Horvitz-Thompson Estimators in Center Sampling

G. Diana, P. Preo, C. Tommasi

2003.14

Dipartimento di Scienze Statistiche
Università degli Studi
Via C. Battisti 241-243
35121 Padova

Luglio 2003

Abstract

The problem of the estimation of the size of an immigrant population is very important in many countries as well as in Italy. *Center sampling* theory has been recently proposed and it is based on the hypotesis that all the individuals use to frequent centers of aggregation. Two approaches to center sampling are developed in literature. In the first people are drawn through simple random sampling without replacement from all the available centers. In the second one centers are drawn through the same scheme and all the individuals in the sampled centers are considered. In this paper the second approach is followed and generalized to a different sampling scheme according to the situation in which all centers has not the same importance: i.e. a big center could have a higher selection probability than a little center. So an unequal selection probabilities scheme is presented.

1 Introduction

Dealing with surveys of immigrant population (regular or irregular), a sampling theory called *center sampling* has been recently introduced in Italy by Blangiardo (1996). Since the population size N is unknown and labelling individuals is not possible, classical finite population sampling theory cannot be used. On the other hand, ethical reasons unable also the use of capture-recapture methods.

Two possible approaches for center sampling have been developed in literature. Both require the hypothesis that each immigrant frequents at least one center of aggregation, in order to have a representative sample.

In the first approach, suggested by Mecatti and Migliorati (2001), a number of immigrants are drawn from each center, according to a simple random sampling without replacement (SRSWOR). Thus an immigrant may be sampled more times, giving an overlapping problem. The authors propose an unbiased estimator for the mean of a characteristic of interest, Y.

In the other approach, developed by Pratesi and Rocco (2002), a number of centers are drawn (using a SRSWOR) and all the individuals of the sampled centers are considered. The aim is to estimate the population size N and the authors suggest an Horvitz-Thompson (H.T.) type estimator, $\hat{N}$. Again the same immigrants may be drawn more times.

In order to solve the overlapping problem, Mecatti and Migliorati (2001) split the population in subsets of units with the same profile (i.e. units who frequents exactly the same centers), while Pratesi and

Rocco (2002) consider subsets of units frequenting the same number of centers.
The first approach deals with all the possible configurations of profiles: it is easy when the number of centers is not large, but otherwise? Furthermore they propose an unbiased estimator for the mean of a characteristic Y and it seems difficult to provide an estimator for the population size (and so for the total of a characteristic Y).
On the contrary this is straightforward in the second approach. Furthermore it may be computationally easier. For these reasons the second approach is followed in this paper.

In Section 2 the estimator proposed by Pratesi and Rocco (2002) is generalized to estimate the total of a characteristic Y: for instance a quantity of interest could be the number of relatives each immigrant has in the native country. It is essential to know the possible immigrants' affluence to a country. When the population mean $\overline{Y}$ is of interest (e.g. the average number of clandestine women's abortions), an obvious estimator is $\hat{\overline{Y}} = \hat{Y}/\hat{N}$ where $\hat{Y}$ and $\hat{N}$ are the H.T. type estimators of Y and N.

In some settings it is unlikely each center has the same importance, thus it could be more convenient to draw centers according to different sampling probabilities. In Section 3 this more general sampling scheme is considered and some algebraic computations are given in the appendix.

A numerical illustration and some conclusions are given in Section 4.

2 Estimation of a population total

In order to estimate a population total a sample of centers is drawn without replacement and all the immigrants of the selected centers are considered. Then an H.T. type estimator is used. By collecting the terms corresponding to all the individuals with the same inclusion probability, the H.T. type estimator takes a specific form. In the same way the expressions for the variance and for the estimated variance of the considered estimators follow from the H.T. estimator general results.

First let us assume that the centers are drawn with the same selection probability, i.e. through a SRSWOR of n centers from the total number of G centers. Thus individuals frequenting exactly g ($g = 1, \ldots, G$) centers have the same probability of inclusion in the sample, whichever the frequented centers are. This assumption is suitable only if all the centers have the same importance.

