STATISTICS IN TRANSITION new series, September 2022
Vol. 23, No. 3, pp. 199–214, DOI 10.2478/stattrans-2022-0038
Received – 24.08.2021; accepted – 26.05.2022

Two-stage cluster sampling with unequal probability sampling in the first stage and ranked set sampling in the second stage

Michael C. Ugwu[1], Mbanefo S. Madukaife[2]

ABSTRACT

In this research work we introduce a new sampling design, namely a two-stage cluster sampling, where probability proportional to size with replacement is used in the first stage unit and ranked set sampling in the second in order to address the issue of marked variability in the sizes of population units concerned with first stage sampling. We obtained an unbiased estimator of the population mean and total, as well as the variance of the mean estimator. We calculated the relative efficiency of the new sampling design to the two-stage cluster sampling with simple random sampling in the first stage and ranked set sampling in the second stage. The results demonstrated that the new sampling design is more efficient than the competing design when a significant variation is observed in the first stage units.

Key words: cluster sampling, population mean estimator, probability proportional to size sampling, ranked set sampling, relative efficiency.

1. Introduction

In scientific research, sample survey to a great extent plays a vital role, most especially in the presence of limited cost. This is because we need not possibly embark on complete enumeration which entails studying the entire population, in order to learn efficiently about the population characteristics of interest. Also in real life situations, there are occasions unlike in element sampling when a list of elements of the population is not available but it is easy (or possible) to obtain a list of segmented groups, known as clusters. Even when such list exists, it is sometimes uneconomical to obtain information from a sample of elements in the population due to the nature of the distribution of the population. In such cases, it becomes ideal to use cluster sampling technique to draw random sample from the population and when this

[1] Department of Statistics, University of Nigeria, Nsukka, Nigeria. ORCID: https://orcid.org/0000-0001-7356-9183

[2] Corresponding author. Department of Statistics, University of Nigeria, Nsukka. Nigeria. E-mail: mbanefo.madukaife@unn.edu.ng. ORCID: https://orcid.org/0000-0003-2823-4223.

technique is carried out in two phases, it becomes two-stage cluster sampling (Okafor, 2002).

In two-stage cluster sampling, the entire units of the population are at first grouped into say *N* clusters, each having M_i, $i = 1, 2, \ldots, N$ elements. Then a random sample of *n* clusters, say, is drawn from the *N* clusters, also known as the first stage units (FSU), as the first stage sample. From each of the *n* selected clusters, each of size M_i, $i = 1, 2, ..., n$ elements, a random sample of cluster elements of size m_i is also selected from M_i second stage units (SSU) as second stage sample. The common motivation of cluster sampling is to reduce cost by increasing sampling efficiency.

A good number of authors have applied two-stage cluster sampling in real life situations in order to enhance sampling efficiency. Some of them include Fears and Gail (2000), Stehman et al. (2009), Phillips et al. (2008), Horney et al. (2010) as well as Galway et al. (2012) and Dilip (2015). The efficiency of the design when applied to real life situations, however depends to great extent on the sampling techniques used in both stages of the design.

It could be recalled that in equal probability sampling, all the population units have equal chances of being selected in the sample regardless of the size of each unit. When units of clusters are of different sizes, it is appropriate to use probability proportional to size (PPS) sampling (Damon, 2018 & Ozturk, 2019). In this sampling plan, the probability of selection of a cluster element is in proportion to its size or measure of size of the element, so that larger clusters have greater chances of being selected than the smaller clusters, provided the sizes of units of clusters in the population are known and also have positive correlation with the variable under study. The choice of PPS scheme in the first-stage of two-stage sampling under variant cluster sizes has also been supported by Innocenti et al. (2019). Such a procedure of sample selection is also known as unequal probability sampling (Okafor, 2002). For a more detailed discussion on selection procedures and estimation in unequal probability sampling, see Shahbaz and Hanif (2010).

