Working Paper Series, N. 2, February 2005

Some Practical Aspects in Multi-Phase Sampling

Giancarlo Diana, Paolo Preo
Department of Statistical Sciences
University of Padua
Italy

Chiara Tommasi
Department of Economics
University of Milan
Italy

Abstract: Multi-phase sampling (M-PhS) scheme is useful when the interest is in the estimation of the population mean of an expensive variable strictly connected with other cheaper (auxiliary) variables. The MSE is an accuracy measure of an estimator. Usually it decreases as the sample size increases. In the practice the sample size cannot become arbitrarily large since there are usually cost constraints. From a practical point of view it would be useful to know the sample sizes which guarantee the greatest accuracy of the estimates for fixed costs. These "optimum" sample sizes can be, in some cases, computable but not admissible. In other cases, they can be neither admissible nor computable. In both the situations the solution is to consider a M-PhS scheme with one or more phases less.

Keywords: auxiliary variables, cost constraints, multi-phase sampling, optimality.

Department of Statistical Sciences
University of Padua
Italy

UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE

Contents

Department of Statistical Sciences
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

Corresponding author:
Giancarlo Diana
tel: +39 049 827 4130
giancarlo.diana@unipd.it
http://www.stat.unipd.it/~diana

Some Practical Aspects in Multi-Phase Sampling

Giancarlo Diana, Paolo Preo
Department of Statistical Sciences
University of Padua
Italy

Chiara Tommasi
Department of Economics
University of Milan
Italy

Abstract: Multi-phase sampling (M-PhS) scheme is useful when the interest is in the estimation of the population mean of an expensive variable strictly connected with other cheaper (auxiliary) variables. The MSE is an accuracy measure of an estimator. Usually it decreases as the sample size increases. In the practice the sample size cannot become arbitrarily large since there are usually cost constraints. From a practical point of view it would be useful to know the sample sizes which guarantee the greatest accuracy of the estimates for fixed costs. These "optimum" sample sizes can be, in some cases, computable but not admissible. In other cases, they can be neither admissible nor computable. In both the situations the solution is to consider a M-PhS scheme with one or more phases less.

Keywords: auxiliary variables, cost constraints, multi-phase sampling, optimality.

1 Introduction

The multi-phase sampling (M-PhS) scheme is useful when the interest is in the estimation of the population mean of an expensive object variable which is strictly connected with other cheaper (auxiliary) variables. Few authors studied the M-PhS, specifically they proposed different estimators where the auxiliary information is used in different ways. See for instance Mukerjee et al. (1987) and Ahmed (2003). In order to unify all the different proposals, Diana et al. (2004) provide a quite general class of estimators and find an optimum estimator in that class.

In sample surveys an accuracy measure of an estimator is its mean square error (MSE), which usually decreases as the sample size increases. However, the sample size cannot become arbitrarily large to get the desired accuracy since usually there is a cost constraint. From a practical point of view it would be useful to know the sample sizes, which guarantee the greatest accuracy of the estimates for fixed costs. From now on these sample sizes are called "optimum".
Both Mukerjee et al. (1987) and Ahmed (2003) make some cost considerations. This paper develops further the results given in Diana et al. (2004) to cope with a cost

constraint. Specifically, two cases are considered. In the first case, called "general", at each phase a new auxiliary variable is recorded and then it is observed at all the subsequent phases. In the second case, called "simplified", each auxiliary variable is observed only twice: at the phase where it is recorded for the first time and at the just subsequent phase. General and simplified cases are described in Section 2 and 3, respectively. For the simplified case the cost condition for using a single phase instead of a two-phase sampling scheme given by Cochran (1977), can be extended. Thus, given a cost constraint it is not always convenient to use a phase more. This matter is carefully investigated for the three-PhS scheme.

In this paper only the M-PhS scheme with dependent samples is investigated since the main aim is to control the costs. When independent samples at a low cost are available it is possible to extend the results here reached, but the algebra is very complex.

