

Hierarchical Estimation in Stratified Surveys

A design-based approach

Andrés Gutiérrez
Universidad Santo Tomás
hugogutierrez@usta.edu.co

JSM - 2011

Contents

- 1 Introduction
- 2 Multipurpose estimation
- 3 Estimation in the first level
 - Proposed approach
 - Indirect sampling
- 4 Simulation study
- 5 Discussion

Contents

- 1 Introduction
- 2 Multipurpose estimation
- 3 Estimation in the first level
 - Proposed approach
 - Indirect sampling
- 4 Simulation study
- 5 Discussion

Contents

- 1 Introduction
- 2 Multipurpose estimation
- 3 Estimation in the first level
 - Proposed approach
 - Indirect sampling
- 4 Simulation study
- 5 Discussion

Contents

- 1 Introduction
- 2 Multipurpose estimation
- 3 Estimation in the first level
 - Proposed approach
 - Indirect sampling
- 4 Simulation study
- 5 Discussion

Contents

- 1 Introduction
- 2 Multipurpose estimation
- 3 Estimation in the first level
 - Proposed approach
 - Indirect sampling
- 4 Simulation study
- 5 Discussion

Idea

Aim:

- To provide a multipurpose approach to the joint estimation of several parameters for different variables in a stratified finite population with two levels.

How?

- Horvitz-Thompson estimator.
- Hajek estimator.
- Indirect sampling.

Idea

Aim:

- To provide a multipurpose approach to the joint estimation of several parameters for different variables in a stratified finite population with two levels.

How?

- Horvitz-Thompson estimator.
- Hajek estimator.
- Indirect sampling.

Idea

Aim:

- To provide a multipurpose approach to the joint estimation of several parameters for different variables in a stratified finite population with two levels.

How?

- Horvitz-Thompson estimator.
- Hajek estimator.
- Indirect sampling.

Idea

Aim:

- To provide a multipurpose approach to the joint estimation of several parameters for different variables in a stratified finite population with two levels.

How?

- Horvitz-Thompson estimator.
- Hajek estimator.
- Indirect sampling.

Framework

- Most of the real applications in survey sampling involve not one, but several characteristics of study (real populations have hierarchical structures).
- The survey methodologist is faced with the estimation of several parameters of interest in different levels of the population and he/she is commanded with the seeking of proper approaches to estimate those parameters as required in the study.
- Although there is a vast research about estimation of hierarchical populations and model-based (or model-assisted) multi-level survey data, the design-based estimation for finite populations with hierarchical structures seems to be omitted by survey statisticians.

Framework

- Most of the real applications in survey sampling involve not one, but several characteristics of study (real populations have hierarchical structures).
- The survey methodologist is faced with the estimation of several parameters of interest in different levels of the population and he/she is commanded with the seeking of proper approaches to estimate those parameters as required in the study.
- Although there is a vast research about estimation of hierarchical populations and model-based (or model-assisted) multi-level survey data, the design-based estimation for finite populations with hierarchical structures seems to be omitted by survey statisticians.

Framework

- Most of the real applications in survey sampling involve not one, but several characteristics of study (real populations have hierarchical structures).
- The survey methodologist is faced with the estimation of several parameters of interest in different levels of the population and he/she is commanded with the seeking of proper approaches to estimate those parameters as required in the study.
- Although there is a vast research about estimation of hierarchical populations and model-based (or model-assisted) multi-level survey data, the design-based estimation for finite populations with hierarchical structures seems to be omitted by survey statisticians.

Clarifying example: establishment survey

- It is of interest to estimate the total sales of the market sections of the stores in detail (sales by toys, grocery, electronics or pharmacy sections) and at the same time it is of interest to estimate the number of employees working in the stores.
- The multipurpose approach is given by the joint inference of two different study variables (sales by market section and number of employees in the stores) but these variables of interest are in different levels of the population: sales are related with the market section level and the number of employees with the store level.
- The set of all **market sections** defines the second level and the set of all **stores** defines the first level.

