BIBLIOTECA DI SCIENZE STATISTICHE
SERVIZIO BIBLIOTECARIO NAZIONALE
BID PUV0850394
ACQ. 484 / '02 INV. 82184
COLL. CLASS. 5-GEO. WP. 2002/9

Optimal estimation for finite population parameters in two phase sampling

G. Diana, C. Tommasi

2002.9

UNIVERSITA' DI PADOVA
BIBLIOTECA DI SCIENZE STATISTICHE
Via C. Battisti, 241 - 35121 PADOVA

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

luglio 2002

Optimal estimation for finite population parameters in two phase sampling

Giancarlo Diana Chiara Tommasi

Abstract

In this paper we propose a general approach for estimating a finite population parameter in double sampling. When two dependent samples are drawn, several estimators were proposed to estimate the population mean, ratio and variance. While there are few proposals in double sampling with independent samples. We treat both cases, i.e. dependent and independent samples, showing that all the proposed estimators can be obtained as particular cases of a unique general class. The minimum variance bound for any estimator in this class is provided (at the first order of approximation). Furthermore, a chain regression type estimator which reaches this minimum is found.

keywords: two-phase sampling – dependent and independent samples – auxiliary variable – regression type estimator.

1 Introduction

The estimation of a parameter of interest when at least one auxiliary variable is available, was widely discussed. It is well known that if there are not information about the population characteristics of the auxiliary variable then a double sampling scheme may be used. We have two possibilities: the second phase sample may be drawn either from the first phase units (i.e. dependent samples) or independent of them from the population (i.e. independent samples). In the first case several estimators were proposed to estimate the population mean and the ratio of two means, while only few papers concern the population variance. For instance, the population mean estimation is treated, among others, by Srivastava *et al.* (1990), Singh and Singh (1991), Upadhyaya *et al.* (1992), Tracy and Singh (1999) and Singh and Espejo (2000). While Chand (1975), Srivastava *et al.* (1988,1989), Khare (1991), Singh *et al.* (1994), Singh and Singh (1994), Prasad *et al.* (1996) are some of

the authors who deal with the estimation of ratio of two means. On the other hand, only Singh (1991) and Gupta *et al.* (1992) work about the population variance in a two-phase sampling.
On the other hand, the double sampling with independent samples was not so fully treated. Only Khare (1991) and Singh and Singh (1994) estimate the ratio of two means using this sampling scheme, when only one auxiliary variable is available.
In this paper we propose a general approach for estimating a finite population parameter in double sampling, both for dependent and independent samples. Specifically, in Section 2 we give a general class of estimators which includes all the previously quoted estimators. Furthermore, we provide the minimum attainable MSE for the estimators in this class, at the first order of approximation. We call it the minimum variance bound, since MSE and variance are the same at this order of approximation. We suggest also a method to get an estimator which reaches this minimum. Such estimator has a nice interpretation when we are estimating a population mean, it is a chain regression type estimator. For this reason we call it the "best" estimator in the class, even if it is not unique. All the results of Section 2 are valid whatever the parameter of interest and the used sampling scheme. Section 3 specifies such general results to the estimation of a population mean, when dependent samples are drawn. Then, in the same context, Section 4 is devoted to the estimation of the ratio and the product of two means and the population variance. These subjects are discussed exhaustively in Diana and Tommasi (2001, 2002a, 2002b) but here they are unified. In fact, we also show that the estimation of these parameters is equivalent to the one of a mean, at least approximatively. Therefore, a central role is played by the mean case.
Finally, in Section 5 we develop the mean case for independent samples. As for the dependent samples case, the estimation of the ratio or product and of the variance follows immediately from the mean estimation. Some general conclusions and remarks are given in Section 6.

2 A general class of estimators

Let $\mathbf{Y}$ be a $m \times 1$ vector of study variables and T_Y be a scalar parameter of interest. Assuming that two auxiliary variables, X and Z, are related with $\mathbf{Y}$ but no information about the population parameters of X are available, then estimation of T_Y can be based on double sampling. Thus, we assume that a preliminary large sample of n' $(n' < N)$ units is drawn by some sampling scheme. At this phase only X and Z are measured. As a second step, a second sample of size n $(n < n')$ is drawn. Without loss of generality, at this

phase all the variables $\mathbf{Y}$, X and Z are observed. From now on $'$ will denote quantities related to the first phase sample only.
If $\mathbf{T_X}$ and $\mathbf{T_Z}$ are $k \times 1$ and $l \times 1$ vectors of population parameters of X and Z and $\mathbf{t_X}$, $\mathbf{t'_X}$ and $\mathbf{t_Z}$, $\mathbf{t'_Z}$ are the corresponding first and second phase unbiased estimates, then the proposed class of estimators is defined by

$$\hat{T}_Y = g\left(t_Y, \mathbf{t}^T\right), \tag{1}$$

where t_Y is, at least approximatively, an unbiased sample estimate of T_Y, $\mathbf{t} = \left(\mathbf{t_X}^T, \mathbf{t_Z}^T, \mathbf{t_X}'^T, \mathbf{t_Z}'^T\right)^T$ and function g is such that

a. $g : \mathcal{S} \to \mathbb{R}$ where $\mathcal{S} \in \mathbb{R}^{2(k+l)+1}$ is a convex and bounded set which contains the point $\left(T_Y, \mathbf{T}^T\right)^T$, where $\mathbf{T} = \mathrm{E}(\mathbf{t}) = \left(\mathbf{T_X}^T, \mathbf{T_Z}^T, \mathbf{T_X}^T, \mathbf{T_Z}^T\right)^T$.

b. It is a continuous and bounded function in $\mathcal{S}$.

c. Its first and second partial derivatives are continuous and bounded in $\mathcal{S}$.

d. $g(t_Y, \mathbf{T}^T) = t_Y$.

