CADERNOS DO IME – Série Estatística
Universidade do Estado do Rio de Janeiro - UERJ
Rio de Janeiro - RJ - Brasil
ISSN impresso 1413-9022 / ISSN on-line 2317-4536 - v.34, p. 33 - 43, 2013

# A RANKED SET SAMPLING MODIFIED RATIO ESTIMATOR

Carlos N. Bouza-Herrera
Facultad de Matemática y Computación
Universidad de La Habana, Cuba
bouza@matcom.uh.cu

**Abstract:**

*We consider the modified ratio estimator proposed by Swain (2013). It is an extension of the classic ratio estimator. The objective of this paper is developing the ranked set counterpart of the new estimator, developed by Swain (2013). We derived a new ranked set sampling ratio estimator. For illustrating the behavior of the proposal a comparison of their approximate mean squared error was developed. The proposed procedure appeared as more accurate. Empirical studies give an insight on the magnitude of the efficiency of the estimator developed.*

**Keywords:** *Ratio Estimators, Ranked Set Sampling, Efficiency, Order Statistics*

## 1. Introduction

Let $U = (U_1, U_2 \ldots U_N)$ be the finite population of size $N$ and take $X$, $Y$ as two characteristics of interest. To each unit we attached $(y_i, x_i)$ and $\rho = \sigma_{yx}/\sigma_y\sigma_x \neq 0$ where:

$$\sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2,\ \sigma_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{X}\right)^2,$$

$$\sigma_{yx} = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)\left(x_i - \bar{X}\right),\ \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i\ ,\ \ \bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

We are interested in the population ratio (1):

$$R = \frac{\bar{Y}}{\bar{X}} \tag{1}$$

There are many ratio type estimators based on Simple Random Sampling (SRS) which has been proposed in the literature. In the sequel we will use the corresponding coefficients of variation of $y$ and $x$, $C_y = \frac{\sigma_y}{\bar{Y}}, C_x = \frac{\sigma_x}{\bar{X}}$ , as well as $C_{yx} = \frac{\sigma_{yx}}{\bar{Y}\bar{X}} = \rho\frac{\sigma_x\sigma_y}{\bar{X}\bar{Y}} = \rho C_y C_x$ .

A sample $s$ of size $|s| = n$ is selected using simple random sampling with replacement (SRSWR) and $(y_i, x_i)$, is measured in each individual i=1, 2 … $n$. Define

1. $\bar{y}$ and $\bar{x}$ as the sample means of $y$ and $x$:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\ ,\ \ \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\ .$$

2. The sample ratio: $r = \frac{\bar{y}}{\bar{x}}$

3. The sample variances and the covariance:

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}, s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}, s_{yx} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{n-1}$$

It is well known that $\bar{y}$ is an unbiased estimator of the population mean $\bar{Y}$ . Its sampling error is its variance:

$$V(\bar{y}) = \frac{\sigma_y^2}{n} = \theta\bar{Y}^2C_y^2, \theta = \frac{1}{n}$$

When we have full information on $X$ it may be used for improving the efficiency of the estimations. The classic ratio estimator is (2):

$$\bar{y}_R = \frac{\bar{y}}{\bar{x}} \bar{X} \tag{2}$$

We have that $x_1, .., x_n$ and $y_1, .., y_n$ are sequences of iid random variables. Let us consider that $g(x_1, .., x_n)$ and $q(y_1, .., y_n)$ are statistics related with the parametric functions represented by:

$$t_n = T + \frac{\delta_T}{n} + \frac{\sum_{i=1}^{n} \tau_0(Z_i)}{n^2} + \frac{\sum_{i=1}^{n} \tau_1(Z_i)}{n^2} + \frac{\sum_{C_2^n} \tau_2(Z_i, Z_j)}{n^2} + \frac{\sum_{C_3^n} \tau_3(Z_i, Z_j, Z_k)}{n^3} + o_P\left(n^{-\frac{3}{2}}\right),$$