Clarifying example: establishment survey

- It is of interest to estimate the total sales of the market sections of the stores in detail (sales by toys, grocery, electronics or pharmacy sections) and at the same time it is of interest to estimate the number of employees working in the stores.
- The multipurpose approach is given by the joint inference of two different study variables (sales by market section and number of employees in the stores) but these variables of interest are in different levels of the population: sales are related with the market section level and the number of employees with the store level.
- The set of all **market sections** defines the second level and the set of all **stores** defines the first level.

Clarifying example: establishment survey

- It is of interest to estimate the total sales of the market sections of the stores in detail (sales by toys, grocery, electronics or pharmacy sections) and at the same time it is of interest to estimate the number of employees working in the stores.
- The multipurpose approach is given by the joint inference of two different study variables (sales by market section and number of employees in the stores) but these variables of interest are in different levels of the population: sales are related with the market section level and the number of employees with the store level.
- The set of all **market sections** defines the second level and the set of all **stores** defines the **first level**.

Possible solutions

- Naive solution: Designing two surveys. (High cost)
- Indirect Sampling: As the population levels are related, it could be proposed to use the Generalized Share Weight Methodology in order to obtain estimates of the variables of interest. (Low cost, low efficiency)
- Our approach: Based on the computation of the first and second order inclusion probabilities, given by the induced sampling design in the first level, by using the principles of the well-known Horvitz-Thompson and Hájek estimators. (Low cost, high efficiency)

Possible solutions

- Naive solution: Designing two surveys. (High cost)
- Indirect Sampling: As the population levels are related, it could be proposed to use the Generalized Share Weight Methodology in order to obtain estimates of the variables of interest. (Low cost, low efficiency)
- Our approach: Based on the computation of the first and second order inclusion probabilities, given by the induced sampling design in the first level, by using the principles of the well-known Horvitz-Thompson and Hájek estimators. (Low cost, high efficiency)

Possible solutions

- Naive solution: Designing two surveys. (High cost)
- Indirect Sampling: As the population levels are related, it could be proposed to use the Generalized Share Weight Methodology in order to obtain estimates of the variables of interest. (Low cost, low efficiency)
- Our approach: Based on the computation of the first and second order inclusion probabilities, given by the induced sampling design in the first level, by using the principles of the well-known Horvitz-Thompson and Hájek estimators. (Low cost, high efficiency)

Notation

- Let $U = \{1, \dots, k, \dots, N\}$ denote the second level finite population of N elements in which a sampling frame is available. Suppose also that U is partitioned into H subsets U_1, U_2, \dots, U_H called strata.
- Each element $k \in U$ in the second level belongs to a unique cluster in the first level. It is supposed that there exist N_I clusters denoted by $U_1, \dots, U_i, \dots, U_{N_I}$. This set of clusters is represented as $U_I = \{1, \dots, i, \dots, N_I\}$.
- This way, the first level population is U_I , the second level population is U and, clearly, the data show a notorious hierarchical structure.

Notation

- Let $U = \{1, \dots, k, \dots, N\}$ denote the second level finite population of N elements in which a sampling frame is available. Suppose also that U is partitioned into H subsets U_1, U_2, \dots, U_H called strata.
- Each element $k \in U$ in the second level belongs to a unique cluster in the first level. It is supposed that there exist N_I clusters denoted by $U_1, \dots, U_i, \dots, U_{N_I}$. This set of clusters is represented as $U_I = \{1, \dots, i, \dots, N_I\}$.
- This way, the first level population is U_I , the second level population is U and, clearly, the data show a notorious hierarchical structure.

Notation

- Let $U = \{1, \dots, k, \dots, N\}$ denote the second level finite population of N elements in which a sampling frame is available. Suppose also that U is partitioned into H subsets U_1, U_2, \dots, U_H called strata.
- Each element $k \in U$ in the second level belongs to a unique cluster in the first level. It is supposed that there exist N_I clusters denoted by $U_1, \dots, U_i, \dots, U_{N_I}$. This set of clusters is represented as $U_I = \{1, \dots, i, \dots, N_I\}$.
- This way, the first level population is U_I , the second level population is U and, clearly, the data show a notorious hierarchical structure.

