ESTIMATION OF THE POPULATION MEAN WIT MISSING OBSERVATIONS USING PRODUCT TYPE ESTIMATORS

Carlos N. Bouza [1]
Facultad de Matematica y Computación
Universidad de La Habana
San Lázaro y L. Habana 10 400, Cuba.

ABSTRACT
This paper deals with the solution of the estimation of the mean using product type estimators when missing observations are present in a survey sampling. The Non Response striatum approach is considered and two estimators are developed. Imputation methods are also developed and two predictors are proposed. They are compared using Monte Carlo experiments.

KEY WORDS: non-response stratum, imputation. expected error., asymptotic unbiasedness, coverage probabilities

MSC 62D05

RESUMEN
En este trabajo se trata de la solución de la estimación de la media usando un estimador del tipo razón cuando hay observaciones perdidas en las encuestas por muestreo. El enfoque del estrato de las no respuestas es considerado y dos estimadores son desarrollados. Métodos de imputación son desarrollados también y se proponen dos predictores. Estos son comparados usando experimentos de Monte Carlo.

1 INTRODUCTION

The usual theory of survey sampling is developed assuming that the finite population $U=\{u_1,\ldots,u_N\}$ is composed by individuals that can be perfectly identified . A sample s of size $n \leq N$ is selected. The variable of interest Y is measured in each selected unit. Real life surveys should deal the existence of missing observations. There are three solutions to cope with this fact: ignore the non respondents, to sub sample the non respondents or to impute the missing values. While to ignore the non responses is a dangerous decision to sub sample is a conservative and costly solution. Imputation is often used to compensate for item non-response. See for discussions on the theme Little and Rubin (1987) Rueda and González. (2004), Särndal and Lundström (2005).

The existence of missing observations invalidates some of the initial assumptions and affects the properties of the statistical models because we can not compute the sample mean

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{1.1}$$

which estimates the population mean μ_Y because the responses are of obtained from a subset of units of the sample (sub sample)

$s_1=\{i \in s \,/\, i$ gives a response at the first visit$\}$

This fact suggests that the population U is divided into two strata: U_1 , where are grouped the units that give a response at the first visit, and U_2 which contains the rest of the individuals. This is the so called

[1] .bouza@matcom.uh.cu

'response strata' model and was first proposed by Hansen-Hurvitz (1946), see Singh (2003). Their proposal was to select a subsample s_2' of size n'_2 among the n_2 non-respondents grouped in the sample

$s_2=s\setminus s_1$.

Then we obtain information on the non-respondent's strata U_2 through a sub sample

$s_2' \subset s_2$.

Product estimators have been thoroughly studied, see Singh**,** Singh and Mangat. (1996). Different recent papers study the use of product type estimators under full response. Agrawal and Sthapit (1997) derived conditions for its asymptotic normality on the finite populations sampling. Singh and Ruiz (2007) proposed a class of ratio-product estimators in two-phase sampling.

In this paper we propose different estimators of the unknown mean, using product type models for coping with nr in survey sampling. We develop estimators of the population mean under the analyzed alternative models. Their errors are obtained and the behavior of them is compared. Section 3 is concerned with the development of these results. In section 4 the use of imputation for compensating for item non response is studied.

2. ESTIMATION OF THE MEAN UNDER THE NR-STRATUM APPROACH AND SRSWR

Non responses may be motivated by a refusal of some units to give the true value of Y or by other causes. They are present in the survey sampling. Hansen-Hurvitz in 1946 proposed to select a sub-sample among the non-respondents, see Cochran (1977). This feature depends heavily on the proposed sub-sampling rule. Alternative sampling rules to Hansen-Hurvitz´s rule have been proposed see for example Srinath (1971) and Bouza (1981).

Theoretically it is a particular double sampling design described as follows:

Step 1: Select a sample s from U using srswr
Step 2: Evaluate Y among the respondents and determine $\{y_i : i \in s_1 \subset U_1,\ /s_1/ = n_1\}$.

$$\text{Compute } \bar{y}_1 = \frac{\sum_{i=1}^{n_1} y_i}{n_1} \tag{2.1}$$

Step 3: Determine $n_2'=n_2/K,\ K>1;\ /s_2/=n_2$ with $s_2=s\setminus s_1$.

Step 4. Select a sub-sample s'_2 of size n_2' from s_2 using srswr.

Step 5. Evaluate Y among the units in s_2' $\{y_i : i \in s_2' \subset s_2,\ s_2 \subset U_2\}$.

$$\text{Compute } \bar{y}'_1 = \frac{\sum_{i=1}^{n'_2} y_i}{n_2} \tag{2.2}$$

Step 6. Compute the estimate of μ

$$\bar{\bar{y}} = \frac{n_1}{n}\bar{y}_1 + \frac{n_2}{n}\bar{y}'_2 = w_1\bar{y}_1 + w_2\bar{y}'_2 \tag{2.3}$$

Note that (2.1) is the mean of a srswr-sample selected from U_1, then its expected value is the mean of Y in the respondent stratum: μ_1. We have that the conditional expectation of (2.2) is:

$$E[\bar{y}'_2 \,|s] = \bar{y}_2 \tag{2.4}$$

as (2.4) is the mean of a srswr-sample selected from U_2

$$EE[\bar{y}'_2 \,|s] = \mu_2 \tag{2.5}$$

and taking into account that for $i=1,2\ E(n_i)=nN_i/N=nW_i$ the unbiasedness of (2.3) is easily derived.

