Inference with Inverse Sampling

L. Greco
Dipartimento di Metodi Quantitativi, Università degli Studi di Siena
greco2@unisi.it

S. Naddeo
Dipartimento di Metodi Quantitativi, Università degli Studi di Siena
naddeo@unisi.it

Abstract. With inverse sampling it is possible to obtain efficient parameter estimators especially when the population of interest is subdivided into two groups, one of which contains only a few units. In this design the selection procedure continues until a prefixed number of units of the rare group is observed in the sample. However, since the subgroup to which a unit belongs is not known until the unit is selected, the sample size is a random variable. In this paper we derive the asymptotic distributions of total estimators of both subgroups as well as of the whole population through the Doeblin and Anscombe theorem.

Keywords. Sequential sampling, rare populations, asymptotic distribution, approximate confidence intervals.

1. Inverse sampling

Frequently in a sample survey estimates are required not just for the entire population but also for a rare subgroup. However, it is not known *a priori* to which subgroup the population units belong. In other situations the population may be considered as subdivided into two subgroups, one of which is rare as, for example, if only a few units display a value of the variable of interest which is highly different from zero, while all the other units show a value equal to or near zero. In these situations inverse sampling is more efficient than classical fixed sample-size designs in parameter estimation, since in this design the selection of units continues until a prefixed number of units with the characteristic of interest is sampled.

Let us consider a population of M units divided into two groups of M_1 and M_2 units respectively, with $M_1+M_2 = M$, and let y_{1i} ($i = 1, 2, \ldots, M_1$) and y_{2i} ($i = 1, 2, \ldots, M_2$) be the values of the variable of interest Y in the first and second group. Moreover, let T_1 and T_2 be the totals of the two groups

and $T=T_1+T_2$ the population total, p_{1i} the selection probability of the *i*-th unit of the first group and p_{2i} the selection probability of the *i*-th unit of the second group. Finally, let $P=\sum_{i=1}^{M_1} p_{1i}$ be the selection probability of the first group and $1-P=\sum_{i=1}^{M_2} p_{2i}$ the selection probability of the second group.

In inverse sampling designs the selection of units continues until k units of the rare group are observed in the sample, so that the sample size N_k is a random variable which can assume the values k, k+1, … . In a previous paper (Greco and Naddeo, 2004) it was shown that, conditionally to $N_k=n$, the selection procedure with inverse sampling is exactly like the selection procedure with stratified sampling, where the two strata correspond to the two subgroups and k and *n*–k units are respectively selected from the first and second stratum. Moreover, it was shown that the sample results in the two groups are independent. For a given *n* the selection probability of the *i*-th unit in the first group is $\frac{p_{1i}}{P}$, while $\frac{p_{2i}}{1-P}$ is the selection probability of the *i*-th unit in the second group.

Let us consider sampling with replacement and let Y_{ji} be the random variable "value of Y" on the *i*-th sample unit from the *j*-th group, so that the estimators of the two subgroup totals may be written as

$$\hat{T}_1=\frac{P}{k}\sum_{i=1}^{k}\frac{Y_{1i}}{p_{1i}}, \tag{1}$$

$$\hat{T}_2=\frac{1-P}{N_k-k}\sum_{i=1}^{N_k-k}\frac{Y_{2i}}{p_{2i}}. \tag{2}$$

Usually, the value of P is not known, but an unbiased estimator is

$$\hat{P}=\frac{k-1}{N_k-1}, \tag{3}$$

as can be derived from the negative binomial distribution of N_k. Hence, if we let

$$W_{ji}=\frac{Y_{ji}}{p_{ji}},$$

the unbiased estimators of the totals are

$$\widetilde{T}_1 = \hat{P}\overline{W}_1, \tag{4}$$

$$\widetilde{T}_2 = \left(1-\hat{P}\right)\overline{W}_2, \tag{5}$$

where the means of $\overline{W}_1$ and $\overline{W}_2$ are respectively

$$\mathrm{E}\left(\overline{W}_1\right) = \mathrm{W}_1 = \frac{\mathrm{T}_1}{\mathrm{P}}$$

and

$$\mathrm{E}\left(\overline{W}_2\right) = \mathrm{W}_2 = \frac{\mathrm{T}_2}{1-\mathrm{P}}.$$

