RANKED SET ESTIMATION WITH IMPUTATION OF THE MISSING OBSERVATIONS: THE MEDIAN ESTIMATOR

Carlos N. Bouza Herrera* and Amer Ibrahim Al-Omari** *
Universidad de La Habana, La Habana, Cuba.
** Al al-Bayt University, Faculty of Sciences, Department of Mathematics, Mafraq, Jordan

ABSTRACT
The existence of missing data in sample surveys reduces the efficiency of the estimators. This paper is concerned with the use of imputation for dealing with this problem. The mean imputation and the ratio methods for the median estimator are used for developing alternatives with respect to the full response case in ranked set sampling.

RESUMEN
La existencia de datos perdidos en una encuesta por muestreo reduce la eficiencia de los estimadores. Este trabajo es dedicado al desarrollo de métodos de imputación para tratar este problema. La imputación mediante la media y la razón para el estimador de la mediana, en el muestreo por conjuntos ordenados, es desarrollada como alternativa respecto al caso en el que no hay no respuestas.

KEY WORDS: mean imputation meted, ratio imputation meted, order statistics.

MSC 62D05

1. INTRODUCTION

Ranked Set sampling (RSS) was proposed by McIntyre (1952) for estimating pastures yields and claimed that it was better than using simple random sampling with replacement (SRSWR). The basic ideas on the method can be obtained in Chen et al. (2004) and in several review papers as Bai- Chen (2003) and Bouza (2005), for example
The theoretical frame that permits to use the rss model is based on the hypothesis

- We wish to enumerate the variable of interest Y.
- The units can be ordered linearly without ties.
- Any sample $s \subset U$ of size m can be enumerated.
- To identify a unit, order the units in s and enumerate them is less costly than to evaluate $\{Y_i, i \in s\}$ or to order U.

The first hypothesis is common to the general sampling problem, the second fixes that the rank can be made without confusions and that any rank is assigned to only one of the sampled units. The third assumption is also common in the applications. The fourth has an economical and a statistical motivation: only if it is cheap to rank rss is a good alternative with respect to rank all the units of U and to stratify.

The basic selection procedure can be described as follows:

Procedure RSS for a ranked set sample of size *n* generation
While $j<m$ do

While $t<m$ do

Select a ssu independently from U using srswr.

Each unit in $s_{(t)}$ is ranked and the os's $Y_{(1:t)},\ldots, Y_{(r(t):t)}$ are determined.

END

End.

Muttlak (1997) suggested median ranked set sampling (MRSS). A sample of size n is selected using SRSWR from a distribution with density function $f(y)$ and distribution function $F(y)$ with expectation $E(Y)=\mu_Y$ and variance σ_Y^2 . Section 2 is devoted to present the results that sustain the median estimator in RSS. The existence of missing data in sample surveys reduces the efficiency of the estimators. Section 3 is concerned with the use of the mean imputation method for the median estimator. In the last section, develops a similar study using ratio imputation. For more about RSS see Jemain and Al-Omari (2006), Samawi and Muttlak (1996), Muttlak (2003).

2. MEDIAN RSS

Muttlak (1997 and 2003) proposed to select the median of $s(j)$ in each ssu. The pdf of Y must have finite mean and variances μ and σ^2. We observe $\{Y_{(1:1),\ldots,}\ Y_{(1:n),}\ Y_{(2:1),\ldots,}\ Y_{(n:n)}\}_m$, $m=1,\ldots,r$. If n is odd the os's measured are $\{Y^*_{(i:med)m}=Y_{([n+1]/2:jm)}, j=1,\ldots,n, m=1,\ldots,r\}$. If n is even is used

$$Y^*_{(i:med)m}=\begin{cases} Y_{\left(\frac{n}{2}:j\right)m} & \text{if} \quad j\le\dfrac{n}{2} \\ Y_{\left(\frac{n+1}{2}:j\right)m} & \text{otherwise} \end{cases}$$