2 Optimality under a cost constraint

Let $\mathcal{U} = \{1, \ldots, j, \ldots, N\}$ be a finite population, Y the study variable and X_i, $i = 1, \ldots, k$, k auxiliary variables taking values Y_j and X_{ij} for the j-th population unit. The intererest is in estimating the population mean of Y under the M-PhS scheme: a first sample of n_1 ($n_1 < N$) units is drawn by a simple random sampling without replacement (SRSWOR), then a sub-sample of size n_2 ($n_2 < n_1$) is drawn by a SRSWOR as well and so on up to the $(k+1)$-th phase where the smallest sub-sample of size n_{k+1} ($n_{k+1} < n_k < \cdots < n_1$) is drawn. At the i-th phase the variables $X_1, \ldots, X_i$, $i = 1, \ldots, k$ are observed while at the last phase all the auxiliary variables as well as Y are measured:

Phase number	1	2	$\cdots$	i	$\cdots$	k	$k+1$
Sample size	n_1	n_2	$\cdots$	n_i	$\cdots$	n_k	n_{k+1}
	X_1	X_1	$\cdots$	X_1	$\cdots$	X_1	X_1
		X_2	$\cdots$	X_2	$\cdots$	X_2	X_2
			$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
				X_i	$\cdots$	X_i	X_i
					$\ddots$	$\vdots$	$\vdots$
						X_k	X_k
							Y

Let $w_{iu} = \overline{x}_i^{(u+1)} - \overline{x}_i^{(u)}$, $i = 1, \ldots, k$, $u = i, \ldots, k$, be the difference between the sample means at two subsequent phases, i.e. $\overline{x}_i^{(u)}$ is the sample mean of X_i at the u-th phase.With this notation, Diana et al. (2004) define a general class of estimators as a function of $\overline{y}$, i.e. the sample mean of Y at the last phase, and w_{iu}, $i = 1, \ldots, k$, $u = i, \ldots, k$. In addition, they find an optimum estimator, i.e. an estimator which reaches the minimum MSE (at the first order of approximation) in the class. This

optimum estimator is

$$\bar{y}_k = \overline{y} + \sum_{u=1}^{k} w_u^T \mathbf{g}_u^*, \tag{1}$$

where $w_u^T = (w_{1u}, w_{2u}, \ldots, w_{uu})$ and $\mathbf{g}_u^* = -S_{uu}^{-1} S_{Yu}$, where S_{Yu} is the $u \times 1$ vector whose r-th element is the population covariance between Y and X_r, $r = 1, \ldots, u$ and S_{uu} is the covariance matrix of $(X_1, \ldots, X_u)^T$, $u = 1, \ldots, k$.

Let $\mathrm{AMSE}^*(\overline{y}_k)$ denote the minimum MSE, at the first order of approximation, in the general class. The aim of this paper is to find the sampling sizes $n_1^* > n_2^* > \cdots > n_{k+1}^*$ which minimize

$$\mathrm{AMSE}^*(\overline{y}_k) = S_Y^2 \left[\frac{\rho_{Y.1}^2}{n_1} + \sum_{i=2}^{k} \frac{\rho_{Y.1,\ldots,i}^2 - \rho_{Y.1,\ldots,i-1}^2}{n_i} + \frac{1 - \rho_{Y.1,\ldots,k}^2}{n_{k+1}} - \frac{1}{N} \right]$$

or equivalently

$$\frac{\mathrm{AMSE}^*(\overline{y}_k)}{S_Y^2} + \frac{1}{N} = \sum_{i=1}^{k+1} \frac{a_i}{n_i} \ , \tag{2}$$

under the following cost constraint

$$C_t = C_0 + \sum_{i=1}^{k} c_i n_i + c_{k+1} n_{k+1} \ . \tag{3}$$

Here, $a_1 = \rho_{Y.1}^2$, $a_i = \rho_{Y.1,\ldots,i}^2 - \rho_{Y.1,\ldots,i-1}^2$, $a_{k+1} = 1 - \rho_{Y.1,\ldots,k}^2$ and $\rho_{Y.1,\ldots,i}^2$ is the multiple correlation coefficient between Y and $X_1, \ldots, X_i$, $i = 1, \ldots, k$. Notice that all the coefficients a_i are positive by definition of multiple correlation coefficient. The quantity S_Y^2 denotes the population variance of Y. Finally, terms c_i and c_{k+1} are the per unit costs for the i-th auxiliary variable and Y, respectively and C_0 is the overhead cost.
In this paper the following ordering