The estimator variances are respectively

$$\mathrm{V}\left(\widetilde{T}_1\right) = \mathrm{W}_1^2 \mathrm{V}_{N_\mathrm{k}}\left(\hat{P}\right) + \frac{\sigma_{1w}^2}{\mathrm{k}} \mathrm{E}_{N_\mathrm{k}}\left(\hat{P}^2\right),$$

$$\mathrm{V}\left(\widetilde{T}_2\right) = \mathrm{W}_2^2 \mathrm{V}_{N_\mathrm{k}}\left(\hat{P}\right) + \frac{\sigma_{2w}^2}{\mathrm{k}-1} \mathrm{E}_{N_\mathrm{k}}\left[\hat{P}\left(1-\hat{P}\right)\right],$$

where

$$\sigma_{jw}^2 = \sum_{i=1}^{\mathrm{M}_j}\left(\frac{\mathrm{y}_{ji}}{\mathrm{p}_{ji}} - \mathrm{W}_j\right)^2 \mathrm{p}_{ji} \qquad j=1,2,$$

while the variance of the general total estimator $\widetilde{T}$ is given by

$$\mathrm{V}\left(\widetilde{T}\right) = \left(\mathrm{W}_1 - \mathrm{W}_2\right)^2 \mathrm{V}_{N_\mathrm{k}}\left(\hat{P}\right) + \frac{\sigma_{1w}^2}{\mathrm{k}} \mathrm{E}_{N_\mathrm{k}}\left(\hat{P}^2\right) + \frac{\sigma_{2w}^2}{\mathrm{k}-1} \mathrm{E}_{N_\mathrm{k}}\left[\hat{P}\left(1-\hat{P}\right)\right].$$

In Greco and Naddeo (2004) the following expressions of the variances estimators are obtained

$$\hat{\mathrm{V}}\left(\widetilde{T}_1\right)=\overline{W}_1^2\,\frac{\hat{P}\left(1-\hat{P}\right)}{N_{\mathrm{k}}-2}+\frac{\hat{S}_{1w}^2}{\mathrm{k}}\hat{P}_{\mathrm{q}}, \tag{6}$$

$$\hat{\mathrm{V}}\left(\widetilde{T}_2\right)=\overline{W}_2^2\,\frac{\hat{P}\left(1-\hat{P}\right)}{N_{\mathrm{k}}-2}+\frac{\hat{S}_{2w}^2}{\mathrm{k}-1}\left(\hat{P}-\frac{\mathrm{k}-1}{\mathrm{k}-2}\hat{P}_q\right), \tag{7}$$

$$\hat{\mathrm{V}}\left(\widetilde{T}\right)=\left(\overline{W}_1-\overline{W}_2\right)^2\,\frac{\hat{P}\left(1-\hat{P}\right)}{N_{\mathrm{k}}-2}+\frac{\hat{S}_{1w}^2}{\mathrm{k}}\hat{P}_q+\frac{\hat{S}_{2w}^2}{\mathrm{k}-1}\left(\hat{P}-\frac{\mathrm{k}-1}{\mathrm{k}-2}\hat{P}_q\right), \tag{8}$$

where $\hat{S}_{jw}^2$ is the unbiased sample variance of the W_{ji}s in the j-th group,

$$\hat{P}_q=\frac{(\mathrm{k}-1)(\mathrm{k}-2)}{(N_{\mathrm{k}}-1)(N_{\mathrm{k}}-2)} \tag{9}$$

is an unbiased estimator of P^2 and finally

$$\hat{\mathrm{V}}\left(\hat{P}\right)=\hat{P}^2-\frac{(\mathrm{k}-1)(\mathrm{k}-2)}{(N_{\mathrm{k}}-1)(N_{\mathrm{k}}-2)}=\frac{\hat{P}\left(1-\hat{P}\right)}{N_{\mathrm{k}}-2}$$

is an unbiased estimator of the variance of $\hat{\mathrm{P}}$.

If the selection probabilities are equal for each population unit, the unbiased estimator of the population total under inverse sampling assumes the form

$$\widetilde{T}=\mathrm{M}\left[\hat{P}\overline{Y}_1+\left(1-\hat{P}\right)\overline{Y}_2\right], \tag{10}$$

where $\overline{Y}_j$ is the sample mean in the j-th subgroup (j=1,2). The variance of $\widetilde{T}$ is

$$\mathrm{V}\left(\widetilde{T}\right)=\mathrm{M}^2\left\{\left(\mu_1-\mu_2\right)^2\mathrm{V}_{N_{\mathrm{k}}}\left(\hat{P}\right)+\frac{\sigma_1^2}{\mathrm{k}}\mathrm{E}_{N_{\mathrm{k}}}\left(\hat{P}^2\right)+\frac{\sigma_2^2}{\mathrm{k}-1}\mathrm{E}_{N_{\mathrm{k}}}\left[\hat{P}\left(1-\hat{P}\right)\right]\right\} \tag{11}$$

where μ_j and σ_j^2 are the mean and the variance of the j-th subpopulation. Expressions (10) and (11) are equivalent to those given by Christman and Lan (2001). Moreover, Salehi and Seber (2001, 2004) obtained expression (10) using Murthy's method.

