Rivista di Statistica Applicata, Vol. 17 N. 4 (Dicembre) 1984

Ppswor sampling from finite populations: a regression approach to estimation

Giovanni Andreatta, Dipartimento di Statistica, Università di Padova

1. Introduction

The problem of estimation in sampling with probabilities proportional to size and without replacement (PPSWOR Sampling) has received increasing attention and a number of estimators have been proposed. Due to the lack of completeness of any non trivial sufficient statistic no unique optimal estimator exists (Cassel *et al.*, 1977). The Horvitz-Thompson (1952) and the Murthy (1957) estimators have proven to be particularly « good » both in terms of theoretical properties and empirical results (Cassel *et al.*, 1977, Rao and Bayless, 1969; Bayless and Rao, 1970).

However the computation required by these estimators rapidly becomes combersome as n, the number of selected units, increases: in practice these estimators have usually been used when $n \leq 2$ and in a few instances when $2 < n \leq 4$.

Thus a variety of approximations have been suggested (e.g. Rosen's approximations; Rosen, 1972), and computationally simpler estimators have often been used at the expenses of some efficiency: for instance the Des-Raj (1956) estimator is a simple unbiased estimator but is less efficient than the Murthy estimator (which also is unbiased), therefore it is not even admissible (Cassel *et al.*, 1977).

On the other hand whenever the choice of wich sampling design to implement is available to the statistician or decision maker, a suitable choice of the sampling design can allow the adoption of estimation procedures which are both efficient and simple: see for instance the Rao-Hartley-Cochran (1962) procedure.

However, in many cases the choice of the sampling design is not available, as for instance in the problem of forecasting energy recoverable resources from a geological play when adopting the model proposed by Kaufman (Barouch and Kaufman, 1976; 1977), see also section 4.

In the present paper the problem of estimation in ppswor sampling from a finite population is considered by taking a regression approach.

The regression approach has been used in problems of estimation in ppswor sampling only when an underlying superpopulation is assumed: this assumption implies the stationarity of the sampling process. When the finite population is such that the previous assumption is not reasonable, the sampling process is not a stationary or linear one; the parameters to be estimated do not suggest any obvious regression curve, and the resulting residuals are not normally distributed (except possibly in trivial cases) nor identically distributed. As a consequence a lot of standard regression techniques are simply not applicable.

Nevertheless the regression approach to estimation in ppswor sampling from finite population can usefully be exploited, even without any superpopulation assumption.

In the sequel a canonical way of associating a meaningful (non linear) regression curve to any population parameter will be described. A new class of estimators will be introduced whose computation can be implemented in a recursive simple way.

These estimators can be particularly useful when the auxiliary prior information is only partially available, which is often the case, as for instance in the forecasting example mentioned above.

2. PPSWOR Sampling

Consider a finite population of N units:

$$\mathcal{U} = \{U_1, U_2, \ldots, U_N\} \tag{2.1}$$

Let X be a real positive-valued stochastic process taking value X_i whenever unit U_i is selected.

X_i will be called the « size » of unit U_i.

Ppswor sampling means that the probability of selecting units $U_1, U_2, \ldots, U_n$ in that order ($n = 1, 2, \ldots, N$) is:

$$P\{(U_1, U_2, \ldots U_n)\} = \Pi_{i=1}^{n} \frac{X_i}{X_i + X_{i+1} + \ldots + X_N} \tag{2.2}$$

Usually, in the ppswor sampling literature all the N values $X_1, X_2, \ldots, X_N$, together with N itself, are assumed known in advance: they are referred to as « auxiliary » or « ancillary » information.

In the present paper the only assumed knowledge prior to any sampling is that of:

$$R = \Sigma_{i=1}^{N} X_i \tag{2.3}$$

Notice that many other estimators are no longer computable when the auxiliary prior information is restricted to knowledge of R only (e.g. Horvitz-Thompson).

Let Y be a stochastic process taking (real) value Y_i whenever unit U_i is selected.

Let:

$$\vartheta = \Sigma_{i=1}^{N} Y_i \tag{2.4}$$

be the quantity to be estimated: the « population total ».

