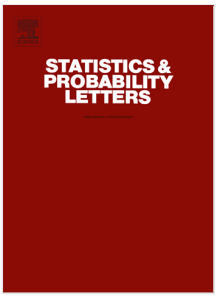
Stratified two-stage sampling in domains: Sample allocation between domains, strata, and sampling stages

Marcin Kozak, Andrzej Zieliński, Sarjinder Singh

Accepted Manuscript



Stratified two-stage sampling in domains: Sample allocation between domains, strata, and sampling stages

Marcin Kozak, Andrzej Zieliński, Sarjinder Singh

PII: S0167-7152(07)00349-5
DOI: 10.1016/j.spl.2007.09.057
Reference: STAPRO 4795

To appear in: *Statistics and Probability Letters*

Received date: 18 April 2006
Revised date: 7 August 2007
Accepted date: 25 September 2007

Please cite this article as: Kozak, M., Zieliński, A., Singh, S., Stratified two-stage sampling in domains: Sample allocation between domains, strata, and sampling stages. *Statistics and Probability Letters* (2007), doi:10.1016/j.spl.2007.09.057

Stratified two-stage sampling in domains: sample allocation between domains, strata, and sampling stages

Marcin Kozak * Andrzej Zieliński Sarjinder Singh

Department of Biometry, Faculty of Agriculture and Biology, Warsaw Agricultural University, Nowoursynowska 159, 02-776 Warsaw, Poland

Department of Biometry, Warsaw Agricultural University, Nowoursynowska 159, 02-776 Warsaw, Poland

Department of Statistics, St. Cloud State University, St. Cloud, MN, 56301, USA

Abstract

In the paper, formulae for optimum sample allocation between domains, strata in the domains, and sampling stages are presented for stratified two-stage sampling in domains under fixed sample size of SSUs from PSUs.

Key words: domain-orientated approach, optimum sample allocation, survey cost.

1 Introduction

Kozak (2005) presented basic concepts of stratified two-stage sampling design, in which a population of primary sampling units is subdivided into strata. He provided formulas for optimum sample allocation between strata and sampling stages under two schemes of the design: (i) in which sample size of secondary sampling units (SSUs) from primary sampling units (PSUs) is fixed, and (ii) with self-weighting design in strata. Kozak and Zieliński (2005), on the other hand, presented basic concepts of a problem of sample allocation between domains and strata in case when domains are subdivided into strata. They considered (i) a so-called domain-orientated approach to the sample allocation, in which one requires precise estimation in all the domains, and (ii) sample allocation orientated towards minimizing total survey cost subject to

* Corresponding author.
Email address: m.kozak@omega.sggw.waw.pl (Marcin Kozak).

fixed level of precision of estimation in the domains. In this paper, we introduce a hybrid of these two designs, namely stratified two-stage sampling in domains. Such a design can be of practical use when a population is subdivided into domains, each domain comprising some number of strata consisting of PSUs. A simple random sample of SSUs is to be taken without replacement from PSUs. We consider a situation in which sample size of SSUs from PSUs is fixed; in such a case, one obtains a sample of fixed size (and fixed total cost). We give formulas for sample allocation between domains, strata and sampling stages (i) under domain-orientated approach, and (ii) orientated towards minimizing a total survey cost.

2 Estimation for domains under stratified two-stage sampling in the domains: basic ideas

Basic concepts of stratified sampling and two-stage sampling, which lay the basis for the design introduced in this section, may be found, e.g., in Särndal et al. (1992) or Singh (2003). Consider a population U comprising N elements. The population is subdivided into D domains U_d $(d = 1, \ldots, D)$; each domain is subdivided into H_d non-overlapping strata U_{dh}; finally, each stratum U_{dh} is subdivided into M_{dh} separate PSUs U_{dhg}. This division can be presented as

$$U = \bigcup_{d=1}^{D} \bigcup_{h=1}^{H_d} \bigcup_{g=1}^{M_{dh}} U_{dhg}, U_d \cap U_{d'} = \emptyset \;\; for \;\; d, d' = 1, \ldots, D, \;\; d \neq d'; \;\; and$$

$$U_{dh} \cap U_{dh'} = \emptyset \;\; for \;\; d = 1, \ldots, D, \;\; h, h' = 1, \ldots, H_d, \;\; h \neq h'; \;\; and$$

$$U_{dhg} \cap U_{dhg'} = \emptyset \;\; for \;\; d = 1, \ldots, D, \;\; h = 1, \ldots, H_d, \;\; g, g' = 1, \ldots, M_{dh}, \;\; g \neq g'.$$

The gth PSU from the hth stratum in the dth domain comprises N_{dhg} SSUs, which are the population elements. Let N_d indicate the number of SSUs in the dth domain and N_{dh} indicate the number of SSUs in the hth stratum of the dth domain.