Optimum sampling methods that are cost friendly have been of great concern in the field of statistics, especially when the cost of measuring the population attribute under study is high. In situations where it is less costly to identify sampling units to be included in the sample and at the same time ranking them accordingly with respect to the attribute of interest than to directly measure the values, a ranked set sample (RSS) yields better efficiency than its simple random sample (SRS) counterpart under the same sample size (McIntyre, 1952 and Chen et al., 2003). Ranked set sampling was introduced by McIntyre (1952) and Halls and Dell (1966) while its theoretical basis was laid by Takahasi and Wakimoto (1968) and Dell and Clutter (1972). Also, it has been applied in real-life situations by a number of researchers including Chen et al. (2003).

In order to improve the efficiency of two-stage sampling, Nematollahi et al. (2008) introduced RSS in the second stage, with the first stage remaining as SRS scheme. They showed that the estimators obtained from the design have significant improvement in efficiency over the dominant case of SRS scheme on both stages. Regardless of the improvement observed in Nematollahi et al. (2008), the problem of sampling from variant cluster sizes in the first stage is not addressed. Innocenti et al. (2021) presented three options, namely: sampling clusters with probability proportional to cluster size, and then sampling the same number of individuals from each selected cluster in the second stage; sampling clusters with equal probability, and then sampling the same percentage of individuals from each sampled cluster in the second stage and sampling clusters with equal probability, and then sampling the same number of individuals per cluster in the second stage. These options, no doubt, addressed the underlying problem only in the first option. In what appears to be an overall improvement so far, in this direction, Ozturk (2019) obtained a frame work for a two-stage cluster sampling where probability proportional to size (PPS) sampling is applied in the first stage as well as RSS applied in the second stage of sampling.

It is well known that PPS can be carried out with or without replacement. However, PPS without replacement (PPSWOR) is more complex in application than PPS with replacement (PPSWR) and that is one of the major advantages of the later over the former. Additionally, when the study population is very large, sampling with replacement is always best suited. In this work therefore, we shall propose a cluster sampling design in two stages where PPSWR is applied in the first stage and RSS in the second stage. Section 2 gives the framework for PPSWR as well as RSS. In section 3, the estimators of population mean and total of the new sampling design as well as the variance of the estimators are derived. Section 4 gives the relative efficiency of the design over the earlier design proposed by Nematollahi et al. (2008) under significantly variant clusters in the first stage of sampling and the paper is concluded in section 5.

2. The new sampling design

In this paper, a two-stage cluster sampling where sampling is done among the first stage units by probability proportional to size sampling with replacement (PPSWR) and ranked set sampling (RSS) among the second stage units is proposed.

2.1. Probability proportional to size sampling with replacement

Suppose $U_1, U_2, U_3, \ldots, U_N$ have measure of sizes $X_1, X_2, X_3, \ldots, X_N$ respectively, where $X_i;\ i = 1, 2, 3, \ldots, N$ is an integer value and $U_i; i = 1, 2, 3, \ldots, N$ is the *i*th first stage unit. In a situation where the $X_i's$ are not integers, they are all multiplied by an appropriate power of 10 to make them integers. Now, suppose a

sample of size *n* units $\{U_i, i = 1,2,3...,n\}$ is to be selected from a population of *N* units, we first form a cumulative aggregate of sizes for each of the first stage units, U_i in the population. Then, the ranges to all the population units are obtained using the cumulative totals. Using a table of random numbers, one is required to select a number d between 1 and $X = \sum_{i=1}^{N} X_i$ inclusive. If the number d falls in the range of U_2, say, then it is selected in the sample. Another random number is drawn between 1 and *X* inclusive, and if the number drawn falls this time in the range of U_i, the unit U_i is selected. In other words, the unit chosen to be included in the sample is the unit whose range contains the drawn random number. The process of drawing a random number is repeated independently until *n* number of units is drawn into the sample. With this selection procedure, the *n* number of units are drawn with PPSWR, and the probability of drawing the *i*th unit from the population is $P_i = X_i / X$ where $\sum_{i=1}^{N} P_i = 1$.