For any given parameter there will be a corresponding stochastic process Y. For instance if, as in section 4, the parameter to be estimated is the total number of units in a given class C, then the corresponding stochastic process Y will take value Y_i as follows:

$$Y_i = \begin{cases} 1 & U_i \in C \\ 0 & U_i \notin C \end{cases}$$

Both X and Y are supposed perfectly observable, in the sense that, posterior to selection of unit $U_i\,(i = 1, 2, \ldots, N)$ knowledge of both the corresponding values X_i and Y_i is available.

3. Regression estimators in PPSWOR Sampling

For clarity sake, let labels $U_1, \ldots, U_N$ be such that U_1 indicates the first unit selected, U_2 the second and so on. The usual step in constructing a regression curve given a particular stochastic process is to consider the curve that fits its expected values and to discover that this is in intimate relation with the parameter(s) to be estimated.

Let:

$$\mathscr{F}_0, \mathscr{F}_1, \ldots \mathscr{F}_N$$

be a sequence of (σ)algebras defined on $\mathscr{U}$ such that:

(i) $\mathscr{F}_0 = \{\varnothing, \mathscr{U}\}$

(ii) $\mathscr{F}_N = \mathscr{P}(\mathscr{U})$ = Collection of all subsets of $\mathscr{U}$

(iii) $\mathscr{F}_i \subset \mathscr{F}_{i+1}$ $(i = 0, 1, \ldots, N-1)$ (3.1)

(iv) $\mathscr{F}_i$ is generated by the first i random selections.

For any given parameter ϑ define the following stochastic process Z:

$$Z = Y/X \tag{3.2}$$

that is Z takes value

$$Z_i = Y_i/X_i \tag{3.3}$$

whenever unit U_i is selected.

Then define:

$$E_n(Z) = E(Z \mid \mathscr{F}_{n-1}) \tag{3.4}$$

that is $E_n(Z)$ is the expected value of Z at the n-th selection conditional upon knowledge of the results from the previous $n - 1$ selections. Thus in the first drawing:

$$E_1(Z) = \Sigma_{i=1}^{N} Z_i \cdot P\{U_i \text{ is selected first}\} = \vartheta/R.$$

Analogously (remember that $U_1, U_2, \ldots, U_{n-1}$ denote the first n — 1 selections):

$$E_n(Z) = \Sigma_{i=n}^{N} Z_i \cdot P\{U_i \mid \mathscr{F}_{n-1}\} = (\vartheta - \Sigma_{i=1}^{n-1} Y_i)/(R - \Sigma_{i=1}^{n-1} X_i)$$

Given that the only unknown in $E_n(Z)$ is the parameter ϑ, let us write explicitely

$$E_n(Z, \vartheta)$$

instead of $E_n(Z)$ in the sequel. Thus:

$$\begin{aligned} E_1(Z; \vartheta) &= \vartheta/R \\ E_n(Z; \vartheta) &= (\vartheta - \Sigma_{i=1}^{n-1} Y_i)/(R - \Sigma_{i=1}^{n-1} X_i) \quad n = 2, \ldots, N \end{aligned} \tag{3.5}$$

Consider now the stochastic process:

$$V(\vartheta) = \{V_i(\vartheta)\} \quad i = 1, 2, \ldots N \tag{3.6}$$

where:

$$V_i(\vartheta) = Z - E_i(Z; \vartheta) \tag{3.7}$$

This process will be called « Innovation ». Obviously the Innovation process has zero mean, and for any $i \neq j$, V_i and V_j are uncorrelated.