The variance of (2.3) is deduced by using the following trick;

$$\bar{\bar{y}} = (w_1\bar{y}_1 + w_2\bar{y}_2) + w_2(\bar{y}'_2 - \bar{y}_2) \tag{2.6}$$

the first term is the mean of s, then its variance is σ^2/n. For the second term we have that

$$V\left(w_2(\bar{y}'_2 - \bar{y}_2)|s\right) = w^2{}_2 E\left(\bar{y}'_2 - \mu_2) - (\bar{y}_2 - \mu_2)|s\right)^2 =$$
$$= \quad w^2{}_2[E\left(\bar{y}'_2 - \mu_2)|s\right)^2 + E\left((\bar{y}_2 - \mu_2)|s\right)^2 \quad - 2E\left(\bar{y}'_2 - \mu_2)((\bar{y}_2 - \mu_2))|s\right)]$$

Conditioning to a fixed n_2 we have that the expectation of the third term is $(\bar{y}_2 - \mu_2)^2$. Then we have that:

$$V\left(w_2(\bar{y}'_2 - \bar{y}_2)|s\right) = w^2{}_2\left(\frac{\sigma_2{}^2}{n'_2} - \frac{\sigma_2{}^2}{n_2}\right) = w^2{}_2\sigma_2{}^2\left(\frac{K}{n_2} - \frac{1}{n_2}\right) \tag{2.7}$$

and

$$EV\left(w_2(\bar{y}'_2 - \bar{y}_2)|s\right) = \frac{W_2(K-1)\sigma_2{}^2}{n} \tag{2.8}$$

Hence the expected error of (2.3) is given by the well known expression

$$EV(\bar{\bar{y}}) = \frac{\sigma^2}{n} + \frac{W_2(K-1)\sigma_2^2}{n} \tag{2.9}$$

Our proposal is to use the Additional information provided by a known variable X for constructing a product type estimator of thee means involved.

3. PRODUCT TYPE ESTIMATORS UNDER NON RESPONSES

3.1. The basics

The product estimator is defined by

$$\bar{y}_p = \frac{\bar{x}\bar{y}}{\mu_X} \tag{3.1}$$

where $\bar{z} = \frac{\sum_{i=1}^{n} z_j}{n}; z = x, y, \qquad \mu_X = \frac{\sum_{i=1}^{N} X_j}{N}$

Its expectation is given by

$$E(\bar{y}_p) = \frac{E(\bar{x}\bar{y})}{\mu_X \mu_Y} = \mu_Y + \frac{\sigma_{XY}}{n\bar{X}}$$

where $\sigma_{XY} = \frac{\sum_{i=1}^{N} (X_j - \mu_X)(Y_j - \mu_Y)}{N}$; $\mu_Y = \frac{\sum_{i=1}^{N} X_j}{N}$

Hence its has as bias $B(\bar{y}_p) = \frac{\sigma_{XY}}{n\mu_X}$

Its variance is

$$V(\bar{y}_p) = \frac{\sigma_Y^2 + R^2\sigma_X^2 + 2R\sigma_{XY}}{N}$$

where

$$R = \frac{\mu_Y}{\mu_X}, \qquad \sigma_Z^2 = \frac{\sum_{j=1}^{N} (Z_j - \mu_Z)^2}{N}, \qquad Z = X, Y$$

A version of it is

$$\bar{y}_{p^*} = \frac{\sum_{i=1}^{n} x_j y_j}{n\mu_X} \tag{3.2}$$

and it has the same bias and variance as (3.1).

These estimators can be used for deriving the estimation of the mean of the nr stratum.

Agrawal and Sthapit (1997) derived the exact formulas for the bias and variance of the product estimator under simple random sampling . Its asymptotic normality was rigorously established under weak and interpretable regularity conditions on the finite populations

3.2. A separate product estimator of μ_Y for non responses

Let us consider

$$\bar{y}_{ps} = \frac{n_1\bar{y}_1 + n_2\bar{y}'_{2p}}{n} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n} + \frac{n_2(\bar{y}'_{2p} - \bar{y}_2)}{n} \tag{3.3}$$

where

$$\bar{y}'_{2p} = \frac{\bar{y}'_2\,\bar{x}_2}{\mu_X}$$

The first member of at the right hand side of (3.3) is the mean of *Y* in *s*. Hence the bias of (3.3) depends on the expectation of the last term. The conditional expectation of it, for a fixed $n'_{2,}$is equal to the product estimator based on the sub sample s_2. Therefore

$$E\left(\frac{n_2(\bar{y}'_{2p}-\bar{y}_2)}{n}\Big|n'_2\right)=\frac{n_2\bar{y}_2\bar{x}_2}{n\mu_X}-\frac{n_2\bar{y}_2}{n}$$

as

$$E\left(\frac{n_2\bar{y}_2\bar{x}_2}{n\mu_X}-\frac{n_2\bar{y}_2}{n}\Big|n_2\right)=\frac{n_2}{n}\left(\frac{\sigma_{2XY}}{n_2\mu_X}\right),$$

where

$$\sigma_{2XY}=\frac{\sum_{j=1}^{N_2}\left(X_{2j}-\mu_{2X}\right)\left(Y_{2j}-\mu_{2Y}\right)}{N_2},\qquad \mu_{2Z}=\frac{\sum_{j=1}^{N_2}Z_{2j}}{N_2},\qquad Z=X,Y$$

Then the bias is equal to

$$B_{ps}=B(\bar{y}_{ps})=\frac{\sigma_{XY}}{n\mu_X}$$

The results obtained previously fix that under the regularity condition

R1:, $$\frac{\sigma_{2ZY}}{n'_2\,\mu_{2Y}\mu_X}\cong\frac{\sum_{j=1}^{n_2}\left(y_{2j}-\bar{y}_2\right)\left(x_{2j}-\bar{x}_2\right)}{n'_2\,\mu_{2Y}\mu_X}$$

we have that

$$E\left(E\left(\bar{y}_{ps}|n'_2\right)n_2\right)\cong\bar{y}+\frac{\mu_{Y2}\rho_{2XY}C_{2Y}C_{2X}}{\vartheta n}$$

The variance of (3.3) is obtained by calculating

$$V\left(E\left(E\left(\bar{y}_{ps}|n'_2,n_e\right)\right)\right)+E\left(V\left(E\left(\bar{y}_{ps}|n'_2,n_e\right)\right)\right)+E\left(E\left(V\left(\bar{y}_{ps}|n'_2,n_e\right)\right)\right)$$

Let us compute the first term

$$V\left(E\left(E\left(\bar{y}_{ps}|n'_2,n_2\right)\right)\right)=V\left(\bar{y}+\frac{C_{2X}C_{2Y}\mu_{2Y}}{n}\right)=V\left(\bar{y}\right)+V\left(\frac{C_{2X}C_{2Y}\mu_{2Y}}{\vartheta n}\right)+2Cov\left(\bar{y},\frac{C_{2X}C_{2Y}\mu_{2Y}}{n}\right)$$

The first term is the variance of the sample mean in srswr

$$V\left(\bar{y}\right)=\frac{\sigma_Y^2}{n}\qquad(3.4)$$

and the second and third ones are equal to zero.