2.1 Total of a characteristic

The H.T. estimator for the total of a characteristic Y in the population is usually defined as

$$\hat{Y} \;=\; \sum_{i=1}^{N} \frac{Y_i I_i}{\pi_i} \,,$$

where Y_i denotes the value of Y for the i-th individual ($i = 1, \dots, N$), I_i is an indicator variable equals 1 if the individual i is in the sample and 0 otherwise, π_i is the inclusion probability of the individual i. Collecting the terms corresponding to individuals who frequent the same number of centers, the estimator becomes

$$\hat{Y} \;=\; \sum_{g=1}^{G} \sum_{i=1}^{N_g} \frac{{}_gY_i \; {}_gI_i}{\pi_g} \,, \tag{1}$$

where ${}_gY_i$ denotes the value of Y for the i-th individual who frequents g centers, N_g is the number of individuals frequenting exactly g centers, ${}_gI_i$ is an indicator variable equals to 1 if the individual i, frequenting g centers, is in the sample and 0 otherwise, and π_g is the first order inclusion probability of any individual who frequents g centers.
This estimator is unbiased.
The variance and an unbiased variance estimator of $\hat{Y}$ are

$$\begin{aligned} V(\hat{Y}) \;&=\; \sum_{g=1}^{G} \sum_{c} \frac{\pi_{gg}^{(c)} - \pi_g^2}{\pi_g^2} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} {}_gY_i \; {}_gY_j \; I_{ijc} \\ &\quad + 2 \sum_{g=1}^{G} \sum_{g'>g}^{G} \sum_{c} \frac{\pi_{gg'}^{(c)} - \pi_g \pi_{g'}}{\pi_g \pi_{g'}} \sum_{i=1}^{N_g} \sum_{j=1}^{N_{g'}} {}_gY_i \; {}_{g'}Y_j \; I_{ijc} \end{aligned} \tag{2}$$

and

$$\begin{aligned} \hat{V}(\hat{Y}) \;&=\; \sum_{g=1}^{G} \sum_{c} \frac{\pi_{gg}^{(c)} - \pi_g^2}{\pi_g^2 \pi_{gg}^{(c)}} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} {}_gY_i \; {}_gI_i \; {}_gY_j \; {}_gI_j I_{ijc} \\ &\quad + 2 \sum_{g=1}^{G} \sum_{g'>g}^{G} \sum_{c} \frac{\pi_{gg'}^{(c)} - \pi_g \pi_{g'}}{\pi_g \pi_{g'} \pi_{gg'}^{(c)}} \sum_{i=1}^{N_g} \sum_{j=1}^{N_{g'}} {}_gY_i \; {}_gI_i \; {}_{g'}Y_j \; {}_{g'}I_j I_{ijc} \end{aligned} \tag{3}$$

where $\pi_{gg'}^{(c)}$ is the second order inclusion probability of couples of units who frequent g and g' centers, respectively and share c centers, while I_{ijc} is a population label equals 1 if units i and j are a couple of units of this kind and 0 otherwise.

2.2 Population size

The estimator for the population size proposed by Pratesi and Rocco (2002) follows from (1), setting ${}_gY_i = 1,\ i = 1, \ldots, N_g\ ,\ g = 1, \ldots, G$

$$\hat{N} = \sum_{g=1}^{G} \sum_{i=1}^{N_g} \frac{{}_gI_i}{\pi_g} .$$

From (2) and (3), the variance and an unbiased variance estimator of $\hat{N}$ are

$$V(\hat{N}) \quad = \sum_{g=1}^{G} \sum_{c} \frac{N_{gg}^{(c)} (\pi_{gg}^{(c)} - \pi_g^2)}{\pi_g^2} + 2 \sum_{g=1}^{G} \sum_{g'>g}^{G} \sum_{c} \frac{N_{gg'}^{(c)} (\pi_{gg'}^{(c)} - \pi_g \pi_{g'})}{\pi_g \pi_{g'}} \tag{4}$$