$$\tau_i = g, q;$$

$$i = 1, 2$$

$$T = G, Q,$$

$$Z = X, Y$$

$\delta_T$ is a bias term. The terms based on single variables have zero expectation:

$$E(\tau_0(Z_i)) = E(\tau_1(Z_i)) = 0$$

We also have:

$$E(\tau_2(Z_i, Z_j)|Z_i) = 0$$

and for the third order cross terms:

$$E(\tau_3(Z_i, Z_j, Z_k| \quad Z_i, Z_j)) = 0$$

When treating with the ratio G/Q, we can use a certain order representation in Taylor Series. This method is used on the sequel.

Accepting that the approximation error of order O(1/n), say AE(O(1/n)), the bias and mean square error (MSE) of the estimator in (2), see standard text books as Cochran (1977), Singh and Deo (2003), are:

$$B(\bar{y}_R) \cong \theta \bar{Y}\left(C_x^2 - C_{yx}\right),$$

$$MSE(\bar{y}_R) \cong \theta \bar{Y}^2 \left( C_y^2 + C_x^2 - 2C_{yx} \right)$$

Well known results are that $\bar{y}_R$ is more efficient than $\bar{y}$ if $\rho \frac{C_y}{C_x} > \frac{1}{2}$ .

The common RSS alternative to $\bar{y}_R$:

$$\bar{y}_{r-rss} = \frac{\bar{y}_{rss}}{\bar{x}_{rss}} \mu_X$$

has been thoroughly studied by Bouza (2001) and Muttlak and Al-Saleh (2000). Using some expansion in Taylor Series of $E\left( \bar{y}_{r-rss} - \mu_Y \right)^2$. The derived MSE is:

$$M(\bar{y}_{r-rss}) \cong \frac{\sigma_y^2 - \sum_{i=1}^{m} \frac{\Delta_{Y(i)}^2}{m} + R^2 \left[ \sigma_x^2 - \sum_{i=1}^{m} \frac{\Delta_{X(i)}^2}{m} \right] - 2R\rho \left( \sigma_x^2 - \sum_{i=1}^{m} \frac{\Delta_{X(i)}^2}{m} \right)^{1/2} \times \left( \sigma_y^2 - \sum_{i=1}^{m} \frac{\Delta_{Y(i)}^2}{m} \right)^{1/2}}{n}$$

$\bar{y}_{r-rss}$ is preferred to $\bar{y}_R$ when

$$\delta_{r-rss} = \frac{\sum_{i=1}^{m} \frac{\Delta_{Y(i)}^2}{m} + R^2 \sum_{i=1}^{m} \frac{\Delta_{Y(i)}^2}{m} - 2R\rho \left[ \sigma_X \sigma_Y - \left( \sigma_x^2 - \sum_{i=1}^{m} \frac{\Delta_{X(i)}^2}{m} \right)^{1/2} \times \left( \sigma_y^2 - \sum_{i=1}^{m} \frac{\Delta_{Y(i)}^2}{m} \right)^{1/2} \right]}{n} > 0$$

The RSS alternative for different ratio type estimators has been studied. Bouza (2013) developed them and established their preference to the estimators belonging to the class:

$$F = \left\{ \begin{array}{l} \bar{y}_\theta = \dfrac{\bar{Y}_{est} + \alpha}{B\bar{X}_{est} + \lambda} \left( B\bar{X} + \lambda \right); \\ \theta = (\alpha, B, \lambda)^T \in A \times B \times L \end{array} \right\}$$

where $\bar{Z}_{est}$ , $Z=X, Y$, estimates the mean and:

$$A = \left\{0, b(\bar{X} - \bar{x}), \sigma_x\right\} = \{\alpha_1, \alpha_2, \alpha_3\}$$
$$B = \{1, B_2(x), C_x, \rho\} = \{B_1, B_2, B_3.B_4\}.$$
$$L = \{0, \rho, B_2(x), C_x\} = \{\lambda_1, \lambda_2, \lambda_3.\lambda_4\}$$

Defining $b = s_{xy}/s_x^2$, $B_2(x)$=kurtosis coefficient of the distribution of $X$. The estimators proposed by Singh-Taylor (2003) and Kadilar-Cingi (2004 y 2005) belong to $F$.

