MISSING OBSERVATION FOR ESTIMATING THE DIFFERENCE: RANKED SET SAMPLING VERSUS SIMPLE RANDOM SAMPLING

Carlos Narciso BOUZA

ABSTRACT: The estimation of the difference of means of two variables is studied when missing observations are present. The model follows the approach proposed in the paper of Bouza and Prabhu-Ajgaonkar [1993, Biometrical J., v.35, p.245-252] when simple random sampling is the design used for selecting samples and sub-samples. In this paper, a ranked set-sampling counterpart is developed. Their sample errors are compared analytically. A numerical comparison is developed for evaluating the magnitude of accuracy gain due to the use of ranked set sampling.

Keywords: Sub-sampling; relative accuracy; double sampling; order statistics.

1 Introduction

McIntire (1952) proposed the ranked set sampling method (rss). He proposed that units may be ranked visually and noticed the existence of a gain in accuracy with respect to the use of the sample mean when simple random sampling with replacement (srswr) was used. Dell and Clutter (1972) and Takahasi and Wakimoto (1968) provided mathematical support to his claims. The following algorithm provides a description of ranked set sampling [rss].

Rss implementation

Input $r$, $m$

$i=0$ and $t=0$

While $t<r+1$ do

While $i<m+1$ do

Select a sample $s_i$ of size $|s_i|=m$ using srswr

Rank the sampled units with respect to the variable of interest $\xi=Y$

Measure $Y$ in the unit with rank $i$ $(\xi_{(i):m})$

$i=i+1$

End

$t=t+1$

End

1Facultad de Matemática y Computación, Universidad de La Habana. San Lázaro y L. CP 10400. Habana, Cuba, E-mail: bouza@matcom.ub.cu

The rss sample is the sequence of order statistics (os) $\xi_{(1)}, \ldots, \xi_{(m)}$, where $(j:h)$ denotes the statistic of order $j$ in the $h$-th sample in the cycle $t=1,\ldots,r$. We have $n=mr$ observations and $r$ of them are of the $i$-th order statistics, $i=1,\ldots,m$. Take $\mu_i$ as the mean of a variable of interest $x$, its rss estimator is

$$\hat{\mu}_{(rss)} = \frac{\sum_{i=1}^{m} \xi_{(i:m)}}{rm}$$

and its variance is given by

$$V(\hat{\mu}_{(rss)}) = \frac{\sum_{i=1}^{m} \sigma_i^2}{rm}$$

where

$$\sigma_i^2 = E[\xi_{(i:m)} - E(\xi_{(i:m)})]^2$$

and

$$\Delta_i = E[\xi_{(i:m)}] - \mu_i$$

The second term of (1.2) is the gain in accuracy due to the use of rss instead of srswr.

The objective of this paper is the study of the behavior of rss in the estimation of the difference of the means of two random variables $D=\mu_X - \mu_Y$ when there are missing observations. In this case, we are involved with two variables of interest $X, Y$. An antecedent is the paper of Pi-Ehr (1971) where the normality of $X$ and $Y$ is assumed. Bouza (1983) extended his results to finite population sampling when srs is used and missing observations are considered as non-responses.

Bouza (2002) developed an rss alternative to that strategy. Bouza and Prabhu-Aigaonkar (1993) reanalyzed this problem using a different approximation technique. They assumed:

1) There is dependence between $X$ and $Y$ only within the units, but not between them.
2) A response is obtained in at least one of the variables in each sampled unit.

This paper is devoted to the development of an rss counterpart of this model. Session 2 presents the sample strategy for estimating $D$ using the $srs$ design and its $rss$ counterpart is developed in session 3. A discussion based on Monte Carlo experiments appears in session 4, while session 5 is concerned with the presentation of a set of conclusions derived from the results of the paper.

