Stratified two-stage sampling in domains: Sample allocation between domains, strata, and sampling stages
Marcin Kozak, Andrzej Zieliński, Sarjinder Singh

To cite this version:

HAL Id: hal-00616535
https://hal.archives-ouvertes.fr/hal-00616535
Submitted on 23 Aug 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
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).

Preprint submitted to Elsevier Science 3 October 2007
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

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

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

\[
U_{dhg} \cap U_{dh'g'} = \emptyset \quad \text{for} \quad d = 1, \ldots, D, \ h = 1, \ldots, H_d, \ g, g' = 1, \ldots, M_{dh}, \ g \neq g'.
\]

The \( g \)th PSU from the \( h \)th stratum in the \( d \)th domain comprises \( N_{dhg} \) SSUs, which are the population elements. Let \( N_d \) indicate the number of SSUs in the \( d \)th domain and \( N_{dh} \) indicate the number of SSUs in the \( h \)th stratum of the \( d \)th domain.

Let a population parameter investigated be the population total of \( Y \), \( Y \) being a characteristic studied. For the \( d \)th domain its estimator is given by

\[
\hat{Y}_d = \sum_{h=1}^{H_d} \hat{Y}_{dh} = \sum_{h=1}^{H_d} \sum_{m=1}^{M_{dh}} \sum_{g=1}^{n_{dhg}} \sum_{i=1}^{n_{dhgi}} y_{dhgi}
\]
where \( \hat{Y}_d \) is the estimator of \( Y_d \), \( Y_d \) being the population total of the variable \( Y \) restricted to the \( d \)th domain; \( \hat{Y}_{dh} \) is the estimator of \( Y_{dh} \), \( Y_{dh} \) being the population total of the variable \( Y \) restricted to the \( h \)th stratum of the \( d \)th domain; \( m_{dh} \) is the sample size of PSUs from the \( h \)th stratum of the \( d \)th domain; \( n_{dhg} \) is the sample size of SSUs from the \( g \)th PSU of the \( h \)th stratum of the \( d \)th domain; and \( y_{dhgi} \) is the \( Y \) value in the \( i \)th SSU (population element) of the \( g \)th PSU from the \( h \)th stratum in \( d \)th 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)

\[
V(\hat{Y}_d) = \frac{H_d}{m_{dh}} \sum_{h=1}^{H_d} \left( m_{dh} - m_{dh} \right) S^2_{1dh} + \frac{1}{n_{dh}} \sum_{g=1}^{M_{dh}} N_{dhg}(N_{dhg} - n_{dh}) S^2_{2dhg} \tag{2}
\]

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

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 \( g \)th PSU of the \( h \)th stratum in the \( d \)th domain. An ordinary unbiased estimator of the variance (2) is obtained by replacing the population quantities \( S^2_{1dh} \) and \( S^2_{2dhg} \) with their sample estimators; the summation in (2) is to be done by sampled PSUs in each \( h \)th stratum from the \( d \)th domain.

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

\[
\delta(\hat{Y}_d) = \frac{\sqrt{V(\hat{Y}_d)}}{\bar{Y}_d}, \quad d = 1, \ldots, D
\]

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

\[
C = C_0 + \sum_{d=1}^{D} \sum_{h=1}^{H_d} m_{dh} \left( c_{1dh} + n_{dh}c_{2dh} \right) \tag{3}
\]
where $C_0$ is the fixed survey cost, $c_{1dh}$ is the cost of selecting one PSU from the $h$th stratum of the $d$th domain, and $c_{2dh}$ is the cost of obtaining the information on $Y$ value in one SSU from the $h$th stratum of the $d$th 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 $\varphi$ 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 $d$th 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 $\varphi$ 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 $\varphi$ of $g_d^{-1}\delta(\hat{Y}_d)$, $d = 1, \ldots, D$, is obtained when for $d = 1, \ldots, D$, $h = 1, \ldots, H_d$,

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

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

\[ Z_{ei} = \sqrt{c_1 ei (M_{ei} S_1^2 - \sum_{k=1}^{M_{ei}} N_{eik} S_2^2)} + \sqrt{c_2 ei \sum_{k=1}^{M_{ei}} N_{eik}^2 S_2^2 eik} \]

and 

\[ v = (v_1, \ldots, v_D)^T \]

is the eigenvector connected with the largest eigenvalue of the matrix 

\[ F = \{(C - C_0)^{-1} AB^T - \text{diag}(E)\}, \]

where 

\[ A = (A_1, \ldots, A_D)^T, \quad B = (B_1, \ldots, B_D)^T \quad \text{and} \quad E = (E_1, \ldots, E_D)^T, \]

provided that 

\[ M_{dh} S_1^2 - \sum_{g=1}^{M_{dh}} N_{dhg} S_2^2 > 0 \] (5)