Assumptions

- Although there is an available sampling frame for U , suppose that it is impossible to obtain a frame for the population of the first level U_1 .
- The requirements of the survey imply the inference of parameters, say population totals or means, for both levels.
- It is supposed there are two variables of interest, say, y in the second level, and z in the first level, and it is requested the estimation of both population totals, defined by

$$t_y = \sum_{k \in U} y_k = \sum_{h=1}^H \sum_{k \in U_h} y_k$$

and

$$t_z = \sum_{i \in U_1} z_i$$

Assumptions

- Although there is an available sampling frame for U , suppose that it is impossible to obtain a frame for the population of the first level U_1 .
- The requirements of the survey imply the inference of parameters, say population totals or means, for both levels.
- It is supposed there are two variables of interest, say, y in the second level, and z in the first level, and it is requested the estimation of both population totals, defined by

$$t_y = \sum_{k \in U} y_k = \sum_{h=1}^H \sum_{k \in U_h} y_k$$

and

$$t_z = \sum_{i \in U_1} z_i$$

Assumptions

- Although there is an available sampling frame for U , suppose that it is impossible to obtain a frame for the population of the first level U_I .
- The requirements of the survey imply the inference of parameters, say population totals or means, for both levels.
- It is supposed there are two variables of interest, say, y in the second level, and z in the first level, and it is requested the estimation of both population totals, defined by

$$t_y = \sum_{k \in U} y_k = \sum_{h=1}^H \sum_{k \in U_h} y_k$$

and

$$t_z = \sum_{i \in U_I} z_i$$

Drawing a sample

- By taking advantage of the sampling frame in the second level, a stratified sample s is drawn. For each $k \in s$, the value of the variable of interest y_k is observed.
- It is supposed that unit k can also provide the information of its corresponding cluster, say U_j . This way, the value of the another variable of interest z_j is recorded.
- Note that for a particular second level sample there exists a corresponding set of units in the first level. In other words, the second level sample s induces a set, contained in the first level population, which will be called the first level sample, denoted by m and given by

$$m = \{i \in U_j \mid \text{at least one unit of the cluster } U_j \text{ belong to } s\}$$

Drawing a sample

- By taking advantage of the sampling frame in the second level, a stratified sample s is drawn. For each $k \in s$, the value of the variable of interest y_k is observed.
- It is supposed that unit k can also provide the information of its corresponding cluster, say U_i . This way, the value of the another variable of interest z_i is recorded.
- Note that for a particular second level sample there exists a corresponding set of units in the first level. In other words, the second level sample s induces a set, contained in the first level population, which will be called the first level sample, denoted by m and given by

$$m = \{i \in U_i \mid \text{at least one unit of the cluster } U_i \text{ belong to } s\}$$

Drawing a sample

- By taking advantage of the sampling frame in the second level, a stratified sample s is drawn. For each $k \in s$, the value of the variable of interest y_k is observed.
- It is supposed that unit k can also provide the information of its corresponding cluster, say U_i . This way, the value of the another variable of interest z_i is recorded.
- Note that for a particular second level sample there exists a corresponding set of units in the first level. In other words, the second level sample s induces a set, contained in the first level population, which will be called the first level sample, denoted by m and given by

$$m = \{i \in U_i \mid \text{at least one unit of the cluster } U_i \text{ belong to } s\}$$

Another clarifying example

Table: Description of a possible hierarchical configuration

	Section 1	Section 2	Section 3	Section 4
Store A	A1	A2	-	A4
Store B	B1	-	B3	-
Store C	-	C2	-	C4
Store D	D1	D2	D3	D4
Store E	E1	E2	E3	E4

If the selected second sample is $s = \{A1, E2, B3, E4\}$, the induced first level sample is $m = \{A, B, E\}$. Note that a store may be selected more than once; however, we omit the repeated information in the first level and carry out the inference by using the reduced sample.