The unbiased estimator of (11) assumes the form

$$\hat{\mathrm{V}}\left(\widetilde{T}\right)=\mathrm{M}^2\left\{\left(\overline{Y}_1-\overline{Y}_2\right)^2\frac{\hat{P}\left(1-\hat{P}\right)}{N_\mathrm{k}-2}+\frac{\hat{S}_1^2}{\mathrm{k}}\hat{P}_q+\frac{\hat{S}_2^2}{\mathrm{k}-1}\left[\hat{P}-\frac{\mathrm{k}-1}{\mathrm{k}-2}\hat{P}_q\right]\right\},$$

where $\hat{S}_j^2$ is the unbiased sample variance in the j-th subgroup (j=1,2).

The estimators of the two subgroup totals and of their variances may be easily obtained from the previous expressions.

3. The asymptotic distributions

In order to obtain the asymptotic distributions of $\widetilde{T}_1$, $\widetilde{T}_2$ and $\widetilde{T}$ for unequal selection probabilities, it is worth noting that the random variable N_k, which has a negative binomial distribution of parameters k and P, is the sum of k independent geometric variables of parameter P. Thus, from the central limit theorem, it turns out that

$$\sqrt{\mathrm{k}}\left(\frac{N_\mathrm{k}}{\mathrm{k}}-\frac{1}{\mathrm{P}}\right)\xrightarrow{d}\mathrm{N}\left(0,\frac{1-\mathrm{P}}{\mathrm{P}^2}\right). \tag{12}$$

Moreover, for any δ>0,

$$\lim_{\mathrm{k}\to\infty}\mathrm{Prob}\left(\left|N_\mathrm{k}-\mathrm{m}_\mathrm{k}\right|\geq\delta\,\mathrm{m}_\mathrm{k}\right)=0, \tag{13}$$

where m_k is the integer value of k/P.

Finally, from (12), by using the delta method, it is easy to obtain the limit distribution of k/N_k and also of $\hat{P}=\dfrac{\mathrm{k}-1}{N_\mathrm{k}-1}$ which is

$$\sqrt{\mathrm{k}}\left(\hat{P}-\mathrm{P}\right)\xrightarrow{d}\mathrm{N}\left[0,\mathrm{P}^2\left(1-\mathrm{P}\right)\right]. \tag{14}$$

As regards estimator (4), observe that

$$\overline{W}_1 = \frac{1}{\mathrm{k}} \sum_{i=1}^{\mathrm{k}} W_{1i} ,$$

where the W_{1i}s are i.i.d., has the following asymptotic distribution

$$\sqrt{\mathrm{k}}\left(\overline{W}_1 - \mathrm{W}_1\right) \xrightarrow{d} \mathrm{N}\left[0, \sigma_{1w}^2\right]. \tag{15}$$

Since $\hat{P}$ and $\overline{W}_1$ are independent random variables, from (14) and (15) the asymptotic distribution of the estimator $\widetilde{T}_1$ turns out to be

$$\sqrt{\mathrm{k}}\left(\widetilde{T}_1 - \mathrm{T}_1\right) \xrightarrow{d} \mathrm{N}\left[0, \mathrm{W}_1^2 \mathrm{P}^2 (1 - \mathrm{P}) + \sigma_{1w}^2 \mathrm{P}^2\right]. \tag{16}$$

As regards estimator (5), observe that

$$\overline{W}_2 = \frac{1}{N_{\mathrm{k}} - \mathrm{k}} \sum_{i=1}^{N_{\mathrm{k}} - \mathrm{k}} W_{2i} ,$$

where the W_{2i}s are i.i.d.

