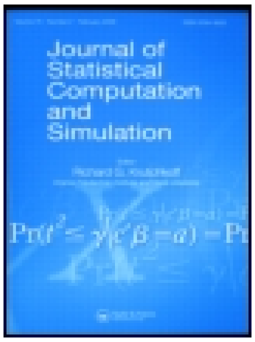


Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: http://www.tandfonline.com/loi/gscs20

Performance of coefficient of variation estimators in ranked set sampling

Cintia Maestreli Consulin, Damiane Ferreira, Idemauro Antonio Rodrigues de Lara, Antonino De Lorenzo, Laura di Renzo & Cesar Augusto Taconeli

To cite this article: Cintia Maestreli Consulin, Damiane Ferreira, Idemauro Antonio Rodrigues de Lara, Antonino De Lorenzo, Laura di Renzo & Cesar Augusto Taconeli (2017): Performance of coefficient of variation estimators in ranked set sampling, Journal of Statistical Computation and Simulation, DOI: 10.1080/00949655.2017.1381959

To link to this article: http://dx.doi.org/10.1080/00949655.2017.1381959

Published online: 29 Sep 2017.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=gscs20

Download by: [CAPES] **Date:** 29 September 2017, At: 09:43

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION, 2017
https://doi.org/10.1080/00949655.2017.1381959

Taylor & Francis
Taylor & Francis Group

Check for updates

Performance of coefficient of variation estimators in ranked set sampling

Cintia Maestreli Consulin[a], Damiane Ferreira[a], Idemauro Antonio Rodrigues de Lara[b], Antonino De Lorenzo[c], Laura di Renzo[c] and Cesar Augusto Taconeli[a]

[a]Department of Statistics, Federal University of Paraná, Curitiba, Brazil; [b]Department of Exact Sciences, Luiz de Queiroz College of Agriculture, University of São Paulo, Piracicaba, Brazil; [c]Division of Clinical Nutrition and Nutrigenomic, Department of Biomedicine and Prevention, University of Rome 'Tor Vergata', Rome, Italy

ABSTRACT

In this paper, we propose and evaluate the performance of different parametric and nonparametric estimators for the population coefficient of variation considering Ranked Set Sampling (RSS) under normal distribution. The performance of the proposed estimators was assessed based on the bias and relative efficiency provided by a Monte Carlo simulation study. An application in anthropometric measurements data from a human population is also presented. The results showed that the proposed estimators via RSS present an expressively lower mean squared error when compared to the usual estimator, obtained via Simple Random Sampling. Also, it was verified the superiority of the maximum likelihood estimator, given the necessary assumptions of normality and perfect ranking are met.

ARTICLE HISTORY
Received 14 June 2017
Accepted 17 September 2017

KEYWORDS
Statistical efficiency; Monte Carlo simulation; parametric estimation; nonparametric estimation

2010 MATHEMATICS SUBJECT CLASSIFICATION
62Dxx; 62F10; 62Fxx; 62Gxx

Downloaded by [CAPES] at 09:43 29 September 2017

1. Introduction

Ranked Set Sampling (RSS) was introduced by McIntyre [1] when estimating the mean yield in pastures. This sampling design represents an efficient alternative to Simple Random Sampling (SRS) when facing the limitations in measuring the variable of interest in large samples. RSS provides more precise estimates than SRS by selecting a smaller number of sampling units [2–4]. It is an appropriate sampling design when it is possible to perform an empirical ranking of sample units, according to the variable of interest, but without its effective measurement, with precision and low cost. The ranking process may be based on the values of a concomitant variable (or a set of them), which is strongly correlated to the variable of interest, or on a personal judgment (usually visual inspection).

The RSS design may be described as follows [1]:

(1) Random selection of n^2 elements from the population under study;
(2) The n^2 selected elements are randomly divided into n samples (sets) of size n;
(3) Ranking of elements within each set based on the established ranking criterion;

CONTACT Cesar Augusto Taconeli taconeli@ufpr.br

© 2017 Informa UK Limited, trading as Taylor & Francis Group

(4) Selection of the element judged to have the ith smallest value for the variable of interest in the ith set, $i = 1, 2, \ldots, n$, to form the ranked set sample.

If desired, this process can be repeated m times (performing m cycles), providing a ranked set sample of size $N = mn$. The final sample is denoted by $\{X_{[1]1}, X_{[2]1}, \ldots, X_{[n]m}\}$, where $X_{[i]r}$ represents the element ranked at the ith position of the ith sample at the rth cycle, with $i = 1, 2, \ldots, n$ and $r = 1, 2, \ldots, m$. Applications of RSS can be found, for example, in [5–8]. A wide list of applications of RSS is available in [3]. Moreover, several studies have been showed the efficiency of RSS relative to SRS, as described hereafter.

McIntyre [1] proposed the mean of a ranked set sample as an unbiased estimator of the population mean. The efficiency of this estimator, relative to the mean of a simple random sample, was verified by Takahasi and Wakimoto [9] and Dell and Clutter [10]. Regarding the estimation of the variance, Stokes [11] presents an asymptotically unbiased estimator with mean squared error smaller than the sample variance under SRS. Yu et al. [12] proposed a modified version of Stokes's estimator [11], which is unbiased under normal distribution. MacEachern et al. [13] proposed an unbiased estimator for the population variance, applicable when there are at least two cycles in the sampling process. Estimators for other population parameters under RSS were also explored. For example, Stokes and Sager [14] present an unbiased estimator for the distribution function based on RSS, which has smaller variance than the one obtained via SRS. Barnett and Moore [15] present the BLUEs – best linear unbiased estimators – for the parameters of of the location-scale family. Other studies about the performance of estimators and tests based on RSS can be found, for example, in [16–20].