For the second term we have the expression

$$E\left(V\left(E\left(\bar{y}_{ps}|n'_2\right)n_2\right)\right)=E\left(V\left(\bar{y}+\frac{n_2}{n}\left(\bar{y}_{2p}-\bar{y}_2\right)n_2\right)\right)=E\left(\left(\frac{n_2}{n}\right)^2V\left(\left(\bar{y}_{2p}-\bar{y}_2\right)n_2\right)\right)$$

Calculating the conditional variance we obtain

$$V\left(\left(\bar{y}_{2p}-\bar{y}_2\right)|n_2\right)=V\left(\left(\bar{y}_{2p}\right)|n_2\right)+V\left(\left(\bar{y}_2\right)|n_2\right)-2Cov\left(\left(\bar{y}_{2p},\bar{y}_2\right)|n_2\right)$$

The first two terms are easily derived as

$$V\left(\left(\bar{y}_{2p}\right)|n_2\right)\cong\frac{\sigma_{2Y}^2+R_2^2\sigma_{2X}^2+2R_2\sigma_{2XY}}{n_2}$$

$$V\left(\left(\bar{y}_2\right)|n_2\right)=\frac{\sigma_{2Y}^2}{n_2}$$

For computing the third term we relay on the properties of the sampling moments enounced by David and Sukhatme (1974). This term can be written as

$$Cov\left(\left(\bar{y}_{2p},\bar{y}_2\right)|n_2\right)=E\left(\frac{\bar{y}_2^2\bar{x}_2}{\mu_X}|n_2\right)-\left(\mu_{2Y}+\frac{\mu_{2Y}\rho_2C_{2X}C_{2Y}}{n_2}\right)\mu_{2Y}$$

As

$$E\left(\bar{y}^2{}_2\bar{x}_2|n_2\right)=\mu_{2Y}^2\mu_{2X}+\frac{2\mu_{2Y}\sigma_{2XY}+\mu_X\sigma^2{}_Y}{n_2}+O(n^{-2})$$

we have that

$$Cov\left(\left(\bar{y}_{2p},\bar{y}_2\right)|n_2\right)\cong\frac{\mu_{2Y}^2}{\mu_X}\left(\mu_{2X}-\frac{\rho_2C_{2X}C_{2Y}}{n_2}\right)+\frac{2\mu_{2Y}\sigma_{2XY}+}{n_2\mu_X}+\frac{\sigma_{2Y}^2-\mu_{2Y}^2}{n_2}$$

Substituting the terms derived previously we have that

$$V\left(\left(\bar{y}_{2p}-\bar{y}_2\right)|n_2\right)\cong\frac{R_2^2\sigma_{2X}^2+2\sigma_{2XY}\left(R_2-\frac{2\mu_{2Y}}{\mu_{2X}}\right)}{n_2}-2\left(\frac{\mu_{2Y}^2}{\mu_X}\left(\mu_{2X}-\frac{\rho_2C_{2X}C_{2Y}}{n_2}\right)-\frac{\mu_{2Y}^2}{n_2}\right)$$

The analyzed variance term is derived by computing the unconditional expectation

$$E\left(\left(\frac{n_2}{n}\right)^2V\left(\left(\bar{y}_{2p}-\bar{y}_2\right)|n_2\right)\right)\cong W_2\left(S(1)+S(2)\right)-2\lambda_{2XY}\qquad(3.5)$$

where

$$S(1)=\frac{R_2^2\sigma_{2X}^2+2\sigma_{2XY}\left(R_2-\frac{2\mu_{2Y}}{\mu_{2X}}\right)}{n}$$

$$S(2)=2\left(\frac{\rho_2C_{2X}C_{2Y}}{n\mu_X}+\frac{\mu_{2Y}^2}{n\mu_X}\right)$$

and

$$\lambda_{2XY}=\left(\frac{\mu_{2Y}^2\mu_{2X}}{n\mu_X}\right)\left(nW_2^2+nW_1W_2\right)$$

The third term of the sampling error is

$$E\left(E\left(V\left(\bar{y}_{ps}|n'_2\right)n_2\right)\right)=E\left(E\left(\left(\frac{n_2}{n}\right)^2 E\left((\bar{y}'_{2p}-\bar{y}_2)^2|n'_2\right)n_2\right)\right)$$

As

$$\bar{y}'_{2p}-\mu_{2Y}=\left(\bar{y}'_{2p}-\bar{y}_2\right)+\left(\bar{y}_2-\mu_{2Y}\right)$$

is derived that

$$E\left((\bar{y}'_{2p}-\bar{y}_2)^2|n'_2\right)=E\left((\bar{y}_2-\mu_{2Y})^2|n'_2\right)-E\left((\bar{y}_2-\mu_{2Y})^2|n'_2\right)=\frac{(1-\theta)\sigma^2_{2Y}}{\theta n_2} \quad (3.6)$$

because the expectation of the cross term is equal to zero. As a consequence

$$E\left(E\left(V\left(\bar{y}_{ps}|n'_2\right)n_2\right)\right)=\frac{W_2(1-\theta)\sigma^2_{2Y}}{n\theta}$$

These results enhance to give as a characterization of the proposed estimator Lemma 3.1

Lemma 3.1 The estimator (3.3)of μ_Y is asymptotically unbiased and its variance is

$$V\left(\bar{y}_{ps}\right)=\frac{\sigma^2_Y}{n}+\frac{W_2\sigma_{ps(2)}}{n}+\frac{W_2(1-\theta)\sigma^2_{2Y}}{n\theta}-2\lambda^*_{2XY}.$$

where

$$\sigma_{ps(2)}\cong R_2^2\sigma^2_{2X}+2\sigma_{2XY}\left(R_2-\frac{2\mu_{2Y}}{\mu_{2X}}\right)+\left(2\left(\frac{\rho_2 C_{2X}C_{2Y}}{\mu_X}+\frac{\mu^2_{2Y}}{\mu_X}\right)\right)$$

$$\lambda^*_{2XY}=\left(\frac{\mu^2_{2Y}\mu_{2X}}{\mu_X}\right)$$

If the regularity condition R1 holds.

Proof.

The first result is derived by fixing that

$$\lim_{n\to\infty}\left(\frac{\sigma_{2XY}}{n\mu_X}\right)=0$$

The expression of the variance is obtained by summing (3.4), (3.5) and (3.6) and doing some algebraic work.