The estimator is:

$$\mu_{rss\ med} = \sum_{j=1}^{n}\sum_{m=1}^{r} Y^*_{(i:med)m} / nr$$

and its expectation is

$$E(\mu_{rss[med]})=\sum_{j=1}^{n}\mu_{(j:med)}/n$$

For n odd $\mu_{(j:med)}=\mu_{([n+1]/2)}$ for any j, then

$$E(\mu_{rss[med]})=\mu_{([n+1]/2)}.$$

If n is even

$$Y^*_{(i:med)m}=\begin{cases} Y_{\left(\frac{n}{2}:j\right)m} & \textit{if} \quad j\le\dfrac{n}{2} \\ Y_{\left(\frac{n+2}{2}:j\right)m} & \textit{otherwise} \end{cases}.$$

Then

$$E\ \mu_{rss[med]} = \frac{1}{2}\left(\mu_{\left(\frac{n}{2}\right)}+\mu_{\left(\frac{n}{2}+1\right)}\right)$$

The variance is

$$V(\mu_{rss[med]})=\sum_{j=1}^{n}\sigma^2_{(j:med)}/n^2r=\sigma^2-\Sigma_{j=1}^{n}\Delta^2_{(j:med),}$$

where

$$\Delta^2_{(j:med)} = \begin{cases} \left(\mu_{\left(\frac{n+1}{2}\right)} - \mu\right)^2 & \text{if } n \text{ is odd} \\ \left(\mu_{\left(\frac{n}{2}\right)} - \mu\right)^2 & \text{if } n \text{ is even and } j \le \frac{n}{2} \\ \left(\mu_{\left(\frac{n+2}{2}\right)} - \mu\right)^2 & \text{if } n \text{ is even and } j > \frac{n}{2} \end{cases}$$

Muttlak´s estimator is unbiased only if the pdf is symmetric with respect to μ and $V[\mu_{rss[med]}] \le V[\mu_{rss}] \le V[\mu_s]$. The relative precision of it (RP) increases with n. For not symmetric pdf's the estimator is still more precise than the arithmetic mean of SRS, μ_s, but it is biased. The RP decreases if $n \ge 6$. The errors in the ranking do not affect seriously these results. Hence the use of median-rss provides a gain in accuracy.

3. MEAN IMPUTATION

Let us consider that there are non responses. First we will study the effect of imputation of the missing observations.

The first imputation method to be analyzed is the mean substitution. Take n odd

$$Y^*_{(i:med)mI(1)} = \begin{cases} Y_{(i:med)m} & \text{if a response is obtained} \\ \frac{1}{n(1)} \sum Y_{(i:med)m\, w(i:m)} & \text{otherwise} \end{cases}$$

Where $w(i:m)$ is a Bernoulli random variable with parameter Q=1-P, P is the probability of response. Hence if the number of responses is $n(1)$

$$\sum_{i=1}^{n} w(i:m) = n(2).$$

The imputation estimator proposed in this case, taking $n(2)=n-n(1)$, is:

$$\mu_{(rss)medI(1)} = \mu_{I(1o)} = \frac{\sum_{i=1}^{n(1)} \sum_{m=1}^{r} Y_{(i:med)m} + \sum_{i=1}^{n} \frac{w(i:m)}{n(1)} \sum_{m=1}^{r} \sum_{j=1}^{n(1)} Y^*_{(j:med)m}}{nr}_{i}.$$

Note that if n is odd

$$E\ \mu_{I(1o)} = \frac{\sum_{i=1}^{n(1)} \sum_{m=1}^{r} \mu_{\left(\frac{n+1}{2}\right)} + \sum_{i=1}^{n(2)} \sum_{m=1}^{r} \mu_{\left(\frac{n+1}{2}\right)}}{nr} = \mu_{\left(\frac{n+1}{2}\right)}.$$