and

$$\hat{V}(\hat{N}) \quad = \sum_{g=1}^{G} \sum_{c} \frac{n_{gg}^{(c)} (\pi_{gg}^{(c)} - \pi_g^2)}{\pi_g^2 \pi_{gg}^{(c)}} + 2 \sum_{g=1}^{G} \sum_{g'>g}^{G} \sum_{c} \frac{n_{gg'}^{(c)} (\pi_{gg'}^{(c)} - \pi_g \pi_{g'})}{\pi_g \pi_{g'} \pi_{gg'}^{(c)}} \tag{5}$$

where $N_{gg'}^{(c)}$ and $n_{gg'}^{(c)}$ are the number of couples of individuals who frequent g and g' centers with c common centers, in the population and in the sample respectively.

2.3 Mean of a characteristic

Following Särndal et al. (1992, pages 176–182) an approximately unbiased estimator for the population mean $\overline{Y}$ is

$$\hat{\overline{Y}} = \frac{\hat{Y}}{\hat{N}} = \frac{\sum_{g=1}^{G} \sum_{i=1}^{N_g} \frac{{}_gY_i\ {}_gI_i}{\pi_g}}{\sum_{g=1}^{G} \sum_{i=1}^{N_g} \frac{{}_gI_i}{\pi_g}} .$$

with the following approximation for the variance

$$V(\hat{\overline{Y}}) \approx \frac{1}{N^2} [V(\hat{Y}) + \overline{Y}^2 V(\hat{N}) - 2\overline{Y} COV(\hat{Y}, \hat{N})] , \tag{6}$$

where $V(\hat{Y})$ and $V(\hat{N})$ are given in (2) and (4) respectively, while

$$COV(\hat{Y}, \hat{N}) = \sum_{g=1}^{G} \sum_{g'=1}^{G} \sum_{c} \frac{\pi_{gg'}^{(c)} - \pi_g \pi_{g'}}{\pi_g \pi_{g'}} \sum_{i=1}^{N_g} \sum_{j=1}^{N_{g'}} {}_gY_i I_{ijc} . \tag{7}$$

A consistent estimator for $V(\hat{\overline{Y}})$ is

$$\hat{V}(\hat{\overline{Y}}) \approx \frac{1}{\hat{N}^2} [\hat{V}(\hat{Y}) + \hat{\overline{Y}}^2 \hat{V}(\hat{N}) - 2\hat{\overline{Y}} \widehat{COV}(\hat{Y}, \hat{N})] , \tag{8}$$

where $\hat{V}(\hat{Y})$ and $\hat{V}(\hat{N})$ are given in (3) and (5) and

$$\widehat{COV}(\hat{Y},\hat{N}) = \sum_{g=1}^{G}\sum_{g'=1}^{G}\sum_{c} \frac{\pi^{(c)}_{gg'} - \pi_g\pi_{g'}}{\pi_g\pi_{g'}\pi^{(c)}_{gg'}} \sum_{i=1}^{N_g}\sum_{j=1}^{N_{g'}} {}_gY_i \; {}_gI_i \; {}_{g'}I_j \; I_{ijc} \ ,$$

which is an unbiased estimator of (7).

3 Unequal selection probabilities

Whenever centers are not assumed to be equally important, to draw samples of centers by unequal selection probabilities (USP) is more suitable. For instance, selection probabilities could be proportional to the centers size, i.e. the number of people can be hosted. In this setting only people who frequent exactly the same centers have equal inclusion probabilities, not people frequenting the same number of centers, as before. Thus, considering all the different combinations of centers, i.e. the profiles, and collecting the terms corresponding to units with the same profile, the specific expression for the H.T. type estimator is got. The different profiles are in number of $2^G - 1$.