3.3. A combined product estimator of μ_Y for non responses

We propose as an alternative the estimator

$$\bar{y}_{pc}=\left(\frac{n_1\bar{y}_1+n_2\bar{y}'_2}{n}\right)\frac{\bar{x}}{\mu_x} \quad (3.7)$$

It uses the combination of the sub samples. As stated in section 3.1 we will consider he structure

$$\bar{y}_{pc}=\left(\frac{n_1\bar{y}_1+n_2\bar{y}_2}{n}\right)\frac{\bar{x}}{\mu_x}+\left(\frac{n_2(\bar{y}'_2-\bar{y}_2)}{n}\right)\frac{\bar{x}}{\mu_x}$$

The first term is the expression of the product estimator i n the original sample. The conditional expectation of the second term is zero. Hence we have that (3.7) is asymptotically unbiased because

$$EEE\left(\bar{y}_{pc}|n'_2, n_2\right) = E\left(\bar{y}_p\right) = \mu_Y + \mu_Y\left(\frac{\rho C_X C_Y)}{n}\right)$$

and the last term (the bias) tends to zero for large sample size values

The unconditional variance of (3.7) is given by

$$V\left(EE\left(\bar{y}_{pc}|n'_2, n_2\right)\right) = V\left(\bar{y}_p\right) = \frac{\sigma_Y^2 + R^2\sigma_X^2 + 2R\sigma_{XY}}{n} = V(1)$$

It is easily derived that

$$E\left(V(E\left(\bar{y}_{pc}|n'_2\right)|\ n_2\right)) = E(V\left(\bar{y}_p|n_2\right) = 0$$

because at the second conditional level we are calculating the variance of a constant.

Let us calculate the last component of the sampling error. Using (3.6) we have

$$V(\bar{y}_{pc}|n'_2) = \left(\frac{n_2\bar{x}}{n\mu_x}\right)^2 E\left((\bar{y}'_2 - \bar{y}_2)^2|n'_2\right) = \left(\frac{n_2\bar{x}}{n\mu_x}\right)^2 \frac{(1-\theta)\sigma_{2Y}^2}{\theta n_2}$$

The expectation conditional to a fixed n_2 is

$$E\left(\left(\frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n}\right)^2|n_2\right) = \left(\frac{n_1}{n}\right)^2\left(\mu_{1X}^2 + \frac{\sigma_{1X}^2}{n_1}\right) + \left(\frac{n_2}{n}\right)^2\left(\mu_{2X}^2 + \frac{\sigma_{2X}^2}{n_2}\right) + 2\left(\frac{n_1 n_2}{n^2}\right)\left(\mu_{1X}\mu_{2X}\right)$$

Calculating $E(n^2_i)$, $I=1,2$, $E(n_1n_2)$, and adding this result to V(1), after grouping we obtain

$$V\left(\bar{y}_{pc}\right) = \frac{\sigma_Y^2 + R^2\sigma_X^2 + 2R\sigma_{XY}}{n} + \frac{(1-\theta)\sigma_{2Y}^2}{\theta\mu_X^2}\left(\mu_X^2 + \frac{W_1W_2\left(\mu_{1X} - \mu_{2X}\right)^2}{n} + \frac{\sum_{i=1}^{2} W_i\sigma_{iX}^2}{n}\right) \tag{3.8}$$

Then we have the following Lemma

Lemma 3. 2 The estimator (3.7)of μ_Y is asymptotically unbiased and its variance is given by (3.8)

4. IMPUTATION USING PRODUCT TYPE ESTIMATORS

There are two types of non-response:, unit non-response and item non-response. Weighting adjustment is often used to compensate for unit non-response. Imputation is usually used to compensate for item non-response. Imputation is widely used in sample surveys to assign values for item non-responses. If the imputed values are treated as if they were observed, then the estimates of the variances of the estimates will generally be underestimations. Methods for imputing missing data under various cases of item non-response. See Rao and Shao (1992), Rao and Sitter [1992) and Singh and Deo [2003).

We propose an imputation procedure based on a product type predictor of the non respondents. The prediction of the mean of the non-respondents is :

$$\bar{y}^*_{2p} = \frac{\sum_{i=1}^{n_2} \frac{x_i}{\mu_X}\bar{y}_1}{n_2} \tag{4.1}$$

for computing the mean of the missing observations. Using the ideas developed in section 3 an estimator is

$$\bar{y}_{IC} = \frac{n_1 \bar{y}_1 + n_2 \bar{y}^*_{2p}}{n} \tag{4.2}$$

Due to the conditional independence between the sub samples we have that

$$E\left(E\left(\bar{y}_{IC}|s\right)\right) = E\left(\frac{n_1\mu_1 + n_2 \frac{\mu_{1Y}\mu_{2X}}{\mu_X}}{n}\right) = W_1\mu_1 + W_2 \frac{\mu_{1Y}\mu_{2X}}{\mu_X}$$

Its bias is

$$B\left(\bar{y}_{IC}\right) = W_2\left(\frac{\mu_{2X}\mu_{1Y}}{\mu_X} - \mu_{2Y}\right)$$

Hence if the population is balanced in the sense

R3: $\mu_{2X} \cong \mu_X$

The bias of (4.1) is equal to the bias obtained when a srswr is the sampling design and the information provided by s_1 is used.

Expressing (4.2) as

$$\bar{y}_{IC} = \bar{y} + \frac{n_2}{n}\left(\bar{y}_{2p} - \bar{y}_2\right)$$

is easily derived that

$$V\left(E\left(\bar{y}_{IC}\{s\}\right)\right) = \left(\frac{\mu_{1Y}\mu_{2X}}{\mu_X} - \mu_{1Y}\right)^2 \frac{W_1 W_2}{n} \tag{4.3}$$

The conditional variance of the estimator is

$$V\left(\bar{y}_{IC}|s\right) = V\left(\bar{y}|s\right) + \left(\frac{n_2}{n}\right)^2 V\left(\bar{y}^*_{2p} - \bar{y}_2|s\right) \tag{4.4}$$

because the cross term is equal to zero. The expectation of the first term is

$$V(IC-1) = E\left(\frac{w_1^2\sigma_{1Y}^2 + w_2^2\sigma_{2Y}^2}{n^2}\right) = \frac{(nW_1^2 + W_1W_2)\sigma_{1Y}^2 + (nW_2^2 + W_1W_2)\sigma_{2Y}^2}{n} \tag{4.5}$$