The conditional variance is

$$V\ \mu_{I(1o)} = \frac{\sum_{i=1}^{n(1)} \sum_{m=1}^{r} V\ Y_{(i:med)m} + \sum_{i=1}^{n} w(i:m)^2 \sum_{m=1}^{r} V\left(\sum_{j=1}^{n(1)} \frac{Y_{(j:med)m}}{n(1)}\right)_{i}}{(nr)^2}$$

$$= \frac{rn(1)\sigma^2_{\left(\frac{n+1}{2}\right)} + \frac{rn(2)\sigma^2_{\left(\frac{n+1}{2}\right)}}{n(1)}}{(nr)^2} = \frac{\left(n(1) + \frac{n(2)}{n(1)}\right)\sigma^2_{\left(\frac{n+1}{2}\right)}}{n^2 r},$$

and accepting the approximation $E(a/b) \cong E(a)/E(b)$ the expected variance is

$$E\left[V\ \ \mu_{I(1o)}\ \right] \cong \frac{\left(nP+\frac{Q}{P}\right)\sigma^2_{\left(\frac{n+1}{2}\right)}}{n^2 r}$$

As in this case the conditional expectation is not random the error of the mean imputation median variance is equal to A.

When n is even we have to consider the no responses obtained in the two sets of samples.

$$Y^*_{(i:med)mI(1)} = \begin{cases} Y_{(i:med)m} & \text{if a response is obtained} \\ \frac{1}{n(11)}\sum_{j=1}^{n} Y_{(i:med)m}\, w((1))\, j:m)\ \ i \text{ does not responed and } i \le \frac{n}{2} \\ \frac{1}{n(12)}\sum_{j=1}^{n} Y_{(i:med)m}\, w((2))\, j:m)\ \ i \text{ does not responed and } i > \frac{n}{2} \end{cases}$$

Now $w((h)i:m)$, h=1,2, is a Bernoulli random variable with parameter $Q(h)$=1-$P(h)$, $P(h)$ is the probability of response in the *h-th* set of samples. Hence if the number of responses are $n(1h)$, h=1, 2, $n(1)$=$n(11)$+$n(12)$ and

$$\sum_{i=1}^{n} w((h)i:m)=n-n((h)1)=n(h2).$$

The imputation estimator proposed in this case is

$$\mu_{(rss)medI(1)} = \mu_{I(1e)} = \frac{\sum_{i=1}^{n(1)}\sum_{m=1}^{r} Y_{(i:med)m} + I(1) + I(2)}{2nr},$$

where

$$I(1) = \sum_{i=1}^{n} \frac{w((1)i:m))}{n(11)} \sum_{m=1}^{r} \left(\sum_{j=1}^{n(11)} Y^*_{(j:med)m} \right)_i,$$

$$I(2) = \sum_{i=1}^{n} \frac{w((2)i:m))}{n(11)} \sum_{m=1}^{r} \left(\sum_{j=1}^{n(11)} Y^*_{(j:med)m} \right)_i.$$

We can divide $\mu_{I(1s)}$ as

$$\mu_{I(11)} = \frac{\sum_{i=1}^{n(11)}\sum_{m=1}^{r} Y_{(i:med)m} + I(1)}{2nr},$$

$$\mu_{I(12)} = \frac{\sum_{i=1}^{n(12)}\sum_{m=1}^{r} Y_{(i:med)m} + I(2)}{2nr}.$$

The expectations of these terms are

$$E\left(\frac{\sum_{i=1}^{n(11)}\sum_{m=1}^{r} Y_{(i:med)m} + I(1)}{nr} \right)^* = \frac{n(11)+n(21)\mu_{\left(\frac{n}{2}\right)}}{n} = \frac{1}{2}\mu_{\left(\frac{n}{2}\right)},$$

$$E\left(\frac{\sum_{i=1}^{n(12)}\sum_{m=1}^{r} Y_{(i:med)m} + I(2)}{nr} \right)^* = \frac{n(12)+n(22)\mu_{\left(\frac{n}{2}+1\right)}}{n} = \frac{1}{2}\mu_{\left(\frac{n}{2}\right)}.$$

Hence

$$E\ \mu_{I(1e)} = \frac{\mu_{\left(\frac{n}{2}\right)} + \mu_{\left(\frac{n}{2}+1\right)}}{2}$$

Consequently the variance of the conditional expectation is zero. The conditional variances are

$$V\ \mu_{I(11)} = \frac{1}{n^2 r}\left(n(11) + \frac{n(21)}{n(21)} \right)\sigma^2_{\left(\frac{n}{2}\right)},$$

$$V\ \mu_{I(12)} = \frac{1}{n^2 r}\left(n(12) + \frac{n(22)}{n(12)} \right)\sigma^2_{\left(\frac{n}{2}+1\right)}.$$