3.1 Total of a characteristic

The H.T. type estimator for the total of a characteristic Y is now

$$\hat{Y}_u = \sum_{g=1}^{G}\sum_{i_g=1}^{\binom{G}{g}} \frac{Y_{i_g} I_{i_g}}{\pi_{i_g}} \ , \tag{9}$$

where Y_{i_g} is the partial total of Y in the subset of people with profile i_g, I_{i_g} is a random variable equals 1 if at least one center of the combination i_g (of g centers) is drawn and 0 otherwise and π_{i_g} is the first order inclusion probability for units with profile i_g.
The variance of $\hat{Y}_u$ is

$$V(\hat{Y}_u) = \sum_{g=1}^{G}\sum_{g'=1}^{G}\sum_{i_g=1}^{\binom{G}{g}}\sum_{i_{g'}=1}^{\binom{G}{g'}} \frac{Y_{i_g}Y_{i_{g'}}}{\pi_{i_g}\pi_{i_{g'}}}(\pi_{i_g,i_{g'}} - \pi_{i_g}\pi_{i_{g'}}) \ . \tag{10}$$

Since $\sum_{g=1}^{G}\sum_{i_g=1}^{\binom{G}{g}} I_{i_g} = 2^G - 2^{G-n}$ is a constant value, another expression for (10) is

$$V(\hat{Y}_u) = \frac{1}{2}\sum_{g=1}^{G}\sum_{g'=1}^{G}\sum_{i_g=1}^{\binom{G}{g}}\sum_{\substack{i_{g'}=1 \\ i_{g'}\neq i_g}}^{\binom{G}{g'}} \left(\pi_{i_g}\pi_{i_{g'}} - \pi_{i_g,i_{g'}}\right)\left(\frac{Y_{i_g}}{\pi_{i_g}} - \frac{Y_{i_{g'}}}{\pi_{i_{g'}}}\right)^2 , \tag{11}$$

see the Appendix for the proof.
An unbiased estimator for $V(\hat{Y}_u)$ is

$$\hat{V}(\hat{Y}_u) = \sum_{g=1}^{G}\sum_{g'=1}^{G}\sum_{i_g=1}^{\binom{G}{g}}\sum_{i_{g'}=1}^{\binom{G}{g}} \frac{Y_{i_g}Y_{i_{g'}}}{\pi_{i_g}\pi_{i_{g'}}}\frac{I_{i_g}I_{i_{g'}}}{\pi_{i_g,i_{g'}}}(\pi_{i_g,i_{g'}} - \pi_{i_g}\pi_{i_{g'}})\ . \tag{12}$$

where $\pi_{i_g,i_{g'}}$ is the second order inclusion probability for couples of units with profiles i_g and $i_{g'}$, respectively. The quantity $\pi_{i_g,i_{g'}}$ is got by summing the selection probabilities of the center samples with at least one center of the combination i_g and another one of the combination $i_{g'}$.

3.2 Population size

A population size estimator follows from (9) setting $Y_{i_g} = 1$, $i_g = 1, \ldots, \binom{G}{g}$,

$$\hat{N}_u = \sum_{g=1}^{G}\sum_{i_g=1}^{\binom{G}{g}} \frac{N_{i_g}I_{i_g}}{\pi_{i_g}} ,$$

where N_{i_g} is the number of subjects who frequent the specific combination i_g of g centers.
In the same way the variance and an unbiased estimator for the variance of $\hat{N}_u$ follow from (10) and (12),

$$V(\hat{N}_u) = \sum_{g=1}^{G}\sum_{g'=1}^{G}\sum_{i_g=1}^{\binom{G}{g}}\sum_{i_{g'}=1}^{\binom{G}{g'}} \frac{N_{i_g}N_{i_{g'}}}{\pi_{i_g}\pi_{i_{g'}}}(\pi_{i_g,i_{g'}} - \pi_{i_g}\pi_{i_{g'}}) \tag{13}$$

and

$$\hat{V}(\hat{N}_u) = \sum_{g=1}^{G}\sum_{g'=1}^{G}\sum_{i_g=1}^{\binom{G}{g}}\sum_{i_{g'}=1}^{\binom{G}{g'}} \frac{N_{i_g}N_{i_{g'}}}{\pi_{i_g}\pi_{i_{g'}}}\frac{I_{i_g}I_{i_{g'}}}{\pi_{i_g,i_{g'}}}(\pi_{i_g,i_{g'}} - \pi_{i_g}\pi_{i_{g'}})\ . \tag{14}$$