Let us compute the second term of (4.4).

$$V\left(\bar{y}^*_{2p} - \bar{y}_2\{s\}\right) = V\left(\bar{y}^*_{2p}\{s\}\right) + V\left(\bar{y}_2\{s\}\right) - 2Cov\left(\bar{y}^*_{2p}, \bar{y}_2\{s\}\right) \tag{4.6}$$

Note that

$$V\left(\bar{y}^*_{2p}|s\right) = \frac{\sum_{i-1}^{n_2} E\left(x_i\ \bar{y}_1\right)^2 - \mu_{1Y}\mu_{2X}\mu_{2Y}}{n_2^2\mu_X^2}$$

As the sub samples are independent the first term in the numerator is the product of the expectation and is equal to

$$\eta(1) = \frac{\left(\mu_{2X}^2 + \sigma_{2X}^2\right)\left(\mu_{1Y}^2 + \frac{\sigma_{1Y}^2}{n_1}\right)}{n_2 \mu_X^2}$$

The expectation of the other terms sum -$\mu_{1Y}\mu_{2X.}$ Then the second term of (4.4) is given b

$$\left(\frac{n_2^2}{n}\right)^2 V\left(\bar{y}^*_{2p}\{s\}\right) = \left(\frac{n_2}{n^2 \mu_X^2}\right)\left(\mu_{1Y}^2 \sigma_{2X}^2 + \mu_{1Y}^2 \mu_{2X}^2 + \frac{\mu_{2X}^2 \sigma_{1Y}^2}{n_1} + \frac{\mu_{1Y}^2 \sigma_{2X}^2}{n_1}\right)$$

Hence the expectation of the second term in the conditional variance (4.4) is

$$V(IC-2) = \left(\frac{W_2}{n\ \mu_X^2}\right)\left(\mu_{1Y}^2 \sigma_{2X}^2 + \mu_{1Y}^2 \mu_{2X}^2\right) + \frac{\mu_{2X}^2 \sigma_{1Y}^2 + \mu_{1Y}^2 \sigma_{2X}^2}{n^2 \mu_X^2} E\left(\frac{n_2}{n_1}\right)$$

Doing some algebraic arrangements have that its expected value is

$$V^* \cong \frac{(nW_1^2 + W_1W_2)\sigma_{1Y}^2 + (nW_2^2 + W_1W_2)\sigma_{2Y}^2}{n} + \frac{W_2(\mu_{1Y}^2 \sigma_{2X}^2 + \mu_{1Y}^2 \mu_{2X}^2)}{n\mu_X^2} + \frac{W_2\left(\mu_{1X}^2 \sigma_{1Y}^2 + \mu_{1Y}\sigma_{2X}^2\right)}{n^2 W_1 \mu_X^2} \quad (4.7)$$

The expected variance second term of (4.3) of the srswr mean of the non respondent sub sample is equal to

$V^{**}=E(n_2\sigma^2_2/n^2)=W_2\sigma^2_2/n$

The development of covariance term leads to accept that it is equal to zero. Then we can state now the following Lemma.

Lemma 4.1. The estimator (4.2) is equivalent to $\bar{y}_1$ if the first order population balancedness R3 holds and its variance is approximately equal to

$$V_{IC} = \left(\frac{\mu_{2X}^2 \mu_{1Y}^2}{\mu_X} - \mu_{1Y}\right)\left(\frac{W_1W_2}{n}\right) + \frac{(nW_1^2 + W_1W_2)\sigma_{1Y}^2 + (nW_2^2 + (W_1+1)W)\sigma_{2Y}^2}{n} + \frac{W_2(\mu_{1Y}^2 \sigma_{2X}^2 + \mu_{1Y}^2 \mu_{2X}^2)}{n\mu_X^2}$$

when $n \rightarrow \infty$ and the second order regularity condition

R4: $E(n_2^t/n_1^q) \cong E(n_2^t/E(/n_1^q)$, $t= 1,..4$, $h=1,..,4$

is satisfied
Proof

The first result is obtained by using the condition posed in R3 and simplifying the derived bias.

Add (4.3) and (4.7). Assuming that R4 holds we have that for n sufficiently large for accepting that the terms of order $O(n^{-2})$ in the variance al negligible we have the stated result

In many occasions the interest of the results is not only to estimate the mean but to predict the response of the individual non responses. The estimator proposed is not longer a solution. Sitter and Rao (1997) proposed the use of a ratio imputation method for the missing values of the variable Y in the non response item 'i' :

$$y_{iI} = \left(\frac{\bar{y}_1}{\bar{x}_1}\right) x_i$$

Liu et al. (2006) proposed to use

$$y_{iII} = \left(\frac{1}{n_1}\sum_{j=1}^{n_1}\frac{y_j}{x_j}\right) x_i$$

We will use the auxiliary information provided by X using the product estimation principle. The result is the imputed value

$$y_i^{**} = \frac{\bar{y}_1 \bar{x}_1}{\mu_X} x_i$$

for the missing observation i. Its expectation is

$$E\left(y_i^{**}|s\right) = \left(\frac{\sigma_{1XY}\bar{x}_1}{n_1} + \mu_{1X}\mu_{1Y}\right)\frac{\mu_{2X}}{\mu_X}$$

Hence if the condition R4 is accepted the mean of the imputed values has as an approximated expected value

$$EE\left(\frac{\sum_{i=1}^{n_2} y_i^{**}}{n_2}|s\right) \equiv \left(\frac{\sigma_{1XY}\bar{x}_1}{nW_1} + \mu_{1X}\mu_{1Y}\right)\frac{\mu_{2X}}{\mu_X}$$

For improving the simplicity in the reasoning let us consider the estimator of μ_Y

$$\bar{y}_{IS} = \frac{n_1\bar{y}_1 + n_2\bar{y}_{2p}^{**}}{n} \qquad (4.8)$$

using the expression

$$\bar{y}_{IS} = \bar{y} + \frac{n_2(\bar{y}_{2p}^{**} - \bar{y}_2)}{n}$$

Its conditional mean is given by

$$E(\bar{y}_{IS}|s) = \frac{w_1\mu_{1Y} + w_2\mu_{2Y}}{n} + \frac{n_2\left(\frac{\mu_{2X}\sigma_{1XY}}{n_1\mu_X} + \frac{\mu_{1X}\mu_{1Y}\mu_{2X}}{\mu_X} - \mu_{2X}\right)}{n} \quad \ldots\ldots\ldots \qquad (4.9)$$