From the foregoing technique according to Hansen and Hurwitz (1943), the unbiased estimator of the population mean is given by:

$$\overline{y}_{PPS} = \frac{1}{nN} \sum_{i=1}^{n} \frac{y_i}{p_i} = \frac{\hat{Y}_{PPS}}{N} \tag{1}$$

where $y_i, i = 1, 2, \ldots, n$ is the value of the variable of interest in the sample, $p_i = \frac{x_i}{X}$ is the probability of drawing the *i*th unit in the sample; x_i is the measure of size of the *i*th sample unit and $\hat{Y}_{pps} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_i}$ is the unbiased estimator of the population total, *Y*. Also, the variance of the sample mean is given by:

$$V(\overline{y}_{pps}) = \frac{1}{nN^2} \left(\sum_{i=1}^{N} \frac{Y_i^2}{P_i} - Y^2 \right) \tag{2}$$

where Y_i is the *i*th cluster total.

2.2. Ranked set sampling (RSS) procedure

The basic premise for RSS is that sampling units are drawn from infinite population or with replacement from a finite population under study and that the sampling units drawn from the population can be ranked by certain means, rather cheaply, devoid of actual measurement of the variable of interest which is either costly or time consuming, or both. It may be considered as a controlled random sampling design. Stokes (1980), Chen et al. (2003) and Al-Omari and Bouza (2014) describe ranked set sampling

procedure as follows: (*i*) Randomly select from the study population sampling units of size m^2 (*ii*) Randomly allot the m^2 units selected into m independent sets where every set is of size m. (*iii*) The units in every set are ranked in line with the information about study variable by visual inspection, concomitant variable or through other methods that cost little or nothing. (*iv*) The samples are chosen for quantification by selecting from the first, second down to the mth set the lowest ranked unit, the second lowest ranked unit, up to the highest ranked unit from the mth set. The entire process from (*i*) to (*iv*) is called a cycle. (*v*) Repeat the cycle, say r times to get a ranked set sample of size rm out of the total of rm^2 units initially selected, see Table 1.

Each cycle of the selection process (i.e. from step *i* to *iv*) will result in measured observations $y_{11}, y_{22}, \ldots, y_{mm}$ into the sample assuming our variable of interest is Y and each of these observations is called judgment order statistic. If $m_1 = m_2 = \ldots = m_m$, that is the set sizes of the independent random samples are equal, the RSS is said to be balanced, else, it is unbalanced. The ranks which the units in the set receive may not necessarily correspond with the numerical layouts of the real values of Y. If they correspond with the numerical layouts, the ranking is said to be perfect, else, it is imperfect. The square brackets [.] are used to denote imperfect ranking in the subscripts of ranked observations while the round brackets (.) are used if the judgment order statistics are perfect.

The efficiency of RSS relies on the sampling allocation, either balanced or unbalanced. In balanced RSS, the rank order statistics has an equal allocation. Takahasi and Wakimoto (1968), Patil (2002) and Al-Omari and Bouza (2014) state that balanced RSS estimator has a variance not greater than its SRS estimator counterpart even in the presence of errors in ranking. This implies that no matter how bad RSS method is, it cannot be worse than SRS method if properly conducted. This no doubt, lies the goodness of the former over the later. Thus, from the measured ranked set sample, we can obtain unbiased estimators of population parameters, such as the population mean and variance.