Let

$$g_0 = \left.\frac{\partial g(t_Y, \mathbf{t}^T)}{\partial t_Y}\right|_{(t_Y, \mathbf{t}^T) = (T_Y, \mathbf{T}^T)}$$

be the partial derivative of g with respect to t_Y and

$$\mathbf{g_t} = (\mathbf{g_X}^T, \mathbf{g_Z}^T, \mathbf{g_{X'}}^T, \mathbf{g_{Z'}}^T)^T = \left.\frac{\partial g(t_Y, \mathbf{t}^T)}{\partial \mathbf{t}}\right|_{(t_Y, \mathbf{t}^T) = (T_Y, \mathbf{T}^T)}$$

be the $2(k+l)$ vector of the partial derivatives of g with respect to each component of $\mathbf{t}$. Specifically, $\mathbf{g_X}$ and $\mathbf{g_{X'}}$ are the $k \times 1$ vectors of the partial derivatives of g with respect to the elements of $\mathbf{t_X}$ and $\mathbf{t_{X'}}$, while $\mathbf{g_Z}$ and $\mathbf{g_{Z'}}$ are $l \times 1$ vectors defined in the same way but referring to $\mathbf{t_Z}$ and $\mathbf{t_{Z'}}$.
From point (d), we have that $g(T_Y, \mathbf{T}^T) = T_Y$ and $g_0 = 1$. Thus, expanding $\hat{T}_Y$ at the point $(T_Y, \mathbf{T}^T)$ in a second order Taylor's series we have

$$\begin{aligned} \hat{T}_Y &\cong t_Y + (\mathbf{t} - \mathbf{T})^T \mathbf{g_t} = t_Y + (\mathbf{g_X} - \mathbf{T_X})^T \mathbf{g_X} + (\mathbf{t_Z} - \mathbf{T_Z})^T \mathbf{g_Z} \\ &+ (\mathbf{t'_X} - \mathbf{T_X})^T \mathbf{g_{X'}} + (\mathbf{t'_Z} - \mathbf{T_Z})^T \mathbf{g_{Z'}}. \end{aligned} \tag{2}$$

Since the population parameters of X are unknown, we have to impose the following constraint, $\mathbf{g_X} = -\mathbf{g_{X'}}$. For computational convenience, it is useful to change parameterization and to rewrite equation (2) as

$$\hat{T}_Y \cong t_Y + (\mathbf{t_V} - \mathbf{t'_V})^T \mathbf{g_V} + (\mathbf{t'_Z} - \mathbf{T_Z})^T \ddot{\mathbf{g}}_Z \tag{3}$$

where $\mathbf{t_V} = \left(\mathbf{t_X^T}, \mathbf{t_Z^T}\right)^T$, $\mathbf{t'_V} = \left(\mathbf{t_X^T}, \mathbf{t_Z'^T}\right)^T$, $\mathbf{g_V} = (\mathbf{g_X^T}, \mathbf{g_Z^T})^T$ and $\ddot{\mathbf{g}}_\mathbf{Z} = (\mathbf{g_Z} + \mathbf{g'_Z})$. The first order approximation for $\mathrm{MSE}(\hat{T}_Y)$ is given by

$$\begin{aligned} \mathrm{MSE}(\hat{T}_Y) \;\cong\; & S^2_{t_Y} + \mathbf{g_V^T}\, \mathbf{S}_{\mathbf{t_V}-\mathbf{t'_V},\mathbf{t_V}-\mathbf{t'_V}}\, \mathbf{g_V} + \ddot{\mathbf{g}}_\mathbf{Z}^T\, \mathbf{S}_{\mathbf{t'_Z},\mathbf{t'_Z}}\, \ddot{\mathbf{g}}_\mathbf{Z} \\ & + \; 2\left(\mathbf{g_V^T}\, \mathbf{S}_{\mathbf{t_V}-\mathbf{t'_V},\mathbf{t_Y}} + \ddot{\mathbf{g}}_\mathbf{Z}^T\, \mathbf{S}_{\mathbf{t'_Z},\mathbf{t_Y}} + \mathbf{g_V^T}\, \mathbf{S}_{\mathbf{t_V}-\mathbf{t'_V},\mathbf{t'_Z}}\, \ddot{\mathbf{g}}_\mathbf{Z}\right). \end{aligned} \tag{4}$$

where

$$S^2_{t_Y} = E\left[(t_Y - T_Y)^2\right], \quad \mathbf{S}_{\mathbf{t}.,\mathbf{t}..} = E\left[(\mathbf{t}. - \mathbf{T}.)(\mathbf{t}.. - \mathbf{T}..)^T\right].$$

This general expression for $\mathrm{MSE}(\hat{T}_Y)$ will be specialized for dependent and independent samples in the next sub-sections.