The asymptotic distribution of $\overline{W}_2$ may be obtained by observing that

$$\begin{aligned} \sqrt{\mathrm{k}}\left(\overline{W}_2 - \mathrm{W}_2\right) &= \frac{\sqrt{\mathrm{k}}}{N_{\mathrm{k}} - \mathrm{k}} \left[\sum_{i=1}^{\mathrm{m_k} - \mathrm{k}} \left(W_{2i} - \mathrm{W}_2\right) + \sum_{i=\mathrm{m_k} - \mathrm{k} + 1}^{N_{\mathrm{k}} - \mathrm{k}} \left(W_{2i} - \mathrm{W}_2\right) \right] = \\ &= \frac{\sqrt{\mathrm{k}} \sqrt{\mathrm{m_k} - \mathrm{k}}}{N_{\mathrm{k}} - \mathrm{k}} \left[\frac{\sum_{i=1}^{\mathrm{m_k} - \mathrm{k}} \left(W_{2i} - \mathrm{W}_2\right)}{\sqrt{\mathrm{m_k} - \mathrm{k}}} + \frac{\sum_{i=\mathrm{m_k} - \mathrm{k} + 1}^{N_{\mathrm{k}} - \mathrm{k}} \left(W_{2i} - \mathrm{W}_2\right)}{\sqrt{\mathrm{m_k} - \mathrm{k}}} \right]. \end{aligned} \tag{17}$$

For $\mathrm{k} \to \infty$, from (13) and the Doeblin and Anscombe theorem the second term inside the last square bracket tends in probability to zero, so that

$$\frac{\sum_{i=1}^{m_k-k}(W_{2i}-W_2)}{\sqrt{m_k-k}} \xrightarrow{d} N\left[0,\sigma_{2w}^2\right]. \tag{18}$$

Moreover, from (12), for $k\to\infty$ it follows that $\frac{\sqrt{k}\sqrt{m_k-k}}{N_k-k}$ converges in probability to $\sqrt{\frac{P}{1-P}}$, so that the asymptotic distribution of $\overline{W}_2$ results in

$$\sqrt{k}\left(\overline{W}_2-W_2\right)\xrightarrow{d} N\left[0,\frac{P}{1-P}\sigma_{2w}^2\right]. \tag{19}$$

Since $\hat{P}$ and (19) are independent random variables, the asymptotic distribution of the estimator $\widetilde{T}_2$ is given by

$$\sqrt{k}\left(\widetilde{T}_2-T_2\right)\xrightarrow{d} N\left[0,W_2^2P^2(1-P)+\sigma_{2w}^2P(1-P)\right]. \tag{20}$$

Finally, from (14), (15) and (19) the asymptotic distribution of the population total estimator $\widetilde{T}$ proves to be

$$\sqrt{k}\left(\widetilde{T}-T\right)\xrightarrow{d} N\left[0,(W_1-W_2)^2P^2(1-P)+\sigma_{1w}^2P^2+\sigma_{2w}^2P(1-P)\right]. \tag{21}$$

Since consistent estimates of the variances of the three estimators are at our disposal, on the basis of (16), (20) and (21) the approximate confidence intervals of $\widetilde{T}_1$, $\widetilde{T}_2$ and $\widetilde{T}$ are respectively

$$\hat{p}\overline{w}_1 \mp u\left[\overline{w}_1^2\frac{\hat{p}(1-\hat{p})}{n-2}+\frac{\hat{s}_{1w}^2}{k}\hat{p}_q\right],$$

$$(1-\hat{p})\overline{w}_2 \mp u\left[\overline{w}_2^2\frac{\hat{p}(1-\hat{p})}{n-2}+\frac{\hat{s}_{2w}^2}{k-1}\left(\hat{p}-\frac{k-1}{k-2}\hat{p}_q\right)\right],$$

$$\hat{p}\overline{w}_1+(1-\hat{p})\overline{w}_2 \mp u\left[(\overline{w}_1-\overline{w}_2)^2\frac{\hat{p}(1-\hat{p})}{n-2}+\frac{\hat{s}_{1w}^2}{k}\hat{p}_q+\frac{\hat{s}_{2w}^2}{k-1}\left(\hat{p}-\frac{k-1}{k-2}\hat{p}_q\right)\right],$$

where n is the observed sample size, lower letters denote estimates and u is the upper $\alpha/2$ point of the normal distribution.

Obviously similar results hold in inverse sampling with equal selection probabilities.

4. A Small Simulation Study

In order to check the performance of the previous asymptotic distributions for moderate sample sizes, a simulation study was performed on 3 artificial populations of 1.000 units subdivided into 2 subgroups with parameter P equal to 0.1, 0.2 and 0.5. The variable Y has a uniform distribution in both subgroups, with different means and variances. From each population 5,000 samples with k=10(10)50 were selected by using equal selection probabilities.