Ranked set sampling (RSS) is a sampling procedure, which is not only cost effective when compared to the commonly used simple random sampling in many situations but more efficient. Mcintyre (1952) proposed the sample mean based on RSS as an estimator of the population mean and established that it allowed using smaller samples. Takahasi and Wakimoto (1968) provided the necessary mathematical theory of RSS. The use of RSS is the theme in a growing number of papers. Some recent ones are Jemain et al. (2007), Al-Hadrami and Al-Omari (2009), Ozturk (2011).

In this paper we make a comparative study of the modified ratio proposed by Swain (2013) and a RSS counterpart where a transformation of the auxiliary variable $X$ is used both for ranking and computing the estimator.

## 2. Swain Generalized Modified Ratio Estimators and its RSS Counterpart

Take the auxiliary variable $X$ and the transformed variable:

$$Z_i = aX_i + (1-a)\bar{X},$$
$$a = \text{constant.}$$

The population mean of $Z$ is:

$$\bar{Z} = \frac{1}{N}\sum_{i \in U} Z_i = \frac{1}{N}\sum_{i=1}^{N} (aX_i + (1-a)\bar{X}) = \bar{X}$$

while the sample Z-mean and Z-ratio are:

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = a\bar{x} + (1-a)\bar{X},$$

$$R_Z = \frac{\bar{Z}}{\bar{z}}$$

Swain (2013) proposed the modified ratio estimator (3):

$$\hat{\hat{Y}}_{GR} = \bar{y}R_Z \tag{3}$$

Using his results for that case of SRSWR, the bias and MSE, in terms of *O(1/n)*, are easily derived . They are:

$$B(\hat{\hat{Y}}_{GR}) = \theta\bar{Y}\left(a^2C_x^2 - aC_{yx}\right), \theta = 1/n,$$

$$MSE(\hat{\hat{Y}}_{GR}) = \theta\bar{Y}^2(C_y^2 + a^2C_x^2 - 2aC_{yx}$$

Let us use *X* for ranking the units. The basic RSS procedure is the following:

**Step 1:** Randomly select $m^2$ units from the target population. These units are randomly allocated into *m* sets, each of size *m*.

**Step2:** The *m* units of each set are ranked visually or by any inexpensive method free of cost, say *X*, with respect to the variable of interest *Y*.

**Step2:** From the first set of *m* units, the smallest ranked unit is measured; from the second set of *m* units the second smallest ranked unit is measured.

**Step 3:** Continue until the *m*th smallest unit (the largest) is measured from the last set.

**Step 4:** Repeat the whole process *r(i)* times (cycles)

**Step 5:** Evaluate the corresponding units.

We can denote it a follows

$$\left.\begin{matrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{im} \end{matrix}\right\}_r \sim > X_{(ii)r} = Y_{(i)r,}\, r = 1, \dots, r(i);\;\; i = 1, \dots, m$$

Let $Y_1$ ,…,$Y_m$ be a sample selected using SRSWR from a probability density function *f(y)*, with mean $\mu_Y$ and variance $\sigma_Y^2$. Considering the selection of *m* independent samples selected using a SRSWR design, each of size *m* each, we have

$Y_{11},..,Y_{1m}$, $Y_{21},\dots,Y_{2m},\dots.Y_{m1},\dots,Y_{mm}$. Let $Y_{i(1m)},..,Y_{i(mm},\dots Y_{i(mm)}$, be the order statistics of the sample $Y_{1i},..,Y_{1m},,\ \dots,Y_{im}$, for $(i=1,2,\dots,m)$.