3.3 Mean of a characteristic

Following the same ideas given in Section 2.3 an estimator for the population mean $\overline{Y}$ is

$$\hat{\overline{Y}}_u = \frac{\hat{Y}_u}{\hat{N}_u} = \frac{\sum_{g=1}^{G}\sum_{i_g=1}^{\binom{G}{g}}\frac{Y_{i_g}I_{i_g}}{\pi_{i_g}}}{\sum_{g=1}^{G}\sum_{i_g=1}^{\binom{G}{g}}\frac{N_{i_g}I_{i_g}}{\pi_{i_g}}} \; .$$

This estimator is asymptotically unbiased. An approximation for its variance is

$$V(\hat{\overline{Y}}_u) \approx \frac{1}{N^2}[V(\hat{Y}_u) + \overline{Y}^2 V(\hat{N}_u) - 2\overline{Y}COV(\hat{Y}_u, \hat{N}_u)] \; , \qquad (15)$$

where $V(\hat{Y}_u)$ and $V(\hat{N}_u)$ are given in (10) and (13) respectively, while

$$COV(\hat{Y}_u, \hat{N}_u) = \sum_{g=1}^{G}\sum_{g'=1}^{G}\sum_{i_g=1}^{\binom{G}{g}}\sum_{i_{g'}=1}^{\binom{G}{g'}}\frac{N_{i_g}Y_{i_{g'}}}{\pi_{i_g}\pi_{i_{g'}}}(\pi_{i_g,i_{g'}} - \pi_{i_g}\pi_{i_{g'}}) \; . \qquad (16)$$

A consistent estimator for $V(\hat{\overline{Y}}_u)$ is

$$\hat{V}(\hat{\overline{Y}}_u) \approx \frac{1}{\hat{N}^2}[\hat{V}(\hat{Y}_u) + \hat{\overline{Y}}_u^2 \hat{V}(\hat{N}_u) - 2\hat{\overline{Y}}_u\widehat{COV}(\hat{Y}_u, \hat{N}_u)] \; . \qquad (17)$$

where $\hat{V}(\hat{Y}_u)$ and $\hat{V}(\hat{N}_u)$ are given in (12) and (14) respectively, while

$$\widehat{COV}(\hat{Y}_u, \hat{N}_u) = \sum_{g=1}^{G}\sum_{g'=1}^{G}\sum_{i_g=1}^{\binom{G}{g}}\sum_{i_{g'}=1}^{\binom{G}{g'}}\frac{N_{i_g}Y_{i_{g'}}}{\pi_{i_g}\pi_{i_{g'}}}\frac{I_{i_g}I_{i_{g'}}}{\pi_{i_g,i_{g'}}}(\pi_{i_g,i_{g'}} - \pi_{i_g}\pi_{i_{g'}}) \; ,$$

which is an unbiased estimator of (16).

4 Numerical illustration

In order to provide an illustration of the given theoretical results the numerical example given in Pratesi and Rocco (2002) is considered. A comparison with their results is also possible.
Specifically a population of 3100 individuals frequenting at least one of four centers A, B, C and D, is considered (see Table 1). The partial totals given in the last column of the Table are randomly generated, while data in the first two columns come from Pratesi and Rocco (2002). A sample of two centers is assumed to be drawn from this population.
In Table 2, sampling distributions of the estimators for N, Y and $\overline{Y}$

are computed under three different sampling schemes: the SRSWOR scheme and two unequal selection probabilities samplings without replacement (USPSWOR) schemes.

The first considered USPSWOR scheme is the Midzuno sampling design with the first unit drawn by probability proportional to size (PPS) of centers. While, in the other USPSWOR scheme all centers are drawn by PPS.

