ESTIMATION OF THE POPULATION MEAN WITH MISSING OBSERVATIONS USING PRODUCT TYPE ESTIMATORS
Carlos N. Bouza
Facultad de Matemática 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 stratum 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

\[
\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \quad (1.1)
\]

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

\( s_1 = \{i \in s / i \text{ gives a response at the first visit}\} \)

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
‘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 SRSSWR

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 \( \{y_i : i \in s_1 \subset U_1, |s_1| = n_1\} \).

Compute

\[
\bar{y}_1 = \frac{\sum_{i=1}^{n_1} y_i}{n_1}
\]

(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 \( s_2' \) \( \{y_i : i \in s_2' \subset s_2, s_2 \subset U_2\} \).

Compute

\[
\bar{y}_2' = \frac{\sum_{i=1}^{n_2'} y_i}{n_2'}
\]

(2.2)

Step 6. Compute the estimate of \( \mu \)

\[
\bar{y} = \frac{n_1}{n} \bar{y}_1 + \frac{n_2}{n} \bar{y}_2' = w_1 \bar{y}_1 + w_2 \bar{y}_2'
\]

(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: \( \mu_1 \). We have that the conditional expectation of (2.2) is:
as \( \bar{y} \) is the mean of a srswr-sample selected from \( U \),

\[
E[\bar{y}^2] = \mu_2
\]  

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

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

\[
\bar{y} = (w_1\bar{y}_1 + w_2\bar{y}_2) + w_2(\bar{y}'_2 - \bar{y}_2)
\]  

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

\[
V(w_2(\bar{y}'_2 - \bar{y}_2) | s) = w_2^2 E((\bar{y}'_2 - \mu_2)^2 - (\bar{y}_2 - \mu_2)^2 | s) =
\]

\[
= w_2^2[E((\bar{y}'_2 - \mu_2)^2 | s)] + E((\bar{y}_2 - \mu_2)^2 | s) - 2E((\bar{y}'_2 - \mu_2)(\bar{y}_2 - \mu_2) | s)
\]

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

\[
V(w_2(\bar{y}'_2 - \bar{y}_2) | s) = w_2^2 \left( \frac{\sigma_2^2}{n_1^2} - \frac{\sigma_2^2}{n_2^2} \right) = w_2^2 \sigma_2^2 \left( \frac{K}{n_2} - \frac{1}{n_2} \right)
\]  

and

\[
EV(w_2(\bar{y}'_2 - \bar{y}_2) | s) = \frac{W_2(K-1)\sigma_2^2}{n}
\]  

(2.7)

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

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

(2.8)

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

**3. PRODUCT TYPE ESTIMATORS UNDER NON RESPONSES**

**3.1. The basics**

The product estimator is defined by

\[
\bar{y}_p = \frac{\bar{z}y}{\mu_X}
\]  

(3.1)

where \( \bar{z} = \frac{1}{n} \sum_{j=1}^n z_j \); \( z = x, y \), \( \mu_X = \frac{1}{N} \sum_{j=1}^N X_j \).

Its expectation is given by
$$E(\overline{y}_p) = \frac{E(\overline{xy})}{\mu_X \mu_Y} = \mu_Y + \frac{\sigma_{XY}}{nx}$$

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

Hence its has as bias

$$B(\overline{y}_p) = \frac{\sigma_{XY}}{n \mu_X}$$

Its variance is

$$V(\overline{y}_p) = \frac{\sigma^2_Y + R^2 \sigma^2_Z + 2R \sigma_{XY}}{N}$$

where

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

A version of it is

$$\overline{y}_{p^*} = \frac{\sum_{i=1}^{n} x_i y_j}{n \mu_X}$$ (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 \( \mu_0 \) for non responses

Let us consider

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

where

$$\overline{y}_{2p} = \frac{\overline{y}_2}{\mu_X}$$

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
Then the bias is equal to

$$B_{ps} = B(\bar{y}_{ps}) = \frac{\sigma_{2XY}}{n\mu_X}$$

The results obtained previously fix that under the regularity condition

$$R1: \quad \frac{\sigma_{2XY}}{n^2\mu_{2Y}\mu_X} \approx \sum_{j=1}^{N_2} \frac{(y_{2j} - \bar{y}_2)(x_{2j} - \bar{x}_2)}{n'_{2} \mu_{2Y} \mu_X}$$

we have that

$$E\left(\frac{\bar{y}_{ps}}{n'_{2}}\right)_{n_2} = \bar{y} + \frac{\mu_{2X} \rho_{2Y} \sigma_{2Y}}{\partial n}$$