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

**Proof.** To prove Theorem 1, a procedure developed by Niemiro and Wesolowski (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:

\[ L = \varphi - \sum_{d=1}^{D} \lambda_d \left[ \frac{1}{Y_d^2} \sum_{h=1}^{H_d} \frac{1}{m_{dh}} (u_{dh} + w_{dh} - x_{dh}) \right] - \frac{1}{Y_d^2} \sum_{h=1}^{H_d} M_{dh} S_1^2 - \sum_{h=1}^{H_d} M_{dh} S_1^2 - \sum_{h=1}^{H_d} M_{dh} S_1^2 - \sum_{h=1}^{H_d} M_{dh} S_1^2 \]

where \( \lambda_d, \alpha \) are the Lagrange multipliers, 

\[ u_{dh} = M_{dh} S_1^2, \quad x_{dh} = M_{dh} S_1^2, \quad w_{dh} = M_{dh} S_1^2, \quad v_{dh} = M_{dh} S_1^2, \quad \text{and} \quad Y_d \]

is the population total of \( Y \) in the \( d \)th domain. Differentiation of (6) with respect to \( m_{dh}, \lambda_d, \alpha \) 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

\[ \begin{align*}
2 & \leq n_{dh} \leq N_{dh}; \quad 2 \leq m_{dh} \leq M_{dh} \quad \text{for} \quad d = 1, \ldots, D, \ h = 1, \ldots, H_d;
\end{align*} \] (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 \( f\{(m_1, m_1), \ldots, (m_D, m_D); \varphi\} = \varphi, \)

where \( n_d = (n_{d1}, \ldots, n_{dH_d})^T \) and \( m_d = (m_{d1}, \ldots, m_{dH_d})^T \) for \( d = 1, \ldots, D \)
subject to:

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

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

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

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

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

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

\( d = 1, \ldots, D, \)

where \( \delta_d \) 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 \( \delta_d \) being fixed), the minimum total survey cost \( C \) is obtained when for \( d = 1, \ldots, D, \ h = 1, \ldots, H_d, \)
\[
\begin{align*}
n_{dh} &= \sqrt{\frac{c_{1dh}}{c_{2dh}}} \left[ \sum_{g=1}^{M_{dh}} N_{d_{gh}}^2 S_{2d_{gh}}^2 - M_{dh} S_{1dh}^2 - \sum_{g=1}^{M_{dh}} N_{d_{gh}} S_{2d_{gh}}^2 \right] \quad \text{for each } d = 1, \ldots, D, \text{ and } h = 1, \ldots, H_d. \\
m_{dh} &= \frac{\sqrt{M_{dh} \left( M_{dh} S_{1dh}^2 - \sum_{g=1}^{M_{dh}} N_{d_{gh}} S_{2d_{gh}}^2 \right)}}{Y_d \sqrt{c_{1dh}}} \cdot \frac{\sum_{i=1}^{H_d} D_i}{\delta_d^2 + \sum_{h=1}^{H_d} M_{dh} S_{1dh}^2} \\
D_i &= \sqrt{c_{2di} M_{di} \left( M_{di} S_{1di}^2 - \sum_{k=1}^{M_{di}} N_{d_{ik}} S_{2d_{ik}}^2 \right) + \sqrt{M_{di} \sum_{k=1}^{M_{di}} N_{d_{ik}} S_{2d_{ik}}^2}} \\
\text{provided that } \\
M_{dh} S_{1dh}^2 - \sum_{g=1}^{M_{dh}} N_{d_{gh}} S_{2d_{gh}}^2 > 0 \\
\text{for each } d = 1, \ldots, D, \text{ and } h = 1, \ldots, H_d.
\end{align*}
\]

Proof. Consider the following Lagrange function

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

where \( \lambda_d \) 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 \( \lambda_d \) 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:

\[
\begin{align*}
\text{minimize } f \left\{ (n_1, m_1), \ldots, (n_D, m_D); \varphi \right\} &= C_0 + \sum_{d=1}^{D} \sum_{h=1}^{H_d} m_{dh} \left( c_{1dh} + n_{dh} c_{2dh} \right), \\
\end{align*}
\]
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:

\[
\frac{1}{Y_d^2} \sum_{h=1}^{H_d} \frac{M_{dh}}{m_{dh}} \left[ \left( M_{dh} - m_{dh} \right) S^2_{1dh} + \frac{1}{n_{dh}} \sum_{g=1}^{M_{dh}} \left( N_{dhg} - n_{dh} \right) S^2_{2dhg} \right] = \delta_d^2 \\
M_{dh} S^2_{1dh} - \sum_{g=1}^{M_{dh}} N_{dhg} S^2_{2dhg} > 0 \quad \text{for each } d = 1, \ldots, D, \text{ and } h = 1, \ldots, H_d
\]

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

References