Another clarifying example

Table: Variables of interest in a possible hierarchical configuration

Y_1	Y_2	Y_3	Y_4	Z
$y_{A1} = 32$	$y_{A2} = 12$	-	$y_{A4} = 51$	$Z_A = 14.12$
$y_{B1} = 18$	-	$y_{B3} = 26$	-	$Z_B = 10.25$
-	$y_{C2} = 36$	-	$y_{C4} = 10$	$Z_C = 17.52$
$y_{D1} = 42$	$y_{D2} = 24$	$y_{D3} = 14$	$y_{D4} = 46$	$Z_D = 22.58$
$y_{E1} = 14$	$y_{E2} = 33$	$y_{E3} = 28$	$y_{E4} = 55$	$Z_E = 24.81$

The parameter of interest in the first level is $t_z = 14.12 + 10.25 + 17.52 + 22.58 + 24.81 = 89.28$ and the parameter of interest in the second level is $t_y = 106 + 105 + 68 + 162 = 441$.

Estimation in the second level

We have that an unbiased estimator of t_y and its variance are given by

$$\hat{t}_{y\pi} = \sum_{h=1}^H \sum_{s_h} \frac{y_k}{\pi_k} = \sum_{h=1}^H \hat{t}_{h\pi} \quad (1)$$

$$V(\hat{t}_{y\pi}) = \sum_{h=1}^H V_h(\hat{t}_{h\pi}) = \sum_{h=1}^H \sum_{k \in U_h} \sum_{l \in U_h} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

where $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$, and $\hat{t}_{h\pi}$ corresponds to the Horvitz-Thompson estimator in the h -th stratum, defined by

$$\hat{t}_{h\pi} = \sum_{s_h} \frac{y_k}{\pi_k}$$

The induced sampling design

Result

The sampling design in the first level induced by the stratified sample s is given by

$$p(m) = \sum_{\{s: s \rightarrow m\}} \prod_{h=1}^H p_h(s_h) \quad (2)$$

where the notation $s \rightarrow m$ indicates that the second level sample s induces the first level sample m .

The induced sampling design

For example, continuing with the population described in Table 1, if the sampling design in the second level is simple random sampling in each stratum such that $N_3 = 3$, $N_1 = N_2 = N_4 = 4$ and $n_h = 1$ for $h = 1, 2, 3, 4$, then in order to compute the selection probability of the particular first level sample $m = \{A, B\}$, it is necessary to find all of the second level samples that induce that specific sample m . Given the data structure, the set $\{s : s \rightarrow m\}$ has only two second level samples; these samples are: $\{A1, A2, B3, A4\}$ and $\{B1, A2, B3, A4\}$. For that m , we have that its selection probability corresponds to

$$\begin{aligned} p(m) &= p(\{A1, A2, B3, A4\}) + p(\{B1, A2, B3, A4\}) \\ &= \prod_{h=1}^4 \frac{1}{N_h} + \prod_{h=1}^4 \frac{1}{N_h} = \frac{2}{192} = 0.0104 \end{aligned}$$

The first order inclusion probabilities

Result

Suppose that the elements of the cluster U_i are denoted by k_1, \dots, k_r , that is, U_i is present in r strata with ($r \leq H$). These strata are denoted by $U_{(1)}, \dots, U_{(r)}$. The first order inclusion probability of the cluster U_i , denoted by π_i , is given by

$$\pi_i = Pr(i \in m) = 1 - \prod_{h=1}^r \pi_{k(h)}^c \quad (3)$$

where $\pi_{k(h)}^c = 1 - \pi_{k(h)}$, is the probability that k_h belongs to $s_{(h)}^c$ with $s_{(h)}^c = U_{(h)} - s_{(h)}$, and $s_{(h)}$ denotes the selected sample in the stratum $U_{(h)}$, for $h = 1, \dots, r$.