Suppose y_i is the value of the variable of interest, Y_i for $i = 1, 2, \ldots, M$, where M is the population size. The set $\{Y_1, Y_2, \ldots, Y_m\}$ is a random sample from Y with pdf $f(y)$, finite mean μ and variance σ^2 and with a set of observed values $\{y_1, y_2, \ldots, y_m\}$. Let $Y_{j1}, Y_{j2}, \ldots, Y_{jm}$; $j = 1, 2, \ldots, m$ be a simple random sample drawn from the population with replacement. In some occasions, it is not an easy task ranking m units for large sample set of m, so we select a ranked set sample with small sample set of m and then replicate this sampling scheme up to r times. If that is well executed, it will turn out to produce r cycles, yielding the judgment order statistics value as it is displayed in Table 1. Let Y_{jl} represent the jth judgment ordered statistic value

from the *j*th sample of size *m* coming from the *l*th cycle of size *r*, $j = 1, 2, \ldots, m; l = 1, 2, \ldots, r$ and with y_{jl} being the value of the observed variable.

According to Takahasi and Wakimoto (1968), based on RSS technique, the unbiased estimator of the mean and its variance are respectively obtained by:

$$\hat{\mu}_{rss} = \frac{1}{mr} \sum_{j=1}^{m} \sum_{l=1}^{r} Y_{jl} \tag{3}$$

and

$$Var(\hat{\mu}_{rss}) = \frac{1}{mr}\left[\sigma^2 - \frac{1}{mr} \sum_{j=1}^{m} \sum_{l=1}^{r} (\mu_{jl} - \mu)^2\right] = \frac{1}{mr}\left[\sigma^2 - \frac{1}{m} \sum_{j=1}^{m} (u_j - \mu)^2\right] \tag{4}$$

Table 1. Display of judgment order statistics (JOS) values from RSS when the cycle is replicated *r* times

Cycle	First JOS	Second JOS	…	m[th] JOS
Cycle 1	$Y_{[1]1}$	$Y_{[2]1}$	…	$Y_{[m]1}$
Cycle 2	$Y_{[1]2}$	$Y_{[2]2}$	…	$Y_{[m]2}$
⋮	⋮	⋮	⋮	⋮
Cycle *r*	$Y_{[1]r}$	$Y_{[2]r}$	…	$Y_{[m]rfy}$

2.3. The proposed two-stage cluster sampling design

Suppose there are *N* first stage units (FSU's) in the population where every *i*th FSU has M_i second stage units (SSU's) with expected value μ_i and variance σ_i^2. Let the sample size from FSU's be represented by n while m_i represents the sample size from SSU's in the *i*th selected FSU. First, a sample of n FSU's is selected from the population using probability proportional to size with replacement (PPSWR) in the first stage. Then from every *i*th selected FSU's, m_i second stage sampling units will be selected by ranked set sampling (RSS) scheme. Assuming RSS procedure where *r* = 1 is the case in the second stage, then out of every *i*th chosen FSU's, we draw m_i units using RSS procedure. The final sample can be displayed in the array of values given by:

$$\begin{matrix} Y_{1[1]} & Y_{1[2]} & \ldots & Y_{1[m_1]} \\ Y_{2[1]} & Y_{2[2]} & \ldots & Y_{2[m_2]} \\ \vdots & \vdots & \ldots & \vdots \\ Y_{n[1]} & Y_{n[2]} & \ldots & Y_{n[m_n]} \end{matrix} \tag{5}$$

This is as illustrated by Nematollahi et al. (2008), where $Y_{i[j]}$ denotes the variable of interest pertaining to the *j*th order of the *j*th random sample in *i*th selected FSU which are independent but not identically distributed. To maintain the attribute of independence of samples in RSS, the units selected from every *i*th FSU drawn for rank ordering in the *j*th sample set is carried by simple random sampling with replacement scheme.

If we consider RSS scheme with replication in the second stage sampling, then from every *i*th FSU selected, $m_i = r_i m_i'$ units will be drawn by RSS method in cycles r_i with fix sample size m'. Going by this, let $Y_{il[j]}$ represent the variable pertaining to the *j*th order of *j*th random sample in *l*th cycle from *i*th drawn FSU. Thus, the observations in (5) will form a random sample in every *i*th selected FSU while m_i and $Y_{i[j]}$ are replaced by m_i' and $Y_{il[j]}$ respectively.

