

SAMPLE ALLOCATION IN ESTIMATION OF PROPORTION IN A FINITE POPULATION DIVIDED AMONG TWO STRATA

Dominik Sieradzki¹, Wojciech Zieliński²

ABSTRACT

The problem of estimating a proportion of objects with a particular attribute in a finite population is considered. The classical estimator is compared with the estimator, which uses the information that the population is divided among two strata. Theoretical results are illustrated with a numerical example.

Key words: survey sampling, sample allocation, stratification, estimation, proportion.

1. Introduction

Consider a population $U = \{u_1, u_2, \dots, u_N\}$ which contains a finite number of N units. In this population we can observe objects which have a given characteristic (property), for example sex, defectiveness, support for a particular candidate in elections, etc. Let M denote an unknown number of units in the population with a given property. We would like to estimate M , or equivalently, a proportion (fraction) $\theta = \frac{M}{N}$. A sample of size n is drawn using simple random sampling without replacement scheme. In the sample the number of objects with a particular attribute is observed. This number is a random variable. To be formal, let ξ be a random variable describing number of units having a certain attribute in the sample. The random variable ξ has hypergeometric distribution (Zieliński 2010) and its statistical model is

$$(\{0, 1, \dots, n\}, \{H(N, \theta N, n), \theta \in \langle 0, 1 \rangle\}), \quad (1)$$

with probability distribution function

$$P_{\theta, N, n} \{\xi = x\} = \frac{\binom{\theta N}{x} \binom{(1-\theta)N}{n-x}}{\binom{N}{n}}, \quad (2)$$

¹Department of Econometrics and Statistics, Warsaw University of Life Sciences.
E-mail: dominik_sieradzki@sggw.pl

²Department of Econometrics and Statistics, Warsaw University of Life Sciences.
E-mail: wojciech_zielinski@sggw.pl

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). Variance of that estimator equals

$$D_{\theta}^2 \hat{\theta}_c = \frac{1}{n^2} D_{\theta}^2 \xi = \frac{\theta(1 - \theta)}{n} \frac{N - n}{N - 1} \text{ for all } \theta. \quad (3)$$

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

2. Stratified estimator

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

$$\theta = w_1 \theta_1 + w_2 \theta_2, \quad (4)$$

where $w_2 = 1 - w_1$. It seems intuitively obvious to take as our estimate of θ ,

$$\hat{\theta}_w = w_1 \frac{\xi_1}{n_1} + w_2 \frac{\xi_2}{n_2}, \quad (5)$$

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

$$\xi_1 \sim H(N_1, \theta_1 N_1, n_1), \quad \xi_2 \sim H(N_2, \theta_2 N_2, n_2). \quad (6)$$

The whole sample size equals $n = n_1 + n_2$. The question now arises: how shall we choose n_1 and n_2 to obtain the best estimate of θ ? This problem concerns sample allocation between strata. One of known approaches to this problem is proportional allocation (Armitage 1943, Cochran 1977). Sample sizes n_1 and n_2 are proportional to w_1 and w_2 ,

$$n_1 = w_1 n \quad \text{and} \quad n_2 = w_2 n. \quad (7)$$

The second approach to sample allocation is Neyman Allocation (Neyman 1934). This method gives values of n_1 and n_2 , which minimize the variance of estimator $\hat{\theta}_w$ for given θ_1 and θ_2 . The values of n_1 and n_2 are as follows

$$n_i = \frac{w_i \sqrt{\theta_i(1 - \theta_i)}}{\sum_i w_i \sqrt{\theta_i(1 - \theta_i)}} n, \quad i = 1, 2. \quad (8)$$

Neyman Allocation requires knowledge of the parameters θ_1 and θ_2 . Those magnitudes would be known exactly when the population were subjected to exhaustive

sampling. Usually values θ_1 and θ_2 are estimated from a preliminary sample. In some cases fairly good estimates of θ_1 and θ_2 are available from past experience (Armitage 1943).

Since our aim is to estimate θ , hence 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

$$\mathcal{A} = \left\{ a_\theta, a_\theta + \frac{1}{N_1}, a_\theta + \frac{2}{N_1}, \dots, b_\theta \right\}, \quad (9)$$

where

$$a_\theta = \max \left\{ 0, \frac{\theta - w_2}{w_1} \right\} \quad \text{and} \quad b_\theta = \min \left\{ 1, \frac{\theta}{w_1} \right\} \quad (10)$$

and let L_θ be cardinality of $\mathcal{A}$.

Theorem. Estimator $\hat{\theta}_w$ is an unbiased estimator of θ .