The literature about RSS is still scarce regarding the estimation of the coeficient of variation, when compared to the estimation of other population parameters. The coefficient of variation is an important measure of relative dispersion, which is dimensionless, allowing to compare the dispersion of measurements in different scales. Additionally, the coefficient of variation is an usual measure of precision in experimentation [21]. Albatineh et al. [22] presents a study on the performance of confidence intervals based on one specific point estimator of the coefficient of variation (which is also considered in our work). The authors showed, for different distributions and sample sizes, the best performance of RSS confidence intervals relative to their SRS counterpart.

In this work, we propose and evaluate the performance of different estimators for the coefficient of variation under RSS. The efficiency of the proposed estimators, relative to the usual SRS estimator, is assessed through a Monte Carlo simulation study. Moreover, the performance of parametric and nonparametric estimators, as well as their corresponding bias, are also compared. The methodological procedures are yet illustrated with an application to anthropometric data from a nutrition study.

2. Estimation of mean and variance via RSS

Here we present estimators for the population mean and variance in the context of RSS.

2.1. Nonparametric estimators for mean and variance

(a) McIntyre's mean estimator

Downloaded by [CAPES] at 09:43 29 September 2017

The population mean estimator proposed by McIntyre [1] is the mean of a ranked set sample, defined as:

$$\hat{\mu}_{\mathrm{MI}} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{r=1}^{n} X_{[i]r}, \tag{1}$$

where n is the number of sets, m is the number of cycles and $X_{[i]r}$ the observation of order i in the rth cycle, with $i = 1, \ldots, n$ and $r = 1, \ldots, m$. Takahasi and Wakimoto [9] verified that this estimator presents smaller variance compared to the correspondent estimator obtained via SRS. Dell and Clutter [10] investigated the effect of ranking errors in the efficiency of $\hat{\mu}_{\mathrm{MI}}$, verifying its superiority relative to the SRS counterpart, unless the ranking process is completely arbitrary, which yields equivalent precision to both designs.

(b) Stokes's variance estimator
The variance estimator proposed by Stokes [11] is an extension of the unbiased estimator usually applied for SRS. It is defined as:

$$\hat{\sigma}^2_{\mathrm{S}} = \frac{1}{mn-1}\sum_{i=1}^{n}\sum_{r=1}^{m}(X_{[i]r} - \hat{\mu})^2, \tag{2}$$

which is an asymptotically unbiased estimator for σ^2.

(c) Modified Stokes's variance estimator
An unbiased version of (2) when the population has normal distribution was proposed by Yu et al. [12]. It is given by:

$$\hat{\sigma}^2_{\mathrm{MS}} = \frac{1}{mn - 1 + \sum_{i=1}^{n}\frac{\alpha_i^2}{n}}\sum_{r=1}^{m}\sum_{i=1}^{n}(X_{[i]r} - \hat{\mu})^2, \tag{3}$$

where α_i is the expected value of the ith order statistic from a simple random sample of size n under standard normal distribution, $i = 1, \ldots, n$.

(d) MacEachern's variance estimator

An unbiased estimator of the population variance based on RSS was proposed by MacEachern et al. [13]. Different from Stokes's estimator [11], this one makes direct use of the positions where each element was ranked, providing a more efficient variance estimator. The MacEachern et al. estimator is defined as:

$$\hat{\sigma}^2_{ME} = \frac{\sum_{i\neq s}\sum_r\sum_j (X_{[i]r} - X_{[s]j})^2}{2m^2n^2} + \frac{\sum_i\sum_r\sum_j (X_{[i]r} - X_{[i]j})^2}{2m(m-1)n^2}. \tag{4}$$

Because this estimator incorporates the variability between observations ranked in the same position, it is only applicable when there are at least two sampling cycles ($m > 2$).

2.2. Parametric estimators for mean and variance

(a) Best linear unbiased estimator
If the distribution belongs to the location-scale family, the best linear unbiased estimators (BLUEs) represent an efficient alternative to estimate different population

Downloaded by [CAPES] at 09:43 29 September 2017

parameters [2,15]. Particularly, if $X \sim N(\mu, \sigma^2)$, the best linear unbiased estimator for the population mean is given by:

$$\hat{\mu}_{\text{BLUE}} = \frac{\sum_{i=1}^{n} \upsilon_i^{-1} \bar{X}_{(i)}}{\sum_{i=1}^{n} \upsilon_i^{-1}}, \tag{5}$$

where υ_i is the variance of the ith order statistic of the standard normal distribution and

$$\bar{X}_{(i)} = \frac{1}{m} \sum_{r=1}^{m} X_{(i)r}. \tag{6}$$

The BLUE is an alternative to the maximum likelihood method, especially when the likelihood maximization is too complicated. However, these estimators are usually less efficient than the MLE competitors.