The variance of (3.3) is obtained by calculating

$$V\left(E\left(\frac{\bar{y}_{ps}}{n'_{2}}\right)_{n_2}, \mu_{2}, n_e \right) + V\left(E\left(\frac{\bar{y}_{ps}}{n'_{2}}\right)_{n_2}, n_e \right) + V\left(E\left(\frac{\bar{y}_{ps}}{n'_{2}}\right)_{n_2}, n_e \right)$$

Let us compute the first term

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

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

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

and the second and third ones are equal to zero.

For the second term we have the expression

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

Calculating the conditional variance we obtain
\[ V(\overline{y}_{2p} - \overline{y}_2) \approx V(\overline{y}_{2p}, n_2) + V(\overline{y}_2) - 2Cov(\overline{y}_{2p}, \overline{y}_2) \]

The first two terms are easily derived as

\[ V(\overline{y}_{2p}, n_2) = \frac{\sigma_{2y}^2 + R_2^2 \sigma_{2x}^2 + 2R_2 \sigma_{2xy}}{n_2} \]

\[ V(\overline{y}_2, n_2) = \frac{\sigma_{2y}^2}{n_2} \]

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

\[ Cov(\overline{y}_{2p}, \overline{y}_2) = E\left( \frac{\overline{y}_{2p} - \overline{x}}{\mu_X} \right) \left( \frac{\overline{y}_2 - \mu_{2y}}{\mu_X} \right) - \left( \frac{\mu_{2y} + \frac{\rho_2 C_{2x} C_{2y}}{n_2}}{n_2} \right) \mu_{2y} \]

As

\[ E\left( \frac{y_{2p}^2 - x^2}{n_2} \right) = \mu_{2y}^2 + \frac{2\mu_{2y} \sigma_{2xy} + \mu_X \sigma_{2y}^2}{n_2} + O(n^{-2}) \]

we have that

\[ Cov(\overline{y}_{2p}, \overline{y}_2) \approx \frac{\mu_{2y}^2}{\mu_X} - \left( \frac{\mu_{2y} - \frac{\rho_2 C_{2x} C_{2y}}{n_2}}{n_2} \right) + \frac{2\mu_{2y} \sigma_{2xy} + \sigma_{2y}^2 - \mu_{2y}^2}{n_2} \]

Substituting the terms derived previously we have that

\[ V(\overline{y}_{2p} - \overline{y}_2) \approx \frac{R_2^2 \sigma_{2x}^2 + 2\sigma_{2xy} \left( \frac{R_2 - 2\mu_{2y}}{\mu_{2x}} \right)}{n_2} - 2\left( \frac{\mu_{2y}^2}{\mu_X} \left( \frac{\mu_{2y} - \frac{\rho_2 C_{2x} C_{2y}}{n_2}}{n_2} \right) \right) \]

The analyzed variance term is derived by computing the unconditional expectation

\[ E\left( \frac{n_2}{n} \right)^2 V(\overline{y}_{2p} - \overline{y}_2, n_2) \equiv W_2 (S(1) + S(2)) - 2\lambda_{2,xy} \]

where

\[ S(1) = \frac{R_2^2 \sigma_{2x}^2 + 2\sigma_{2xy} \left( \frac{R_2 - 2\mu_{2y}}{\mu_{2x}} \right)}{n} \]

\[ S(2) = 2\left( \frac{\rho_2 C_{2x} C_{2y}}{n\mu_X} + \frac{\mu_{2y}^2}{n\mu_X} \right) \]

and

\[ \lambda_{2,xy} = \left( \frac{\mu_{2y}^2}{n\mu_X} \right) \left( nW_2^2 + nW_1W_2 \right) \]