Due to the unbiasedness of:

$$\bar{\xi}_{rss} = \frac{1}{n}\sum_{t=1}^{r}\sum_{i=1}^{m}\xi_{(i:i)t}, \qquad \xi = y,x,z$$

$$\bar{z}_{rss} = \frac{1}{n}\sum_{t=1}^{r}\sum_{i=1}^{m}z_{(i:i)t} = a\bar{x}_{rss} + (1-a)\bar{X}$$

is unbiased. As we are dealing with order statistics:

$$E\left(\xi_{(i:i)t}\right) = \mu_{\xi_{(i)}},\ V(\xi_{(i:i)t}) = \sigma^2_{\xi_{(i)}} = \sigma^2_{\xi} - \Delta^2_{\xi_{(i)}},$$

where $\left|\Delta_{\xi_{(i)}}\right| = \left|\mu_{\xi_{(i)}} - \mu_{\xi}\right|$.

We propose the RSS-GR class of modified ratio estimator of the mean:

$$\bar{y}_{GR(rss)} = \bar{y}_{rss}R_{Z(rss),}\ R_{Z(rss)} = \frac{\bar{Z}}{\bar{z}_{rss}} \tag{4}$$

As we deal with RSS we have the RSS-covariation coefficients $e_{1(rss)} = \frac{\bar{x}_{rss}-\bar{X}}{\bar{X}}, e_{0(rss)} = \frac{\bar{y}_{rss}-\bar{Y}}{\bar{Y}}$. It is easy to prove that $E(e_{1rss}) = E(e_{0rss}) = 0$. The variances are:

$$V(e_{1rss}) = E\left(\frac{\bar{x}_{rss}-\bar{X}}{\bar{X}}\right)^2 = \frac{V(\bar{x}_{rss})}{\bar{X}^2} = \frac{1}{\bar{X}^2}\left(\frac{\sigma^2_x}{n} - \sum_{i=1}^{m}\frac{\Delta^2_{x(i)}}{mn}\right) = \theta\left(C^2_x - \sum_{i=1}^{m}\frac{\Delta^2_{x(i)}}{m\bar{X}^2}\right)$$

$$V(e_{0rss}) = E\left(\frac{\bar{y}_{rss}-\bar{Y}}{\bar{Y}}\right)^2 = \frac{V(\bar{y}_{rss})}{\bar{Y}^2} = \frac{1}{\bar{Y}^2}\left(\frac{\sigma^2_y}{n} - \sum_{i=1}^{m}\frac{\Delta^2_{y(i)}}{mn}\right) = \theta\left(C^2_y - \sum_{i=1}^{m}\frac{\Delta^2_{y(i)}}{m\bar{Y}^2}\right)$$

accepting that $O(1/n)$ in the Taylor series expansion of $\bar{y}_{GR(rss)}$ is sustained by the validity of using:

$$\bar{y}_{GR(rss)} \cong \bar{Y} + \bar{Y}\left(e_{0(rss)} - ae_{1(rss)} + a^2e^2_{1(rss)} - ae_{0(rss)}e_{1(rss)}\right)$$

Hence,

$$E(\bar{y}_{GR(rss)}) - \bar{Y} \cong \bar{Y}\left(a^2 E\left(e_{1(rss)}^2\right) - aE\left(e_{0(rss)}e_{1(rss)}\right)\right)$$
$$= \theta a^2\left(\bar{Y}a^2C_x^2 - \sum_{i=1}^{m}\frac{\Delta_{y(i)}^2}{m\bar{Y}}\right) - \frac{a\theta\bar{Y}C_{x(rss)y(rss)}}{\bar{X}} = B(\bar{y}_{GR})$$

as

$$E\left(e_{1(rss)}e_{0(rss)}\right) = E\left(\frac{(\bar{x}_{rss} - \bar{X})}{\bar{X}}\frac{(\bar{y}_{rss} - \bar{Y})}{\bar{Y}}\right) = \frac{Cov(\bar{x}_{rss}, \bar{y}_{rss})}{\bar{X}\bar{Y}} = \frac{C_{x(rss)y(rss)}}{n\bar{X}\bar{Y}}$$