(b) Maximum likelihood estimators

The maximum likelihood estimators in RSS are obtained, in general, using numerical methods, once the maximization process by analytical methods is usually difficult. The estimation via maximum likelihood in RSS is considered in [2,17,23], amongst others. Let us consider again $X \sim N(\mu, \sigma^2)$, such that:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \quad \text{and} \quad F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right), \tag{7}$$

where $\Phi(\cdot)$ is the accumulated distribution function of the standard normal distribution $(\mu = 0, \sigma^2 = 1)$, $x \in (-\infty, \infty)$, $\mu \in (-\infty, \infty)$ and $\sigma > 0$. The maximum likelihood estimation, considering a ranked set sample $x_{(1)1}, x_{(1)2}, \ldots, x_{(n)m}$ from a normal distribution, under perfect ranking, is based on the following likelihood function:

$$L(\mu, \sigma^2; x_{(1)1}, x_{(1)2}, \ldots, x_{(n)m}) = \prod_{r=1}^{m} \prod_{i=1}^{n} g_i(x_{(i)r}; \mu, \sigma^2), \tag{8}$$

where $g_i(x_{(i)\cdot}; \mu, \sigma^2)$ is the distribution of the ith order statistic of a simple random sample of $X \sim N(\mu, \sigma^2)$ [24]. Hence:

$$\begin{aligned} L(\mu, \sigma^2; x_{(1)1}, x_{(1)2}, \ldots, x_{(n)m}) = \prod_{r=1}^{m} \prod_{i=1}^{n} & \frac{n!}{(i-1)!(n-i)!} \times [F(x_{(i)r})]^{i-1} \\ & \times [1 - F(x_{(i)r})]^{n-i} f(x_{(i)r}), \end{aligned} \tag{9}$$

where $f(\cdot)$ and $F(\cdot)$ are defined in Equation (7). The log-likelihood function is then given by:

$$\begin{aligned} & l(\mu, \sigma^2; x_{(1)1}, x_{(1)2}, \ldots, x_{(n)m}) \\ & \quad = \sum_{r=1}^{m} \sum_{i=1}^{n} \{(i-1)\log(F(x_{(i)r})) + (n-i)\log(1 - F(x_{(i)r})) + \log(f(x_{(i)r}))\}. \end{aligned} \tag{10}$$

The maximum likelihood estimators for μ and σ result from the maximization of Equation (10), and are denoted here, respectively, by $\hat{\mu}_{\text{ML}}$ and $\hat{\sigma}^2_{\text{ML}}$.

Downloaded by [CAPES] at 09:43 29 September 2017

Downloaded by [CAPES] at 09:43 29 September 2017

3. Proposed estimators for the coefficient of variation

Based on the presented estimators for the mean and variance under RSS, we proposed the following estimators for the coefficient of variation.

(a) Stokes's/McIntyre's estimator
It is a nonparametric estimator, defined by the ratio between the square root of the Stokes's estimator (2) and McIntyre's estimator for the mean (1). In this way, we have an immediate extension to the usual SRS estimator:

$$\hat{\tau}_1 = \frac{\sqrt{\hat{\sigma}_S^2}}{\hat{\mu}_{\mathrm{MI}}}. \tag{11}$$

(b) MacEachern's/McIntyre's estimator
In this case, the Stokes's estimator of the variance is replaced by the unbiased MacEachern et al.'s estimator (4):

$$\hat{\tau}_2 = \frac{\sqrt{\hat{\sigma}_{ME}^2}}{\hat{\mu}_{\mathrm{rep}}}. \tag{12}$$

(c) Stokes's modified estimator/BLUE
Assuming a population with normal distribution, the Stokes's modified estimator for the variance (3) is an unbiased and more precise version than the original Stokes's estimator (2). As a third alternative for the estimation of the coefficient of variation, we consider this estimator conjugated with the BLUE for the mean. This proposed estimator is given by:

$$\hat{\tau}_3 = \frac{\sqrt{\hat{\sigma}_{\mathrm{MS}}^2}}{\hat{\mu}_{\mathrm{BLUE}}}. \tag{13}$$

In our study, $\hat{\sigma}_{\mathrm{BLUE}}$ was not considered because this estimator may produce negative estimates for σ.

(d) Maximum likelihood estimator

Finally, given the already proven efficiency of the maximum likelihood estimators, as well as their desired properties (sufficiency, efficiency for big samples, invariance principle and asymptotic normality, among others [25]), we propose the maximum likelihood estimation of the coefficient of variation which, by the property of invariance of this class of estimators, can be represented by:

$$\hat{\tau}_4 = \frac{\sqrt{\hat{\sigma}_{\mathrm{ML}}^2}}{\hat{\mu}_{\mathrm{ML}}}. \tag{14}$$

4. Simulation study

To assess the performance of the proposed estimators and compare them with the usual SRS' estimator ($\hat{\tau}_0 = s/\bar{x}$), we carried out a Monte Carlo simulation study. In our study, we

Table 1. Relative efficiency between the RSS estimators and the SRS estimator for $\tau = 5\%$.

Size of sets (n)	Number of cycles (m)	Relative efficiency $RE(\hat{\tau}_1, \hat{\tau}_0)$	$RE(\hat{\tau}_2, \hat{\tau}_0)$	$RE(\hat{\tau}_3, \hat{\tau}_0)$	$RE(\hat{\tau}_4, \hat{\tau}_0)$
3	2	1.0617	1.1931	1.1476	1.3464
	3	1.0625	1.1504	1.1087	1.3077
	5	1.0241	1.0897	1.0625	1.2500
	8	1.1224	1.1458	1.1458	1.3415
5	2	1.2762	1.3814	1.3400	1.5765
	3	1.3235	1.3846	1.3636	1.5789
	5	1.2927	1.3590	1.3250	1.5588
	8	1.3200	1.3750	1.3750	1.6500
8	2	1.5849	1.7143	1.7143	2.0488
	3	1.5882	1.6364	1.6364	1.9286
	5	1.6500	1.6500	1.6500	2.0625
	8	1.6667	1.6667	1.6667	2.0000
10	2	1.7568	1.8571	1.8571	2.2414
	3	1.8750	1.9565	1.9565	2.3684
	5	1.8571	1.8571	1.8571	2.3636
	8	1.7778	1.7778	1.7778	2.2857

simulated data from a normal distribution, considering different values for the coefficient of variation ($\tau = 5\%, 10\%, 20\%, 33\%$ and 50%); set sizes ($n = 3, 5, 8$ and 10) and number of cycles ($m = 2, 3, 5$ and 8).