The second order inclusion probabilities

Result

Suppose that the elements of the cluster U_i are denoted by k_1, \dots, k_r , that is, U_i is presented in r strata with $r \leq H$, and the elements of the cluster U_j are l_1, \dots, l_q , that is, U_j is presented in q strata with $q \leq H$. The second order inclusion probability for any pair of clusters U_i, U_j is given by

$$\pi_{ij} = P(i, j \in m) = \pi_i + \pi_j + \prod_{h=1}^p \pi_{k_{(h)}^c}^{I_{(h)}^c} \prod_{h=p+1}^r \pi_{k_{(h)}^c} \prod_{h=r+1}^{q-p+r} \pi_{l_{(h)}^c} - 1 \quad (4)$$

With $\pi_{k_{(h)}^c}^{I_{(h)}^c} = Pr(k_h \text{ y } l_h \notin s_{(h)})$.

HT

Once these inclusion probabilities are computed, it is possible to estimate t_z by means of the well known Horvitz-Thompson estimator given by

$$\hat{t}_{z\pi} = \sum_{i \in m} \frac{z_i}{\pi_i} \quad (5)$$

with variance given by

$$V(\hat{t}_{z\pi}) = \sum_{i \in U_1} \sum_{j \in U_1} \Delta_{ij} \frac{z_i}{\pi_i} \frac{z_j}{\pi_j}$$

Since the first level sample is induced by the second level sample, the size of m is random, even when the stratified sample design of the second level is of fixed size.

Hajek

In order to avoid extreme estimates, sometimes obtained with the previous estimator, and having into account that N_I is known, we propose to use the expanded sample mean estimator (denoted in this research as Hajek estimator) given by

$$\tilde{t}_z = N_I \frac{\hat{t}_{z\pi}}{\hat{N}_{I,\pi}} \quad (6)$$

Where $\hat{N}_{I,\pi} = \sum_{i \in m} \frac{1}{\pi_i}$. And its approximate variance is given by

$$V(\tilde{t}_z) = \sum_{i \in U_I} \sum_{j \in U_I} \Delta_{ij} \frac{z_i - \bar{z}_{U_I}}{\pi_i} \frac{z_j - \bar{z}_{U_I}}{\pi_j} \quad (7)$$

With $\bar{z}_{i \in U_I} = \sum_{U_I} z_i / N_I$.

Example: STSI

The first order inclusion probability for a cluster U_i , which is presented in r strata, is given by

$$\pi_i = 1 - \prod_{h=1}^r \left(1 - \frac{n_h}{N_h}\right) \quad (8)$$

The product in the last expression is defined over the strata where the store is present (not necessary all of the H strata).

Example: STSI

The second order inclusion probability for clusters U_i and U_j is given by

$$\begin{aligned} \pi_{ij} = & 1 - \prod_{h=1}^r \left(1 - \frac{n_h}{N_h}\right) - \prod_{h=1}^q \left(1 - \frac{n_h}{N_h}\right) \\ & + \prod_{h=1}^p \frac{(N_h - n_h)(N_h - n_h - 1)}{N_h(N_h - 1)} \prod_{h=p+1}^r \left(1 - \frac{n_h}{N_h}\right) \prod_{h=r+1}^{q-p+r} \left(1 - \frac{n_h}{N_h}\right) \end{aligned} \quad (9)$$

Example: STSI

Following the finite population in Table 1, the first inclusion probabilities of the store A and store B are given by

$$\pi_{store(A)} = 1 - \left(1 - \frac{n_1}{N_1}\right) \left(1 - \frac{n_2}{N_2}\right) \left(1 - \frac{n_4}{N_4}\right)$$

$$\pi_{store(B)} = 1 - \left(1 - \frac{n_1}{N_1}\right) \left(1 - \frac{n_3}{N_3}\right)$$

And the second order inclusion probability for these two stores is given by

$$\begin{aligned} \pi_{store(A),store(B)} &= 1 - \left(1 - \frac{n_1}{N_1}\right) \left(1 - \frac{n_2}{N_2}\right) \left(1 - \frac{n_4}{N_4}\right) - \left(1 - \frac{n_1}{N_1}\right) \left(1 - \frac{n_3}{N_3}\right) \\ &\quad + \frac{(N_1 - n_1)}{N_1} \frac{(N_1 - n_1 - 1)}{(N_1 - 1)} \frac{(N_3 - n_3)}{N_3} \frac{(N_3 - n_3 - 1)}{(N_3 - 1)} \left(1 - \frac{n_4}{N_4}\right) \end{aligned}$$

WSGM

This kind of situations can also be handled by using the indirect sampling approach. It is assumed that the first level population U_1 is related to the second level population U through a link matrix representing the correspondence between the elements of U_1 and U . Since there is no available a sampling frame for U_1 , an estimate for t_z can be obtained indirectly using a sample from U and the existing links between the two populations.