For each sampling scheme and for each estimator, the exact or the approximated variances and the exact bias are provided.

The exact variances for the estimator of N and Y are computed by (4), (2), (13) and (10), while the exact variance and the bias for the estimator of $\overline{Y}$ are computed by the sampling distribution. The approximated variances (AV) for the estimator of $\overline{Y}$ are given by the terms in the right-hand side of (6) and (15).

The results given in Table 2 show that choosing an USPSWOR with PPS of centers instead of a SRSWOR, there is a gain in terms of efficiency for the estimators of the parameters N, Y and $\overline{Y}$. However, from other more general simulations it seems there is a gain for the estimators of N and of Y, but not alwaysfor the estimator of $\overline{Y}$.

So when the interest is on the total of a characteristic, Y, it may

Table 1: Distribution of immigrants in the centers

PROFILES (i_g)					
A	B	C	D	FREQUENCIES (N_{i_g})	PARTIAL TOTALS (Y_{i_g})
1	0	0	0	200	3240
0	1	0	0	300	4550
0	0	1	0	400	6119
0	0	0	1	300	4957
0	1	1	0	200	4960
0	0	1	1	300	4476
1	0	1	0	300	4954
0	1	0	1	400	6607
1	1	0	0	100	3111
1	0	0	1	100	2938
1	1	1	0	200	4487
1	0	1	1	50	2525
1	1	0	1	100	2310
0	1	1	1	100	2742
1	1	1	1	50	2340
				3100	**60316**

Table 2: Estimators for N, Y, $\bar{Y}$

SRSWOR

Sampled Centers	$\hat{N}$	$\hat{Y}$	$\hat{\bar{Y}} = \hat{Y}/\hat{N}$	Sample probability
AB	2820	57068,0	20,236	0,167
AC	2900	57648,8	19,878	0,167
AD	2940	57301,2	19,490	0,167
BC	3460	64671,6	18,691	0,167
BD	3020	59928,4	19,843	0,167
CD	3460	65278,0	18,866	0,167
$V(\cdot)$	68266,7	11749234,7	0,314	
$AV(\cdot)$			0,328	
$B(\cdot)$			0,044	

USPSWOR: Midzuno scheme

Sampled Centers	$\hat{N}_u$	$\hat{Y}_u$	$\hat{\bar{Y}}_u = \hat{Y}_u/\hat{N}_u$	Sample probability
AB	2839,92	57420,99	20,219	0.153
AC	2886,77	57498,91	19,918	0.162
AD	2965,92	57789,09	19,484	0.150
BC	3403,21	63792,40	18,745	0.183
BD	3000,74	59652,92	19,879	0.171
CD	3410,74	64513,49	18,915	0.180
$V(\cdot)$	56357,88	8992931,12	0,302	
$AV(\cdot)$			0,307	
$B(\cdot)$			0,040	

USPSWOR: probabilities proportional to size of centers

Sampled Centers	$\hat{N}_u$	$\hat{Y}_u$	$\hat{\bar{Y}}_u = \hat{Y}_u/\hat{N}_u$	Sample probability
AB	2875,49	58055,94	20,190	0.135
AC	2889,47	57639,24	19,948	0.152
AD	3008,26	58575,80	19,472	0.129
BC	3330,41	62672,01	18,818	0.208
BD	2968,98	59204,09	19,941	0.177
CD	3347,48	63537,86	18,981	0.199
$V(\cdot)$	40881,54	5607841,00	0,282	
$AV(\cdot)$			0,278	
$B(\cdot)$			0,033	

be convenient (in terms of efficiency) to choose a USPSWOR design instead of a SRSWOR scheme.

On the contrary the SRSWOR seems a better choice if the quantity of interest is the mean of a characteristic, $\overline{Y}$: using USPSWOR there is a higher complexity and not always a gain in terms of efficiency.

APPENDIX

In this Appendix the equivalence between the expression (10) and the expression (11) for $V(\hat{\overline{Y}}_u)$ is proved.