The τ values were achieved by fixing the standard deviation ($\sigma = 1$) and varying the mean ($\mu = 20, 10, 5, 3$ and 2, respectively). The number of simulations was $10,000$, based on the results of a previous study, in which we verified that from approximately 3000 simulations, the change in the variance of the estimators becomes negligible. So, for each combination of τ, n and m, we simulated 10,000 samples, and computed, for each simulated sample, the estimates of the coefficient of variation provided by each one of the proposed estimators.

To assess the performance of the estimators, we calculated the relative bias (RB) and mean squared error (MSE). Let $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_{10,000}$ be the estimates obtained by the simulation study for a particular estimator and a combination of m, n and θ, the true value of the parameter. Then,

$$\mathrm{RB} = \frac{1}{10,000} \sum_{j=1}^{10,000} \frac{\hat{\theta}_j - \theta}{\theta}; \quad \mathrm{MSE} = \frac{1}{10,000} \sum_{j=1}^{10,000} (\hat{\theta}_j - \theta)^2. \tag{15}$$

Afterwards, we computed the relative efficiencies based on the ratios of mean squared errors. Each proposed estimator had its mean squared error compared to that relative to $\hat{\tau}_0 = s/\bar{x}$, the SRS' estimator for samples of size $N = mn$, by means of:

$$\mathrm{RE}(\hat{\tau}_k, \hat{\tau}_0) = \frac{\mathrm{MSE}(\hat{\tau}_0)}{\mathrm{MSE}(\hat{\tau}_k)}, \quad k = 1, 2, 3, 4. \tag{16}$$

Tables 1– 5 present the relative efficiencies provided by the simulation study. Each table refers specifically to one of the fixed values for τ. Based on these results it is possible conclude that:

Downloaded by [CAPES] at 09:43 29 September 2017

Downloaded by [CAPES] at 09:43 29 September 2017

Table 2. Relative efficiency between the RSS estimators and the SRS estimator for $\tau = 10\%$.

Size of sets (n)	Number of cycles (m)	Relative efficiency $RE(\hat{\tau}_1, \hat{\tau}_0)$	$RE(\hat{\tau}_2, \hat{\tau}_0)$	$RE(\hat{\tau}_3, \hat{\tau}_0)$	$RE(\hat{\tau}_4, \hat{\tau}_0)$
3	2	1.0654	1.2000	1.1506	1.3602
	3	1.0653	1.1503	1.1171	1.3053
	5	1.0390	1.0881	1.0679	1.2628
	8	1.1206	1.1554	1.1378	1.3434
5	2	1.2820	1.3836	1.3559	1.5819
	3	1.3200	1.3855	1.3647	1.5991
	5	1.3110	1.3522	1.3438	1.5809
	8	1.3434	1.3711	1.3711	1.6024
8	2	1.5991	1.7035	1.6950	2.0545
	3	1.5971	1.6692	1.6567	2.0367
	5	1.6543	1.6962	1.6962	2.0615
	8	1.5800	1.6122	1.6122	1.9750
10	2	1.7852	1.8732	1.8601	2.2735
	3	1.8958	1.9783	1.9783	2.3947
	5	1.8750	1.9091	1.9091	2.3333
	8	1.8056	1.8571	1.8571	2.2414

Table 3. Relative efficiency between the RSS estimators and the SRS estimator for $\tau = 20\%$.

Size of sets (n)	Number of cycles (m)	Relative efficiency $RE(\hat{\tau}_1, \hat{\tau}_0)$	$RE(\hat{\tau}_2, \hat{\tau}_0)$	$RE(\hat{\tau}_3, \hat{\tau}_0)$	$RE(\hat{\tau}_4, \hat{\tau}_0)$
3	2	1.0830	1.2225	1.1764	1.3979
	3	1.0821	1.1714	1.1394	1.3333
	5	1.0637	1.1179	1.0970	1.2952
	8	1.1446	1.1790	1.1658	1.3672
5	2	1.3140	1.4226	1.3949	1.6303
	3	1.3545	1.4249	1.4079	1.6481
	5	1.3481	1.3914	1.3830	1.6250
	8	1.3857	1.4100	1.4065	1.6686
8	2	1.6498	1.7632	1.7546	2.1257
	3	1.6462	1.7185	1.7122	2.1029
	5	1.7066	1.7538	1.7538	2.1269
	8	1.6520	1.6850	1.6850	2.0549
10	2	1.8492	1.9482	1.9415	2.3598
	3	1.9617	2.0344	2.0344	2.4569
	5	1.9264	1.9690	1.9778	2.4185
	8	1.8699	1.8958	1.8958	2.3136

- The relative efficiency is larger than 1 in all simulated scenarios. The higher efficiency of the RSS estimators, relative to the SRS counterpart, was already expected, due to the additional information used in the ranking process. We could observe relative efficiency values ranging between 1.02 and 2.97;
- Fixing the value of the coefficient of variation, the relative efficiency increases as the set size (n) is larger for any of the four estimators under RSS. However, when fixing the value of the coefficient of variation and set size, an increase in the number of cycles (m) does not provide evidence of an increase in relative efficiency;
- For all the proposed estimators, we verified higher relative efficiency for the largest values for τ. Hence, the efficiency obtained by using RSS instead of SRS increases with a higher relative dispersion of the data;

Table 4. Relative efficiency between the RSS estimators and the SRS estimator for $\tau = 33\%$.