Let a population parameter investigated be the population total of Y, Y being a characteristic studied. For the dth domain its estimator is given by

$$\hat{Y}_d = \sum_{h=1}^{H_d} \hat{Y}_{dh} = \sum_{h=1}^{H_d} \frac{M_{dh}}{m_{dh}} \sum_{g=1}^{m_{dh}} \frac{N_{dhg}}{n_{dhg}} \sum_{i=1}^{n_{dhg}} y_{dhgi} \tag{1}$$

where $\hat{Y}_d$ is the estimator of Y_d, Y_d being the population total of the variable Y restricted to the dth domain; $\hat{Y}_{dh}$ is the estimator of Y_{dh}, Y_{dh} being the population total of the variable Y restricted to the hth stratum of the dth domain; m_{dh} is the sample size of PSUs from the hth stratum of the dth domain; n_{dhg} is the sample size of SSUs from the gth PSU of the hth stratum of the dth domain; and y_{dhgi} is the Y value in the ith SSU (population element) of the gth PSU from the hth stratum in dth domain. In both cases (i.e., when sampling PSUs from strata and when sampling SSUs from PSUs), simple random sample is to be taken without replacement.

Let us consider a two-stage sampling scheme in which we deal with a fixed size of sample of SSUs. In our design, it consists in sampling the same number n_{dh} ($n_{dhg} = n_{dh}$ for each combination of $d = 1, \ldots, D$ and $h = 1, \ldots, H_d$) of SSUs from PSUs in each section *domain* $d \times$ *stratum* h. Under such a design, the variance of the estimator (1) is given by (Kozak, 2005)

$$V(\hat{Y}_d) = \sum_{h=1}^{H_d} \frac{M_{dh}}{m_{dh}} \Big[(M_{dh} - m_{dh}) S_{1dh}^2 + \frac{1}{n_{dh}} \sum_{g=1}^{M_{dh}} N_{dhg}(N_{dhg} - n_{dh}) S_{2dhg}^2 \Big] \quad (2)$$

where $S_{1dh}^2 = (M_{dh} - 1)^{-1} \sum_{g=1}^{M_{dh}} (Y_{dhg} - \overline{Y}_{dh})^2$, $Y_{dhg} = \sum_{j=1}^{N_{dhg}} y_{dhgj}$, $\overline{Y}_{dh} = N_{dh}^{-1} \sum_{g=1}^{M_{dh}} \sum_{j=1}^{N_{dhg}} y_{dhgj}$, $S_{2dhg}^2 = (N_{dhg} - 1)^{-1} \sum_{j=1}^{N_{dhg}} (y_{dhgj} - \overline{Y}_{dhg})^2$, $\overline{Y}_{dhg} = N_{dhg}^{-1} \sum_{j=1}^{M_{dhg}} y_{dhgj}$.

Note again that sample sizes n_{dh} in the variance (2) refer to sample sizes n_{dhg}, which are assumed to be the same for all $g = 1, \ldots, N_{dh}$ in a particular section *domain* $d \times$ *stratum* h. Hence, for the sake of convenience, we write n_{dh} instead of n_{dhg}, keeping in mind that n_{dh} is the sample size of SSUs from every gth PSU of the hth stratum in the dth domain. An ordinary unbiased estimator of the variance (2) is obtained by replacing the population quantities S_{1dh}^2 and S_{2dhg}^2 with their sample estimators; the summation in (2) is to be done by sampled PSUs in each hth stratum from the dth domain.

The coefficient of variation of the estimator $\hat{Y}_d$, say $\delta(\hat{Y}_d)$, is

$$\delta(\hat{Y}_d) = \frac{\sqrt{V(\hat{Y}_d)}}{Y_d}, \quad d = 1, \ldots, D$$

In this paper, we understand optimum conditions of a design as the ones for which some function of $\delta(\hat{Y}_d)$ is minimum. Let the overall survey cost C be

$$C = C_0 + \sum_{d=1}^{D} \sum_{h=1}^{H_d} m_{dh} \Big(c_{1dh} + n_{dh} c_{2dh} \Big) \quad (3)$$

where C_0 is the fixed survey cost, c_{1dh} is the cost of selecting one PSU from the hth stratum of the dth domain, and c_{2dh} is the cost of obtaining the information on Y value in one SSU from the hth stratum of the dth domain.