2.1 Dependent Samples

When the second phase sample is drawn from the first phase one, then the first order approximation for $\mathrm{MSE}(\hat{T}_Y)$ is

$$\begin{aligned} \mathrm{MSE}(\hat{T}_Y) \;\cong\; & S^2_{t_Y} + \mathbf{g_V^T}\left(\mathbf{S}_{\mathbf{t_V},\mathbf{t_V}} - \mathbf{S}_{\mathbf{t'_V},\mathbf{t'_V}}\right)\mathbf{g_V} + \ddot{\mathbf{g}}_\mathbf{Z}^T \mathbf{S}_{\mathbf{t'_Z},\mathbf{t'_Z}} \ddot{\mathbf{g}}_\mathbf{Z} \\ & + \; 2\left(\mathbf{g_V^T} \mathbf{S}_{\mathbf{t_V}-\mathbf{t'_V},\mathbf{t_Y}} + \ddot{\mathbf{g}}_\mathbf{Z}^T \mathbf{S}_{\mathbf{t'_Z},\mathbf{t_Y}}\right). \end{aligned} \tag{5}$$

Minimizing (5) with respect to $\mathbf{g_V}$ and $\ddot{\mathbf{g}}_\mathbf{Z}$ we obtain the optimum values

$$\mathbf{g_V^*} = -\left(\mathbf{S}_{\mathbf{t_V},\mathbf{t_V}} - \mathbf{S}_{\mathbf{t'_V},\mathbf{t'_V}}\right)^{-1} \mathbf{S}_{\mathbf{t_V}-\mathbf{t'_V},\mathbf{t_Y}} \quad \text{and} \quad \ddot{\mathbf{g}}_\mathbf{Z}^* = -\mathbf{S}_{\mathbf{t'_Z},\mathbf{t'_Z}}{}^{-1}\, \mathbf{S}_{\mathbf{t'_Z},\mathbf{t_Y}}, \tag{6}$$

and replacing them in (5) we get the minimum first order approximation of $\mathrm{MSE}(\cdot)$, denoted by $\mathrm{MSE}^*(\cdot)$. It is the minimum variance bound for all the estimators based on the auxiliary vector $\mathbf{t}$ and has a simple form,

$$\mathrm{MSE}^*\left(\hat{T}_Y\right) = S^2_{t_Y}\left(1 - \rho^2_{t_Y,t_V-t'_V} - \rho^2_{t_Y,t'_Z}\right), \tag{7}$$

where for any one random vector $\mathbf{U}$,

$$\rho^2_{t_Y,U} = \frac{\mathbf{S}_{\mathbf{U},\mathbf{t_Y}}{}^T\, \mathbf{S}_{\mathbf{U},\mathbf{U}}{}^{-1}\, \mathbf{S}_{\mathbf{U},\mathbf{t_Y}}}{S^2_{t_Y}}.$$

Finally, replacing $\mathbf{g_V^*}$ and $\ddot{\mathbf{g}}_\mathbf{Z}^*$ in (3) we get an optimum estimator. For all the estimation problems that will be considered, it is a regression type estimator. Thus, we will call it "the best" estimator in the class g.

2.2 Independent Samples

When both the first and the second phase samples are drawn independently from the population then the first order approximation for $\mathrm{MSE}(\hat{T}_Y)$, given in (4), becomes

$$\begin{aligned}\mathrm{MSE}(\hat{T}_Y) &\cong S^2_{t_Y} + \mathbf{g}_{\mathbf{V}}^T \left(\mathbf{S}_{\mathbf{t_V},\mathbf{t_V}} + \mathbf{S}_{\mathbf{t'_V},\mathbf{t'_V}}\right) \mathbf{g_V} + \ddot{\mathbf{g}}_{\mathbf{Z}}^T \mathbf{S}_{\mathbf{t'_Z},\mathbf{t'_Z}} \ddot{\mathbf{g}}_{\mathbf{Z}} \\ &+ 2\left(\mathbf{g}_{\mathbf{V}}^T \mathbf{S}_{\mathbf{t_V},\mathbf{t_Y}} - \mathbf{g}_{\mathbf{V}}^T \mathbf{S}_{\mathbf{t'_V},\mathbf{t'_Z}} \ddot{\mathbf{g}}_{\mathbf{Z}}\right). \end{aligned} \tag{8}$$

Minimizing (8) with respect to $\mathbf{g_V}$ and $\ddot{\mathbf{g}}_{\mathbf{Z}}$ we obtain the optimum values

$$\begin{aligned}\mathbf{g}^*_{\mathbf{V}} &= -\left(\mathbf{S}_{\mathbf{t_V},\mathbf{t_V}} + \mathbf{S}_{\mathbf{t'_V},\mathbf{t'_V}|\mathbf{t'_Z}}\right)^{-1} \mathbf{S}_{\mathbf{t_V},\mathbf{t_Y}} = -\mathbf{S}^{-1}_{\mathbf{t_V}+\mathbf{t'_V},\mathbf{t_V}+\mathbf{t'_V}|\mathbf{t'_Z}} \mathbf{S}_{\mathbf{t_V}+\mathbf{t'_V},\mathbf{t_Y}|\mathbf{t'_Z}} \\ \ddot{\mathbf{g}}^*_{\mathbf{Z}} &= \mathbf{S}_{\mathbf{t'_Z},\mathbf{t'_Z}}{}^{-1} \mathbf{S}^T_{\mathbf{t'_V},\mathbf{t'_Z}} \mathbf{g}^*_{\mathbf{V}}, \end{aligned} \tag{9}$$

where, if $\mathbf{U}_i$, $i = 1, 2, 3$, are any three random vectors, then

$$\mathbf{S_{U1,U2|U3}} = \mathbf{S_{U1,U2}} - \mathbf{S_{U1,U3}} \mathbf{S^{-1}_{U3,U3}} \mathbf{S_{U3,U2}}.$$