Size of sets (n)	Number of cycles (m)	Relative efficiency $\mathrm{RE}(\hat{\tau}_1, \hat{\tau}_0)$	$\mathrm{RE}(\hat{\tau}_2, \hat{\tau}_0)$	$\mathrm{RE}(\hat{\tau}_3, \hat{\tau}_0)$	$\mathrm{RE}(\hat{\tau}_4, \hat{\tau}_0)$
3	2	1.1347	1.2868	1.2440	1.4954
	3	1.1354	1.2321	1.2029	1.4116
	5	1.1228	1.1814	1.1613	1.3697
	8	1.1982	1.2350	1.2235	1.4280
5	2	1.3945	1.5143	1.4920	1.7417
	3	1.4393	1.5191	1.5065	1.7551
	5	1.4387	1.4873	1.4834	1.7328
	8	1.4736	1.5017	1.5017	1.7630
8	2	1.7712	1.8962	1.8986	2.2862
	3	1.7590	1.8394	1.8417	2.2440
	5	1.8241	1.8749	1.8809	2.2619
	8	1.7778	1.8118	1.8182	2.2034
10	2	2.0000	2.1101	2.1139	2.5534
	3	2.1024	2.1821	2.1942	2.6321
	5	2.0650	2.1128	2.1260	2.5841
	8	1.9833	2.0121	2.0268	2.4572

Table 5. Relative efficiency between the RSS estimators and the SRS estimator for $\tau = 50\%$.

Size of sets (n)	Number of cycles (m)	Relative efficiency $\mathrm{RE}(\hat{\tau}_1, \hat{\tau}_0)$	$\mathrm{RE}(\hat{\tau}_2, \hat{\tau}_0)$	$\mathrm{RE}(\hat{\tau}_3, \hat{\tau}_0)$	$\mathrm{RE}(\hat{\tau}_4, \hat{\tau}_0)$
3	2	1.2955	1.4788	1.4439	1.7676
	3	1.2793	1.3935	1.3710	1.6146
	5	1.2461	1.3141	1.2973	1.5239
	8	1.2995	1.3415	1.3338	1.5435
5	2	1.5870	1.7330	1.7218	2.0043
	3	1.6207	1.7171	1.7156	1.9848
	5	1.6116	1.6689	1.6758	1.9376
	8	1.6352	1.6681	1.6783	1.9398
8	2	2.0335	2.1843	2.2078	2.6330
	3	1.9893	2.0839	2.1037	2.5302
	5	2.0449	2.1032	2.1282	2.5224
	8	2.0020	2.0404	2.0631	2.4603
10	2	2.3073	2.4428	2.4701	2.9456
	3	2.3756	2.4683	2.5095	2.9710
	5	2.3282	2.3836	2.4221	2.8904
	8	2.2075	2.2412	2.2759	2.7179

- When comparing the performance of the four estimators, we observe that the maximum likelihood estimator ($\hat{\tau}_4$) dominates the other estimators, presenting higher efficiency in all the simulated scenarios;
- Within the nonparametric estimators, it is clear that $\hat{\tau}_2$, based on the variance estimator proposed by MacEachern et al., presents better performance than its opponent ($\hat{\tau}_1$). Moreover, $\hat{\tau}_2$ presents very similar performance to $\hat{\tau}_3$, based on the BLUE for the mean and on the modified (unbiased) version of Stokes's estimator.

Another important property to be assessed is the bias of the proposed estimators. Figures 1 and 2 present their relative bias, calculated as described in Equation (15). We

Downloaded by [CAPES] at 09:43 29 September 2017

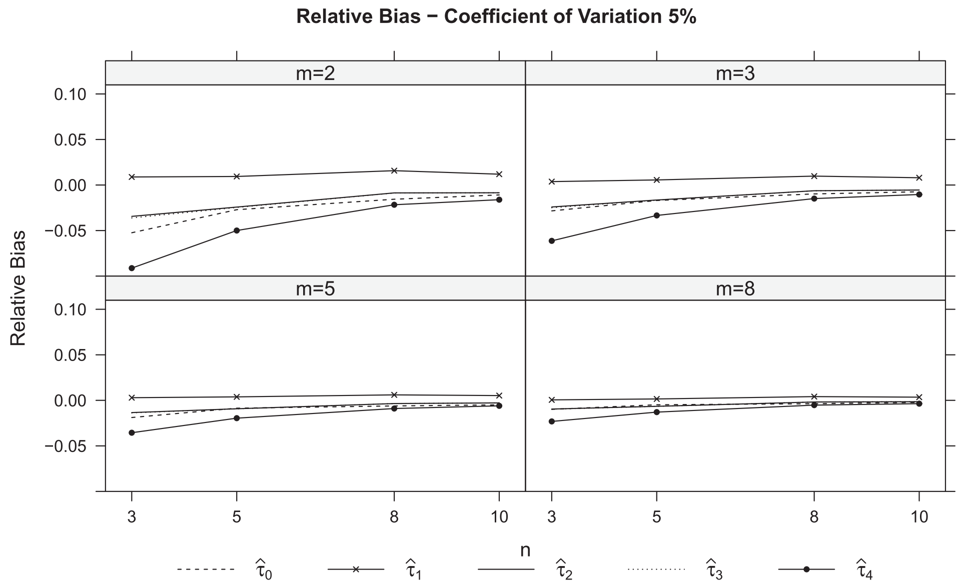


Figure 1. Relative bias of proposed estimators for $\tau = 5\%$.