2 The SRS model for estimating $D$

The model considers that the population is stratified: $U=U_1 \cup U_2 \cup U_3$, $U_j \cap U_j' = \emptyset, \forall i \neq j, j'=1,2,3$. When sampled, the units in stratum $U_j$ report $X$ and $Y$, the units in $U_2$ only report $X$ while $Y$ is the variable reported by the units of $U_j$ at the first visit. The units in sample $s$ can be denoted as $s=s_j \cup s_2 \cup s_j', s_j \cap s_j' = \emptyset, \forall i \neq j, j'=1,2,3$. The sample size is $|s|=n=rm$. Therefore, the units in $s_j$ give information on $X$ and $Y$, but we have missing information of $Y$ of those in $s_2$ and in $X$ by the respondents in $s_j'$. Without losing in generality we can rearrange the units in $s$ and denoting the sample size of $s_j$ by $|s_j|$.
Each subsample $s_t$ belongs to a stratum $U_t$, $t=1, 2, 3$. The information provided by $s$ permits to calculate

$$
\bar{x}_1 = \frac{\sum_{i=1}^{n_1} x_i}{n_1}, \quad \bar{y}_1 = \frac{\sum_{i=1}^{n_1} y_i}{n_1}, \quad \bar{d}_1 = \frac{\sum_{i=1}^{n_1} (x_i - y_i)}{n_1}, \quad \bar{x}_2 = \frac{\sum_{i=n_1+1}^{n_1+n_2} x_i}{n_2}, \quad \bar{y}_2 = \frac{\sum_{i=n_1+1}^{n_1+n_2} y_i}{n_2}, \quad \bar{x}_3 = \frac{\sum_{i=n_1+n_2+1}^{n} x_i}{n_3}, \quad \bar{y}_3 = \frac{\sum_{i=n_1+n_2+1}^{n} y_i}{n_3}
$$

The need to obtain information from the non-respondents establishes that subsamples $s_j$, $j=2, 3$, should be resampled for obtaining it. This decision is reasonable when we expect that the means and variances of the variables in the strata are very different.

Denote by $s'_j \subset s_j$ the corresponding sub-sample of size $n'_j$, $j=2, 3$ selected from the corresponding $s_j$. Using the notation of Bouza and Prabhu-Ajgaonkar (1993) and using the rule of Hansen-Hurwitz, see Hansen and Hurwitz (1953), where $n'_j = n_j/K_j, K_j > 1:$

$$\bar{x}^* = \frac{\sum_{i=1}^{n_1} x_i + n_2 \bar{x}'_2}{n} = \frac{\sum_{i=1}^{n_1} x_i + n_2 \bar{x}'_2}{n}, \quad \bar{y}^* = \frac{\sum_{i=1}^{n_1} y_i + n_3 \bar{y}'_3}{n} = \frac{\sum_{i=1}^{n_1} y_i + n_3 \bar{y}'_3}{n}
$$

where

$$\bar{y}'_3 = \frac{\sum_{i=n_1+1}^{n_1+n_2} y_i}{n'_3}, \quad \bar{x}'_2 = \frac{\sum_{i=n_1+1}^{n_1+n_2} x_i}{n'_2}, \quad \bar{x}'_2 = \frac{\sum_{i=n_1+1}^{n_1+n_2} x_i}{n'_2}$$

and $w_i = n_i/n$.

We will establish its properties in the following proposition

**Proposition 2.1.** [Bouza and Prabhu-Ajgaonkar (1993)]. The estimator of $D$

$$d_{SRS} = \bar{x}^* - \bar{y}^* \tag{2.1}$$

With

\[ z^* = \sum_{i=1}^{3} w_i \bar{z}_j + w_j (\bar{z}_j - \bar{z}) \text{,} \quad j = 2, 3 \quad \text{if} \quad Z = Y(X) \]

is unbiased if srs is used for selecting the samples and sub-samples. Its expected variance under Hansen-Hurwitz rule \( n'_j = n/K_j \), \( K_j > 1 \), \( j = 2, 3 \) is

\[ \overline{V} = \frac{\sigma^2_j + W_2 (K_2 - 1)\sigma^2_{2y} + W_3 (K_3 - 1)\sigma^2_{3y}}{n} \]

where \( W_j = \frac{N_j}{N} = \frac{\text{number of units in } U_j}{N} \) and \( N = N_1 + N_2 + N_3 \)

**Proof:** We have that the conditional expectations of the subsample means are

\[ E(\bar{z}_j | s) = \bar{z}_j \quad j = 2, 3 \quad Z = Y(X) \text{ then} \quad E[E(d_{r,s} | s)] = E(\bar{x} - \bar{y}) = \Omega \]