Replacing $\mathbf{g}^*_{\mathbf{V}}$ and $\ddot{\mathbf{g}}^*_{\mathbf{Z}}$ in (8) we have the minimum first order approximation of $\mathrm{MSE}(\hat{T}_Y)$,

$$\mathrm{MSE}^*\left(\hat{T}_Y\right) = S^2_{t_Y}\left(1 - \rho^2_{t_Y,t_V+t'_V|t'_Z}\right), \tag{10}$$

where we have set

$$\rho^2_{t_Y,t_V+t'_V|t'_Z} = \frac{\mathbf{S}^T_{\mathbf{t_V}+\mathbf{t'_V},\mathbf{t_Y}|\mathbf{t'_Z}} \mathbf{S}^{-1}_{\mathbf{t_V}+\mathbf{t'_V},\mathbf{t_V}+\mathbf{t'_V}|\mathbf{t'_Z}} \mathbf{S}_{\mathbf{t_V}+\mathbf{t'_V},\mathbf{t_Y}|\mathbf{t'_Z}}}{S^2_{t_Y}}.$$

As in the previous sub-section, replacing $\mathbf{g}^*_{\mathbf{V}}$ and $\ddot{\mathbf{g}}^*_{\mathbf{Z}}$ in (3) we get an optimum estimator.

3 Estimation of a finite population mean

All the results of the previous section are valid whatever the used sampling scheme and the involved auxiliary statistics. If both t_Y and all the auxiliary statistics are, at least at the first order of approximation, sample means, then the computations of Section 2 can be developed further, when a specific sampling scheme is chosen. For this reason, the estimation of a finite population mean plays a central role. This case is treated in this section for dependent samples and then the obtained results are used in Section 4, where the estimation of ratio and product of two population means and the estimation of the population variance are dealt with. It will be seen that the

sample estimates of such parameters may be approximated by sample means, at least at the first order of approximation.
Henceforth we assume that both the first and the second phase samples are drawn by a simple random sample without replacement (SRSWOR). Let $\mathcal{U} = \{1, \ldots, i, \ldots, N\}$ be a finite population. $\mathbf{U}_i$ denotes the value of any one observable $h \times 1$ random vector $\mathbf{U}$, for the i-th population unit. On the other hand, the i-th sampled value will be denoted by the lowercase letter $\mathbf{u}_i$. The population mean and covariance matrix of $\mathbf{U}$ are $\overline{\mathbf{U}} = \sum_{i=1}^{N} \mathbf{U}_i/N$ and $\mathbf{S_{UU}} = \{S_{U_r,U_s}\}$,respectively, where $S_{U_r,U_s} = \sum_{i=1}^{N}(U_{ri} - \overline{U}_r)(U_{si} - \overline{U}_s)/(N-1)$. While $\overline{\mathbf{u}}'$, $\mathbf{s}'_{\mathbf{UU}} = \{s'_{U_r,U_s}\}$ and $\overline{\mathbf{u}}$, $\mathbf{s_{UU}} = \{s_{U_r,U_s}\}$ are the first and second phase sample means and covariance matrices of $\mathbf{U}$. Specifically, $s_{U_r,U_s} = \sum_{i=1}^{n}(u_{ri} - \overline{u}_r)(u_{si} - \overline{u}_s)/(n-1)$ and s'_{U_r,U_s} is defined in the same way with n' instead of n. When $U_r = U_s$ we will write $S^2_{U_r}$, $s^2_{U_r}$ and $s'^2_{U_r}$ for S_{U_r,U_r}, s_{U_r,U_r} and s'_{U_r,U_r}, respectively.
Now we are interested in $T_Y = \overline{Y}$, thus $t_Y = \overline{y}$. Furthermore we consider the following auxiliary vectors, $\mathbf{t_X} = (\overline{x}, s^2_X)^T$, $\mathbf{t_Z} = (\overline{z}, s^2_Z)^T$, $\mathbf{t}'_{\mathbf{X}}$ and $\mathbf{t}'_{\mathbf{Z}}$. With this notation equation (3) becomes

$$\overline{y}_g \cong \overline{y} + (\mathbf{t_V} - \mathbf{t}'_{\mathbf{V}})^T \mathbf{g_V} + (\mathbf{t}'_{\mathbf{Z}} - \mathbf{T_Z})^T \ddot{\mathbf{g}}_{\mathbf{Z}}. \tag{11}$$