WSGM

It is important to remark that even though the resulting inferences of indirect sampling from the GSWM are defined for the first level population, they are directly induced by the probability measure of the sampling design in the second level $p(s)$. However, the inferences from our proposal approach are given directly by the induced sampling design of the first level $p(m)$.

Setup

We compare the performances of the two proposed estimators with the indirect sampling estimator. We simulate several stratified populations with hierarchical structure where all clusters are presented in each stratum, that is, $N_h = N_l$ in all strata. The values of the variables of interest y and z are generated from different gamma distributions.

In each stratum, a simple random sample of equal size n is selected, then the two proposed estimators and the indirect sampling estimator are computed in order to estimate t_z . The process was repeated $G = 1000$ times with $N_l = 20, 50, 100, 400$ clusters, and $H = 5, 5, 10, 50$ for each of these values of N_l . The performance of an estimator \hat{t} of the parameter t was tracked by the Relative Efficiency (RE).

Results

Table: *Ratio of MSE of HT and Hájek estimators to indirect sampling estimator for $H = 5$ strata and $N_l = 20$ clusters*

Sample size per stratum	HT	Hájek
$n=1$	0,08	1,06
$n=5$	0,03	1,84
$n=10$	0,05	5,50
$n=15$	0,52	73,75

Results

Table: *Ratio of MSE of HT and Hájek estimators to indirect sampling estimator for $H = 5$ strata and $N_l = 50$ clusters*

Sample size per stratum	HT	Hájek
n=1	0,12	1,02
n=5	0,03	1,29
n=10	0,02	1,57
n=20	0,02	3,24
n=40	1,06	175,83

Results

Table: *Ratio of MSE of HT and Hájek estimators to indirect sampling estimator for $H = 10$ strata and $N_j = 100$ clusters*

Sample size per stratum	HT	Hájek
n=1	0,09	1,03
n=10	0,02	1,83
n=20	0,02	3,64
n=50	0,44	101,47

Results

Table: *Ratio of MSE of HT and Hájek estimators to indirect sampling estimator for $H = 50$ strata and $N_I = 40$ clusters*

Sample size per stratum	HT	Hájek
n=1	0,02	1,98
n=5	0,77	110,25
n=10	Inf	Inf
n=20	Inf	Inf

Conclusions

The reduction in the variability of our proposal may be explained because different second level samples may induce the same first level sample m . In this case, the estimates obtained by applying the GWSM principles will be generally different because the vector of weights \mathbf{w} , that depends on the inclusion probabilities of the selected elements in s , differs from sample to sample in the second level. Then we will have different estimates for the same induced sample m . This feature is not present if we follow the approach proposed in this research, since $\hat{t}_{z,\pi}$ and \tilde{t}_z remain constant for different second level samples that induce the same first level sample m .

Further work

Further work could be focused in the development of a general methodology that conducts to the joint estimation in more than two levels when the sampling frame is only available in the last level of the hierarchical population. Besides, noting that the current formulation of the GWSM does not admit auxiliary information in its functional form, the proposed approach could be easily extended in some situations where there is a suitable auxiliary variable (continuous or discrete) that helps to improve the efficiency of the resulting estimators. Also, if we can prove that the induced sampling design $p(m)$ belongs to the exponential family, then we can use a suitable approximation of the variance of the HT and Hajek estimators without involving second order probabilities of inclusion.

Acknowledgments

Thank you !!!

<http://predictive.wordpress.com>