Proof. Note that for a given θ there are L_θ values of θ_1 and θ_2 giving θ . Hence, averaging with respect to θ_1 is made assuming the uniform distribution of θ_1 on the set $\{a_\theta, \dots, b_\theta\}$. We have

$$\begin{aligned} E_\theta \hat{\theta}_w &= E_\theta \left(w_1 \frac{\xi_1}{n_1} + w_2 \frac{\xi_2}{n_2} \right) = \frac{1}{L_\theta} \sum_{\theta_1 \in \mathcal{A}} \left(\frac{w_1}{n_1} E_{\theta_1} \xi_1 + \frac{w_2}{n_2} E_{\frac{\theta - w_1 \theta_1}{w_2}} \xi_2 \right) \\ &= \frac{1}{L_\theta} \sum_{\theta_1 \in \mathcal{A}} \left(\frac{w_1}{n_1} \frac{\theta_1 N_1 n_1}{N_1} + \frac{w_2}{n_2} \frac{\frac{\theta - w_1 \theta_1}{w_2} N_2 n_2}{N_2} \right) \quad (11) \\ &= \theta \end{aligned}$$

for all θ .

Averaged variance of estimator $\hat{\theta}_w$ equals:

$$\begin{aligned} D_\theta^2 \hat{\theta}_w &= D_\theta^2 \left(w_1 \frac{\xi_1}{n_1} + w_2 \frac{\xi_2}{n_2} \right) = \\ &= \frac{1}{L_\theta} \sum_{\theta_1 \in \mathcal{A}} \left(\left(\frac{w_1}{n_1} \right)^2 D_{\theta_1}^2 \xi_1 + \left(\frac{w_2}{n_2} \right)^2 D_{\frac{\theta - w_1 \theta_1}{w_2}}^2 \xi_2 \right) = \\ &= \frac{1}{L_\theta} \sum_{\theta_1 \in \mathcal{A}} \left[\frac{w_1^2}{n_1} \theta_1 (1 - \theta_1) \frac{N_1 - n_1}{N_1 - 1} + \frac{w_2^2}{n_2} \frac{\theta - w_1 \theta_1}{w_2} \left(1 - \frac{\theta - w_1 \theta_1}{w_2} \right) \frac{N_2 - n_2}{N_2 - 1} \right]. \quad (12) \end{aligned}$$

Let $f = \frac{n_1}{n}$ denote the contribution of the first strata in the sample. For $0 < \theta < w_1$

variance of $\hat{\theta}_w$ equals $\left(a_\theta = 0 \text{ and } b_\theta = \frac{\theta}{w_1}\right)$:

$$\begin{aligned} & \frac{h(f)}{-6(N_1 - 1)(N_2 - 1)Nf(1 - f)n} \theta \\ & + \frac{(N_2 - 1)N_1 - (N(n + 1) - 2(N_1 + n))f + (N - 2)nf^2}{3(N_1 - 1)(N_2 - 1)f(1 - f)n} \theta^2, \end{aligned} \quad (13)$$

where

$$\begin{aligned} h(f) = & N_1(N_2 - 3N_1(N_2 - 1) - 1) \\ & + (3N_1^2(N_2 - 1) + 3N_2^2 + 2n + N_1(6N_2n - 3N_2^2 - 4n + 1) - N_2(4n + 1))f \\ & + 2(N_1(2 - 3N_2) + 2N_2 - 1)nf^2 \end{aligned} \quad (14)$$

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

$$\begin{aligned} & \frac{(N_2 - (1 - f)n)}{(N_2 - 1)(1 - f)n} \theta(1 - \theta) + \\ & - \frac{N_1(2(N + 1)f^2 + (3NN_2 + N_2 - N_1 - 2n(N + 1))f - N_1(N_2 - 1))}{6N^2(N_2 - 1)nf(1 - f)} \end{aligned} \quad (15)$$

To obtain explicit formula for variance of $\hat{\theta}_w$ for $1 - w_1 < \theta < 1$ it is sufficient to replace θ by $1 - \theta$ in (13). Observe that variance $D_\theta^2 \hat{\theta}_w$ depends on size n of the sample, size N of the population, contribution w_1 of the first strata in population and contribution f of the first strata in the sample. In Figure 1 variances of $\hat{\theta}_w$ and $\hat{\theta}_c$ are drawn against θ , for $N = 100000$, $n = 100$, $w_1 = 0.4$ and $f = 0.3$.