Let us define $P(t)$ as the probability of obtaining a response of the statistics of order $n/2$ if t=1 and $P(2)$ when t=2 and the order is $1+n/2$. Then

$$E\left[V\ \mu_{I(1e)}\ \right] = \frac{1}{4n^2 r}\left[\left(nP(1) + \frac{Q(1)}{P(1)} \right)\sigma^2_{\left(\frac{n}{2}\right)} + \left(nP(2) + \frac{Q(2)}{P(2)} \right)\sigma^2_{\left(1+\frac{n}{2}\right)} \right]$$

4. RATIO IMPUTATION

We are going to consider ratio imputation methods. Let μ_X be the population mean of the auxiliary variable X, and σ_X^2 its population variance. X is a known variable and $\bar{x}$ is the mean of it in the nr observations. Take the ratio of the RSS means as

$$r_{(rss)} = \frac{\bar{y}_{(rss)}}{\bar{x}_{(rss)}}$$

and

$$\bar{y}_{r(rss)} \cong r_{(rss)}\mu_X$$

is the estimated mean. Take Q_1 and Q_3 as the first and third quartiles of the distribution of X, respectively. One of them is selected and used for estimating. Using the Taylor approximation, see Al-Omari, et al. (2008).

$$\bar{y}_{r(rss)} \cong \bar{y}_{(rss)} - Q_1\ \bar{x}_{(rss)} - \mu_X\ + Q_2\ \bar{x}_{(rss)} - \mu_X\ ^2 - Q_3\ \bar{x}_{(rss)} - \mu_X\ \ \bar{y}_{(rss)} - \mu_Y$$

where

$$Q_1 = \frac{\mu_Y}{\mu_X + q_i},\ Q_2 = \frac{\mu_Y}{\mu_X + q_i}\frac{1}{\mu_X + q_i},\ Q_3 = \frac{1}{\mu_X + q_i}\ \bar{x}_{(rss)} - \mu_X\ \ \bar{y}_{(rss)} - \mu_Y$$

The MSE is approximated, considering, that the Taylor Series with terms $O(n^{-2})$ is a good approximation

$$MSE\ \bar{y}_{(rss)}\ \cong Var\ \bar{y}_{(rss)}\ + Q_2^2 Var\ \bar{x}_{(rss)}\ - 2Q_1 Cov\ \bar{x}_{(rss)}, \bar{y}_{(rss)}\ ,$$

where

$$Cov\ \bar{x}_{(rss)}, \bar{y}_{(rss)}\ = E\ \bar{x}_{(rss)} - \mu_X\ \ \bar{y}_{(rss)} - \mu_Y\ .$$

Consider the median RSS estimator when there are missing observations and define the Bernoulli random variable $w^*(i:m)$ with probability of success $P(1)$. If we obtain a response at the ith-sample in the cycle m then $w(i:m)$=1. The number of responses is

$$n(1)=\sum_{i=1}^{n}\sum_{m=1}^{r} w(i:m)=\sum_{i=1}^{n} r(i).$$

In the case of full response $r(j)=r$ for any j and $n(1)=n$. The mean of the responses to Y is

$$\mu^*_{Y(rss)}=\sum_{i=1}^{n}\sum_{m=1}^{r} Y_{(i:med)m}\, w(i:m)/n(1).$$

Its expectation is

$$E\ \mu^*_{Y\ rss}\ = \sum_{i=1}^{n}\sum_{m=1}^{r}\mu_{Y\ n+1\ /2)} w\ i:m\ /n(1) = \mu_{Y(\ n+1\ /2)}.$$

When non responses are present we propose to use the ratio of mean of the responses to Y for the mean in the sample of the auxiliary variable X

$$r^*_{(rss)} = \frac{\mu_{*Y(rss)}}{\mu_{*X(rss)}},$$

and

$$\bar{y}_{(rss)med(R)} \cong r^*_{(rss)}\ \mu_X$$

is the ratio imputation estimator of the mean. The existence of missing.