$$c_1 < c_2 < \ldots < c_k < c_{k+1} \tag{4}$$

is assumed for the per unit costs. That is, X_1 is the cheapest auxiliary variable, X_2 the second cheapest one and so on up to X_k, while Y is the most expensive variable. Minimizing

$$(C_t - C_0) \sum_{i=1}^{k+1} \frac{a_i}{n_i}$$

gives, by the Cauchy-Schwartz inequality,

$$n_i^* \propto \sqrt{\frac{a_i}{c_i}}, \quad i = 1, \ldots, k+1 \ .$$

Using constraint (3),

$$n_i^* = \frac{C_t - C_0}{D} \sqrt{\frac{a_i}{c_i}}, \quad i = 1, \ldots, k+1 \tag{5}$$

where $D = \sum_{i=1}^{k+1} \sqrt{a_i\, c_i}$. The minimum $\mathrm{AMSE}^*(\overline{y}_k)$ under the cost constraint is

$$\mathrm{AMSE}_o^*(\overline{y}_k) = S_Y^2 \left[\frac{D^2}{C_t - C_0} - \frac{1}{N} \right] . \tag{6}$$

The n_i^*'s are admissible only if they satisfy the ordering $n_1^* > n_2^* > \cdots > n_{k+1}^*$, i.e. only if $c_{i+1}/c_i > a_{i+1}/a_i$ for any $i = 1, \ldots, k$. When this is not the case then at least one auxiliary variable should be dropped, thus the M-PhS scheme will have at least one phase less.
So far the number of phases to be used, $k+1$, was given, but actually it is unknown. The best choice would be to use so many phases as to achieve a fixed threshold for $\mathrm{AMSE}_o^*(\cdot)$
The following step by step procedure may be used.
Let k denote the number of auxiliary variables.

step 1. Set $k = 0$ and observe the study variable Y.
Compute $\mathrm{AMSE}_o^*(\overline{y}_0)$ ($\overline{y}_0 = \overline{y}$), if it achieves the threshold the procedure stops and a 1-PhS scheme is used, otherwise go to step 2.

step 2. Set $k = k + 1$ and observe another auxiliary variable.

step 3. Compute $\mathrm{AMSE}_o^*(\overline{y}_k)$.
If it is greater than $\mathrm{AMSE}_0^*(\overline{y}_{k-1})$ than the procedure stops and a $k-$PhS scheme is used. On the contrary, when $\mathrm{AMSE}_o^*(\overline{y}_k) < AMSE_o^*(\overline{y}_{k-1})$, if $\mathrm{AMSE}_o^*(\overline{y}_k)$ reaches the fixed threshold then the procedure stops and a $(k+1)-$PhS scheme is used, otherwise go back to step 2.

Remark. Usually the population variances and covariances which appare in expression (1) are unknown. However, replacing suitable estimates of such quantities a new estimator which is equivalent (at the first order of approximation) to $\overline{y}_k$ may be got.

3 A simplified case

When the number of phases becomes large, then many coefficients $\mathbf{g}_u^*$ must be computed in order to find the optimal estimator (1). Sometimes, to overcome this problem the auxiliary variable X_i is measured only at the i-th and $(i+1)$-th phases, with $i = 1, \ldots, k$. Thus, some information is ignored but only k coefficients, instead of $k(k+1)/2$, are computed. With this simplification the optimum estimator is

$$_s\bar{y}_k = \bar{y} + \sum_{u=1}^{k} w_{uu} g_u^*$$

where index "s" stands for "simplified case". Here, $w_{uu} = \bar{x}_u^{(u+1)} - \bar{x}_u^{(u)}$ and $g_u^* = -S_{Yu}/S_u^2$, where S_{Yu} is the population covariance between Y and X_u and S_u^2 the population variance of X_u, $u = 1, \ldots, k$. Expressions for n_i^* and $\mathrm{AMSE}_o^*(_s\bar{y}_k)$ are given again by (5) and (6), but now $a_i = \rho_{Y.i}^2 - \rho_{Y.i-1}^2$, $i = 2, \ldots, k$ and $a_{k+1} =$