3 Optimizing a design under domain-orientated approach

Here we apply a domain-orientated approach to the design presented in previous section. It aims at precise estimation for each domain U_d of the population U (Kozak and Zieliński, 2005). Let $g = (g_1, \ldots, g_D)^T$ be a vector of important weights of the domains. Following Kozak and Zieliński (2005), the optimum design is the one under which the smallest common value φ of $g_d^{-1}\delta(\hat{Y}_d)$, $d = 1, \ldots, D$, is obtained. Thus, we require coefficient of variation of the estimator $\hat{Y}_d$ of the population total in the dth domain to satisfy the condition (Kozak and Zieliński, 2005)

$$\delta(\hat{Y}_d) = g_d\varphi, \quad d = 1, \ldots, D \tag{4}$$

Then, our aim is to find optimum values of n_{dh} and m_{dh} ($d = 1, \ldots, D$, $h = 1, \ldots, H_d$) under fixed overall survey cost (3) equal C (given c_{1dh} and c_{2dh}) so that the condition (4) is satisfied and the common value φ is minimum. We will optimize the design based on the assumption that the survey variable is the same as the auxiliary variable used to allocate the survey cost. Of course, in practice, it is an untrue situation; instead of the population values, the quantities originating from recent censuses or previous/pilot surveys are used.

Theorem 1. When a population U is subdivided into D domains and stratified two-stage sampling with fixed sample size of secondary sampling units from primary sampling units is to be applied within the domains, under a cost function (3), given survey costs C, C_0, c_{1dh} and c_{2dh}, the smallest common value φ of $g_d^{-1}\delta(\hat{Y}_d)$, $d = 1, \ldots, D$ is obtained when for $d = 1, \ldots, D$, $h = 1, \ldots, H_d$,

$$n_{dh} = \sqrt{\frac{c_{1dh}}{c_{2dh}}}\sqrt{\frac{\sum_{g=1}^{M_{dh}} N_{dhg}^2 S_{2dhg}^2}{M_{dh}S_{1dh}^2 - \sum_{g=1}^{M_{dh}} N_{dhg}S_{2dhg}^2}}$$

$$m_{dh} = \frac{(C - C_0)v_d\sqrt{M_{dh}\left(M_{dh}S_{1dh}^2 - \sum_{g=1}^{M_{dh}} N_{dhg}S_{2dhg}^2\right)}}{Y_d\sqrt{c_{1dh}}\sum_{e=1}^{D} v_e Y_e^{-1}\sum_{i=1}^{H_e}\sqrt{M_{ei}}Z_{ei}}$$

where $Z_{ei} = \left[\sqrt{c_{1ei}\left(M_{ei}S_{1ei}^2 - \sum_{k=1}^{M_{ei}} N_{eik}S_{2eik}^2\right)} + \sqrt{c_{2ei}\sum_{k=1}^{M_{ei}} N_{eik}^2 S_{2eik}^2}\right]$ and $\mathbf{v} = (v_1, \ldots, v_D)^T$ is the eigenvector connected with the largest eigenvalue of the matrix

$$\mathbf{F} = \Big\{(C - C_0)^{-1}\mathbf{A}\mathbf{B}^T - diag(\mathbf{E})\Big\},$$

where $\mathbf{A} = (A_1, \ldots, A_D)^T$, $\mathbf{B} = (B_1, \ldots, B_D)^T$ and $\mathbf{E} = (E_1, \ldots, E_D)^T$, provided that

$$M_{dh}S_{1dh}^2 - \sum_{g=1}^{M_{dh}} N_{dhg}S_{2dhg}^2 > 0 \tag{5}$$

for each $d = 1, \ldots, D$, and $h = 1, \ldots, H_d$.

Proof. To prove Theorem 1, a procedure developed by Niemiro and Wesołowski (2001) may be used. It was recently applied in sample allocation between domains and strata by Kozak and Zieliński (2005). Consider the following Lagrange function:

$$\begin{aligned} L = \varphi - \sum_{d=1}^{D} \lambda_d \left[\frac{1}{Y_d^2}\sum_{h=1}^{H_d}\frac{1}{m_{dh}}\left(u_{dh} + \frac{w_{dh}}{n_{dh}} - x_{dh}\right) - \frac{1}{Y_d^2}\sum_{h=1}^{H_d} M_{dh}S_{1dh}^2 - g_d^2\varphi^2\right] \\ -\alpha\left[\sum_{d=1}^{D}\sum_{h=1}^{H_d} m_{dh}(c_{1dh} + n_{dh}c_{2dh}) - (C - C_0)\right] \end{aligned} \tag{6}$$