For n odd its expected value is

$$\begin{aligned} E\ \bar{y}_{(rss)med(R)}\ &\cong \mu_{(\frac{n+1}{2})} - E\Big[Q_1\ \bar{x}_{(rss)med} - \mu_X\ + Q_2\ \bar{x}_{(rss)med} - \mu_X\ ^2 - Q_3\ \bar{x}_{(rss)med} - \mu_X\ \ \bar{y}_{(rss)med} - \mu_Y\ \Big] \\ &= \mu_{\ n+1\ /2} - Q_1 E\ A\ 1\ + Q_2 E\ A\ 2\ - Q_3 E\ A\ 3 \end{aligned}$$

where

$$E\ A\ 1\ = \Delta_{X\ n+1\ /2},$$

$$E\ A\ 2\ = \Big[\sigma^2_{\ n+1\ /2} + \Delta^2_{X\ n+1\ /2)}\Big] / n^2 r^2 = \sigma^2 / nr - \ nr-1\ \Delta^2_{X\ n+1\ /2)} / nr,$$

Taking

$$A(3) = A(3) \pm \mu_{X\ \ n+1\ /2}\ \mu_{Y\ \ n+1\ /2}\ ,$$

$$E\ A\ 3\ = Cov\ x_{(rss)}, y_{(rss)}\ - \ \mu_y \Delta_{X\ \ n+1\ /2} + \mu_X \Delta_{Y\ n\ n+1\ /2}$$

When n is even we have that

$$n(1) = \sum_{i=1}^{n/2}\sum_{m=1}^{r} w\ i:m\ + \sum_{i=1+n/2}^{n}\sum_{m=1}^{r} w\ i:m\ = n(11) + n(12).$$

The expression of the ratio estimator is

$$\mu^*_{Y(rss)} = \frac{1}{2}\Big[\sum_{i=1}^{n/2}\sum_{m=1}^{r} Y_{i:n/2\ m} w\ i:m\ /n\ 11\ + \sum_{i=1+n/2}^{n}\sum_{m=1}^{r} Y_{i:1+n/2\ m}\mathrm{w}\ i:m\ /n\ 12\ \Big],$$

and

$$EE\ \mu^*_{Y(rss)} / s\ = \frac{1}{2}\ \mu_{\ n/2\ } + \mu_{1+n/2}$$

The conditional variances of the terms of $\mu^*_{Y(rss)}$ are

$$V\left(\sum_{i=1}^{n/2}\sum_{m=1}^{r} Y_{i:n/2\ m}\ w\ i:m\ /n\ 11\ /s\right)=\sum_{i=1}^{n/2}\sum_{m=1}^{r}\sigma^2_{n/2}\ w^2\ \ i:m\ /n\ 11\ ^2$$

$$V\left(\sum_{i=1+n/2}^{n/2}\sum_{m=1}^{r}\mu_{1+n/2}\ w\ i:m\ /n\ 12\ /s\right)=\sum_{i=1+n/2}^{n/2}\sum_{m=1}^{r}\sigma^2_{1+n/2}\ w^2\ \ i:m\ /n\ 12\ ^2\ .$$

Then

$$E\ \ \mu^*_{Y(rss)}/s\ \ =\ \sigma^2_{n/2}+\sigma^2_{1+n/2}\ \ /2nP(1),$$

and

$$V\left[E\ \ \mu^*_{Y(rss)}/s\ \right]=\frac{1}{2}\left[\mu^2_{(n/2)}rQ(1)+\mu^2_{1+n/2}\ rQ(1)\right]$$

Hence the error of the missing observation estimator is

$$\varepsilon(\mu^*_{Y\ rss})\approx\frac{1}{2}\left[\mu^2_{n/2}\ rQ(1)+\mu^2_{1+n/2}\ rQ(1)\right]+\left[\sigma^2_{n/2}\ +\mu^2_{1+n/2}\ \right]/2nP(1)\ .$$