Calculating the expected value of this last expression we have that

$$E\left(E(\bar{y}_{IS}|s)\right) = \mu_Y + W_2\left(\frac{\mu_{1X}\mu_{1Y}\mu_{2X}}{\mu_X} - \mu_{2X}\right) + \frac{\mu_{2X}\sigma_{1XY}}{n\mu_X}E\left(\frac{n_2}{n_1}\right)$$

The two last terms are equal to the bias of (4.8). Note that that it is considerably larger than the bias of the combined product estimator. If R4 is accepted we have an approximation to the expectation of (4.8) given by

$$E\left(E(\bar{y}_{IS}|s)\right) \equiv \mu_Y + W_2\left(\frac{\mu_{1X}\mu_{1Y}\mu_{2X}}{\mu_X} - \mu_{2X}\right) + \frac{W_2\mu_{2X}\sigma_{1XY}}{nW_1\mu_X}$$

It is also larger than the bias of (4.2).

The calculation of error of the imputed mean using the separated principle is very cumbersome. We will give the main results in the sequel.

First we consider the variance of (4.9). Accepting the R4 is valid it is

$$V\left(E\left(\bar{y}_{IS}|s\right)\right) \cong \frac{W_1W_2\left(\mu_{1Y}^2 + \mu_{2Y}^2\right)}{n} + \left(\frac{\mu_{2X}\sigma_{1XY}}{n\mu_X}\right)^2 = V(1IPS)$$

(4.10)

Secondly we will consider the conditional variance of (4.8). Using its alternative expression we have

$$V\left(\bar{y}_{IS}|s\right) = V\left(\bar{y}|s\right) + \left(\frac{n_2}{n}\right)^2 V\left(\bar{y}^{**}_{2P} - \bar{y}_2|s\right) + 2\left(\frac{n_2}{n}\right)Cov\left(\bar{y}, \left(\bar{y}^{**}_{2P} - \bar{y}_2|s\right)\right)$$

The first term is equal to zero and

$$V_{2P|s} = V\left((\bar{y}^{**}_{2P} - \bar{y}_2|s)\right) = E(\bar{y}^{**}_{2P}\Big|s)^2 E(\bar{y}_2\Big|s)^2 - 2E(\bar{y}^{**}_{2P}\bar{y}_2\Big|s) - \left(E(\bar{y}^{**}_{2P} - \bar{y}_2|s)\right)^2$$

We computed the terns and arrange the similar terms. Afterwards the unconditional expectation was calculated. Assuming that the regularity conditions R4 and

R5: $W_2^t \cong 0$, for $t \geq 3$

Hold we have that

$$EV_{2P|s} \cong V(1ps) + V(2ps) + V(3ps) + V(4ps) = V(2IPS) \qquad (4.11)$$

where

$$V(1ps) \cong W_2\left(2\mu_{1X}\mu_{1Y}\mu_{2X}\mu_{2Y} + \mu_{1X}\mu_{1Y}\mu_{2X}\mu_Y + \frac{\sigma_{2X}^2}{n} + \frac{\mu_{1X}\mu_{1Y}\sigma_{2XY}}{n\mu_X}\right)$$

$$V(2ps) \cong W^2{}_2\left(\left(\frac{\mu_{1X}\mu_{1Y}\mu_{2X}}{\mu_X}\right)^2 + W_1\mu_{2Y}{}^2 - \frac{2(\mu_{1X}\mu_{1Y} + \mu_{1X}\mu_{1Y}\mu_{2X}\mu_Y}{nW_1\mu_X}\right)$$

$$V(3ps) \cong \left(\frac{\sigma_{1XY}\mu_{21Y}}{\mu_X}\right)^2 + \frac{2W_2\sigma_{1XY}\mu_{1X}\mu_{1Y}\mu^2{}_{2X}}{nW_1\mu^2{}_X}$$

$$V(4ps) \cong \frac{W_1\left(\sigma_{1XY} + \mu_{1X}\mu_{1Y}\mu_{2X}\right)}{n\mu_X} + \left(\frac{\mu_{1X}\mu_{1Y}\mu_{2X}\mu_{2Y}}{n\mu_X}\right)$$

Lemma 4.2 Fixes the behavior of the separate imputation product estimator.

Lemma 4.2. The estimator (4.8) is biased. We should prefer $\bar{y}_1$ in terms of the bias and variance.

Proof:

Comparing the biases the first affirmation is evident. On the other hand, as the approximate variance of (4.8) is

$V(ISP)=V(1IPS)+V(2IPS)$

We have the second thesis by comparing it to *V(IC).*

5. VARIANCE ESTIMATION

To study the properties of imputation based estimator, are often considered through the consideration of a super population model, the sampling mechanism generating the sample , the variable response mechanism and the imputation mechanism. The properties of the variance estimators rely, among others, on the assumption.

C.1: the complete-sample point estimator θ^*_n satisfies

$E(\theta^*_n)= \theta+O(n^{-1})$*:*

It is not accomplished by the (4.2) and (4.8). Hence to develop an estimator of the variances of the proposed estimators must cope with this disadvantage. The posed statistical problem is to obtain an interval $I(\theta)$ of minimum volume for a fixed probability π. Usually the methods are supported by a particular Central Limit Theorem that must establish that when $m\rightarrow\infty$:

$Prob\ (\theta)\in\{I^*(\theta)=(\theta(F_m)-z_{1-\alpha/2}\ \sigma_m(\theta^*_m),\ (\theta(F_m)+z_{1-\alpha/2}\ \sigma_m\ (\theta(F_m))\}\geq\pi$

$\theta(F_m)$ is the estimator (predictor) of the parameter, $z_{1-\alpha/2}$ is the percentile of the Standard Normal and $\sigma_m(\theta(F_m))$ is the standard deviation estimator of $\sigma(\theta(F))$. The robustness of $\theta(F_m)$ and $\sigma_m(\theta(F_m))$ play a key role in the validity that π be close to the coverage probability.