where λ_d and α are the Lagrange multipliers, $w_{dh} = M_{dh}\sum_{g=1}^{M_{dh}} N_{dhg}^2 S_{2dhg}^2$, $u_{dh} = M_{dh}^2 S_{1dh}^2$, $x_{dh} = M_{dh}\sum_{g=1}^{M_{dh}} N_{dhg}S_{2dhg}^2$, and Y_d is the population total of Y in the dth domain. Differentiation of (6) with respect to m_{dh}, n_{dh}, λ_d, and α and solving the obtained equations yield the results presented in Theorem 1. A detailed proof may be obtained from the authors upon request.

Remark 1. If any of the conditions (5) or any of the following conditions

$$2 \le n_{dh} \le N_{dh};\ 2 \le m_{dh} \le M_{dh}\ for\ d = 1, \ldots, D, h = 1, \ldots, H_d, \tag{7}$$

is not fulfilled, the values of n_{dh} and m_{dh} from Theorem 1 are not real numbers, so they are not optimum. In such a case, the optimum n_{dh} and m_{dh} are the solution of the following numerical problem:

minimize $f\Big\{(\mathbf{n}_1, \mathbf{m}_1), \ldots, (\mathbf{n}_D, \mathbf{m}_D); \varphi\Big\} = \varphi$,

where $\mathbf{n}_d = \Big(n_{d1}, \ldots, n_{dH_d}\Big)^T$ and $\mathbf{m}_d = \Big(m_{d1}, \ldots, m_{dH_d}\Big)^T$ for $d = 1, \ldots, D$

subject to:

$$\frac{1}{Y_d^2}\sum_{h=1}^{H_d}\frac{M_{dh}}{m_{dh}}\left[\Big(M_{dh}-m_{dh}\Big)S_{1dh}^2+\frac{1}{n_{dh}}\sum_{g=1}^{M_{dh}}N_{dhg}\Big(N_{dhg}-n_{dh}\Big)S_{2dhg}^2\right]=g_d^2\varphi^2$$

$$\sum_{d=1}^{D}\sum_{h=1}^{H_d}m_{dh}(c_{1dh}+n_{dh}c_{2dh})=C-C_0$$

$$M_{dh}S_{1dh}^2-\sum_{g=1}^{M_{dh}}N_{dhg}S_{2dhg}^2>0\quad for\ each\ d=1,\ldots,D,\ and\ h=1,\ldots,H_d$$

$$2\le n_{dh}\le N_{dh};\ 2\le m_{dh}\le M_{dh}\quad for\ d=1,\ldots,D,\ h=1,\ldots,H_d.$$

4 Optimizing a design subject to constraints connected with domain precisions

Here we consider a question dual to the problem presented in previous section. We aim at minimizing a total survey cost C given in (3) subject to

$$\frac{1}{Y_d^2}\sum_{h=1}^{H_d}\frac{M_{dh}}{m_{dh}}\left[\Big(M_{dh}-m_{dh}\Big)S_{1dh}^2+\frac{1}{n_{dh}}\sum_{g=1}^{M_{dh}}N_{dhg}\Big(N_{dhg}-n_{dh}\Big)S_{2dhg}^2\right]=\delta_d^2,\quad d=1,\ldots,D,\tag{8}$$

where δ_d is the fixed value of coefficient of variation of $\hat{Y}_d$. Thus, this time we consider a design in which we look for optimum values of n_{dh} and m_{dh} for which the constraint (8) is fulfilled and the cost (3) is minimum.

Theorem 2. When a population U is subdivided into D domains and stratified two-stage sampling with fixed sample size of secondary sampling units from primary sampling units is to be applied within the domains, under a cost function (3), given survey costs C_0, c_{1dh} and c_{2dh}, and under the condition (8) (for δ_d being fixed), the minimum total survey cost C is obtained when for $d=1,\ldots,D$, $h=1,\ldots,H_d$,

$$n_{dh} = \sqrt{\frac{c_{1dh}}{c_{2dh}}} \sqrt{\frac{\sum_{g=1}^{M_{dh}} N_{dhg}^2 S_{2dhg}^2}{M_{dh} S_{1dh}^2 - \sum_{g=1}^{M_{dh}} N_{dhg} S_{2dhg}^2}}$$

