Quantitative Methods In Economics
Volume XIX, No. 1, 2018, pp. 71 – 77

Received: 17.03.2018
Accepted: 05.04.2018

SAMPLE ALLOCATION IN ESTIMATION OF PROPORTION IN A FINITE POPULATION DIVIDED INTO TWO STRATA: AN EXAMPLE OF APPLICATION

Dominik Sieradzki (ORCID: 0000-0001-5843-3413)
Wojciech Zieliński (ORCID: 0000-0003-0749-8764)
Faculty of Applied Informatics and Mathematics
Warsaw University of Life Sciences - SGGW, Poland
e-mail: dominik_sieradzki@sggw.pl, wojciech_zielinski@sggw.pl

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 the proportion which uses proposed sample allocation is compared to variance of the standard one. In the paper an application of sample allocation described in Sieradzki & Zieliński [2017] is presented.

Keywords: survey sampling, sample allocation, stratification, estimation, election poll

JEL classification: C83, C99

INTRODUCTION

In the last years some of the election polls disappointed in their accuracy - the recent American Presidential Elections are the perfect example for that. Election polls are very important not only for the candidates, political party or media, but they can really make a serious impact on voters' decisions. Most of the voters use election polls to take one candidate's side. Moreover, some of them use election polls to decide whether even go to the elections or not! To prevent the feeling of guilt and the common view that the election polls do not mean a thing, it is very important, that they are the most accurate and precise as they can be and the quality of the standard way of the election polls are exquisite. Only this way we will be able to use the election polls as an Academic (scientific) tool.

Consider a problem of the estimation of the support for political parties or a particular candidate in the elections. We would like to know as accurately as possible

DOI: 10.22630/MIBE.2018.19.1.7

a real value of unknown support for a particular candidate in the elections. This magnitude would be known exactly if the society was subjected to exhaustive polling. In practice the easiest and the standard way is to take a sample of size n, count the "yes" answers and divide them number by a sample size. Therefore it could be dealt with sampling error and non-sampling error. The size of sampling error depends on the population variance and can be controlled by the sample size [Hansen et al. 1953]. Non-sampling error is associated with the non-response problem. We distinguish four types of non-response: non-coverage, not-at-homes, unable to answer and the "hard core" [Cochran 1977]. In the next part we are focused on sampling error only.

Consider a population $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ which contains finite number of N people. In this population we could observe people (units, objects) which support for a one political option. So we can call this people as units or objects with a particular attribute. Let M denote an unknown number of units in population, which support a particular party in elections. We would like to estimate M, or equivalently, fraction $\theta = \frac{M}{N}$. Sample of size n is drawn due to simple random sampling without replacement scheme. In the sample number of objects which support a particular party in elections is observed. To estimate θ we count people in sample which support particular party and divide by size of sample. The number with certain attribute in the sample is a random variable. To be formal, let ξ be random variable describing number of units having a certain characteristic in the sample. The random variable ξ has hypergeometric distribution [Zieliński 2010] and its statistical model is

$$\left(\{0, 1, \ldots, n\}, \{H\left(N, \theta N, n\right), \theta \in \langle 0, 1\rangle\}\right),$$

with probability distribution function

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

for integer x from interval $\langle \max\{0, n - (1-\theta)N\}, \min\{n, \theta N\}\rangle$. Unbiased estimator with minimal variance of the parameter θ is $\hat{\theta}_c = \frac{\xi}{n}$ [Bracha 1998, Cochran 1977, Steczkowski 1995, Wywiał 2010]. It is the standard way to estimate unknown value of θ. Variance of that estimator equals

$$D^2_\theta \hat{\theta}_c = \frac{1}{n^2} D^2_\theta \xi = \frac{\theta(1-\theta)}{n} \frac{N-n}{N-1} \quad \text{for all } \theta.$$

It is easy to check that variance $D^2_\theta \hat{\theta}_c$ takes on its maximal value at $\theta = \frac{1}{2}$.

STRATIFICATION

The sample is drawn due to simple random sampling without replacement scheme, so when the support for a party is strongly variable and depends on region, gender of voters etc, it is possible that a part of population would be represented too often, while another part too rarely: the sample may contain only people which support the party or only people which do not support the party. To avoid this, let's divide our population into two disjoint strata $\mathcal{U}_1$ and $\mathcal{U}_2$, $\mathcal{U} = \mathcal{U}_1 \cup \mathcal{U}_2$ of N_1 and

N_2 units, respectively. For example, support in elections may depend on a gender or on dominant political option at the time. In each strata proportions of distinguished objects are θ_1 and θ_2, respectively. We are still interested in estimation the overall proportion θ, not θ_1 and θ_2. The question is, does the information of this division of the population into two strata improve estimation of the unknown proportion θ? We could answer the question, if we consider stratified estimator of the proportion θ.