3. Estimators of the population mean and total in the new sampling design

The mean estimator for two-stage cluster sampling with probability proportional to size sampling with replacement in the first stage units and RSS design in the second stage units is given by:

$$\bar{y}_{ppsrss} = \frac{1}{n\sum_{i=1}^{N} M_i} \sum_{i=1}^{n} \frac{M_i \bar{y}_i}{P_i} = \frac{1}{n\sum_{i=1}^{N} M_i} \sum_{i=1}^{n} \frac{\hat{Y}_i}{P_i} = \frac{\hat{Y}_{ppsrss}}{\sum_{i=1}^{N} M_i} \tag{6}$$

where $\bar{y}_i = \frac{1}{r_i m_i} \sum_{l=1}^{r_i} \sum_{j=1}^{m_i'} Y_{il[j]}$ is the sample mean of the variable pertaining to the *j*th ordered value from the *j*th random sample in *l*th cycle of sampling in *i*th selected FSU and

$$\hat{Y}_{ppsrss} = \frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{r_i} \sum_{j=1}^{m_i'} \frac{M_i}{r_i m_i' p_i} Y_{il[j]} \tag{7}$$

is the unbiased estimator of the population total *Y*. $p_i = \frac{x_i}{X}$ is the probability of selecting the *i*th unit in the first stage sample; x_i is the measure of size of the *i*th sample unit. In the case of the measure of size used in this work, $p_i = \frac{m_i}{\sum_{i=1}^{n} M_i}$

It is straight forward to show that the estimator in (6) is an unbiased estimator of the population mean. This is because we have:

$$E\left(\overline{y}_{ppsrss}\right) = E_1E_2\left(\frac{1}{n\sum_{i=1}^{N} M_i}\sum_{i=1}^{n}\frac{M_i\overline{y}_i}{P_i}\right) = E_1\left(\frac{1}{\sum_{i=1}^{N} M_i}\sum_{i=1}^{n}\frac{M_i}{nP_i}E_2\left(\overline{y}_i\right)\right)$$

$$= \frac{1}{n\sum_{i=1}^{N} M_i}\sum_{i=1}^{n}\frac{\hat{Y}_i}{P_i} = \frac{\hat{Y}_{ppsrss}}{\sum_{i=1}^{N} M_i}$$

$$E\left(\overline{y}_{ppsrss}\right) = E_1\left(\frac{1}{\sum_{i=1}^{N} M_i}\sum_{i=1}^{n}\frac{\hat{Y}_i}{nP_i}\right) = \frac{Y_{ppsrss}}{\sum_{i=1}^{N} M_i} = \overline{Y} \tag{8}$$

The variance of the unbiased estimator of the population mean is given by:

$$V(\overline{y}_{ppsrss}) = \frac{1}{nM_0^{\,2}}\sum_{i=1}^{N} P_i\left(\frac{Y_i}{P_i} - Y\right)^2 + \frac{1}{nM_0^{\,2}}\sum_{i=1}^{N}\frac{M_i^2}{P_i}\frac{\sigma_i^2}{m_i} - \frac{1}{n^2M_0^{\,2}}E_1\left[\sum_{i=1}^{n}\sum_{l=1}^{r_i}\sum_{j=1}^{m_i'}\frac{M_i^2}{P_i^2m_i^2}\left(\mu_{i[j]} - \mu_i\right)^2\right] \tag{9}$$

The result in (9) is derived as follows:

Without loss of generality, let the number of cycles in each FSU be one such that $r_1 = r_2 = \ldots = r_n = 1$. Hence, $m_i = m_i'$. Then,

$$V(\overline{y}_{ppsrss}) = V_1E_2(\overline{y}_{ppsrss}) + E_1V_2(\overline{y}_{ppsrss}) \tag{10}$$