4. NUMERICAL COMPARISSONS

		Uniform (0,1)		Normal (0,1)		Laplace(0,1)	
n	*r*	*mimp*	*rimp*	*mimp*	*rimp*	*mimp*	*rimp*
2	2	2.29	2.35	1.73	1.62	2.35	1..41
3	2	2.28	2.34	1.72	1..67	2.35	1.97
4	2	2.42	2.34	1.73	1.59	1.73	1.66
5	2	2.23	2.35	1.71	1..62	1.70	1.51
2	3	2.22	2.32	1.90	1.90	1..63	1.43
3	3	2.03	2.29	2..23	2.73	1..69	1..58
4	3	2.28	2.35	1.71	1..62	1..66	1.58
5	3	2.22	2.36	1.72	1.72	1..98	1.63
2	4	2.32	2.37	1.65	1..25	1..73	1.65
3	4	2.24	2.35	1.72	1..62	1..66	1.71
4	4	2.56	2.42	1.77	1.17	1..46	1.45
5	4	2.29	2.43	1.88	1..68	1..99	1.51
2	5	2.80	2.35	1.72	1..62	1.66	1.61
3	5	2.37	2.36	1.72	1.72	1..68	1.71
4	5	2.44	2.37	1.70	1.20	1..85	1.71
5	5	2.39	2.35	1.72	1..62	1..66	1.65

Table 1. Efficiency of the developed imputation estimators versus the corresponding full response

We will consider the efficiency of the proposals with respect to the corresponding full response models. 1000 samples of size 100 and a 10% of non-responses were generated with median friendly distributions.

Defining $e(I, i)$ and $e(i)$ as the estimator using the imputation and the full response one. The efficiency measure used was

$$\xi(e(I, i))=\sum /e(I,i)-\mu_Y /_h/\sum /e(I,i)-\mu_Y/_h$$

The same distribution was used for describing the behaviour of X and Y.

The results are given in Table 1. Note that the imputation works very well for the normal and the Laplace distributions. For the uniform it doubles the error. The behavior of the ratio imputation was not so different from the use of the mean imputation.

RECEIVED JULY 2010
REVISED NOVEMBER 2010

REFERENCES

[1] AL-OMARI, A.I. and K. JABER, and AL-OMARI, A. (2008): Modified ratio-type estimators of the mean using extreme ranked set sampling, **Journal of Mathematics and Statistics**, 4, 150-155.

[2] BAI, Z. D. and Z. CHEN *(*2003): On the theory of ranked set sampling *and* its ramifications, **Journal of Statistical Planning and Inference**, 109, 81-99

[3] BOUZA, C.N. (2001): Model assisted ranked survey sampling**. Biometrical Journal**, 36, 753-764.

[4] CHEN, Z., BAI, Z. and SINHA, B.K, (2004): **Ranked set sampling: theory and applications**. Lectures Notes in Statistics, 176. Springer, N. York.

[5] COCHRAN, W.G. (1977): **Sampling Techniques**. 3rd edition, Wiley and Sons. New York.

[6] JEMAIN, A.A. and A.I. AL-OMARI, (2006): Multistage median ranked set samples for estimating the population mean, **Pakistan Journal of Statistics**, 22, 195–207.

[7] MCINTYRE, G.A. (1952): A method for unbiased selective sampling using ranked sets. **Australian Journal of Agricultural Research**, 3, 385–390.

[8] MUTTLAK, H.A. (1997): Median ranked set sampling, **Journal of Applied Statistical Sciences**, 6, 245-255.

[9] MUTTLAK, H.A. (2003): Investigating the use of quartile ranked set samples for estimating the population mean**. Journal of Applied Mathematics and Computation**, 146, 437-443.

[10] SAMAWI. H.M. and H.A. MUTTLAK, (1996): Estimation of ratio using rank set sampling. **Biometrical Journal**, 36, 753–764.

[11] SINGH, S. and, DEO, B. (2003): Imputation by power transformation. **Statistical Papers**, 555-579.

[12] TOUTENBURG, V, SRIVASTAVA, K. and SHALABH. (2008): Amputation versus imputation of missing values through ratio method in sample surveys. **Statistical Papers**, 49, 237-247.