Let contribution of the first strata be w_1, i.e $w_1 = N_1/N$. Hence, the overall proportion θ equals

$$\theta = w_1\theta_1 + w_2\theta_2,$$

where $w_2 = 1 - w_1$. Let n_1 and n_2 denote sample sizes from the first and the second strata, respectively. The whole sample size equals $n = n_1 + n_2$. Now we have two random variables describing number of units with property in samples drawn from each strata:

$$\xi_1 \sim H\left(N_1, \theta_1 N_1, n_1\right), \quad \xi_2 \sim H\left(N_2, \theta_2 N_2, n_2\right).$$

Values θ_1, θ_2 and θ are unknown. Since $\theta \in \langle 0, 1\rangle$, hence

$$\theta_1 \in \left\langle \max\left\{0, \frac{\theta - w_2}{w_1}\right\}, \min\left\{1, \frac{\theta}{w_1}\right\}\right\rangle$$

[Zieliński 2016]. Note that θ_1 is a rationale of type M_1/N_1.

Denote left end of the above interval by a_θ and its right end by b_θ, i.e.

$$a_\theta = \max\left\{0, \frac{\theta - w_2}{w_1}\right\}, \quad b_\theta = \min\left\{1, \frac{\theta}{w_1}\right\}$$

and let $L_\theta = b_\theta - a_\theta + 1$. Consider the estimator

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

The estimator $\hat{\theta}_w$ is unbiased estimator of unknown parameter θ [Sieradzki & Zieliński 2017]. Hence it is necessary to compare variances of estimators $\hat{\theta}_w$ and $\hat{\theta}_c$. The estimator with smaller variance would be more efficient. For given θ there are many θ_1 and θ_2 such that $\theta = w_1\theta_1 + w_2\theta_2$. We are not interested in estimating θ_1 and θ_2, hence we apply averaging with respect to θ_1 (parameter θ_1 is considered as a nuisance one). In such approach variance of estimator $\hat{\theta}_w$ equals:

$$\begin{aligned} D^2_\theta\hat{\theta}_w =& D^2_\theta\left(w_1\frac{\xi_1}{n_1} + w_2\frac{\xi_2}{n_2}\right) \\ =& \frac{1}{L_\theta}\sum_{\theta_1=a_\theta}^{b_\theta}\left(\left(\frac{w_1}{n_1}\right)^2 D^2_{\theta_1}\xi_1 + \left(\frac{w_2}{n_2}\right)^2 D^2_{\frac{\theta - w_1\theta_1}{w_2}}\xi_2\right) \\ =& \frac{1}{L_\theta}\sum_{\theta_1=a_\theta}^{b_\theta}\left[\frac{w_1^2}{n_1}\theta_1(1-\theta_1)\frac{N_1 - n_1}{N_1 - 1} + \right. \\ & \left. + \frac{w_2^2}{n_2}\frac{\theta - w_1\theta_1}{w_2}\left(1 - \frac{\theta - w_1\theta_1}{w_2}\right)\frac{N_2 - n_2}{N_2 - 1}\right]. \end{aligned}$$

Let $f=\frac{n_1}{n}$ denote the contribution of first strata in the sample. For $0<\theta<w_1$ variance of $\hat{\theta}_w$ equals ($a_\theta=0$ and $b_\theta=\frac{\theta}{w_1}$):

$$\begin{aligned}&\frac{h(f)}{-6(N_1-1)(N_2-1)Nf(1-f)n}\theta\\&\qquad+\frac{(N_2-1)N_1-(N(n+1)-2(N_1+n))f+(N-2)nf^2}{3(N_1-1)(N_2-1)f(1-f)n}\theta^2,\end{aligned}\tag{$*$}$$

where

$$\begin{aligned}h(f)=&\,N_1(N_2-3N_1(N_2-1)-1)\\&+\left(3N_1^2(N_2-1)+3N_2^2+2n+N_1\left(6N_2n-3N_2^2-4n+1\right)-N_2(4n+1)\right)f\\&+2\left(N_1(2-3N_2)+2N_2-1\right)nf^2.\end{aligned}$$

For $w_1\le\theta\le 1-w_1$ variance of $\hat{\theta}_w$ equals ($a_\theta=0$ and $b_\theta=1$):

$$\begin{aligned}&\frac{(N_2-(1-f)n)}{(N_2-1)(1-f)n}\theta(1-\theta)+\\&\qquad-\frac{N_1\left(2(N+1)f^2+(3NN_2+N_2-N_1-2n(N+1))f-N_1(N_2-1)\right)}{6N^2(N_2-1)nf(1-f)}.\end{aligned}$$

To obtain explicit formula for variance of $\hat{\theta}_w$ for $1-w_1<\theta<1$ it is enough to replace θ by $1-\theta$ in $(*)$.

Detailed analysis of variance of estimator $\hat{\theta}_w$ could be found in Sieradzki & Zieliński [2017]. We would like to find "the worst" situation, i.e. the value of θ for which variance $D^2_\theta\hat{\theta}_w$ takes on its maximal value and then find optimal f which minimizes this maximal variance. General formula for the optimal f is unobtainable, because of complexity of symbolic computation. Nevertheless numerical solution is easy to obtain. In the next section we will considered an example of application.