The third term of the sampling error is...
\[
E(E(V(\bar{y}_{2p}\mid n'_2)\mid n_2)) = E\left(E\left(\left(\frac{n_2}{n}\right)^2 E\left(\left(\bar{y}'_{2p} - \bar{y}_2\right)^2\mid n'_2\right)\mid n_2\right)\right)
\]

As

\[
\bar{y}'_{2p} - \mu_{2Y} = (\bar{y}'_{2p} - \bar{y}_2) + (\bar{y}_2 - \mu_{2Y})
\]

is derived that

\[
E\left(\left(\bar{y}'_{2p} - \bar{y}_2\right)^2\mid n'_2\right) = E\left((\bar{y}_2 - \mu_{2Y})^2\mid n'_2\right) - E\left((\bar{y}_2 - \mu_{2Y})\mid n'_2\right) = \frac{(1 - \theta)\sigma_{2Y}^2}{\theta n_2}
\]

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

\[
E(E(V(\bar{y}_{2p}\mid n'_2)\mid n_2)) = \frac{W_2 (1 - \theta)\sigma_{2Y}^2}{n \theta}
\]

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

**Lemma 3.1** The estimator (3.3) of \( \mu_Y \) is asymptotically unbiased and its variance is

\[
V(\bar{y}_{2p}) = \frac{\sigma_Y^2}{n} + \frac{W_x \sigma_{p(2)}}{n} + \frac{W_2 (1 - \theta)\sigma_{2Y}^2}{n \theta} - 2\lambda_{2XY}.
\]

where

\[
\sigma_{p(2)} = R_x^2 \sigma_{2X}^2 + 2\sigma_{2XY} \left( R_x^2 - \frac{2\mu_{2Y}}{\mu_{2X}} \right) + \left( \frac{\rho_{2X} C_{2X} - C_{2Y} + \mu_{2Y}^2}{\mu_{2X}} \right)
\]

\[
\lambda_{2XY} = \left( \frac{\mu_{2Y}^2}{\mu_{2X}} \right)
\]

If the regularity condition R1 holds.

**Proof.**

The first result is derived by fixing that

\[
\lim_{n \to \infty} \left( \frac{\sigma_{2XY}}{n \mu_{2X}} \right) = 0
\]

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 \( \mu_Y \) for non responses

We propose as an alternative the estimator

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

(3.7)

It uses the combination of the sub samples. As stated in section 3.1 we will consider the 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 in the original sample. The conditional expectation of the second term is zero. Hence we have that (3.7) is asymptotically unbiased because

\[ EEE(\overline{y}_{pc} | n_1, n_2) = E(\overline{y}_p) = \mu_Y + \mu_Y \left( \frac{\rho C \cdot C_Y}{n} \right) \]

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

The unconditional variance of (3.7) is given by

\[ V(EE(\overline{y}_{pc} | n_1, n_2)) = V(\overline{y}_p) = \sigma_Y^2 + R^2 \sigma_X^2 + 2R \sigma_X \sigma_{XY} \]

It is easily derived that

\[ E[V(E(\overline{y}_{pc} | n_2)] = E(V(\overline{y}_p | n_2) = 0 \]

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

\[ V(\overline{y}_{pc} | n_2) = \left( \frac{n_2 \sigma^2_Y}{n \mu_Y} \right)^2 E(\overline{y}_{2} - \overline{y}_2)^2 | n_2) = \left( \frac{n_2 \sigma^2_Y}{n \mu_Y} \right)^2 \left( (1-\theta) \sigma_{XY}^2 \right) \]

The expectation conditional to a fixed \( n_2 \) is

\[ E\left( \left( \frac{n_1 \sigma X + n_2 \sigma^2_Y}{n} \right)^2 | n_2 \right) = \left( \frac{n_1}{n} \right)^2 \left( \mu^2_Y + \frac{\sigma^2_Y}{n_1} \right) + \left( \frac{n_2}{n} \right)^2 \left( \mu^2_X + \frac{\sigma^2_X}{n_2} \right) + 2 \left( \frac{n_1 n_2}{n^2} \right) \mu_X \mu_{XY} \]