For a better interpretation of the results, it is convenient to define a random vector $\mathbf{V} = (\tilde{\mathbf{X}}^T, \tilde{\mathbf{Z}}^T)^T$ where, for any variable U, $\tilde{\mathbf{U}} = (U, \delta_U)^T$ and $\delta_U = \left(U - \overline{U}\right)^2$. In addition, let $\mathbf{U_1}$ and $\mathbf{U_2}$ be any two random vector of size $h \times 1$ and $k \times 1$. Then $\mathbf{S_{U_1,U_2}} = \{S_{U_{1r},U_{2s}}\}$ is the $h \times k$ matrix of the covariances among the elements of $\mathbf{U_1}$ and $\mathbf{U_2}$.
Setting

$$\theta_1 = \left(\frac{1}{n'} - \frac{1}{N}\right), \quad \theta_2 = \left(\frac{1}{n} - \frac{1}{N}\right) \quad \text{and} \quad \theta = \left(\frac{1}{n} - \frac{1}{n'}\right)$$

we have that, at least up to terms of order $O(n^{-1})$,

$$S^2_{t_Y} \cong \theta_2 S^2_Y, \quad \mathbf{S_{t_V,t_V}} \cong \theta_2 \mathbf{S_{V,V}}, \quad \mathbf{S_{t'_V,t'_V}} \cong \theta_1 \mathbf{S_{V,V}}, \quad \mathbf{S_{t'_Z,t'_Z}} \cong \theta_1 \mathbf{S_{\tilde{Z},\tilde{Z}}},$$

$$\mathbf{S_{t_V,t_Y}} \cong \theta_2 \mathbf{S_{V,Y}}, \quad \mathbf{S_{t'_V,t_Y}} \cong \theta_1 \mathbf{S_{V,Y}} \quad \text{and} \quad \mathbf{S_{t'_Z,t_Y}} \cong \theta_1 \mathbf{S_{\tilde{Z},Y}}.$$

With these mean values, from (6) we have that the "best" values $\mathbf{g}^*_{\mathbf{V}}$ and $\ddot{\mathbf{g}}^*_{\mathbf{Z}}$ are the partial regression coefficients of Y on $\mathbf{V}$ and $\tilde{\mathbf{Z}}$, respectively, that is

$$\mathbf{g}^*_{\mathbf{V}} = -\mathbf{S}^{-1}_{\mathbf{V},\mathbf{V}} \mathbf{S_{V,Y}} = -\begin{bmatrix} \beta_{Y,X|\delta_X,\tilde{Z}} \\ \beta_{Y,\delta_X|X,\tilde{Z}} \\ \beta_{Y,Z|\tilde{X},\delta_Z} \\ \beta_{Y,\delta_Z|\tilde{X},Z} \end{bmatrix} \quad \text{and} \quad \ddot{\mathbf{g}}^*_{\mathbf{Z}} = -\mathbf{S}^{-1}_{\tilde{\mathbf{Z}},\tilde{\mathbf{Z}}} \mathbf{S_{\tilde{Z},Y}} = -\begin{bmatrix} \beta_{Y,Z|\delta_Z} \\ \beta_{Y,\delta_Z|Z} \end{bmatrix}$$

Replacing such optimal values for $\mathbf{g_V}$ and $\ddot{\mathbf{g}}_\mathbf{Z}$ in (5) we get

$$\mathrm{MSE}^*(\overline{y}_g) \cong \theta_2 S_Y^2 \left(1 - \frac{\theta}{\theta_2}\rho_{Y,V}^2 - \frac{\theta_1}{\theta_2}\rho_{Y,Z}^2\right), \tag{12}$$

where $\rho_{.,.}$ denotes the multiple correlation coefficient between the specified variables. Up to terms of $O(n^{-1})$, $\mathrm{MSE}^*\left(\overline{y}_g\right)$ is the minimum variance bound for all the estimators based on the auxiliary vector $\mathbf{t}$, which summarizes now the first and second phase sample means and variances of X and Z.
Replacing $\mathbf{g}_\mathbf{V}^*$ and $\ddot{\mathbf{g}}_\mathbf{Z}^*$ in (11) we get the following chain regression type estimator

$$\begin{aligned}\overline{y}_{reg} &= \overline{y} - \beta_{Y,X|\delta_X,\tilde{Z}}\left(\overline{x} - \overline{x}'\right) - \beta_{Y,\delta_X|X,\tilde{Z}}\left(s_X^2 - s_X'^2\right) - \beta_{Y,Z|\tilde{X},\delta_Z}\left(\overline{z} - \overline{z}'\right) \\ &- \beta_{Y,\delta_Z|\tilde{X},Z}\left(s_Z^2 - s_Z'^2\right) - \beta_{Y,Z|\delta_Z}\left(\overline{z}' - \overline{Z}\right) - \beta_{Y,\delta_Z|Z}\left(s_Z'^2 - S_Z^2\right), \end{aligned} \tag{13}$$

whose bias is of order n^{-1}. Expression (13) shows that the construction of $\overline{y}_{reg}$ is based on two steps. At the first step a regression type estimator for the unknown first phase sample mean of Y is provided. Of course, this estimator is based on X and Z conditionally to the first phase sample. Then such estimator is improved through another regression type estimator which uses the first phase sample information about Z. Actually, a similar interpretation was provided by Sahoo and Sahoo (1999) in a less general context.

Remark 1. Usually the partial regression coefficients of Y on $\mathbf{V}$ or $\tilde{\mathbf{Z}}$ are unknown but replacing suitable estimates of such coefficients in (13) we get a new estimator which is equivalent to $\overline{y}_{reg}$, at the first order of approximation.