Data sets	I	II	III	IV	V
$\rho_{Y.1}$	0.97	0.890	0.92	0.988	0.941
$\rho_{Y.2}$	0.99	0.920	0.99	0.995	0.915
$\rho_{Y.12}$	0.99	0.922	0.99	0.995	0.946

Table 1: Correlation coefficients for five data sets

$1-\rho^2_{Y.k+1}$, while a_1 is unchanged.
From equation (5), the sample sizes n_i^*'s exist only if a_i are positive for any $i = 1,\ldots,k+1$. If this is not the case at least one auxiliary variable should be dropped and the sampling scheme is a M-PhS with at least one phase less.

Cochran (1977) provides a condition for preferring a single phase against a double phase sampling scheme. This condition can be generalized for preferring a k-PhS scheme against the $(k+1)$-phase one. The proof is straightforward. Let all the a_i's be positive and the n_i^*'s follow a decreasing order. If the per unit costs of X_i and X_{i+1} are such that

$$\sqrt{\frac{c_{i+1}}{c_i}} < \frac{1}{\sqrt{a_i}}\left(\sqrt{a_i + a_{i+1}} + \sqrt{a_{i+1}}\right) \quad i = 1,\ldots,k \tag{7}$$

then, the k-PhS scheme got by dropping the i-th auxiliary variable is preferred to the $(k+1)$-PhS scheme, since

$$\mathrm{AMSE}_o^*(\overline{y}_{k-}) < \mathrm{AMSE}_o^*(\overline{y}_k),$$

where $\overline{y}_{k-}$ denotes the optimum estimator under the k-phase sampling scheme got by dropping the variable X_i.
Notice that the step by step procedure described at the end of the previous section works well if condition (7) is not satisfied at each step.

4 2-PhS vs 3-PhS: an example

In this section, for explanatory purposes only the simple case of 2-PhS scheme vs 3-PhS scheme is analyzed. Thus, the previous step by step procedure is used for choosing between a 2-PhS scheme, i.e. to observe only Y and X_2, and a 3-PhS scheme, i.e. to observe X_1 too. The analysis is based on a population of size $N = 10.000$ with the correlation coefficients given in Table 1. The five data sets are taken from Mukerjee et al. (1987).
In addition, $C_0 = 10$ and the following conditions on the ratio between the per unit costs are imposed

$$\frac{c_1}{c_2} = \frac{c_2}{c_3} = r_c\ , \quad r_c \in (0,1)\ .$$

Let

$$\mathrm{Eff}(r_c) = \frac{\mathrm{AMSE}_o^*(\overline{y}_1)}{\mathrm{AMSE}_o^*(\overline{y}_2)}$$

be a measure of efficiency of the 3-PhS scheme with respect to the 2-PhS one.

Figure 1 shows Eff(r_c) only for the first four data sets of Table 1.
For the fifth data set $a_2 = \rho^2_{Y.2} - \rho^2_{Y.1}$ is negative since $\rho^2_{Y.2} < \rho^2_{Y.1}$, thus n_2^* cannot