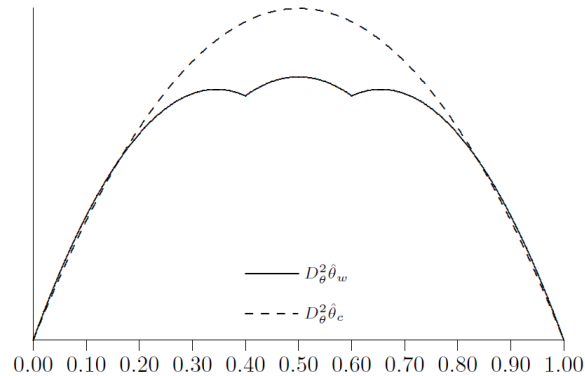


Figure 1. Variances of $\hat{\theta}_c$ and $\hat{\theta}_w$ for $w_1 = 0.4$ and $f = 0.3$

It is easy to note that $D_{\theta}^2 \hat{\theta}_w = D_{1-\theta}^2 \hat{\theta}_w$ and $D_0^2 \hat{\theta}_w = 0$.

Maximum of variance $D_{\theta}^2 \hat{\theta}_w$ determines for which value of unknown parameter θ estimation of θ is the worst one. After the analysis of variance of $\hat{\theta}_w$, it is seen that the maximal variance may be in the one of the intervals: $(0, w_1)$, $(w_1, 1 - w_1)$ or $(1 - w_1, 1)$. It depends on the values of w_1 and f . In Figures 2, 3, 4 and 5 variance of $\hat{\theta}_w$ as well as variance of $\hat{\theta}_c$ is drawn for $N = 100000$, $n = 100$, $w_1 = 0.4$ and $f = 0.2, 0.4, 0.6, 0.9$.

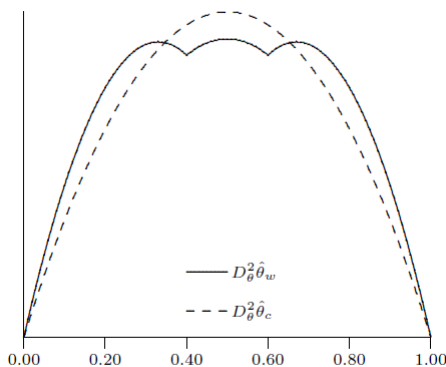


Figure 2. Variances of $\hat{\theta}_c$ and $\hat{\theta}_w$ for $w_1 = 0.4$ and $f = 0.2$

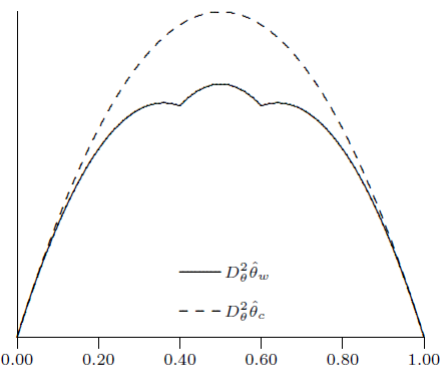


Figure 3. Variances of $\hat{\theta}_c$ and $\hat{\theta}_w$ for $w_1 = 0.4$ and $f = 0.4$

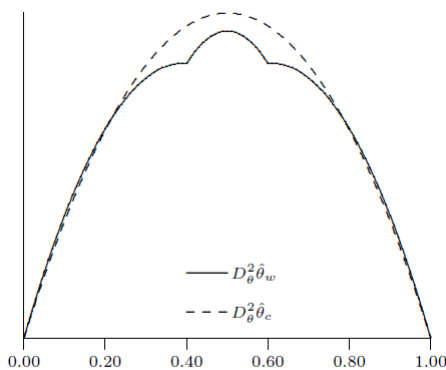


Figure 4. Variances of $\hat{\theta}_c$ and $\hat{\theta}_w$ for $w_1 = 0.4$ and $f = 0.6$

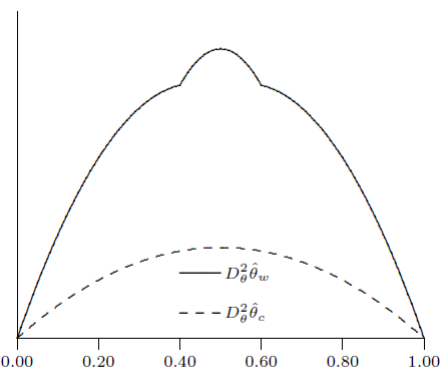


Figure 5. Variances of $\hat{\theta}_w$ and $\hat{\theta}_c$ for $w_1 = 0.4$ and $f = 0.9$

Source: Own calculations.