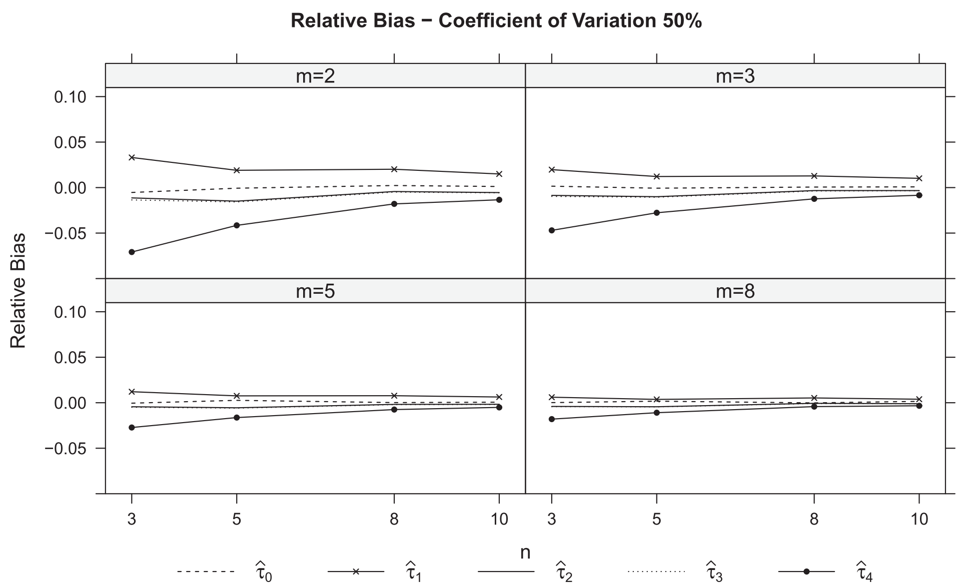
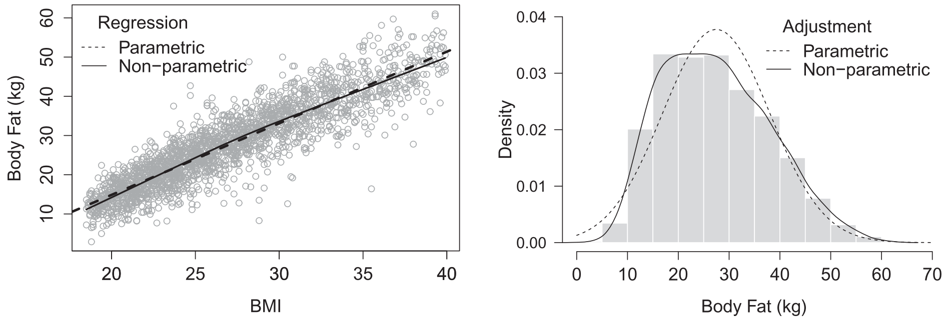


Figure 2. Relative bias of proposed estimators for $\tau = 50\%$.

observe in Figure 1 a trend to underestimate the true value of the coefficient of variation, except for $\hat{\tau}_1$, which presents small positive bias. The maximum likelihood estimator ($\hat{\tau}_4$) presents a larger negative bias than the other estimators. The negative bias of $\hat{\tau}_0$, estimator based on SRS, is a known result, and there are correction factors to be used. Finally, for all studied estimators, in general the bias becomes smaller as the sample size ($N = mn$) increases.

Downloaded by [CAPES] at 09:43 29 September 2017



(a) Scatter plot of the variables Body Fat and BMI. (b) Histogram for the variable Body Fat.

Figure 3. Correlation analysis between the variable of interest Body Fat and the concomitant variable BMI and descriptive analysis of the Body Fat. (a) Scatter plot of the variables Body Fat and BMI and (b) Histogram for the variable Body Fat.

Downloaded by [CAPES] at 09:43 29 September 2017

For $\tau = 50\%$, as observed in Figure 2, in general there is a smaller bias regarding that observed for $\tau = 5\%$ (Figure 1). Besides that, a larger negative bias is again observed for the maximum likelihood estimator ($\hat{\tau}_4$) and it reduces as $N = mn$ increases (indicating absence of bias, asymptotically). Moreover, $\hat{\tau}_1$, based on the variance estimator proposed by Stokes [11], presented positive bias, even under scenarios involving a smaller sample size.

5. Simulation based on real data

To complement this study, a real data set was used as a reference population for a new simulation study. Our objective was to analyse the performance of the estimators when applied in a real population. Additionally, it was possible to assess the impact generated by the imperfect ranking of the sets. We used data from a cross-sectional study carried out between 2002 and 2009 by Human Unity from University of Rome 'Tor Vergata', regarding the nutritional status of apparently healthy adults from the Center-South region of Italy [26]. Height, weight, body mass index, fat and lean mass indices, among others, are available. The sample comprises records obtained from 2089 individuals, composing the reference population for this application.

Although there are some techniques capable of measuring body fat mass with high precision, it is known that this may involve the use of very expensive equipments, with difficult access. However, there is a strong correlation between fat mass and other anthropometric measures, like the body mass index (BMI), as it can be seen in Figure 3(a). Once the BMI is easily obtained simply by measuring height and weight, RSS is potentially applicable as a sampling design for inference regarding fat mass distribution, taking BMI as the concomitant variable in the ranking process.