Let ϑ^* be the true value of the parameter ϑ. It can be easily seen that:

$$E[V_i^2(\vartheta)] = E[V_i^2(\vartheta^*)] + [E_i(Z; \vartheta^*) - E_i(Z; \vartheta)]^2 \tag{3.8}$$

Therefore:

$$E(\Sigma_{i=1}^{n} V_i^2(\vartheta)) = \Sigma_{i=1}^{n} E(V_i^2(\vartheta^*)) + \Sigma_{i=1}^{n} (E_i(Z; \vartheta^*) - E_i(Z; \vartheta))^2 \tag{3.9}$$

Notice that the true value ϑ^* of the parameter ϑ is the only one that minimizes the quantity in (3.9). This fact suggest the following estimator $\hat{\vartheta}$ of ϑ^*:

$$\hat{\vartheta} = \text{Argmin}\{\Sigma_{i=1}^{n} v_i^2(\hat{\vartheta})\} \tag{3.10}$$

where lower case denote actual observed values. An explicit solution for $\hat{\vartheta}$ is given by:

$$\hat{\vartheta} = [\Sigma_{i=1}^{n}(R - \Sigma_{k=1}^{i-1} x_k)^{-2}]^{-1} \cdot$$

$$\cdot \Sigma_{i=1}^{n} \left\{ \left(\frac{y_i}{x_i}(R - \Sigma_{k=1}^{i-1} x_k) + \Sigma_{k=1}^{i-1} y_k \right) \cdot (R - \Sigma_{k=1}^{i-1} x_k)^{-2} \right\} \quad (3.11)$$

The proof is straightforward and thus omitted. Notice that $\hat{\vartheta}$ is a linear estimator with respect to y_i: the characteristics of interest.

Furthermore notice that if $Y = X$ then the estimator will be umbiased and with zero variance: this fact is trivial, but it is not true of some other estimators (e.g. Horvitz-Thompson). In general however the estimator will not be unbiased.

4. Estimation of North Sea Oil reservoirs

The Oil Industry estimates for the total amount of recoverable oil from the North Sea (see Smith; 1979) range from a minimum of 35000 to maximum of 67000 (milion barrel). The same global amount of oil could in principle be spread out into many subeconomical fields or concentrated into a few giant reservoirs. The estimation procedure described in the previous section has been applied to forecast, for different size classes, the total number of (discovered and undiscovered) North Sea oil reservoirs.

The data used are the amount of recoverable oil («size ») of the first 99 discoveries declared in the North Sea before 1979 as reported in by Smith (1979). He classified the observed sizes into 7 different categories C_j and, adopting a ppswor sampling scheme to model the oil discovery process as was first suggested by Kaufman *et al.* in (1976, 1977), he made estimates of the total number of reservoirs in each size class. Letting:

x_i: = amount of recoverable oil from the i-th discovered reservoir

$$y_i: = \begin{cases} 1 & \text{if } x_i \in C_j \\ 0 & \text{otherwise} \end{cases}$$

and assuming $R = 48250$ in order to make a comparison with Smith estimates, we obtained the results reported in Tab. I. These results, obtained by using the present simple procedure are surprisingly close to those obtained by Smith using a sophisticated (local) Maximum Likelihood (MLE) procedure which required hours of computer time.

Tab. I - Reservoir Size Classification* and Estimates.

Class	Class bounds	Estimated number of reservoirs Smith (MLE)	Andreatta (Regression)
1	0-75	208	206.7
2	75-150	36	32.3
3	150-300	30	30.5
4	300-500	16	16.2
5	500-1000	9	9.2
6	1000-2000	4	4.1
7	2000-5000	3	3.0

(* reservoir size measured in million barrel).

5. Some remarks

A particularly difficult problem in ppswor sampling is, for any proposed estimator to find a corresponding computable and reliable variance estimator. Usually the variance estimators are much more complicated than the original estimators. Often, even when variance estimators exist and are umbiased many doubts remain about their reliability: for instance the variance estimator associated with the Horvitz-Thompson estimator is sometimes negative.

Here, the Innovation process is not a stationary one, not even in Weak sense, and certainly it is not a Gaussian process, thus the well known standard techniques are of no use in evaluating the efficiency of the estimator. Application of robust techniques like the Bootstrap or Jackknife methods should prove to be quite effective to this purpose. Notice that the estimator of ϑ obtained by disregarding the m-th observation will simple be:

$$\vartheta_{.m} = y_m + (\Sigma^{n}_{\substack{i=1 \\ i\neq m}} (R - x_m - \Sigma^{i-1}_{\substack{k=1 \\ k\neq m}} x_k)^{-2})^{-1} \cdot$$

$$\cdot \Sigma^{n}_{\substack{i=1 \\ i\neq m}} \left\{ \left(\frac{y_i}{x_i} (R - x_m - \Sigma^{i-1}_{\substack{k=1 \\ k\neq m}} x_k) + \Sigma^{i-1}_{\substack{k=1 \\ k\neq m}} y_k \right) / (R - x_m - \Sigma^{i-1}_{\substack{k=1 \\ k\neq m}} x_k)^2 \right\} \tag{5.1}$$

Thus it will be easy to implement an estimation of the Jackknife type.