Considering $V_1E_2(\overline{y}_{ppsrss})$ gives the result:

$$V_1E_2(\overline{y}_{ppsrss}) = V_1E_2\left(\frac{1}{nM_0}\sum_{i=1}^{n}\frac{M_i\overline{y}_i}{P_i}\right),$$

where $\overline{y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} Y_{i[j]} = \hat{\mu}_{i[j]}$

$$V_1E_2(\overline{y}_{ppsrss}) = V_1\left[\frac{1}{nM_0}\sum_{i=1}^{n}\frac{M_i}{P_i}E_2(\overline{y}_i)\right] = V_1\left[\frac{1}{nM_0}\sum_{i=1}^{n}\frac{M_i}{P_i}\mu_i\right]$$

$$= \frac{1}{M_0^{\,2}}V_1\left(\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{P_i}\right)$$

$$= \frac{1}{M_0^{\,2}}\left[\frac{1}{n}\sum_{i=1}^{N} P_i\left(\frac{Y_i}{P_i} - Y\right)^2\right] \tag{11}$$

Also considering $E_1V_2(\overline{y}_{ppsrss})$ in (10) gives the result:

$$E_1V_2(\overline{y}_{ppsrss}) = E_1V_2\left(\frac{1}{nM_0}\sum_{i=1}^{n}\frac{M_i\overline{y}_i}{P_i}\right)$$

$$= E_1\left[V_2\left(\frac{1}{nM_0}\sum_{i=1}^{n}\frac{M_i\overline{y}_i}{P_i}\right)\right] = E_1\left[\frac{1}{n^2M_0{}^2}\sum_{i=1}^{n}\frac{M_i^2}{P_i^2}V_2(\overline{y}_i)\right]$$

But Takahasi and Wakimoto (1968) have obtained that

$$V_2(\overline{y}_i) = \frac{1}{m_i}\left[\sigma_i^2 - \frac{1}{m_i}\sum_{j=1}^{m_i}\left(\mu_{i[j]} - \mu_i\right)^2\right]$$

where σ_i^2 is the variance of the variable of interest Y in the ith FSU and $\mu_{i[j]}$ is the expected value of $Y_{i[j]}$. Hence,

$$E_1V_2(\overline{y}_{ppsrss}) = E_1\left[\frac{1}{n^2M_0{}^2}\sum_{i=1}^{n}\frac{M_i^2}{P_i^2}\frac{1}{m_i}\left(\sigma_i^2 - \frac{1}{m_i}\sum_{j=1}^{m_i}\left(\mu_{i[j]} - \mu_i\right)^2\right)\right]$$

$$= E_1\left[\frac{1}{n^2M_0{}^2}\sum_{i=1}^{n}\frac{M_i^2}{P_i^2}\frac{1}{m_i}\sigma_i^2\right] - E_1\left[\frac{1}{n^2M_0{}^2}\sum_{i=1}^{n}\frac{M_i^2}{P_i^2m_i^2}\sum_{j=1}^{m_i}\left(\mu_{i[j]} - \mu_i\right)^2\right]$$

$$= \frac{1}{nM_0{}^2}\sum_{i=1}^{N}\frac{M_i^2}{P_im_i}\sigma_i^2 - \frac{1}{n^2M_0{}^2}E_1\left[\sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{M_i^2}{P_i^2m_i^2}\left(\mu_{i[j]} - \mu_i\right)^2\right] \tag{12}$$

Adding (11) and (12) gives the variance of $\overline{y}_{ppsrss}$ as:

$$\frac{1}{nM_0{}^2}\sum_{i=1}^{N}P_i\left(\frac{Y_i}{P_i} - Y\right)^2 + \frac{1}{nM_0{}^2}\sum_{i=1}^{N}\frac{M_i^2}{P_i}\frac{\sigma_i^2}{m_i} - \frac{1}{n^2M_0{}^2}E_1\left[\sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{M_i^2}{P_i^2m_i^2}\left(\mu_{i[j]} - \mu_i\right)^2\right] \tag{13}$$

Now, if the number of cycles is r_i instead of one, (13) would have turned out to be (9).