4 Other parameter estimation

The estimation of ratio or product of two population means, i.e. $R = \overline{Y}_0/\overline{Y}_1$ or $P = \overline{Y}_0\overline{Y}_1$, and the population variance, $S_Y^2 = \sum_{i=1}^N (Y_i - \overline{Y})^2/(N-1)$, follow immediately from the mean case. In fact for all these parameters t_Y is approximatively a sample mean.
Let us first consider the ratio estimation, thus $T_Y = R$ and $t_Y = \hat{R} = \overline{y}_0/\overline{y}_1$. Expanding $\hat{R}$ at the point $(\overline{Y}_0, \overline{Y}_1)$ in a second order Taylor's series, we have

$$\hat{R} \cong R\left(1 + \frac{\overline{y}_0}{\overline{Y}_0} - \frac{\overline{y}_1}{\overline{Y}_1}\right) = R\left(1 + \overline{d}\right),$$

where $\overline{d}$ is the second phase sample mean of $D = Y_0/\overline{Y}_0 - Y_1/\overline{Y}_1$. Since $\hat{R}$ can be considered as a convenient sample mean (at the first order of approximation), we can use all the results of the previous section replacing Y

UNIVERSITA' DI PADOVA
BIBLIOTECA DI SCIENZE STATISTICHE
Via C. Battisti, 241 - 35121 PADOVA

with $R(1+D)$. Specifically, denoting now the general class (1) by $\hat{R}_g$, the minimum first order approximation of $\mathrm{MSE}(\hat{R}_g)$ is

$$\begin{aligned}\mathrm{MSE}^*\left(\hat{R}_g\right) &= \theta_2 R^2 S_D^2 \left(1 - \frac{\theta}{\theta_2}\rho_{D,V}^2 - \frac{\theta_1}{\theta_2}\rho_{D,\tilde{Z}}^2\right) + \\ &= \mathrm{MSE}'\left(\hat{R}\right) - \theta_2 R^2 S_D^2 \left(\frac{\theta}{\theta_2}\rho_{D,V}^2 + \frac{\theta_1}{\theta_2}\rho_{D,\tilde{Z}}^2\right),\end{aligned}$$

where $\mathrm{MSE}'\left(\hat{R}\right) = \theta_2 R^2 S_D^2$ is the first order approximation of the MSE of the standard estimator $\hat{R}$. For giving the chain regression type estimator, $\hat{R}_{reg}$, which reaches $\mathrm{MSE}^*\left(\hat{R}_g\right)$ we need to introduce more notation. Let U_1 and U_2 be any two auxiliary variables (where U_2 can be a random vector) and let us denote the partial regression coefficient of Y_j on U_1 given U_2 as $\beta_{j,U_1|U_2}$, $j = 0, 1$. If $\hat{\beta}_{D,U_1|U_2}$ is a suitable estimate of $\beta_{D,U_1|U_2} = \beta_{0,U_1|U_2}/\overline{Y}_0 - \beta_{1,U_1|U_2}/\overline{Y}_1$, then the best estimator is given by

$$\begin{aligned}\hat{R}_{reg} &= \hat{R}\left[1 - \hat{\beta}_{D,X|\delta_X,\tilde{Z}}\left(\overline{x} - \overline{x}'\right) - \hat{\beta}_{D,\delta_X|X,\tilde{Z}}\left(s_X^2 - s_X'^2\right) - \hat{\beta}_{D,Z|\tilde{X},\delta_Z}\left(\overline{z} - \overline{z}'\right)\right. \\ &- \left.\hat{\beta}_{D,\delta_Z|\tilde{X},Z}\left(s_Z^2 - s_Z'^2\right) - \hat{\beta}_{D,Z|\delta_Z}\left(\overline{z}' - \overline{Z}\right) - \hat{\beta}_{D,\delta_Z|Z}\left(s_Z'^2 - S_Z^2\right)\right].\end{aligned}$$

If the interest is in the population product $P = \overline{Y}_0\,\overline{Y}_1$, all the results for the ratio estimation can be used. It is enough to replace R and $\hat{R}$ with P and $\hat{P}$ respectively, where now $\hat{P} = \overline{y}_0\,\overline{y}_1$ and $D = Y_0/\overline{Y}_0 + Y_1/\overline{Y}_1$.

Finally when we want to estimate a population variance, i.e. $T_Y = S_Y^2$, we have that $t_Y = s_Y^2$ is the sample mean of $\delta_Y = (Y - \overline{Y})^2$, except for terms of order n^{-1}. In fact

$$s_Y^2 = \frac{n}{n-1}\left[\tilde{s}_Y^2 - (\overline{y} - \overline{Y})^2\right] \cong \tilde{s}_Y^2,$$

where $\tilde{s}_Y^2 = \sum_{i=1}^n \delta_{Yi}/n$. Therefore replacing Y with δ_Y, we can use again all the results of Section 4, getting

$$\mathrm{MSE}^*\left(\hat{S}_Y^2\right) = \theta_2 S_{\delta_Y}^2 \left(1 - \frac{\theta}{\theta_2}\rho_{\delta_Y,V}^2 - \frac{\theta_1}{\theta_2}\rho_{\delta_Y,\tilde{Z}}^2\right)$$

and

$$\begin{aligned}\hat{S}_{reg,Y}^2 &= s_Y^2 - \beta_{\delta_Y,X|\delta_X,\tilde{Z}}(\overline{x} - \overline{x}') - \beta_{\delta_Y,\delta_X|X,\tilde{Z}}\left(s_X^2 - s_X'^2\right) - \beta_{\delta_Y,Z|\tilde{X},\delta_Z}(\overline{z} - \overline{z}') \\ &- \beta_{\delta_Y,\delta_Z|\tilde{X},Z}\left(s_Z^2 - s_Z'^2\right) - \beta_{\delta_Y,Z|\delta_Z}\left(\overline{z}' - \overline{Z}\right) - \beta_{\delta_Y,\delta_Z|Z}\left(s_Z'^2 - S_Z^2\right),\end{aligned}$$

where $\hat{S}_Y^2$ denotes the general class (1) and $\hat{S}_{reg,Y}^2$ is the best estimator in the class.