For many practical purposes it will be sufficient to have some kind of indication about the stability (or instability) of the estimator being used. A simple indication of stability can be obtained by looking at the

variation in the estimator itself as n, the number of selected units (observations), varies.

Denoting by $\hat{\vartheta}_i$ the estimator based on the first i selections ($i = 1, 2, \ldots, n$), a possible choice of stability index could be the following:

$$S^2 = \Sigma_{i=1}^{n} (\hat{\vartheta}_i - \hat{\vartheta}_n)^2/n \tag{5.2}$$

A second remark is that the estimator (3.11) is of the form:

$$\hat{\vartheta} = \Sigma_{i=1}^{n} w_i T_i \tag{5.3}$$

where T_i is defined by:

$$T_i = \Sigma_{k=1}^{i-1} y_k + \frac{y_i}{x_i} (R - \Sigma_{k=1}^{i-1} x_k) \tag{5.4}$$

and:

$$w_i = (R - \Sigma_{k=1}^{i-1} x_k)^{-2} / \Sigma_{j=1}^{n} (R - \Sigma_{k=1}^{j=1} x_k)^{-2} \tag{5.5}$$

Des-Raj (1956) has proved that any estimator of the form (5.3) is unbiased for any choice of weights w_i ($i = 1, 2, \ldots, n$; $\Sigma_{i=1}^{n} w_i = 1$). However the estimator (3.11) is not necessarily unbiased because the weights are themselves stochastic and not deterministic.

Notice that, for any $j > i$, the estimator (3.11) giver more weight to T_j than to T_i, which makes sense: as extreme case observe that $T_N = \vartheta$, thus T_N is a much better estimator than T_1, and analogously we can expect that T_j is a « better » estimator than T_i.

Usually, a standard method of choosing the weights is to consider a variance-covariance matrix and then make use of its inverse (or pseudo-inverse): this is not too difficault (notice that for $i \neq j$, T_i and T_j are uncorrelated), however, we decided not to pursue this way, for many reasons. First, there is no « normal theory » to help us in obtaining a « nice » interpretation of the results. Furthermore, computation of the expected variance-covariance matrix involves knowledge of some extra parameter, which is not the case in the applications we are interested in.

Finally, in our admittedly limited numerical experience, we observed a more unstable behaviour of the estimator.

As a last remark notice that the proposed estimator corresponds to a particular criterion of goodness of fit: namely minimization of the so called « norm in L^2 ». Using the same approach with other criteria will lead to a whole class of different estimators which would be worth studying.

6. A recursive scheme for computing regression estimators in PPSWOR Sampling

A recursive scheme for computing the estimator (3.11) can easily be set up. Define, for any $n = 1, 2, \ldots, N$

$$x_n^c = \Sigma_{i=1}^n x_i \tag{6.1}$$

$$y_n^c = \Sigma_{i=1}^n y_i \tag{6.2}$$

$$K_n = \Sigma_{i=1}^n \left\{ \left(y_{i-1}^c + \frac{y_i}{x_i} (R - x_{i-1}^c) \right) \cdot (R - x_{i-1}^c)^{-2} \right\} \tag{6.3}$$

$$G_n = \Sigma_{i=1}^n (R - x_{i-1}^c)^{-2} \tag{6.4}$$

The following recursion relations are obvious:

$$x_{n+1}^c = x_n^c + x_{n+1} \tag{6.5}$$

$$y_{n+1}^c = y_n^c + y_{n+1} \tag{6.6}$$

$$K_{n+1} = K_n + (R - x_n^c)^{-2} \left[y_n^c + \frac{y_{n+1}}{x_{n+1}} (R - x_n^c) \right] \tag{6.7}$$