$$m_{dh} = \frac{\sqrt{M_{dh}\Big(M_{dh} S_{1dh}^2 - \sum_{g=1}^{M_{dh}} N_{dhg} S_{2dhg}^2\Big)}}{Y_d \sqrt{c_{1dh}}} \cdot \frac{\sum_{i=1}^{H_d} D_i}{\delta_d^2 + Y_d^{-2} \sum_{h=1}^{H_d} M_{dh} S_{1dh}^2}$$

$$D_i = \sqrt{c_{2di} M_{di}\Big(M_{di} S_{1di}^2 - \sum_{k=1}^{M_{di}} N_{dik} S_{2dik}^2\Big)} + \sqrt{M_{di} \sum_{k=1}^{M_{di}} N_{dik}^2 S_{2dik}^2}$$

provided that

$$M_{dh} S_{1dh}^2 - \sum_{g=1}^{M_{dh}} N_{dhg} S_{2dhg}^2 > 0 \tag{9}$$

for each $d = 1, \ldots, D$, and $h = 1, \ldots, H_d$.

Proof. Consider the following Lagrange function

$$\begin{aligned} L = C_0 &+ \sum_{d=1}^{D} \sum_{h=1}^{H_d} m_{dh}(c_{1dh} + n_{dh} c_{2dh}) \\ &+ \sum_{d=1}^{D} \lambda_d \left[\frac{1}{Y_d^2} \sum_{h=1}^{H_d} \frac{1}{m_{dh}} \Big(u_{dh} + \frac{w_{dh}}{n_{dh}} - x_{dh}\Big) - \frac{1}{Y_d^2} \sum_{h=1}^{H_d} M_{dh} S_{1dh}^2 - \delta_d^2 \right] \end{aligned} \tag{10}$$

where λ_d are the Lagrange multipliers and u_{dh}, w_{dh}, and x_{dh} are the same as defined in previous section. Differentiating of (10) with respect to m_{dh}, n_{dh}, and λ_d and solving the obtained equations lead to the results presented in Theorem 1. A detailed proof may be obtained from the authors upon request.

Remark 2. If any of the conditions (9) or any of the conditions (7) is not fulfilled, the values of n_{dh} and m_{dh} from Theorem 2 are not real numbers, so they are not optimum. In such a case, the optimum n_{dh} and m_{dh} are the solution of the following numerical problem:

$$\text{minimize } f\Big\{(\mathbf{n}_1, \mathbf{m}_1), \ldots, (\mathbf{n}_D, \mathbf{m}_D); \varphi\Big\} = C_0 + \sum_{d=1}^{D} \sum_{h=1}^{H_d} m_{dh}\Big(c_{1dh} + n_{dh} c_{2dh}\Big),$$

where $\mathbf{n}_d = \left(n_{d1}, \ldots, n_{dH_d}\right)^T$ and $\mathbf{m}_d = \left(m_{d1}, \ldots, m_{dH_d}\right)^T$ for $d = 1, \ldots, D$

subject to:

$$\frac{1}{Y_d^2} \sum_{h=1}^{H_d} \frac{M_{dh}}{m_{dh}} \left[\left(M_{dh} - m_{dh} \right) S_{1dh}^2 + \frac{1}{n_{dh}} \sum_{g=1}^{M_{dh}} N_{dhg} \left(N_{dhg} - n_{dh} \right) S_{2dhg}^2 \right] = \delta_d^2$$

$$M_{dh} S_{1dh}^2 - \sum_{g=1}^{M_{dh}} N_{dhg} S_{2dhg}^2 > 0 \quad for\ each\ d = 1, \ldots, D,\ and\ h = 1, \ldots, H_d$$

$$2 \leq n_{dh} \leq N_{dh};\ 2 \leq m_{dh} \leq M_{dh} \quad for\ d = 1, \ldots, D,\ h = 1, \ldots, H_d.$$

References

Niemiro, W., Wesołowski, J. (2001), Fixed precision optimal allocation in two-stage sampling, Applicationes Mathematicae **23**, 73-82.

Kozak, M. (2005), On stratified two-stage sampling: Optimum stratification and sample allocation between strata and sampling stages. Model Assisted Statistics and Applications **1**(1), 23-29.

Kozak, M., Zieliński, A. (2005). Sample allocation between domains and strata. International Journal of Applied Mathematics and Statistics **3**, 19-40.

Särndal, C. E., Swensson, B., Wretman, J. (1992), Model Assisted Survey Sampling (Springer-Verlag, New York).

Singh, S. (2003), Advanced Sampling Theory with Applications. How Michael *"Selected" Amy* (Kluwer Academic Publishers, The Netherlands).