The Bootstrap, introduced by Efron (1979), is a powerful tool for nonparametric estimation of sampling distributions and standard errors. It may be described as follows. Let $Z = (Z_1;Z_2; : : : ;Z_m)$ be a random sample from an unknown distribution F, and let $T_m = T_m(\mathbf{Z}; F)$ be a statistic of interest. Let F_m be the empirical distribution function of the random sample. An independent random sample from F_m, $\mathbf{Z}_b$, is called a Bootstrap sample. We can use the Bootstrap method for estimating the distribution of T_m through the conditional distribution of $T_{b(m)} = T_m (\mathbf{Z}; F_m)$, given $Z_1 ;Z_2; : : : ;Z_m$. The method works by drawing B Bootstrap samples selected by using simple random samples of size m, selected with replacement from the original sample.

The Bootstrap distribution is denoted by $F^*_{B(m)}$ and $T^*_m=T(F_{B(m)}^*)=T(Z^*_1 ,..,Z^*_m)$ estimates $T(F_m)$. Due to the definitions, the conditional independence is supported and $Prob(Z_i^* =Z_t\ |F_m)=1/m\ ,\ \forall t=1,...,m,\ i=1,\ldots,\ m.$ Each sample $s(b)\in S(BS)$, $S(BS)$ the space of the Boostrap samples, is drawn with a probability $1/m^m$, hence

$E(T^*(F^*_{B(m)})|F_m)= m^{-m}\ \sum_{s(b)\in S(BS)}\ T(Z^*_{1,\ldots,}Z^*_m)_b= m^{-m}\sum_{s(b)\in S(BS)}\ T_{B(m).}$

Its conditional error is $E(T^*(F^*_m)-T_m\ |F_m)^2= m^{-m}\ \sum_{s(b)\in S(BS)}\ (T_{B(m)}-T_m)^2$. It converges to σ_T^2 if $n\rightarrow\infty$. In practice we select B random samples independently from $S(BS)$ and T_{nb}, is calculated for $s(b), b=1,\ldots,B$. The Boostrap estimator of the variance is

$V^*_{B(m)}=(mB)^{-1}\sum_{b=1}^{B}\ (T_{B(m)}-T_m)^2))=\sigma^2_{B(m)}$

It is expected , if the functional is smooth, that the limit of $\sigma^2_{B(m)}$ is the true variance of the estimator (predictor). A Central Limit Theorem supports that

$Prob\ (\theta)\in\{I^*(\theta)=(\theta(F_m)-z_{1-\alpha/2}\ \sigma_{B(m)}(\theta^*_m),\ (\theta(F_m)+z_{1-\alpha/2}\ \sigma_{B(m)}\ (\theta(F_m))\}\geq\pi$

Note that the accuracy of θ^*_m may be measured using its distribution function by estimating the confidence limits based on

$L(Z_1 ,\ldots, Z_m)=L_m =Sup\ \{t|\ F_\theta (z)\geq t\}$

$U(Z_1 ,\ldots, Z_m)=U_m =Inf\ \{t|\ F_\theta (z)\leq t\}$

The interval (L_n, U_n) has random bounds and the coverage probability of θ, π is such that

$Prob_\theta\{T(F)=\theta\in(L_m ,U_m)\}\geq\pi$, for any θ.

Usually $\pi=1-\alpha$ is fixed and close to *1*.

An alternative confidence interval, see Parr (1983) and Babu and Singh (1983) for example, is obtained by defining the parameter as the functional $\theta(F)$, $F\in\Upsilon$, and to denote the confidence interval from the relationship

$Prob_F\{\theta(F)\in(L_m ,U_m)=I(\theta)\ |\ F\in\Upsilon\}\geq\pi$

The Bootstrap distribution allows to calculate directly the quantiles

$F^*_m (t)= B^{-1}\ \Phi_{b=1}^{B}\ I((T_{B(m)}-T_m)m^{-1/2}\leq t),\quad t\in\Re$

They converge, under weak regularity conditions, see Jurečkovà-Sen (1996), $\sigma^2_{B(m)}\rightarrow\sigma_T^2$ and the quantiles of F^*_m to those of the true distribution function of the data G, whenever, for $m\rightarrow\infty$

$P_F\{(T(F_m)-T(F)m^{-1/2}\leq t\}\longrightarrow G(t)$

The first intervals will be called *normalized Bootstrap* (parametric) and the second ones *Bootstrap quantiles* (non parametric) confidence intervals.

We evaluate the behavior of the estimators proposed by computing the percent of samples in which the mean is included in the confidence intervals

$$I(\mu_Y)_q=\left(\mu_{Y\upsilon(q)}-\varepsilon_{Yp\upsilon(q)},\mu_{Yp\upsilon(q)}+\varepsilon_{Yp\upsilon(q)}\right)$$

where q identifies the criteria used for constructed confidence interval as follows

$q=1$ if the normal approximation is accepted

$q=2$ if the Parametric Boostrap is used

$q=3$ if the Non-parametric Bootstrap is used

υ=separate product estimator, combined product estimator, separate imputation predictor, combined imputation predictor.

$\varepsilon_{Yp\upsilon(q)}$ is the semi-amplitude of the interval calculated using the corresponding method q for the estimator υ with $\alpha=0,05$.

Experiment 1

We compared the different proposals developed in this paper using a data base provided from an experiment where the results for obtaining a recombinant protein production using fermentation in *786* samples. They are considered as a population and we identified the total protein in the liquid as the auxiliary variable X. The measured content of a protein is considered as Y. The non responses were considered for the samples which were re-evaluated due to technical problems. The results of interest for the estimation are given in Table 1.

Strata	W_i	Mean of the auxiliary variable	Variance of the auxiliary variable	
1	*0,682*	*66,39*	*58,9*	
2	*0,372*	*131,83*	*16,2*	

Table 1. Parameters of the strata

1 000 samples of size *80* were selected independently and the behavior of the estimations are in Table 2. The results establishes that to sub sample is better than to impute being the use of the Non Parametric Bootstrap the best alternative. The separate estimator is more reliable. The use of imputation using the separate criteria has a considerable better behavior.

	q=1 Normal Approximation	*q=2 , B=20* Parametric Boostrap	*q=3, B=20* Non-parametric Bootstrap
$\bar{y}_{ps}$	*0,80*	*0,85*	*0,91*
$\bar{y}_{pC}$	*0,60*	*0,70*	*0,75*
$\bar{y}_{Is}$	*0,41*	*0,59*	*0,74*
$\bar{y}_{IC}$	*0,75*	*0,79*	*0,74*

Table 2. Percent of inclusion of the mean in *1 000* samples generated from a population of measurements of total and recombinant protein in fermentation experiments. .