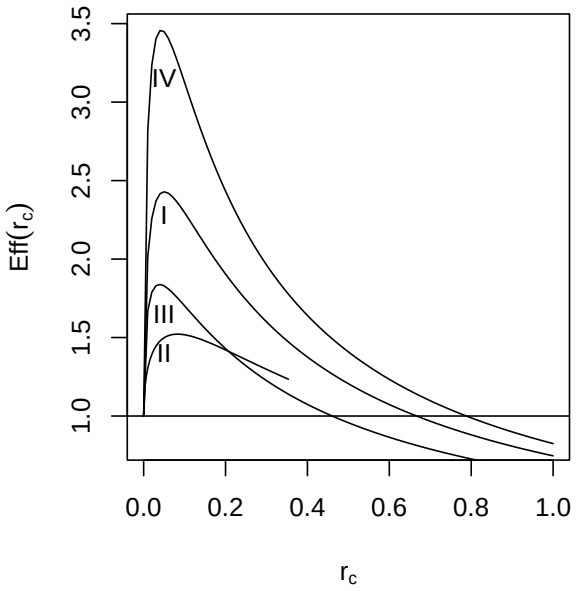

Figure 1: Efficiency of three-PhS vs two-PhS. Simplified case

be computed. In this case, the optimization problem (2) under the cost constraint (3) leads to a 2-PhS scheme. This problem will be treated more in detail in the next section.

For the other data sets there is a threshold ${}_0r_c$, such that for r_c greater than ${}_0r_c$ condition (7) is satisfied and so the 2-PhS scheme is preferred to the 3-PhS one. For instance, for the third data set ${}_0r_c = 0.462$. Thus, when r_c is greater than 0.462, Eff(r_c) is less than 1 and so the 2-PhS scheme is more efficient than the 3-PhS one.

A different case is the second data set. Here, when r_c is greater than 0.354 the optimal solutions n_i^*'s are not well ordered and so, as stressed at the end of Section 2, at least one phase should be dropped. In other words, when r_c is greater than 0.354 no optimal 3-phase sampling scheme exists.

In the general case described in Section 2 the analysis can be done for all the five data sets since values a_i's are always positive. However, a figure for Eff(r_c) in the general case is not given. It would be like Figure 1 since only for the second data set $\rho_{Y.12}$ is greater than $\rho_{Y.2}$ and even in this case the difference is very small (0.002). Of course, the shape of Eff(r_c) changes from the general case to the simplified one, as $\rho_{Y.12}$ is further away from $\rho_{Y.2}$. This change in Eff(r_c) is shown only for the second data set. Figure 2 gives Eff(r_c) in the general case for increasing values of $\rho_{Y.12}$.

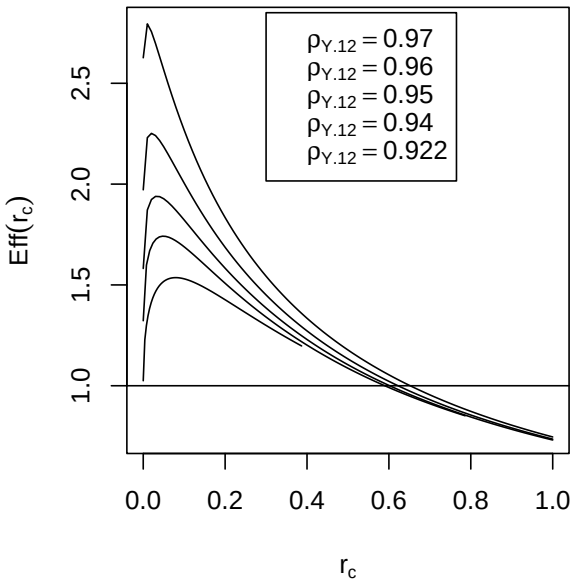

Figure 2: Efficiency of three-PhS vs two-PhS for data set II in general case

5 Which variable should be dropped?

In the simplified case the optimum sample sizes exist only if all the coefficients a_i, $i = 1, \ldots, k+1$, are positive. Moreover they are admissible if they satisfy the decreasing order $n_1^* > n_2^* > \cdots > n_{k+1}^*$. Sometimes one of the previous conditions can be unsatisfied. In these cases one or more auxiliary variables should be dropped and so a M-PhS scheme has a lower number of phases. In the 3-PhS, useful conditions for deciding which one between X_1 and X_2 should be dropped, may be given. Then, the 2-PhS scheme which minimizes the AMSE is found.
Two possible cases are discussed:

a) coefficient $a_2 < 0$;

b) coefficient $a_2 > 0$ but $n_1^* > n_2^*$ and $n_2^* < n_3^*$.