4. Relative Efficiency

Relative efficiency of a sampling design ξ_1 over another ξ_2 based on an estimator $\hat{\theta}$ of a population parameter θ is a measure of relative overall quality of the designs

evidenced in their estimators. Algebraically, the relative efficiency of ξ_1 over ξ_2, based on $\hat{\theta}$ is obtained by:

$$RE(\xi_1 | \xi_2) = \frac{\text{var}\left(\hat{\theta}_{\xi_2}\right)}{\text{var}\left(\hat{\theta}_{\xi_1}\right)} \tag{14}$$

where var(.) is a measure of variability of the estimators obtained from the two designs. Using (14), ξ_1 will be adjudged a more efficient design if $RE(\xi_1 | \xi_2)$ is greater than 1 and less efficient if otherwise.

The proposed sampling design and its associated estimators are applied to the greenhouses data obtained in the 2003 agricultural survey conducted in Iran as adopted from Nematollahi et al. (2008). The provinces or a set of provinces are considered as first stage units (FSU's) and greenhouses as second stage units (SSU's). For us to estimate the mean value of the greenhouses products and subsequently compare our proposed sample mean in (6) with the mean estimator ($\hat{\mu}^r_{TSCRSS}$) proposed by Nematollahi et al. (2008) for relative efficiency, a simulation study is carried out on this data. The sampling units are ranked based on the values of the greenhouses in the frame, and the ranking is assumed to be flawless. The study variable is also the same as the greenhouses values in our simulation survey, consequently, the sizes of the second stage units M_i are used as our measure of sizes.

4.1. Layout of the data selection

In this study, there are $N = 25$ first stage units (FSU's) or provinces in the frame. And every ith province contains a total of $M_i; i = 1, \ldots, N$ greenhouses that are regarded as second stage units as they appeared in Table 2. For the sake of demonstration of the methodology for the proposed estimator of the mean, a random sample of size $n = 5$ first stage units are selected from the population of $N = 25$ clusters, using unequal probability sampling (PPSWR). The FSU's selected in the first stage of sampling via PPSWR are marked asterisks (*) in Table 3. Out of every ith selected province, $m = rm'$ greenhouses (SSU's) were selected by RSS. This paper considers where $r = 4$ and $m' = 3$ to get a ranked set sample of size 12 units each in the second stage sampling.

Similarly, for the estimator according to Nematollahi et al. (2008), a random sample of size $n = 5$ is also selected from the population by simple random sampling without replacement. Out of every ith chosen FSU, a sample of SSU's, $m = rm'$ is selected by RSS. Also, $r = 4$ and $m' = 3$ are considered to get a ranked set sample of size 12 units each in the second stage sampling.

Table 2. The number of secondary sampling units in the first stage units

FSU's	M_i	**FSU's**	M_i	**FSU's**	M_i	**FSU's**	M_i	**FSU's**	M_i
1	42	6	61	11	27	16	750	21	30
2	169	7	680	12	26	17	32	22	26
3	538	8	936	13	14	18	275	23	18
4	38	9	167	14	40	19	14	24	14
5	33	10	20	15	93	20	20	25	84

Table 3. Cumulative table for selection of 5 provinces by PPSWR

FSU's	M_i	Cum. of M_i's	Prob (M_i)	**FSU's**	M_i	Cum. of M_i's	Prob (M_i)
1	42	42	0.010127803	14	40	2791	0.009645527
2	169	211	0.040752351	15	93	2884	0.022425850
3	538	749	0.129732337	16*	750	3634	0.180853629
4	38	787	0.009163251	17*	32	3666	0.007716422
5	33	820	0.007957560	18*	275	3941	0.066312997
6	61	881	0.014709429	19	14	3955	0.003375934
7*	680	1561	0.163973957	20	20	3975	0.004822763
8*	936	2497	0.225705329	21	30	4005	0.007234145
9	167	2664	0.040270075	22	26	4031	0.006269592
10	20	2684	0.004822763	23	18	4049	0.004340487
11	27	2711	0.006510731	24	14	4063	0.003375934
12	26	2737	0.006269592	25	84	4147	0.020255606
13	14	2751	0.003375934				

