.000232	5.13%
70	45	17.12%	0.0002263	7.45%
80	39	17.11%	0.0002189	10.49%

Table 5. Possible results for ξ=200, $\hat{v}_c(200) = 0.000387547$, c_1=1, c_2=3, C=1200, n_1=405, n_2=265, n_c=582, $\hat{\theta}_w$=34.36%.

ξ_1	ξ_2	support	variance	reduction
10	157	32.28%	0.0002645	25.31%
20	152	32.46%	0.0002805	20.79%

30	146	32.44%	0.0002955	16.56%
40	140	32.43%	0.0003088	12.82%
50	134	32.41%	0.0003203	9.58%
60	128	32.4%	0.00033	6.82%
70	122	32.38%	0.000338	4.56%
80	116	32.36%	0.0003443	2.79%
90	110	32.35%	0.0003488	1.52%
100	104	32.33%	0.0003516	0.74%
110	98	32.32%	0.0003526	0.45%
120	92	32.3%	0.0003519	0.65%
130	86	32.28%	0.0003494	1.35%
140	80	32.27%	0.0003452	2.54%
150	75	32.45%	0.000341	3.73%
160	69	32.44%	0.0003334	5.86%

Table 6. Possible results for ξ=300, $\hat{v}_c(300) = 0.000429142$, c_1=1, c_2=3, C=1200, n_1=405, n_2=265, n_c=582, $\hat{\theta}_w$=51.55%.

ξ_1	ξ_2	**support**	**variance**	**reduction**
10	254	51.49%	0.0000548	87.23%
20	248	51.48%	0.0000886	79.36%
30	242	51.46%	0.0001206	71.89%
40	237	51.64%	0.0001479	65.54%
50	231	51.63%	0.0001766	58.85%
60	225	51.61%	0.0002036	52.56%
70	219	51.6%	0.0002289	46.67%
80	213	51.58%	0.0002524	41.2%
90	207	51.56%	0.0002741	36.13%
100	201	51.55%	0.0002941	31.46%
110	195	51.53%	0.0003124	27.21%
120	189	51.52%	0.0003289	23.36%
130	183	51.5%	0.0003437	19.92%
140	177	51.49%	0.0003567	16.88%
150	171	51.47%	0.000368	14.25%
160	165	51.45%	0.0003775	12.03%

Summary

In the article an example of application of averaged sample allocation was presented. The classical estimator and stratified estimator were compared with respect to their estimated variances. In the numerical study it was shown that estimated variance of $\hat{\theta}_w$ is smaller than estimated variance of $\hat{\theta}_c$. Reduction is up to 90%. For such sample allocation there is no need to estimate an unknown fraction in each strata by preliminary sample.

References

Barnett, V. (1974). Elements of Sampling Theory, *The English Universities Press Ltd*.

Bartkowiak, R. (2003). Historia myśli ekonomicznej, *Polskie Wydawnictwo Ekonomiczne*, Warszawa.

Cochran, W., G. (1977). Sampling Techniques (3rd ed.), New York: John Wiley.

Hansen, M., H., & Hurwitz, W., N. (1946). The problem of non-response in sample surveys, *Journal of the American Statistical Association*, 41, 236, 517-529.

Roszkowska-Mądra, B., & Mańkowski D., R. (2010). Determinanty decyzji rolników o korzystaniu z funduszy unii europejskiej i kredytów na działalność rolniczą: przykład dla rolnictwa z rozwiniętym systemem produkcji mlecznej w województwie podlaskim, *Roczniki Nauk Rolniczych. Seria G, Ekonomika Rolnictwa*, 97, 14–27.

Sieradzki, D., & Zieliński W. (2017). Sample allocation in estimation of proportion in a finite population divided into two strata. *Statistics in Transition new series*, 18(3), 541-548, 10.21307.

Sieradzki, D., & Zieliński W. (2018). Cost Issue in Estimation of Proportion in a Finite Population Divided Among Two Strata. *submitted*.

Steczkowski, J. (1995). Metoda reprezentacyjna w badaniach zjawisk ekonomiczno-społecznych. *PWN*, Warszawa.

Szreder M. (2010). Metody i techniki sondażowych badań opinii. *Wydawnictwo UEK*, Kraków

Wywiał, J. (2010). Wprowadzenie do metody reprezentacyjnej. *Wydawnictwo Akademii Ekonomicznej w Katowicach*, Katowice.

Zieliński, W. (2010). Estymacja wskaźnika struktury. *Wydawnictwo SGGW*.