EXAMPLE

Suppose we want to estimate support for a political party (it will be referred to as a party "A") in Poland. In Poland there is more than 30000000 people who may vote (due to official statistics, in 2011 there were $N=30762931$ voters[1]). The standard way of estimation is to take a sample of size $n=1000$ due to the scheme of simple sampling without replacement. Let ξ denote the number of "yes, I will vote on party A" answers. The standard estimator of the support is $\frac{\xi}{n}$.

In 2011 the party "A" won in 27 out of 41 regions. In those regions there were 20222414 people who may vote, while in the rest of regions there were 10540517 voters. To improve estimation of the support for party "A" we divide Poland into two strata: the first one of the weight $w_1=10540517/30762931=0.342636955$ and the second one of the weight $w_2=20222414/30762931=0.657363045$. The optimal f for

[1] http://wybory2011.pkw.gov.pl/wyn/pl/000000.html#tabs-1

Table 1. Possible results for $\xi = 200$, $\hat{v}_c(200) = 0.0001599948$

ξ_1	ξ_2	*variance*	*reduction*
25	175	0.000109763	5.23%
50	150	0.000158481	0.94%
75	125	0.000159807	0.11%
100	100	0.000155599	2.74%
125	75	0.000145855	8.83%
150	50	0.000130577	18.38%
175	25	0.000109763	31.39%

Source: own calculations

Table 2. Possible results for $\xi = 300$, $\hat{v}_c(300) = 0.0002099932$

ξ_1	ξ_2	*variance*	*reduction*
25	275	0.000183173	12.77%
50	250	0.000197636	5.88%
75	225	0.000206565	1.63%
100	200	0.000209959	0.01%
125	175	0.000207817	1.03%
150	150	0.000200141	4.69%
175	125	0.000186929	10.98%
200	100	0.000168183	19.91%
225	75	0.000143901	31.47%
250	50	0.000114085	45.67%
275	25	0.000078733	62.50%

Source: own calculations

this numerical case could be find. Finding the optimal f is equivalent to finding the optimal division (n_1, n_2) of the sample. After some calculations (in a mathematical software, for example Mathematica) we obtain optimal $f = 0.343$, hence $n_1 = 343$ and $n_2 = 657$.

Suppose that in the whole sample 200 “yes” answers were obtained. The point estimate of the support equals $\hat{\theta}_c = 0.2$ and its variance may be estimated as $\hat{v}_c(200) = 0.000159995$. If in the sample of size n_1 from the first stratum there were 25 “yes” answers and in the sample of the size n_2 from the second stratum there were 175 “yes” answers, then the point estimate of the support is $\hat{\theta}_w = 0.2$ and its variance may be estimated as $\hat{v}_w(25, 175) = 0.000109763$. Note that the stratified estimator has smaller variance than the one based on the non stratified sample. The relative reduction of variance equals

$$\textit{reduction} = \left(1 - \frac{\hat{v}_w(25, 175)}{\hat{v}_c(200)}\right) \cdot 100\% = 5.23\%.$$

Table 1 shows other possible results of the pool, assuming that the overall “yes” answers equal to 200.

In the first and in the second column possible results of the pool are given. Values of estimated variances are given in the third column. The last column shows the relative reduction of variance, i.e. of how many percent estimator $\hat{\theta}_w$ is better than

Table 3. Possible results for $\xi = 400$, $\hat{v}_c(400) = 0.0002399922$

ξ_1	ξ_2	*variance*	*reduction*
25	375	0.0001842856	23.21%
50	350	0.0002063514	14.01%
75	325	0.0002228821	7.12%
100	300	0.0002338778	2.54%
125	275	0.0002393385	0.27%
150	250	0.0002392641	0.30%
175	225	0.0002336546	2.64%
200	200	0.0002225102	7.28%
225	175	0.0002058306	14.23%
250	150	0.0001836161	23.49%
275	125	0.0001558665	35.05%

Source: own calculations

estimator $\hat{\theta}_c$. In Tables 2 and 3 there are given possible results assuming, that the overall positive answers is 300 and 400 respectively. It is seen, that whatever results of pool are in strata the stratified estimator is better than the standard one.

CONCLUSIONS

In the paper an example of application of estimation of unknown fraction in population divided into two strata was presented. Estimators $\hat{\theta}_c$ and $\hat{\theta}_w$ were compared with respect to their variances. In that example for optimal allocation between strata it was shown that variance of stratified estimator is always smaller than variance of classical estimator. Hence in practice, it is recommended to use the information of the division of the population into two strata, because quality of stratified estimator is better than the quality of the classical one.

REFERENCES

Bracha Cz. (1998) Metoda reprezentacyjna w badaniach opinii publicznej i marketingu. PWN, Warszawa (in Polish).

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

Hansen M. H., Hurwitz W. N., Madow W. G. (1953) Sample Survey Methods and Theory. New York: John Wiley.

Sieradzki D. (2016) Estimation of proportion in finite population divided into two strata. master thesis, WZIiM SGGW Warszawa (in Polish).

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.

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

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

Zieliński W. (2010) Estymacja wskaźnika struktury. Wydawnictwo SGGW, Warszawa (in Polish).

Zieliński W. (2016) A remark on estimating defectiveness in sampling acceptance inspection. Colloquium Biometricum, 46, 9-14.