The point at which $D_{\theta}^2 \hat{\theta}_w$ takes on the maximal value may be located in interval $(0, w_1)$ or in interval $(w_1, 1 - w_1)$. Hence, to find the global maximum due to θ , we have to find local maximum in both intervals. Denote by θ^* a local maximum point in interval $(0, w_1)$ (local maximum point in interval $(1 - w_1, 1)$ is $1 - \theta^*$). In an interval $(w_1, 1 - w_1)$ local maximum is achieved at $\theta = 1/2$. Let $\tilde{\theta}$ denote a global

maximum point, i. e. $\tilde{\theta} = 1/2$ or $\tilde{\theta} = \theta^*$, hence

$$\max_{\theta \in (0,1)} D_{\theta}^2 \hat{\theta}_w = \max \{D_{0.5}^2 \hat{\theta}_w, D_{\theta^*}^2 \hat{\theta}_w\}. \quad (16)$$

Regardless of which point is the global maximum point ($1/2$ or θ^*), the maximum of the variance $D_{\theta}^2 \hat{\theta}_w$ depends on size n of the sample, size N of the population, contribution w_1 of the first strata in the population and the contribution f of the first strata in the sample. Values N, n, w_1 are treated as given. It may be seen that for given w_1 , variance $D_{\theta}^2 \hat{\theta}_w$ may be smaller as well as greater than $D_{\theta}^2 \hat{\theta}_c$. We would like to find optimal f , which minimizes maximal variance $D_{\theta}^2 \hat{\theta}_w$.

3. Results

A general formula for the optimal f is unobtainable, because of complexity of symbolic computation. But for given N, w_1 and n numerical solution is easy to obtain. Table 1 shows some numerical results for $N = 100000$ and $n = 100$.

Table 1. Maximal variances $D_{\theta}^2 \hat{\theta}_w$

w_1	f^{opt}	n_1^{opt}	$D_{\theta}^2 \hat{\theta}_w$	$D_{0.5}^2 \hat{\theta}_c$	$\left(1 - \frac{D_{\theta}^2 \hat{\theta}_w}{D_{0.5}^2 \hat{\theta}_c}\right) \cdot 100\%$
0.05	0.018	2	0.0004645	0.0025	81%
0.10	0.041	4	0.0008404	0.0025	66%
0.15	0.071	7	0.0011328	0.0025	55%
0.20	0.111	11	0.0013493	0.0025	46%
0.25	0.166	17	0.0015004	0.0025	40%
0.30	0.250	25	0.0015984	0.0025	36%
0.35	0.350	35	0.0017045	0.0025	32%
0.40	0.400	40	0.0017982	0.0025	28%
0.45	0.450	45	0.0018544	0.0025	26%
0.50	0.500	50	0.0018731	0.0025	25%

Source: Own calculations.

In the first column of Table 1. the values of w_1 are given. In the second column, optimal contribution of the first strata in the sample is shown. It is a value f , which gives minimum of $D_{\theta}^2 \hat{\theta}_w$. Column n_1^{opt} shows optimal sample size from the first strata (called averaged sample allocation). The values of minimal (maximal) variances $D_{\theta}^2 \hat{\theta}_w$ are given in the fourth column. The next column contains maximal variance $D_{0.5}^2 \hat{\theta}_c$. The last column shows how much estimator $\hat{\theta}_w$ is better than $\hat{\theta}_c$.

4. Summary

In the paper a new approach to the sample allocation between strata was proposed. Two estimators of an unknown fraction θ in the finite population were considered: standard estimator $\hat{\theta}_c$ and stratified estimator $\hat{\theta}_w$. It was shown that both estimators are unbiased. Their variances were compared. It appears that for a given sample size there exists its optimal allocation between strata, i.e. the allocation for which variance of $\hat{\theta}_w$ is smaller than variance of $\hat{\theta}_c$. Since a theoretical comparison seems to be impossible, hence a numerical example was presented. In that example it was shown that variance of the stratified estimator may be smaller at least 25% with respect to variance of the classical estimator. For such an approach there is no need to estimate unknown θ_1 and θ_2 by preliminary sample. It will be interesting to generalize the above results to the case of more than two "subpopulations". Work on the subject is in progress.

REFERENCES

- ARMITAGE, P., (1947). A Comparison of Stratified with Unrestricted Random Sampling from a Finite Population, *Biometrika*, 34, 3/4 , pp. 273–280.
- BRACHA, CZ., (1998). *Metoda reprezentacyjna w badaniach opinii publicznej i marketingu*. PWN, Warszawa.
- COCHRAN, W. G., (1977). *Sampling Techniques* (3rd ed.), New York: John Wiley.
- NEYMAN, J., (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection, *Journal of the Royal Statistical Society*, 97, pp. 558–606.
- SIERADZKI, D., (2016). Estimation of proportion in finite population divided into two strata, master thesis, WZliM SGGW Warszawa (in polish).
- ZIELIŃSKI, W., (2010). *Estymacja wskaźnika struktury*, Wydawnictwo SGGW, Warszawa.
- ZIELIŃSKI, W., (2016). A remark on estimating defectiveness in sampling acceptance inspection, *Colloquium Biometricum*, 46, pp. 9–14.