Calculating \( E(n^2), i=1,2, E(n_1 n_2) \), and adding this result to \( V(1) \), after grouping we obtain

\[ V(\overline{y}_{pc}) = \frac{\sigma^2_Y + R^2 \sigma^2_X + 2R \sigma_X \sigma_{XY}}{n} + \frac{(1-\theta) \sigma_{XY}^2}{\theta \sigma^2_X} \left( \mu^2_Y + \frac{W_1 W_2 (\mu_{1X} - \mu_{2X})^2}{\sum_{i=1}^{2} W_i \sigma^2_i} \right) \]

Then we have the following Lemma

**Lemma 3.2** The estimator (3.7) of \( \mu_Y \) 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:
\[ \bar{y}^*_p = \frac{\sum_{i=1}^{n_2} x_i \bar{y}_i}{n_2} \]  
\[ (4.1) \]

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

\[ \bar{y}^* = \frac{n_1 \bar{y}_1 + n_2 \bar{y}^*_p}{n} \]
\[ (4.2) \]

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

\[ E(\bar{y}^*_IC|s) = E \left( \frac{n_1 \mu_1 + n_2 \mu_{2X}}{n} \mu_X \right) = W_1 \mu_1 + W_2 \mu_{2X} \]

Its bias is

\[ B(\bar{y}^*_IC) = W_2 \left( \frac{\mu_{2X} \mu_{1Y}}{\mu_X} - \mu_{2Y} \right) \]

Hence if the population is balanced in the sense

\[ R3: \mu_{2X}=\mu_X \]

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

\[ \bar{y}^* = \bar{y} + n_2 \left( \bar{y}^*_{2p} - \bar{y}^*_2 \right) \]

is easily derived that

\[ V(E(\bar{y}^*_IC|s)) = \left( \frac{\mu_{1Y} \mu_{2X}}{\mu_X} - \mu_{1Y} \right)^2 \frac{W_1 W_2}{n} \]
\[ (4.3) \]

The conditional variance of the estimator is

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

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

\[ V(\bar{y}^*_IC|s) = E \left( \frac{w_1^2 \sigma_{1Y}^2 + w_2^2 \sigma_{2Y}^2}{n^2} \right) = \frac{(nW_1^2 + W_1 W_2) \sigma_{1Y}^2 + (nW_2^2 + W_1 W_2) \sigma_{2Y}^2}{n^2} \]
\[ (4.5) \]

Let us compute the second term of (4.4).

\[ V(\bar{y}^*_{2p} - \bar{y}^*_2|s) = V(\bar{y}^*_{2p}|s) + V(\bar{y}^*_2|s) - 2 Cov(\bar{y}^*_{2p}, \bar{y}^*_2|s) \]
\[ (4.6) \]

Note that
\[ V(\bar{y}_{2p}^*) = \sum_{i=1}^{n_2} \frac{E(x_i y_i) - \mu_{1Y} \mu_{2X} \mu_{2Y}}{n_2^2 \mu_X^2} \]

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

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

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

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

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

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

Doing some algebraic arrangements have that its expected value is

\[ V^* = \frac{(nW_1^2 + W_1W_2) \sigma_{1Y}^2 + (nW_2^2 + W_1W_2) \sigma_{2Y}^2 + W_2 \left( \mu_{1Y}^2 \sigma_{1X}^2 + \mu_{2Y}^2 \mu_{2X}^2 \right)}{n \mu_X^2} + \frac{W_2 \left( \mu_{1Y}^2 \sigma_{1X}^2 + \mu_{2Y}^2 \mu_{2X}^2 \right)}{n^2 \mu_X^2} \]

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

\[ V^{**} = E(n_2^2/n^2) = W_2 \sigma_{2Y}^2/n \]

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

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

when \( n \to \infty \) and the second order regularity condition

**R4:** \( E(n_1^2/n^4) \equiv E(n_1^2/E(n_1^4)), t = 1,..,4, h = 1,..,4 \)

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 are 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 any 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 \( 'i' \):

\[
y_d = \left( \frac{y_1}{x_1} \right) x_i
\]

Liu et al. (2006) proposed to use

\[
y_{il} = \left( \frac{1}{n_1} \sum_{j=1}^{n_1} \frac{y_j}{x_j} \right) x_i
\]