The unbiasedness of the estimator sustains that the sampling error is the variance. Note that the first term in the proposed estimator of the mean of \( Z \) is the sample mean then, for a fixed \( s \) we have

\[ V(\bar{z}^* | s) = V(\bar{z}) + w_j V(\bar{z}_j - \bar{z}) = \frac{\sigma^2}{n} + \frac{w_j (K_j - 1)\sigma^2_{jy}}{n} \quad (2.3) \]

The conditional variance of (2.1) is function of the variances and covariance of \( X \) and \( Y \) in the units that give responses. Therefore

\[ V(d_{r,s} | s) = V(\bar{x}^* | s) + V(\bar{y}^* | s) - 2Cov(\bar{x}^*, \bar{y}^* | s) \]

where

\[ Cov(\bar{x}^*, \bar{y}^* | s) = Cov(\bar{x}, \bar{y} | s) = \rho \sigma_x \sigma_y \]

As \( s_j \) is sample selected from \( U_j \), we have that \( E(w_j) = N_j/N = W_j \) and

\[ V(\bar{d}) = \frac{\sum_{j=2}^{3} \sigma^2_j + \sigma^2_y - 2\rho \sigma_x \sigma_y}{n} \]

Adding this result to the expectations of the variances of the estimators of the mean of the variables \( X \) and \( Y \), we easily obtain that the expected variance of (2.1) is (2.2) \( \blacksquare \)
3 Rss estimation of D

One of the practical problems that arise when rss is applied is to obtain an accurate and cheap ranking procedure. If it is not accurate, we will have errors in the rankings. Dell and Clutter (1972) established that in such case we obtain not the \( i \)-th os, but the \('ith – judgmental\) one. The use of a concomitant variable is supported by the existence of a correlation between the true variable and the ranking variable. Stokes (1977) and Patil et. al (1995a) studied the use of concomitant variables and used them for ranking.

First we consider the case of sampling using rss for estimating \( D \). From the rss theory the following proposition is easily obtained.

Proposition 3.1 [Bouza, 2002]. An unbiased estimator of \( D \), when rss is used is

\[
d_{rss} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} d_{(im)j}}{rm}
\]

with

\[
V(d_{rss}) = \frac{\sigma_d^2}{rm} \frac{\sum_{i=1}^{m} \Delta_d^2(\im)}{rm^2} = \frac{\sigma_d^2}{n} \frac{\Delta_d^2}{n}
\]

as variance.

Proof: It is easily obtained from (1.2) considering \( \xi_i = d_i \).

The positiveness of \( \Delta^2 \) grants that the rss design, \( d_{rss} \) is more accurate than its srs counterpart. Bouza (2001) studied the effect of using rss for subsampling the non-respondent (NR) stratum. We will extend them to the estimation of the difference. Bouza (2002) studied this problem considering the use of two-phase sampling for stratification under assumptions on the sub-sample sizes and the expectation of some their function. In this paper, we take the usual non-response stratum approach and consider that:

H1. There is dependence between \( X \) and \( Y \) only within the units but not between them.
H2. A response is obtained in at least one of the order statistics in each sampled unit.

Let us take \( r \) as fixed, \( n_{j}=rm_{j} \), then the subsampling size among the non-respondents of stratum \( U_j \). As \( n'_{j}=rm'_{j}=rm/K_j \). It makes sense to use as a concomitant variable \( X (Y) \) if \( j=2(3) \). Considering the conditional unbiasedness of

\[
\bar{z}_{jrss} = \frac{\sum_{i=1}^{m_j} \sum_{i=1}^{rm_j} z_{(im)j}'}{rm_j'}, \quad j = 2(3), \quad if \quad Z = Y (X)
\]
Mimicking the structure of \( dsrs \) we consider the estimator.

\[
\bar{d}_{rs} = w_1 d_{1rs} + w_2 (x_{2rs} - \bar{y}_{2rs}) + w_3 (x_{3rs} - \bar{y}_{3rs})
\]

where

\[
d_{1rs} = \frac{\sum_{j=1}^{m_1} \sum_{i=1}^{n_1} d_{ijy}}{n_1}
\]

(3.3)

Let us prove the unbiasedness of (3.3).