The mean square is approximately:

$$E(\bar{y}_{GR(rss)} - \bar{Y})^2$$
$$\cong \theta\bar{Y}^2\left(C_y^2 + a^2C_x^2\right)$$
$$- \theta\left(R^2a^2\sum_{i=1}^{m}\frac{\Delta_{x(i)}^2}{m} + \sum_{i=1}^{m}\frac{\Delta_{y(i)}^2}{m} + 2a\bar{Y}^2C_{x(rss)y(rss)}\right)$$
$$= MSE(\bar{y}_{GR(rss)})$$

$$MSE(\hat{\hat{Y}}_{GR}) = \theta\bar{Y}^2(C_y^2 + a^2C_x^2 - 2aC_{yx})$$

Hence the use of RSS generates a gain in accuracy:

$$\zeta(GR(rss), GR) = MSE(\hat{\hat{Y}}_{GR}) - MSE(\bar{y}_{GR(rss)})$$
$$= 2a\theta\bar{Y}^2\left(C_{x(rss)y(rss)} - 2aC_{yx}\right) - \theta\left(R^2a^2\sum_{i=1}^{m}\frac{\Delta_{x(i)}^2}{m} + \sum_{i=1}^{m}\frac{\Delta_{y(i)}^2}{m}\right)$$

The optimum value of $a$ is obtained by minimizing $MSE(\bar{y}_{GR(rss)})$. Differentiating $\zeta(GR(rss), GR)$ and equating to zero we derive that the solution of that optimization problem is:

$$a_{0(rss)} = argminMSE(\bar{y}_{GR(rss)}) = \frac{C_{x(rss)y(rss)}}{R\left(\sigma_x^2 - \sum_{i=1}^{m}\frac{\Delta_{x(i)}^2}{m}\right)}$$

## 3. Numerical Comparisons

We compared the behavior of our proposed RSS method with the proposal of Swain (2013) when SRSWR is used using data from three populations. Their description is given as follows:

**Population 1.** A set of *244* accounts was considered. The balance of each of them in the previous semester was $X$ and $Y$ was produced by an auditory.

**Population 2.** The evaluation of radiographies provided values of $X$ in *350* patients with cancer. $Y$ was the size of an extirpated tumor.

**Population 3.** The height of 1270 pigs provided the information on $X$ in the population. Y is the pig's weight reported by the butchers.

The values of *r* and *m* were fixed conveniently for obtaining a sample of size *24*. The means and variances of the os's involved were determined by forming all the possible samples and computing them. The relative gain in accuracy due to the use of RSS was measured by (5):

$$\varpi = \zeta(GR(rss), GR)/MSE(\hat{\hat{Y}}_{GR}) \quad (5)$$

for $m$=*3, 4, 6*. The results are given in Table 1. They sustain that the use of RSS provides gains of accuracy larger than 20%.

Table 1. Gain in accuracy due to the use of RSS in three populations

| **Population** | ***m=3*** | ***m=4*** | ***m=6*** |
|---|---|---|---|
| Balance of accounts | *0,334* | *0,293* | *0,240* |
| Size of tumors | *0,286* | *0,251* | *0,233* |
| Height of pigs | *0,322* | *0,424* | *0,360* |

A similar study was developed by generating a sample of 240 values of $X$ and determining (6):

$$Y = 5 + 2X + \varepsilon \quad (6)$$

$\varepsilon$ was generated using the same distribution. The results are given in Table 2. Note that generally the gain in efficiency is larger when the underlying distribution is symmetric. The best results are derived when $m$=*4*.