5 Estimating the population mean with independent samples

In this section we consider the estimation of the population mean when the first and second phase samples are independent.
Using the same notation given in Section 3 and after some algebra, from the general results of Section 2.2 follow that

$$\mathbf{g}_{\mathbf{V}}^{*} = \begin{bmatrix} -\frac{\theta_2}{\theta_{12}} \mathbf{S}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}|\tilde{\mathbf{Z}}}^{-1} \mathbf{S}_{\tilde{\mathbf{X}}\mathbf{Y}|\tilde{\mathbf{Z}}} \\ -\frac{\theta_1}{\theta_{12}} \mathbf{S}_{\tilde{\mathbf{Z}},\tilde{\mathbf{Z}}}^{-1} \mathbf{S}_{\tilde{\mathbf{Z}},\mathbf{Y}} - \frac{\theta_2}{\theta_{12}} \mathbf{S}_{\tilde{\mathbf{Z}},\tilde{\mathbf{Z}}|\tilde{\mathbf{X}}}^{-1} \mathbf{S}_{\tilde{\mathbf{Z}},\mathbf{Y}|\tilde{\mathbf{X}}} \end{bmatrix} \quad \text{and} \quad \ddot{\mathbf{g}}_{\mathbf{Z}}^{*} = -\mathbf{S}_{\tilde{\mathbf{Z}},\tilde{\mathbf{Z}}}^{-1} \mathbf{S}_{\tilde{\mathbf{Z}},\mathbf{Y}},$$

where $\theta_{12} = \theta_1 + \theta_2$. In this case too, these optimum values depend on some coefficient regressions. For instance, the quantity $\mathbf{S}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}|\tilde{\mathbf{Z}}}^{-1} \mathbf{S}_{\tilde{\mathbf{X}}\mathbf{Y}|\tilde{\mathbf{Z}}}$ is the regression coefficient of Y on the residuals from the regression of $\tilde{\mathbf{X}}$ on $\tilde{\mathbf{Z}}$. The minimum first order approximation for $\mathrm{MSE}(\overline{y}_g)$ is now

$$\mathrm{MSE}^{*}(\overline{y}_g) = \theta_2 \mathrm{S}_{\mathrm{Y}}^2 \left(1 - \frac{\theta_2}{\theta_{12}} \rho_{YV}^2 - \frac{\theta_1}{\theta_{12}} \rho_{Y\tilde{Z}}^2\right).$$

From the comparison between double sampling with the dependent and the independent samples, i.e. comparing the last expression for $\mathrm{MSE}^{*}(\overline{y}_g)$ with (12), we have that the sampling scheme with independent samples is always more efficient. It is more expensive, however, and thus it is less used in practice.

As for the case of dependent samples, the estimation of a ratio or product of two means and of the variance follow immediately from the mean case.

6 Conclusions

The main result of this paper is that, given an auxiliary vector $\mathbf{t}$, any estimator for a parameter T_Y, belongs to general class $g(t_Y, \mathbf{t})$. Furthermore we provide the minimum MSE for the class and a regression type estimator which reaches that minimum. No estimator based on the same $\mathbf{t}$ can be more efficient than the best one. If an estimator is, however, equivalent to the best one, at the first order of approximation, it is optimum as well.

Among the parameters of interest a special role is played by the mean. The general results for this case let us to treat the estimation of a ratio or product of two means and the variance estimation. This is, actually, possible whenever the used estimator t_Y is a linear combination of samples means, at least at the first order of approximation.

Finally, we stress that in this paper some general results about the double sampling with independent samples are provided. This sampling scheme was not widely treated in literature, even if it gives more efficient estimators. From a practical point of view it is more expensive, however.

7 Appendix

In this appendix we provide another expression for $\mathbf{g}_{\mathbf{V}}^*$ and $\ddot{\mathbf{g}}_{\mathbf{Z}}^*$ given in Section 2.2. With these new expression it is easier to derive $\mathbf{g}_{\mathbf{V}}^*$ and $\ddot{\mathbf{g}}_{\mathbf{Z}}^*$ for the mean case.