**Proposition 3.2.** Under the described \( rss \) design and if \( H1-H2 \) holds the estimator (3.3) is unbiased and its expected variance is

\[
E[V(\bar{d}_{rs})] = \frac{\sum_{j=1}^{m_1} W_j [K_j - 1] \Delta^2 Y_{(j,m_1)}}{m_2} + \frac{\sum_{j=1}^{m_1} W_j [K_j - 1] \Delta^2 X_{(j,m_1)}}{m_3}
\]

(3.4)

**Proof:** We can rewrite the estimator as:

\[
\bar{d}_{rs} = w_1 d_{1rs} + \sum_{j=2}^{3} w_j (x_{jrs} - \bar{y}_{jrs}) + w_2 (x_{2rs} - \bar{y}_{2rs}) + w_3 (x_{3rs} - \bar{y}_{3rs})
\]

(3.5)

and

\[
E(\bar{d}_{rs} | s) = \bar{y}_{jrs}, \quad j = 2(3) \quad if \quad Z = Y(X)
\]

The sum of the two first terms of (3.5) is equal to \( d_{rs} \), hence the conditional variance with expectation equal to (3.2). Let us analyze the second term.

\[
E(\bar{y} - \bar{y}_{2rs} | s) = E(\bar{y} - \bar{y}_{2rs} | s) \cdot \sigma_{2y}^2
\]

because

\[
\bar{y}_{2rs} - \mu_{2y} = (\bar{y}_{2rs} - \bar{y}_{2rs}) + (\bar{y}_{2rs} - \mu_{2y})
\]

The cross expectations are zero. In this case, \( rss \) is balanced and we can express the variance of \( os \) as a function of the variance of \( Y \) in \( U_z \) and the gains in accuracy measured by the \( \Delta^2_{2Y(i,m_2)} \) as

\[
V(\bar{y} - \bar{y}_{2rs} | s) = \sigma_{2y}^2 \left( \frac{1}{n_2} - \frac{1}{n_2} \right) - \sum_{i=1}^{m_2} \frac{\Delta^2_{2Y(i)}}{n_2 m_2}
\]

Substituting \( n_2 = rm_2/K_2 \) we obtain:
A similar reasoning with the last term yields
\[ V\left(\bar{y}_{rss} - \bar{y}_{srss}\right) = \frac{\sigma_{2Y}}{r} \left( \frac{K_{z} - 1}{m_{2}} \right) - \sum_{i=1}^{m} \frac{\Delta^{2Y(i,m_{2})}}{rm_{2}} \left( \frac{K_{z} - 1}{m_{2}} \right) = V_{2} \]

A similar reasoning with the last term yields
\[ V\left(\bar{y}_{3rss} - \bar{y}_{3srss}\right) = \frac{\sigma_{3X}}{r} \left( \frac{K_{z} - 1}{m_{3}} \right) - \sum_{i=1}^{m} \frac{\Delta^{3X(i,m_{3})}}{rm_{3}} \left( \frac{K_{z} - 1}{m_{3}} \right) = V_{3} \]

Now we have that
\[ V\left(\bar{d}_{rss}\right) = V_{1} + w_{2}^{2}V_{2} + w_{3}^{2}V_{3} \]

as \( w_{j}^{2}/rm_{j}=n/n^{2} \) is a Binomial random variable with expectation \( nW_{j} = nN/N \) we have that
\[ E\left[V\left(\bar{d}_{rss}\right)\right] = \frac{\sigma_{2Y}}{n} + \frac{W_{2}(K_{z} - 1)\sigma_{2Y}}{n} + \frac{W_{3}(K_{z} - 1)\sigma_{3X}}{n} - \left( \frac{\Delta^{2}_{j}}{n} + \Psi \right) \]

where
\[ \Psi = E \left( \frac{\sum_{i=1}^{m} W_{2}[K_{z} - 1] \Delta^{2Y(i,m_{2})}}{m_{2}} + \frac{\sum_{i=1}^{m} W_{3}[K_{z} - 1] \Delta^{3X(i,m_{3})}}{m_{3}} \right) \]

which completes the proof.  

The last term of the expression is always positive and represents the gain in accuracy due to rss. Hence the use of rss for subsampling the non respondents when \( D \) is estimated is a better alternative than the srswr strategy.