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

\[
y_{ii}^* = \frac{y_1 x_1}{\mu_X} x_j
\]

for the missing observation \( i \). Its expectation is

\[
E(y_{ii}^*) = \left( \frac{\sigma_{1XY} x_1}{n_1} + \mu_{1X} \mu_{1Y} \right) \frac{\mu_2X}{\mu_X}
\]

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

\[
E \left( \frac{\sum_{i=1}^{n_2} y_{ii}^*}{n_2} \right) = \left( \frac{\sigma_{1XY} x_1}{nW_1} + \mu_{1X} \mu_{1Y} \right) \frac{\mu_2X}{\mu_X}
\]

For improving the simplicity in the reasoning let us consider the estimator of \( \mu_Y \)

\[
\bar{y}_{IS} = \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2^*}{n}
\]

using the expression

\[
\bar{y}_{IS} = \bar{y} + \frac{n_2 (\bar{y}_2^* - \bar{y})}{n}
\]

Its conditional mean is given by

\[
E(\bar{y}_{IS} | s) = \frac{w_1 \mu_Y + w_2 \mu_{2Y}}{n} + \frac{n_2 \left( \frac{\mu_{2X} \sigma_{1XY}}{n_1 \mu_X} + \frac{\mu_{1X} \mu_{1Y} \mu_{2X}}{\mu_X} - \mu_{2X} \right)}{n}
\]

\[\text{(4.8)}\]
Calculating the expected value of this last expression we have that

\[ E(\tilde{E}_{\text{IM}}(s)) = \mu_Y + W_2 \left( \frac{\mu_{1X} \mu_{1Y} \mu_{2X} - \mu_{2X}}{\mu_X} \right) + \frac{\mu_{2X} \sigma_{1XY}}{n \mu_X} E \left( \frac{n_2}{n_1} \right) \]

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

\[ E(\tilde{E}_{\text{IM}}(s)) = \mu_Y + W_2 \left( \frac{\mu_{1X} \mu_{1Y} \mu_{2X} - \mu_{2X}}{\mu_X} \right) + \frac{W_2 \mu_{2X} \sigma_{1XY}}{n W_1 \mu_X} \]

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

\[ V(\tilde{E}_{\text{IM}}(s)) \equiv \frac{W_1 W_2}{n} \left( \frac{n_2}{n} \right)^2 \left( \frac{\mu_{1Y}^2 + \mu_{2Y}^2}{n \mu_X^2} \right)^2 = V(1 IPS) \]

(4.10)

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

\[ V(\tilde{E}_{\text{IM}}(s)) = V(\tilde{E}(s)) + \left( \frac{n_2}{n} \right)^2 V(\left( \tilde{E}^{**P} - \tilde{E}_2 \right)) + 2 \left( \frac{n_2}{n} \right)^2 \text{Cov}(\tilde{E}(s), \left( \tilde{E}^{**P} - \tilde{E}_2 \right)) \]

The first term is equal to zero and

\[ V(2P) = V(\tilde{E}^{**P} - \tilde{E}_2(s)) = E(\tilde{E}^{**P} | s)^2 E(\tilde{E}^{**P} | s)^2 - 2E(\tilde{E}^{**P} | \tilde{E}_2(s)) - E(\tilde{E}^{**P} - \tilde{E}_2 | s)^2 \]

We computed the terms and arrange the similar terms. Afterwards the unconditional expectation was calculated. Assuming that the regularity conditions R4 and R5: \( W_i \equiv 0 \) for \( i \geq 3 \)

R5: \( W_i \equiv 0 \) for \( i \geq 3 \)

Hold we have that

\[ EV_{2P} \equiv V(1 ps) + V(2 ps) + V(3 ps) + V(4 ps) = V(1 IPS) \]

where

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

\[ V(2 ps) \equiv W_2 \left( \mu_{1X} \mu_{1Y} \frac{\mu_{2X}}{\mu_X} \right)^2 + W_1 \mu_{2Y} \left( \frac{\mu_{1X} \mu_{1Y} \mu_{2X}}{\mu_X} \right)^2 - \frac{2(\mu_{1X} \mu_{1Y} + \mu_{1X} \mu_{1Y} \mu_{2X})}{n W_1 \mu_X} \]

\[ V(3 ps) \equiv \left( \frac{\sigma_{1XY} \mu_{21Y}}{\mu_X} \right)^2 + \frac{2 W_2 \sigma_{1XY} \mu_{1X} \mu_{1Y} \mu_{2X}}{n W_1 \mu_X^2} \]