Case a
$\rho_{Y.2}^2 < \rho_{Y.1}^2$ thus $a_2 < 0$. In this case, the optimum sample size n_2^* cannot be computed. The solution is given by dropping one of the two auxiliary variables. There are two possibilities:

1. to drop X_1, the variable with the largest correlation with Y;

2. to drop X_2, the variable with the smallest correlation with Y.

The best solution is to drop X_2 if one of the following conditions is satisfied

$$\text{I.} \quad \sqrt{\frac{a_1+a_2}{a_1}} < \sqrt{\frac{c_1}{c_2}} < 1 \quad \text{and} \quad \sqrt{\frac{c_2}{c_3}} < \frac{\sqrt{a_3}-\sqrt{a_2+a_3}}{\sqrt{a_1\dfrac{c_1}{c_2}}-\sqrt{a_1+a_2}} \ ,$$

$$\text{II.} \quad \sqrt{\frac{c_1}{c_2}} < \sqrt{\frac{a_1+a_2}{a_1}}.$$

If neither I. nor II. is satisfied the solution is to drop X_1, which has the largest correlation with Y and is the cheapest variable!

Case b
In this case $a_2 > 0$ and so X_2 is the most correlated variable with Y, but the sample sizes are not admissible. Again the best solution is to drop X_1, the variable with the smallest correlation with Y, when

$$\sqrt{\frac{c_2}{c_3}} < \frac{\sqrt{a_2+a_3}-\sqrt{a_3}}{\sqrt{a_1+a_2}-\sqrt{a_1\dfrac{c_1}{c_2}}} \ .$$

However, if the above condition is not satisfied the best solution is given by dropping X_2!

Remark: in both cases the best choice could be to keep the variable which has the smallest correlation with Y.

Data sets II and V given in the previous section are examples of case b and case a, respectively. In both cases the above conditions on the per unit costs are satisfied and so the variable with the largest correlation with Y is maintained.

6 Conclusion

In the present paper the M-PhS scheme is analyzed, specifically the general and a simplified case are considered. When there is a cost constraint it would be useful to compute the optimum sample size at each phase, but it is not easy to reach this goal. In the general case the optimum sample sizes are always computable but they may be unadmissible. In the simplified case these optimum sample sizes could be neither admissible nor computable. In both cases the solution is to consider a M-PhS scheme with one or more phases less. The number of phases (or variables) to drop depends on how many sample sizes are not in decreasing order, in both the general and the simplified case. For the simplified case, it depends also on how many sample sizes are not computable.

For the 3-PhS scheme, useful conditions on the per unit costs for choosing which variable should be dropped are available, see the previous section. Of course with more than three phases everything becomes more difficult. The more the phases, the less likely the sample sizes are computable and/or admissible. Furthermore, the conditions on the per unit costs for deciding which phases should be dropped become

very complex. Thus, when the sample sizes are not computable and/or admissible, for choosing which phases should be dropped, the advice is: compute and compare directly the AMSE_o^* corresponding to the different eliminations of the phases and take the M-PhS scheme with the least AMSE_o^*.
From these short notes, it is not always convenient to add more and more phases: take into consideration the trade off between the efficiency gain and the computational effort.

References

Ahmed, M. S. (2003). General chain estimators under multi phase sampling. *J. Applied Statist. Science*, 17:to appear.

Cochran, W. G. (1977). *Sampling Techniques*. John wiley and Sons, New York. 3^{rd} edition.

Diana, G., Tommasi, C., and Preo, P. (2004). Estimation for finite population mean under multi-phase sampling. *Atti della XLII Riunione Scientifica SIS, Bari*, pages 525–528.

Mukerjee, R., Rao, T. J., and Vijayan, K. (1987). Regression type estimators using multiple auxiliary information. *Austral. J. Statist.*, 29 (3):244–254.

Working Paper Series
Department of Statistical Sciences, University of Padua

You may order paper copies of the working papers by emailing wp@stat.unipd.it
Most of the working papers can also be found at the following url: http://wp.stat.unipd.it

UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE

Department of Statistical Sciences
University of Padua
Italy