$$G_{n+1} = G_n + (R - x_n^c)^{-2} \tag{6.8}$$

Then the regression estimator based on $n + 1$ observations, ϑ_{n+1} can be computed by:

$$\hat{\vartheta}_{n+1} = K_{n+1}/G_{n+1} \tag{6.9}$$

Thus in order to compute ϑ_{n+1} it is not necessary to record all the past information: $\{R, x_1, y_1, \ldots x_n, y_n\}$ but just five numbers:

$$\{R,\ x_n^c,\ y_n^c,\ K_n,\ G_n\} \tag{6.10}$$

together with the current information (x_{n+1}, y_{n+1}), and then simple update those quantities using formulas 6.5, 6.6, 6.7 and 6.8.

Therefore (6.10) can be considered as a sufficient statistic. Analogously, in order to compute the suggested Stability index S^2, all what is needed to record and update are just two numbers:

$$\{\Sigma_{i=1}^n \hat{\vartheta}_i^2,\ \Sigma_{i=1}^n \hat{\vartheta}_1\} \tag{6.11}$$

as S^2 can be expressed by:

$$S^2 = (\Sigma_{i=1}^n \hat{\vartheta}_i^2 - 2\,\hat{\vartheta}_n \Sigma_{i=1}^n \hat{\vartheta}_i + n\,\hat{\vartheta}_n^2)/n \tag{6.12}$$

It is evident that due to its computational simplicity this recursive scheme can easily be implemented and the estimator here proposed can have

an edge over more sophisticated but much more cumbersome estimators, particularly when dealing with a very large number of selected units.

The author is deeply indebted to G.M. Kaufman for the many helpful discussion had during the period he spent al MIT as a visiting scholar.

References

Cassel C.M., Särndal C.E. and Wretman J.H., *Foundations of Inference in Survty.* Wiley, New York, 1977.

Horvitz D.G. and Thompson D.J., A generalization of sampling without replacement from a finite universe, *Jour. Amer. Statist. Assoc.*, 47, 663-685, 1952.

Murthy M.N., « Ordered and unordered estimators in sampling without replacement ». *Sankhyà*, 18, 378-390, 1957.

Rao J.N.K. and Bayless D.L., An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling of wo units per stratum. *Jour. Amer. Stat. Assoc.*, 64, 540-559, 1969.

Bayless D.L. and Rao J.N.K., An empirical study of stabilities of estimators and variance stimators in unequal probability sampling (n = 3 or 4). *Jour. Amer. Stat. Assoc.*, 65, 1645-1667, 1970.

Rosen B., Asymptotic theory for successive sampling with varying probabilities without replacemente, I. *The Ann. of Math. Stat.*, 43, 373-397, 1972.

Des Raj, Some estimators in sampling with varying probabilities without replacement. *Jour. Amer. Stat. Assoc.*, 51, 269-284, 1956.

Rao J.N.K., Hartley H.O. and Cochran W.G., A simple procedure of unequal probability sampling without replacement. *Jour. Roy. Stat. Soc.*, B24, 482-491, 1962.

Barouch E. and Kaufman G., Oil and Gas Discovery Modelled as Sampling Proportional to Random Size. Sloan School of Management, Working Paper, WP888-76, Dec. 1976.

Barouch E. and Kaufman G., Estimation of Undiscovered Oil and Gas. *Proc. Sym. in Appl. Math.*, Vol. XXI, American Math. Society, pp. 77-91, 1977.

Smith J.L., A probabilistim model of oil discovery. Working paper No 1/1979, Center for Applied Research, Norvegian School of Economics and Business Administration, Bergen (Norway) 1979.

Summary

Many estimators have been proposed for sampling designs with probabilities proportional to size and without replacement (PPSWOR Sampling). Due to the lack of completeness of any non trivial sufficient statistic no unique optimal estimator exists.

Many of the existing « good » estimators are quite cumbersome to compute. In this paper a new estimator, based on a regression approach, is presented, together with a simple, easy to implement, recursive scheme of computation.