Table 2. Gain in accuracy due to the use of RSS of six populations: *n*=240* and *K=0,10*

| **Distribution** | ***m=3*** | ***m=4*** | ***m=6*** |
|---|---|---|---|
| Uniform (0,1) | 0,207 | 0,247 | 0,182 |
| Normal (0,1) | 0,223 | 0,228 | 0,173 |
| Logistic (0,1) | 0,207 | 0,310 | 0,244 |
| Laplace (0,1) | 0,149 | 0,169 | 0,104 |
| Exponential (1) | 0,191 | 0,156 | 0,147 |
| Gamma (2,1) | 0,104 | 0,131 | 0,084 |
| Weibull (1,3) | 0,219 | 0,266 | 0,195 |
| Beta (7,4) | 0,109 | 0,128 | 0,104 |

## 4. Conclusions

The accuracy of the proposed method in the considered real life problems moved between 0,233 and 0,424. Hence the sample size to be used can be diminished seriously for obtaining a fixed efficiency. Then we consider that our proposal is better than the SRSWR method. When $G_{rss}$ is analyzed using *m=4* seems to be the best choice.

The results with probabilistic models suggest that the derived RSS estimator is also better. Its worst gain in accuracy was obtained for the Beta distribution (0,104) and the best with the Logistic (0,311). Using m=4 appeared as the best choice in 87,5% of the cases.

**Acknowledgments**: The present version is an improvement on the original paper due to the useful suggestions made by unknown referees.

## References

AL-HADHRAMI, S.; AL-OMARI, A.I. Bayesian Inference on the Variance of Normal Distribution using Moving Extremes Ranked Set Sampling. **Journal of Modern Applied Statistical Methods**, 8(1), 273-281, 2006.

BOUZA, C. N. Model Assisted Ranked Survey Sampling; **Biometrical J**., 43, 249-259, 2001.

BOUZA, C. N. Una Clase de Estimadores basados en una Razón: Muestreo Simple Aleatorio y Muestreo por Conjuntos Ordenados, Accepted by **Rev. Inv. Operacional**, 2013.

DELL, T. R.; CLUTTER, J. L. Ranked Set Sampling Theory with Order Statistics Background, **Biometrics** 28, 545-555, 1972.

JEMAIN, A. A.; AL-OMARI, A. I.; IBRAHIM, K. Multistage Extreme Ranked Set Sampling for Estimating the Population Mean, **Journal of Statistical Theory and Applications**, 6, 456-471, 2007.

KADILAR, C.; CINGI, H. Ratio Estimators in Simple Random Sampling. **Applied Math. and Computation,**151, 893-902, 2004.

KADILAR, C.; CINGI, H. A New Ratio on Some Modified Ratio and Product Type Estimators-Revisited Estimator in Stratified Random Sampling. **Comm. in Stat:. Theory and Methods**, 34, 597-602, 2007.

MCINTYRE, G. A. A Method of Unbiased Selective Sampling using Ranked Sets. **Australian J. Agricultural Research**. 3, 385-390, 1952.

MUTTLAK. H. A.; AL-SALEH, M. F. Recent Developments in Ranked Set Sampling, **J. Applied Stat. Sc.** 10, 269-290, 2002.

OZTURK, O. Parametric Estimation of Location and Scale Parameters in Ranked Set **Sampling. Journal of Statistical Planning and Inference**, 141, 1616-1622, 2011.

SINGH, H. P.; TAYLOR, R. Use of Known Correlation Coefficient in Estimating Finite Population Means, **Statistics In Transition**, 6, 555-560, 2003.

SWAIN, A. K. P. C. On Some Modified Ratio And Product Type Estimators-Revisited **Rev. Inv. Operacional**, 34, 35-57, 2013.

TAKAHASI K.; WAKIMOTO, K. On Unbiased Estimates of the Population Mean based on Sample Stratified by Means of Ordering, **Annals of the Inst. of Statistical Mathematics**. 20, 1-31, 1968.

VOCK, M.; BALAKRISHNAN, N. A Jonckheere–Terpstra-Type Test for Perfect Ranking in Balanced Ranked Set Sampling. **Journal of Statistical Planning and Inference**, 141, 624-630, 2011.