$$\begin{aligned}
\mathbf{g}_{\mathbf{V}}^* &= \begin{bmatrix} \mathbf{g}_{\mathbf{X}}^* \\ \mathbf{g}_{\mathbf{Z}}^* \end{bmatrix} = \begin{bmatrix} -\left(\mathbf{S}_{\mathbf{t_X},\mathbf{t_X}|\mathbf{t_Z}} - \mathbf{S}_{\mathbf{t}'_{\mathbf{X}},\mathbf{t}'_{\mathbf{X}}|\mathbf{t}'_{\mathbf{Z}}}\right)^{-1} \mathbf{S}_{\mathbf{t_Y},\mathbf{t_X}|\mathbf{t_Z}} \\ -\mathbf{S}_{\mathbf{t_Z},\mathbf{t_Z}}^{-1}\mathbf{S}_{\mathbf{t_Y},\mathbf{t_Z}} - \mathbf{S}_{\mathbf{t_Z},\mathbf{t_Z}}^{-1}\mathbf{S}_{\mathbf{t_X},\mathbf{t_Z}}^{T}\mathbf{g}_{\mathbf{X}}^* \end{bmatrix} \\
\ddot{\mathbf{g}}_{\mathbf{Z}}^* &= -\mathbf{S}_{\mathbf{t_Z},\mathbf{t_Z}}^{-1}\mathbf{S}_{\mathbf{t_Y},\mathbf{t_Z}} - \mathbf{S}_{\mathbf{t_Z},\mathbf{t_Z}}^{-1}\mathbf{S}_{\mathbf{t_X},\mathbf{t_Z}}^{T}\mathbf{g}_{\mathbf{X}}^* + \mathbf{S}_{\mathbf{t}'_{\mathbf{Z}},\mathbf{t}'_{\mathbf{Z}}}^{-1}\mathbf{S}_{\mathbf{t}'_{\mathbf{X}},\mathbf{t}'_{\mathbf{Z}}}^{T}\mathbf{g}_{\mathbf{X}}^*.
\end{aligned}$$

Specifically, $\mathbf{g}_{\mathbf{X}}^*$ and $\ddot{\mathbf{g}}_{\mathbf{Z}}$ for the mean case follow immediately from the above expression. While some algebra leads to the equivalence of the two expressions for $\mathbf{g}_{\mathbf{Z}}^*$.

References

Chand L. (1975), *Some Ratio-Type Estimators based on Two or More Auxiliary Variates* , (Unpublished Ph.D. Dissertation, Iowa State University, Ames, IOWA, USA).

Diana G., Tommasi C. (2001), *Estimation for Finite Population Variance in Double Sampling*, Working Paper #2001.22, University of Padova, Padova, Italy.

Diana G., Tommasi C. (2002a), *Optimal Estimation for Finite Population Mean in Two Phase Sampling*, Working Paper #2002.3, University of Padova, Padova, Italy.

Diana G., Tommasi C. (2002b), *Estimation for Ratio of Two Population Means in Double Sampling*, Working Paper #2002.5, University of Padova, Padova, Italy.

Gupta R.K., Singh S., Mangat N.S. (1992). Some chain ratio type estimators for estimating finite population variance, *Aligarh J. Statist.*, 12-13, 65-69.

Khare B.B. (1991), Determination of Sample Sizes for a Class of Two-Phase Sampling Estimators for Ratio and Product of two Population Means Using Auxiliary Character, *Metron*, 49, 185-197.

Prasad B., Singh R.S., Singh H.P. (1996), Some Chain Ratio-Type Estimators for Ratio of two Population Means Using Two Auxiliary Characters in Two Phase Sampling, *Metron*, 54, 95–113.

Sahoo J., Sahoo L.N. (1999), An Alternative Class of Estimators in Double Sampling Procedures, *Cal. Statist. Assoc. Bull.*, 49, 193–194.

Singh S. (1991), Estimation of finite population variance using double sampling, *Aligarh J. Statist.*, 11, 53-56.

Singh H.P., Espejo Ruiz M. (2000), An Improved Class of Chain Regression Estimator in Two-Phase Sampling, *Statistics & Decisions*, 18, 205–218.

Singh V.K., Singh G.N. (1991), Chain Type Regression Estimators with Two Auxiliary Variables under Double Sampling Scheme, *Metron*, 49, 279–289.

Singh V.K., Singh H. P. (1994), Estimation of Ratio and Product of two Finite Population Means in Two Phase Sampling, *Journ. Statist. Plann. Infer.*, 41, 163–171.

Singh V.K., Singh Hari P., Singh Housila P., Shukla D. (1994), A General Class of Chain Estimators for Ratio and Product of two Means in Finite Population, *Commun. Statist. Theory & Meth.*, 23(5), 1341–1355.

Srivastava R., Khare B.B., Srivastava S.R. (1990), A Generalized Chain Ratio Estimator for Mean of Finite Population, *J. Indian Soc. Agric. Statist.*, 42, 108–117.

Srivastava S. Rani, Khare B.B., Srivastava S.R. (1988), On Generalized Chain Estimators Ratio and Product of two Population Means Using Axiliary Characters, *Assam. Stat. Review*, 2(1), 21–29.

Srivastava S. Rani, Srivastava S.R., Khare B.B. (1989), Chain Ratio-Type Estimators for Ratio of two Population Means Using Auxiliary Characters, *Commun. Statist. Theory & Meth.*, 18(10), 3917–3926.

Tracy D.S., Singh H.P. (1999), A General Class of Chain Regression Estimators in Two-Phase Sampling, *J. Appl. Statist. Sci.*, 8, 205–216.

Upadhyaya L.N., Dubey S.P., Singh H.P. (1992), A Class of Ratio-in-Regression Estimators Using Two Auxiliary Variables in Double Sampling, *Journ. Scient. Res.*, 42, 127–134.

UNIVERSITA' DI PADOVA
BIBLIOTECA DI SCIENZE STATISTICHE
Via C. Battisti, 241 - 35121 PADOVA