Since the number of profiles intersecting the selected center sample is a constant value,

$$\nu(s) = \nu = \sum_{g=1}^{G} \sum_{i_g=1}^{\binom{G}{g}} I_{i_g} = 2^G - 2^{G-n} \ ,$$

where n is the number of sampled centers, it follows (Cicchitelli et al., 1997) that

$$\sum_{g=1}^{G} \sum_{i_g=1}^{\binom{G}{g}} E(I_{i_g}) = \sum_{g=1}^{G} \sum_{i_g=1}^{\binom{G}{g}} \pi_{i_g} = \nu \ ,$$

and so

$$\sum_{g'=1}^{G} \sum_{\substack{i_{g'}=1 \\ i_{g'} \neq i_g}}^{\binom{G}{g'}} \pi_{i_{g'}} = \nu - \pi_{i_g} \ .$$

In a similar way,

$$\begin{aligned}
\sum_{g'=1}^{G} \sum_{\substack{i_{g'}=1 \\ i_{g'} \neq i_g}}^{\binom{G}{g'}} \pi_{i_g, i_{g'}} &= \sum_{g'=1}^{G} \sum_{\substack{i_{g'}=1 \\ i_{g'} \neq i_g}}^{\binom{G}{g'}} E(I_{i_g} \cdot I_{i_{g'}}) = \sum_{g'=1}^{G} \sum_{\substack{i_{g'}=1 \\ i_{g'} \neq i_g}}^{\binom{G}{g'}} \sum_{s} \Big(I_{i_g} \cdot I_{i_{g'}} p(s) \Big) \\
&= \sum_{s} I_{i_g} \sum_{g'=1}^{G} \sum_{\substack{i_{g'}=1 \\ i_{g'} \neq i_g}}^{\binom{G}{g'}} I_{i_{g'}} p(s) = \sum_{s} I_{i_g} \sum_{g'=1}^{G} \sum_{i_{g'}=1}^{\binom{G}{g'}} \Big(I_{i_{g'}} - I_{i_g} \Big) p(s) \\
&= \sum_{s} I_{i_g} (\nu - I_{i_g}) p(s) = \nu \sum_{s} I_{i_g} p(s) - \sum_{s} I_{i_g}^2 p(s) \\
&= \nu E(I_{i_g}) - E(I_{i_g}) = (\nu - 1) E(I_{i_g}) = (\nu - 1) \pi_{ig} \ .
\end{aligned}$$

Developing (11),

$$\begin{aligned} V(\hat{Y}_u) &= \frac{1}{2}\sum_{g=1}^{G}\sum_{g'=1}^{G}\sum_{i_g=1}^{\binom{G}{g}}\sum_{\substack{i_{g'}=1\\ i_{g'}\neq i_g}}^{\binom{G}{g'}}\left(\frac{Y_{i_g}^2}{\pi_{i_g}^2}+\frac{Y_{i_{g'}}^2}{\pi_{i_{g'}}^2}\right)(\pi_{i_g}\pi_{i_{g'}}-\pi_{i_g,i_{g'}}) \\ &+\sum_{g=1}^{G}\sum_{g'=1}^{G}\sum_{i_g=1}^{\binom{G}{g}}\sum_{\substack{i_{g'}=1\\ i_{g'}\neq i_g}}^{\binom{G}{g'}}\frac{Y_{i_g}Y_{i_{g'}}}{\pi_{i_g}\pi_{i_{g'}}}(\pi_{i_g,i_{g'}}-\pi_{i_g}\pi_{i_{g'}}) \end{aligned}$$

is equivalent to expression (10) since

$$\frac{1}{2}\sum_{g=1}^{G}\sum_{g'=1}^{G}\sum_{i_g=1}^{\binom{G}{g}}\sum_{\substack{i_{g'}=1\\ i_{g'}\neq i_g}}^{\binom{G}{g'}}\left(\frac{Y_{i_g}^2}{\pi_{i_g}^2}+\frac{Y_{i_{g'}}^2}{\pi_{i_{g'}}^2}\right)(\pi_{i_g}\pi_{i_{g'}}-\pi_{i_g,i_{g'}})=\sum_{g=1}^{G}\sum_{i_g=1}^{\binom{G}{g}}\frac{Y_{i_g}^2}{\pi_{i_g}^2}(\pi_{i_g}-\pi_{i_g}^2)\,.$$