Experiment 2

The other set of experiments consisted in the generation of *1 000* variables distributed according with the distributions normal, lognormal and exponential . Rueda et al.. (2005) developed a similar experience for evaluating the behavior of some estimators of the mean when some observations were missing. We use the same parameters for generating variables distributed Normal and a lognormal variables with mean *4,9* and standard deviation *0,586*. For the exponential the parameter was $\lambda = 4.9$. Once a variable was generated a Bernoulli experiment with parameter $W_2 = 0{,}372$ was performed . If the generated variable took the value one it was considered as a nr. The Monte Carlo procedure was used for evaluating the behavior of the estimators.

N(4,9 0,586)	*q=1* Normal Approximation	*q=2 , B=100* Parametric Boostrap	*q=3, B=100* Non-parametric Bootstrap
$\bar{y}_{ps}$	*0,83*	*0,89*	*0,93*
$\bar{y}_{pC}$	*0,71*	*0,81*	*0,87*
$\bar{y}_{Is}$	*0,44*	*0,52*	*0,58*
$\bar{y}_{IC}$	*0,49*	*0,57*	*0,68*
logN(4,9 0,586)	*q=1* Normal Approximation	*q=2 , B=20* Parametric Boostrap	*q=3, B=20* Non-parametric Bootstrap
$\bar{y}_{ps}$	*0,87*	*0,88*	*0,94*
$\bar{y}_{pC}$	*0,81*	*0,85*	*0,89*
$\bar{y}_{Is}$	*0,72*	*0,77*	*0,80*
$\bar{y}_{IC}$	*0,66*	*0,70*	*0,79*

Exp(4,9)	*q=1* Normal Approximation	*q=2 , B=20* Parametric Boostrap	*q=3, B=20* Non-parametric Bootstrap
$\bar{y}_{ps}$	*0,74*	*0,79*	*0,92*
$\bar{y}_{pC}$	*0,67*	*0,71*	*0,89*
$\bar{y}_{Is}$	*0,53*	*0,65*	*0,71*
$\bar{y}_{IC}$	*0,45*	*0,56*	*0,67*

Table 3 Percent of inclusion of the mean in *1 000* samples generated from continuous variables

Acknowledgements: This research was partially supported by a grant of AUIP which allowed to develop this paper while the author was a visiting professor of Universidad de Granada, Spain..

Received September 2008
Revised December 2008

REFERENCES

[1] AGRAWAL M. C. and STHAPIT A. B (1997):Hierarchic predictive ratio-based and product-based estimators and their efficiency. **Journal of Applied Statistics**, 24,. 97-104.

[2] BABU, C.J. y K. SINGH (1983): Non parametric inferences on means using bootstrap**. Ann. Statist**. 11, 999-1003

[3] BOUZA, C.N. (1981): Sobre el problema de la fracciòn de muestreo para el caso de las no respuestas. **Trabajos de Estadística**. 21, 18-24

[4] EFRON, B. (1979): Bootstrap methods : another look to the jacknife**. Ann. Statistic**., 7, 1-26

[5] HANSEN, M.H.. and HURWITZ, W.N. (1946): The problem of non responses in sample surveys. **J. American. Statistical Association**. 41, 517-529

[6] DAVID, I. P. AND SUKHATME, B. V. (1974): On the bias and mean square error of the ratio estimator. **J. American Statistical. Assoc**. 69, 464-466

[7] JUREČKOVÀ, J. and P. K. SEN (1996): **Robust Statistical Procedures: Asymptotics and Int**errelations. Wiley, N. York.

[8] LIU, LI, YUJUAN TUA, YINGFU Lib and GUOHUA ZOU ,(2006): Imputation for missing data and variance estimation when auxiliary information is incomplete**. Model Assisted Statistics and Applications**. 1 , 83–94.

[9] LITTLE, R.J.A. and RUBIN D.B. (1987) **Statistical Analysis with Missing Data**. Wiley, N. York

[10] PARR, W.C. (1983): The boostrap: some large sample theory and connections with robustness**. Sta**tist. **and Probability Letters**. 3, 97-100

[11] RAO J.N.K and J. SHAO (1992): Jackknife variance estimation with survey data under hot deck imputation**. Biometrika** 79 , 811–822.

[12] RAO, J.N. K and R.R. SITTER (1995): Variance estimation under two-phase sampling with application to imputation for missing data. **Biometrika** 82, 453–460.

[13] RUEDA, M. and GONZÁLEZ, S. (2004) Missing data and auxiliary information in surveys, **Computational. Statistics**. 10, 559-567.

[14] RUEDA, M,. MARTÍNEZ, S. MARTÍNEZ H. and ARCOS, A. (2006): Mean estimation with calibration techniques in presence of missing data. **Computational. Statistics. And Data An**alysis, 50, 3263-3277.

[15] SÄRNDAL, KARL-ERIK and SIXTEN LUNDSTRÖM (2005**) Estimation in Surveys with Nonresponse**. Wiley, Chichester.

[16] SINGH, S. and B. DEO (2003):, Imputing with power transformation. **Statistical Pap**ers. 44 , 555–579.

[17] SINGH, HOUSILA and MARIANO RUIZ Espejo (2007): Double Sampling Ratio-product Estimator of a Finite Population Mean in Sample Surveys. **Journal of Applied Statistics**, 34, 71-85.

[18] SINGH, S. (2003): **Advanced Sampling Theory with Applications**. Kluwer Academic Publishers, Dordrecht, Amsterdam,..

[19] SINGH**,** R**.,** SINGH and MANGAT**,** N. (1996):**Elements of Survey Sampling**. Springer Series Texts in the Mathematical Sciences. N. York

[20] SITTER R.R and J.N.K. RAO (1997): Imputation for missing values and corresponding variance estimation. **The Canadian Journal of Statistics** 25, 61–73.

[21] SRINATH, K. P. (1971): Multi-phase sampling in non-response problems. **J. American Statistical**. **Association**. 66, 583-589