The expected variance derived in Proposition 3.2 is similar to the results of Bouza (2002), except that the factors of the stratum variances \( \sigma_{jz}^{2} \) and gains in accuracy \( \Delta^{2}_{jz} \) for, \( Z=X, (Y) \) if \( j=2 \) (3), \( K_{z}-I \) should be replaced by \( K_{j} \). The results of Bouza (2002) were based on the assumptions:

A1. \( E[n_{j}^{1/4}] \approx E[n_{j}^{1/4}], j=1, 2, 3. \)

A2. \( n_{j} \propto W_{j}, l=1, 2, 3. \)

A3. \( N \) is large.

Note that A1-A3 may be more restrictive than H1-H2. For example if we are studying the effect of a medication in a population of patients we can assume independence of the results before (X) and after (Y) among them, but it is logical to accept that there is a correlation between the variables within patients. \( N \) is not necessarily large and the number of non-respondents does not need to be proportional to the stratum size.
Then we can use the approach developed in this paper instead of the strategy of Bouza (2002).

4 Analysis of a Monte Carlo experiment.

We will compare the accuracy of the proposed rss sampling strategy with its srs counterpart. The data used were obtained from a national inquiry developed for determining the effect of AcM murino of isotope IgG2a developed by Centro de Inmunología Molecular of Cuba which should have an effect on the reduction of the area affected by psoriasis. A sample of 200 patients was selected and a longitudinal survey was developed, the patients were evaluated in 13 occasions, see Viada et.al. (2004) for details. The index called PASI (Psoriasis Severity Index), see Lauprakis et.al. (1988), was determined in each visit. We considered:

\[ X = \text{Value of the index at the first visit} \]
\[ Y = \text{value of the index at the end of the treatment}. \]

The set of measurements of PASI constituted an artificial population. It was partitioned using the non-responses at the second \((U_2)\) and third visits \((U_3)\). We selected a sample from \(s\) and the subsamples were determined by classifying a selected patient in \(U_i\) if he/she assisted to the evaluation on occasions 1, 2 and 3, in \(U_2\) if the second visit failed and in \(U_3\) if the third one failed. The last evaluation was made for all the patients.

100 samples were generated and with sample fractions of 0.10, 0.20 and 0.50 using \(srs\) and \(rss\). \(D\) was computed and estimated using \(d_{srs}\) and \(d_{rss}\). The relative accuracy:

\[ G_b = \sum_{i=1}^{100} \frac{h_i - \bar{y}}{100D} \quad b = srs, rss \]

was used for measuring the behavior of the alternative estimators. The results appear in Table 4.1 for \(K_j = K = 2, j = 2, 3\).

Table 4.1 - Relative mean Accuracy in Percent

<table>
<thead>
<tr>
<th>Sample fraction</th>
<th>100(G_{srs})</th>
<th>100(G_{rss})</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>13.12</td>
<td>4.45</td>
</tr>
<tr>
<td>0.20</td>
<td>9.03</td>
<td>2.24</td>
</tr>
<tr>
<td>0.50</td>
<td>8.17</td>
<td>1.95</td>
</tr>
</tbody>
</table>

That \(rss\) provided more accurate estimations than \(srs\) were expected. These results give an idea of how large the gains are. They are increased with the increase in the sample fractions. A similar result is expected for other sets of values of the sub-sampling fractions.

Conclusions

We can mention some theoretical results which should be taken into account for using the proposed model:

1) The use of the proposed rss model is more accurate than the classic srs one.
2) The proposed model relies on assumptions on the behavior of the independence of the involved variables and not on the size of the population or on properties of the subsample sizes.

The use of \( r_{ss} \) in the estimation of a difference with non-responses seems to be a natural approach because having information on one of the variables allows to rank without an extra effort by the non-respondents leading to an increase in the accuracy of the estimators.

Acknowledgements

This paper was developed within a project with the Institute of Mathematics of China supported by TWAS. This version of the paper has been benefited thanks to the comments and suggestions of two anonymous referees.


PALAVRAS-CHAVE: Subamostras; aproximação relativa; dupla amostragem; estatísticas de ordem.

References


Aprovado após revisão em 08.02.2006.