In fact

$$\begin{aligned} &\frac{1}{2}\sum_{g=1}^{G}\sum_{g'=1}^{G}\sum_{i_g=1}^{\binom{G}{g}}\sum_{\substack{i_{g'}=1\\ i_{g'}\neq i_g}}^{\binom{G}{g'}}\left(\frac{Y_{i_g}^2}{\pi_{i_g}^2}+\frac{Y_{i_{g'}}^2}{\pi_{i_{g'}}^2}\right)(\pi_{i_g}\pi_{i_{g'}}-\pi_{i_g,i_{g'}})= \\ &=\sum_{g=1}^{G}\sum_{g'=1}^{G}\sum_{i_g=1}^{\binom{G}{g}}\sum_{\substack{i_{g'}=1\\ i_{g'}\neq i_g}}^{\binom{G}{g'}}\frac{Y_{i_g}^2}{\pi_{i_g}^2}(\pi_{i_g}\pi_{i_{g'}}-\pi_{i_g,i_{g'}})= \\ &=\sum_{g=1}^{G}\sum_{g'=1}^{G}\sum_{i_g=1}^{\binom{G}{g}}\frac{Y_{i_g}^2}{\pi_{i_g}^2}\left(\pi_{i_g}\sum_{\substack{i_{g'}=1\\ i_{g'}\neq i_g}}^{\binom{G}{g'}}\pi_{i_{g'}}-\sum_{\substack{i_{g'}=1\\ i_{g'}\neq i_g}}^{\binom{G}{g'}}\pi_{i_g,i_{g'}}\right) \\ &=\sum_{g=1}^{G}\sum_{i_g=1}^{G}\frac{Y_{i_g}^2}{\pi_{i_g}^2}\left(\pi_{i_g}\sum_{g'=1}^{G}\sum_{\substack{i_{g'}=1\\ i_{g'}\neq i_g}}^{\binom{G}{g'}}\pi_{i_{g'}}-\sum_{g'=1}^{G}\sum_{\substack{i_{g'}=1\\ i_{g'}\neq i_g}}^{\binom{G}{g'}}\pi_{i_g,i_{g'}}\right) \\ &=\sum_{g=1}^{G}\sum_{i_g=1}^{\binom{G}{g}}\frac{Y_{i_g}^2}{\pi_{i_g}^2}\left(\pi_{i_g}(\nu-\pi_{i_g})-(\nu-1)\pi_{i_g}\right) \\ &=\sum_{g=1}^{G}\sum_{i_g=1}^{\binom{G}{g}}\frac{Y_{i_g}^2}{\pi_{i_g}^2}\left(\pi_{i_g}-\pi_{i_g}^2\right)\ . \end{aligned}$$

References

Blangiardo, G. C. 1996. Il campionamento per centri o ambienti di aggregazione nelle indagini sulla presenza straniera. *Studi in onore di Giampiero Landenna* Giuffré, Milano.

Cicchitelli, G., A. Herzel, and G. E. Montanari. 1997. *Il campionamento statistico*. Il Mulino, Bologna.

Mecatti, F., and S. Migliorati. 2001. Center sampling: theory and estimation. Technical Report 01-06, Dipartimento di Statistica, Università degli Studi di Milano-Bicocca.

Pratesi, M., and E. Rocco. 2002. Centre sampling for estimating elusive population size. *Working Paper 2002/15*, Dipartimento di Statistica "Giuseppe Parenti", Università degli Studi di Firenze.

Särndal, C. E., B. Swensson, and J. Wretman. 1992. *Model assisted survey sampling*. Springer–Verlag, New York.