Hence, considering fat mass quantity as the variable of interest and the BMI as the concomitant variable, we repeated the simulation study presented in Session 4 for this pair of variables, taking the same combinations of n and m previously considered. Simulations were carried out in two distinct ways: under perfect ranking (sampling units ranked directly by fat mass values) and imperfect ranking (using BMI to rank the sets). The simulation under perfect ranking has the objective of identifying the maximum efficiency

Table 6. Relative efficiency between the RSS estimators and the SRS estimator with perfect ranking.

Size of sets (n)	Number of cycles (m)	Relative efficiency			
		RE($\hat{\tau}_1, \hat{\tau}_0$)	RE($\hat{\tau}_2, \hat{\tau}_0$)	RE($\hat{\tau}_3, \hat{\tau}_0$)	RE($\hat{\tau}_4, \hat{\tau}_0$)
3	2	1.1552	1.3380	1.2416	1.4146
	3	1.1596	1.2916	1.2133	1.3690
	5	1.1654	1.2511	1.1903	1.3447
	8	1.1481	1.2100	1.1609	1.2810
5	2	1.3936	1.5817	1.4438	1.6111
	3	1.4125	1.5552	1.4277	1.5333
	5	1.4663	1.5722	1.4488	1.5019
	8	1.4399	1.5215	1.3983	1.3641
8	2	1.8476	2.0893	1.8476	1.9433
	3	1.7719	1.9465	1.7256	1.7292
	5	1.7951	1.9343	1.6978	1.5128
	8	1.8063	1.9242	1.6618	1.3076
10	2	1.9947	2.2602	1.9697	1.9068
	3	1.9796	2.1849	1.8852	1.7150
	5	1.9599	2.1193	1.7948	1.4103
	8	1.9677	2.1037	1.7340	1.1478

provided by RSS, while under imperfect ranking we aim to assess the impact of ranking errors on the efficiency of the proposed estimators. The coefficient of variation for fat mass in the reference population is $\tau = 38\%$.

We observe in Figure 3(b) that the fat mass distribution presents slight asymmetry to the right, which allows us to assess the performance of the proposed estimators under non-normal distribution. Moreover, we observe that fat mass and BMI present strong linear correlation, such that the Pearson correlation coefficient for these data is $r = 0.84$.

Perfect ranking

Table 6 presents the simulation results based on the anthropometric data under perfect ranking. We observe that the proposed estimators obtained better performance than this SRS counterpart in all simulated scenarios. The relative efficiency of the four proposed estimators ranges between 1.14 and 2.26. The superiority of the maximum likelihood estimator relative to the other RSS estimators is no longer noted, with this estimator being less efficient than their competitors in several scenarios. This fact may be attributed to the non-normality for the reference distribution. Moreover, $\hat{\tau}_2$, based on MacEachern et al.'s variance estimator was, in most cases, more efficient than the other estimators.

Imperfect ranking

Table 7 presents the simulation results based on anthropometric data under imperfect ranking. As expected, we observe smaller efficiency of the estimators based on RSS when compared to the results obtained under perfect ranking. Also, $\hat{\tau}_2$ consistently presents higher efficiency when compared to the others estimators. Additionally, the maximum likelihood estimator ($\hat{\tau}_4$) is less efficient than the SRS estimator in most of the simulated scenarios. This results from two factors: the fact that we are not simulating from a normal distribution and the imperfect ranking of the sets. The latter implies on the sampling observations no more matching to order statistics, and the likelihood function presented in Equation (9) is no longer valid. Even so, the results presented previously for this estimator

Downloaded by [CAPES] at 09:43 29 September 2017

Table 7. Relative precision between the RSS estimators and the SRS estimator with imperfect ranking.

Size of sets (n)	Number of cycles (m)	Relative precision $RP(\hat{\tau}_1, \hat{\tau}_0)$	$RP(\hat{\tau}_2, \hat{\tau}_0)$	$RP(\hat{\tau}_3, \hat{\tau}_0)$	$RP(\hat{\tau}_4, \hat{\tau}_0)$
3	2	1.1508	1.2855	1.2252	1.1754
	3	1.1244	1.2107	1.1644	1.0863
	5	1.1261	1.1869	1.1410	0.9700
	8	1.0965	1.1397	1.1023	0.8400
5	2	1.3478	1.4689	1.3587	1.1144
	3	1.2790	1.3444	1.2610	0.9228
	5	1.3177	1.3653	1.2944	0.7017
	8	1.1856	1.2351	1.1472	0.4790
8	2	1.6222	1.7168	1.5623	0.8775
	3	1.4846	1.5169	1.4025	0.6590
	5	1.5067	1.5333	1.4264	0.4175
	8	1.2252	1.2983	1.1133	0.2140
10	2	1.6600	1.7312	1.5636	0.6734
	3	1.5911	1.5964	1.4738	0.5279
	5	1.6624	1.6791	1.5606	0.3110
	8	1.1587	1.2436	1.0055	0.1385

serve as reference for the maximum efficiency reached by $\hat{\tau}_4$. Therefore, $\hat{\tau}_2$ is a robust alternative to $\hat{\tau}_4$ under imperfect ranking and non-normality.

6. Conclusions

In this work, we have verified, through a simulation study, that different proposed estimators for the coefficient of variation under RSS, considering the normal distribution, present higher efficiency than the usual SRS estimator. Considering perfect ranking, the maximum likelihood estimator presented the best performance. Regarding nonparametric estimators, that composed by the ratio of MacEachern et al.'s variance estimator and McIntyre's estimator for the mean was the one that presented the highest efficiency.