The results show that the distributions of the estimators $\widetilde{T}_1$ and $\widetilde{T}$ are positively skewed, while the distribution of $\widetilde{T}_2$ is negatively skewed, with an obvious decreasing asymmetry for increasing sample size. Accordingly, the resulting coverage of the confidence intervals is lower than the nominal one. The following tables show the results obtained from the 3 populations for a probability level $1-\alpha = 0.95$, where the column "Total" contains the results obtained for $\widetilde{T}$, while "Group A" and "Group B" report results for $\widetilde{T}_1$ and $\widetilde{T}_2$ respectively. Moreover, the column "under" contains the proportions of confidence intervals with an upper bound which is lower than the true parameter, while the column "over" shows the proportions of intervals with a lower bound which is higher than the true parameter.

Table 4.1
Coverages of the approximate confidence intervals. P=0.1

k	Total			Group A			Group B		
	coverage	under	over	coverage	under	over	coverage	under	over
10	.9472	.0378	.0150	.9138	.0824	.0038	.9386	.0284	.0330
20	.9482	.0334	.0184	.9318	.0606	.0076	.9454	.0236	.0310
30	.9432	.0366	.0202	.9354	.0556	.0090	.9520	.0210	.0270
40	.9482	.0314	.0204	.9450	.0454	.0096	.9500	.0198	.0302
50	.9556	.0258	.0186	.9412	.0456	.0132	.9490	.0230	.0280

Table 4.2
Coverages of the approximate confidence intervals. P=0.2

k	Total			Group A			Group B		
	coverage	under	over	coverage	under	over	coverage	under	over
10	.9364	.0500	.0136	.9116	.0842	.0042	.9402	.0278	.0320
20	.9340	.0412	.0158	.9328	.0586	.0086	.9442	.0254	.0304
30	.9472	.0388	.0140	.9388	.0530	.0082	.9486	.0246	.0268
40	.9496	.0348	.0156	.9442	.0458	.0100	.9496	.0232	.0272
50	.9502	.0320	.0178	.9442	.0430	.0128	.9470	.0236	.0294

Table 4.3
Coverages of the approximate confidence intervals. P=0.5

k	Total			Group A			Group B		
	coverage	under	over	coverage	under	over	coverage	under	over
10	.9305	.0561	.0134	.9174	.0696	.0130	.9278	.0480	.0242
20	.9392	.0434	.0174	.9370	.0492	.0138	.9308	.0450	.0242
30	.9440	.0388	.0172	.9368	.0470	.0162	.9340	.0420	.0240
40	.9446	.0352	.0202	.9420	.0416	.0164	.9386	.0384	.0230
50	.9454	.0332	.0214	.9472	.0370	.0158	.9400	.0372	.0228

As regards group A, the results are similar for each value of P, with a low coverage and a rather high underestimate for moderate k values, both due to the positive skewness of the distribution of $\widetilde{T}_1$. If the value of P is low, the average sample size of units from group B is high, so that the asymptotic distribution of $\widetilde{T}_2$ is nearly symmetric and the total coverage is quite satisfactory even if the overestimates are a bit higher than the underestimates for moderate P, because of the slightly negative skewness of the estimator. Instead, when the population is equally subdivided into the 2 subgroups, $\widetilde{T}_1$ and $\widetilde{T}_2$ have quite similar performance. Obviously, the results obtained from the two subgroups reflect on the general total estimate, so that the underestimate is always higher than the overestimate and the coverage gets slightly better as k increases.

References

Christman, M.C. and Lan, F. (2001). Inverse Adaptive Cluster Sampling. *Biometrics* **57**, 1096-1105.

Greco, L. and Naddeo, S. (2004). Inverse sampling with unequal selection probabilities. Atti del Convegno "Disegni campionari per le indagini ambientali, economiche e sociali: aspetti teorici e pratici", Siena.

Salehi M. M. and Seber G.A.F. (2001). A new proof of Murthy's estimator which applies to sequential sampling. *Australian and New Zealand Journal of Statistics* **43**, 281-286.

Salehi M. M. and Seber G.A.F. (2004). A General Inverse Sampling Scheme and its Application to Adaptive Cluster Sampling, *Australian and New Zealand Journal of Statistics* **46**, 483-494.