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

Lemma 4.2 Fixes the behavior of the separate imputation product estimator.
Lemma 4.2. The estimator (4.8) is biased. We should prefer $\bar{y}$ 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 $\theta^*_n$ 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 $\pi$. Usually the methods are supported by a particular Central Limit Theorem that must establish that when $m \to \infty$

$$\text{Prob} \left( \theta \in I(\theta) = (\theta(\hat{F}_m) - z_{1-\alpha/2} \sigma_m(\theta(\hat{F}_m)), (\theta(\hat{F}_m) + z_{1-\alpha/2} \sigma_m(\theta(\hat{F}_m))) \right) \geq \pi$$

$\theta(\hat{F}_m)$ is the estimator (predictor) of the parameter, $z_{1-\alpha/2}$ is the percentile of the Standard Normal and $\sigma_m(\theta(\hat{F}_m))$ is the standard deviation estimator of $\sigma(\theta(F))$. The robustness of $\theta(\hat{F}_m)$ and $\sigma_m(\theta(\hat{F}_m))$ play a key role in the validity that $\pi$ 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 $Z = (Z_1, Z_2, \ldots, Z_m)$ be a random sample from an unknown distribution $F$, and let $T_m = T_m(Z, F)$ be a statistic of interest. Let $F_m$ be the empirical distribution function of the random sample. An independent random sample from $F_m$, $Z^*$, is called a Bootstrap sample. We can use the Bootstrap method for estimating the distribution of $T_m$ through the conditional distribution of $T_m(Z_m) = T_m(Z, F_m)$, given $Z_1, Z_2, \ldots, Z_m$. 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 $T^*_m = T(F^*_{B(m)}) = T(Z^*_1, \ldots, Z^*_m)$ estimates $T(F_m)$. Due to the definitions, the conditional independence is supported and

$$\text{Prob}(Z^*_1 = Z_1 | F_m) = 1/m, \ \forall t = 1, \ldots, m, \ i = 1, \ldots, m.$$ Each sample $s(b) \in S(BS)$, $S(BS)$ the space of the Bootstrap samples, is drawn with a probability $1/m$, hence

$$E(T^*_B(m) | F_m) = m^{-m} \sum_{s(b) \in S(BS)} T(Z^*_1, \ldots, Z^*_m) = m^{-m} \sum_{s(b) \in S(BS)} T(B(m)),$$

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

$$V^*_B(m) = \left( \sum_{b=1}^B (T(B(m)) - T_m)^2 \right)^{1/2} = \sigma^*_B(m)$$

It is expected , if the functional is smooth, that the limit of $\sigma^*_B(m)$ is the true variance of the estimator (predictor). A Central Limit Theorem supports that...
Prob \( (\theta \in I^*(\theta) = (\theta(F_m) - z_{1-\alpha/2} \cdot \sigma_{B(m)}(\theta(F_m)), \theta(F_m) + z_{1-\alpha/2} \cdot \sigma_{B(m)}(\theta(F_m))) \geq \pi \)

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

\[ L(Z_1, ..., Z_m) = L_m = \text{Sup} \{ t \mid F_{\theta}(z) \geq t \} \]
\[ U(Z_1, ..., Z_m) = U_m = \text{Inf} \{ t \mid F_{\theta}(z) \leq t \} \]

The interval \((L_m, U_m)\) has random bounds and the coverage probability of \( \theta, \pi \) is such that

\[ \text{Prob}_{(\theta \in (L_m, U_m))} \geq \pi, \text{ for any } \theta. \]

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

\[ \text{Prob} \{ \theta(F) \in (L_m, U_m) = I(\theta) \mid F \in \Upsilon \} \geq \pi \]

The Bootstrap distribution allows to calculate directly the quantiles

\[ F_m^* (t) = B^{-1} \sum_{b=1}^{B} I((T_{B(m)} - T_m)m^{-1/2} \leq t), \quad t \in \mathbb{R} \]