When carrying out a new simulation study from a real data set, we conclude that the superiority of the estimators provided by RSS remains, even under the imperfect ranking scenario, however with lower relative efficiency. However, in this scenario, the maximum likelihood estimator has its efficiency drastically reduced, even presenting lower efficiency than SRS estimator. Maximum likelihood estimation of the coefficient of variation under imperfect ranking requires additional and specific studies. In this work, we have approached the maximum gain reached by it (under perfect ranking).

The relative efficiencies of the estimators for the coefficient of variation under RSS are higher than those obtained by variance estimators, as we can see in [11,13,27]. Hence, based on our simulation results, the estimation of the coefficient of variation as a measurement of relative dispersion via RSS is strongly recommended, with potential use in several applications in statistics.

Disclosure statement

No potential conflict of interest was reported by the authors.

Downloaded by [CAPES] at 09:43 29 September 2017

References

[1] McIntyre GA. A method for unbiased selective sampling, using ranked sets. Crop Pasture Sci. 1952;3(4):385–390.
[2] Chen Z, Bai Z, Sinha B. Ranked set sampling: theory and applications. New York (NY): Springer Science & Business Media; 2003.
[3] Al-Omari AI, Bouza CN. Review of ranked set sampling: modifications and applications. Rev Investig Oper. 2014;35(3):215–240.
[4] Sinha AK. Ranked set sampling: as a cost-effective and more efficient data collection method. Stat J Intern Assoc Off Stat. 2016;32(4):607–611.
[5] Cobby JM, Ridout MS, Bassett PJ, et al. An investigation into the use of ranked set sampling on grass and grass-clover swards. Grass Forage Sci. 1985;40(3):257–263.
[6] Husby CE, Stasny EA, Wolfe DA. An application of ranked set sampling for mean and median estimation using usda crop production data. J Agric Biol Environ Stat. 2005;10(3): 354–373.
[7] Tiwari N, Pandey GS. Application of ranked set sampling design in environmental investigations for real data set. Thail Stat. 2013;11(2):173–184.
[8] Arslan M, Afzal M, Amin I, et al. Nutrients can enhance the abundance and expression of alkane hydroxylase cyp153 gene in the rhizosphere of ryegrass planted in hydrocarbon-polluted soil. PloS one. 2014;9(10):e111208.
[9] Takahasi K, Wakimoto K. On unbiased estimates of the population mean based on the sample stratified by means of ordering. Ann Inst Statist Math. 1968;20(1):1–31.
[10] Dell TR, Clutter JL. Ranked set sampling theory with order statistics background. Biometrics. 1972;28(2):545–555.
[11] Stokes SL. Estimation of variance using judgment ordered ranked set samples. Biometrics. 1980;36(1):35–42.
[12] Yu PLH, Lam K, Sinha BMK. Estimation of normal variance based on balanced and unbalanced ranked set samples. Env Ecol Stat. 1999;6(1):23–46.
[13] MacEachern SN, Öztürk Ö, Wolfe DA, et al. A new ranked set sample estimator of variance. J R Stat Soc Ser B Stat Methodol. 2002;64(2):177–188.
[14] Stokes SL, Sager TW. Characterization of a ranked-set sample with application to estimating distribution functions. J Amer Statist Assoc. 1988;83(402):374–381.
[15] Barnett V, Moore K. Best linear unbiased estimates in ranked-set sampling with particular reference to imperfect ordering. J Appl Stat. 1997;24(6):697–710.
[16] Terpstra JT, Nelson EJ. Optimal rank set sampling estimates for a population proportion. J Statist Plann Inference. 2005;127(1):309–321.
[17] Ozturk O. Parametric estimation of location and scale parameters in ranked set sampling. J Statist Plann Inference. 2011;141(4):1616–1622.
[18] Strzalkowska-Kominiak E, Mahdizadeh M. On the Kaplan–Meier estimator based on ranked set samples. J Stat Comput Simul. 2014;84(12):2577–2591.
[19] Zamanzade E, Vock M. Variance estimation in ranked set sampling using a concomitant variable. Statist Probab Lett. 2015;105:1–5.
[20] Noughabi AH. Efficiency of ranked set sampling in tests for normality. J Stat Comput Simul. 2016;87(5):956–965.
[21] Montgomery D. Design and analysis of experiments. 7th ed. New York (NY): John Wiley & Sons; 2008.
[22] Albatineh AN, Kibria BMG, Wilcox ML, et al. Confidence interval estimation for the population coefficient of variation using ranked set sampling: a simulation study. J Appl Stat. 2014;41(4):733–751.
[23] Stokes L. Parametric ranked set sampling. Ann Inst Statist Math. 1995;47(3):465–482.
[24] Balakrishnan N, Cohen AC. Order statistics & inference: estimation methods. San Diego (CA): Academic Press; 1991.
[25] Casella G, Berger RL. Statistical inference. 2nd ed. Pacific Grove (CA): Duxbury; 2002.

Downloaded by [CAPES] at 09:43 29 September 2017

[26] De Lorenzo A, Bianchi A, Maroni P, et al. Adiposity rather than bmi determines metabolic risk. Int J Cardiol. 2013;166(1):111–117.

[27] Garcia CA, Barreto MCM. Precisão relativa para estimadores da variância de uma distribuiç ao normal sob o delineamento por conjuntos ordenados [Relative precision for estimators of the variance of normal distribution under ranked set sampling]. Rev Mat Est. 2005;23(1): 67–80.

Downloaded by [CAPES] at 09:43 29 September 2017