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

\[ \text{Prob} \{ (T(F_m) - T(F)m^{-1/2} \leq t \} \rightarrow G(t) \]

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

\[ I(\mu_Y)_q = (\hat{\mu}_{Y_{(q)}} - \varepsilon_{Y_{(q)}}, \hat{\mu}_{Y_{(q)}} + \varepsilon_{Y_{(q)}}) \]

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 Bootstrap is used

\( q=3 \) if the Non-parametric Bootstrap is used

\( v=q \) separate product estimator, combined product estimator, separate imputation predictor, combined imputation predictor.

\( \varepsilon_{Y_{(q)}} \) is the semi-amplitude of the interval calculated using the corresponding method \( q \) for the estimator \( v \) with \( \alpha=0.05 \).

Experiment 1

<table>
<thead>
<tr>
<th>Strata</th>
<th>( W_j )</th>
<th>Mean of the auxiliary variable</th>
<th>Variance of the auxiliary variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.682</td>
<td>66.39</td>
<td>58.9</td>
</tr>
<tr>
<td>2</td>
<td>0.372</td>
<td>131.83</td>
<td>16.2</td>
</tr>
</tbody>
</table>

Table 1. Parameters of the strata
We compared the different proposals developed in this paper using a database 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.

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.

<table>
<thead>
<tr>
<th>$q=1$ Normal Approximation</th>
<th>$q=2, B=20$ Parametric Bootstrap</th>
<th>$q=3, B=20$ Non-parametric Bootstrap</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{y}_{ps}$</td>
<td>0.80</td>
<td>0.85</td>
</tr>
<tr>
<td>$\bar{y}_{psC}$</td>
<td>0.60</td>
<td>0.70</td>
</tr>
<tr>
<td>$\bar{y}_{ls}$</td>
<td>0.41</td>
<td>0.59</td>
</tr>
<tr>
<td>$\bar{y}_{IC}$</td>
<td>0.75</td>
<td>0.79</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$q=1$ Normal Approximation</th>
<th>$q=2, B=100$ Parametric Bootstrap</th>
<th>$q=3, B=100$ Non-parametric Bootstrap</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N(4.9, 0.586)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\bar{y}_{ps}$</td>
<td>0.83</td>
<td>0.89</td>
<td>0.93</td>
</tr>
<tr>
<td>$\bar{y}_{psC}$</td>
<td>0.71</td>
<td>0.81</td>
<td>0.87</td>
</tr>
<tr>
<td>$\bar{y}_{ls}$</td>
<td>0.44</td>
<td>0.52</td>
<td>0.58</td>
</tr>
<tr>
<td>$\bar{y}_{IC}$</td>
<td>0.49</td>
<td>0.57</td>
<td>0.68</td>
</tr>
<tr>
<td>$logN(4.9, 0.586)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\bar{y}_{ps}$</td>
<td>0.87</td>
<td>0.88</td>
<td>0.94</td>
</tr>
<tr>
<td>$\bar{y}_{psC}$</td>
<td>0.81</td>
<td>0.85</td>
<td>0.89</td>
</tr>
<tr>
<td>$\bar{y}_{ls}$</td>
<td>0.72</td>
<td>0.77</td>
<td>0.80</td>
</tr>
<tr>
<td>$\bar{y}_{IC}$</td>
<td>0.66</td>
<td>0.70</td>
<td>0.79</td>
</tr>
<tr>
<td>$Exp(4.9)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\bar{y}_{ps}$</td>
<td>0.74</td>
<td>0.79</td>
<td>0.92</td>
</tr>
<tr>
<td>$\bar{y}_{psC}$</td>
<td>0.67</td>
<td>0.71</td>
<td>0.89</td>
</tr>
<tr>
<td>$\bar{y}_{ls}$</td>
<td>0.53</td>
<td>0.65</td>
<td>0.71</td>
</tr>
<tr>
<td>$\bar{y}_{IC}$</td>
<td>0.45</td>
<td>0.56</td>
<td>0.67</td>
</tr>
</tbody>
</table>

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

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