4.2. Computation of estimated means for the two competing designs

In order to obtain the estimated means and totals using the Nematollahi et al. (2008) estimators and the new proposed estimators, their mean estimators and computations are presented as follows:

$$\hat{\mu}^{r}_{TSCRSS} = \frac{1}{n\bar{M}} \sum_{i=1}^{n} \sum_{l=1}^{r_i} \sum_{j=1}^{m_i'} \frac{M_i}{r_i m_i'} Y^{(j)}_{il[j]} = \frac{1}{n\bar{M}} \sum_{i=1}^{n} M_i \hat{\mu}_i \tag{15}$$

where $\bar{M} = \sum_{i=1}^{N} \frac{M_i}{N}$ and $\hat{\mu}_i = \frac{1}{r_i m_i'} \sum_{l=1}^{r_i} \sum_{j=1}^{m_i'} Y_{il[j]}^{(j)}$. The terms of the mean estimator in (15) are computed and presented in Table 4. Using the computed terms, the mean is estimated as $\hat{\mu}_{TSCRSS}^{r}$ = 28.25125.

Table 4. Calculation of the estimated population mean in Nematollahi et al. (2008)

FSU's	M_i	$m_i = r_i m_i'$	$\hat{\mu}_i$	$M_i \hat{\mu}_i$
13	38	12	14.0000	532.0000
8	938	12	12.0000	11256.0000
10	20	12	13.8333	276.6667
16	750	12	12.5833	9437.5000
2	169	12	11.4167	1929.4167

Also, the terms contained in the new proposed estimator of the mean in (6) are computed for the sample in Table 5 and the mean is estimated as $\bar{y}_{ppsrss}$ = 20.2204.

Table 5. Calculation of the estimated population mean using the new mean estimator

FSU's	M_i	p_i	$m = rm'$	y_i	$\frac{y_i}{p_i}$	$\bar{y}_i$	$\frac{M_i \bar{y}_i}{p_i}$
7	680	0.162213740	12	160	986.3529	13.33	55893.33
16	750	0.178912214	12	140	782.5067	11.67	48906.67
8	936	0.223282443	12	170	761.3675	14.17	59386.67
17	32	0.007633588	12	158	200698.0000	13.17	55194.67
18	275	0.065601145	12	154	2347.5200	12.83	53797.33

The entire process of sampling and computation is carried out using appropriate packages in the *R* statistical software.

It has been shown that the mean estimator proposed in this paper is unbiased. As a result, the appropriate measure of its variability to be used in this section is the variance. However, to ensure uniformity of computation with the estimator due to Nematollahi et al. (2008), the mean squared error (MSE) is used. Now, MSE of the two competing estimators are obtained empirically using 10000 replications of samples and computations. Precisely, the MSE of each estimator of the mean is obtained by:

$$MSE\left(\bar{y}_b\right) = \frac{1}{10000} \sum_{a=1}^{10000} \left(\bar{y}_{ab} - \mu\right)^2 ;\ b = 1, 2 \tag{16}$$

where $\bar{y}_{a1}$ and $\bar{y}_{a2}$ denote $\bar{y}_{ppsrss}$ and $\hat{\mu}^{r}_{TSCRSS}$ respectively in ath replication of the sample, $a = 1, 2, \ldots, 10000$. The results for first stage sample sizes n = 4, 5, 6 and 8 are presented in Table 6 for the new estimator as MSE_1 and Nematollahi et al. (2008) estimator as MSE_2.

Table 6. Mean square errors corresponding to each mean estimator $\bar{y}_{ppsrss}$ and $\hat{\mu}^{r}_{TSCRSS}$

FSU Sample Size	r = 3, $m' = 2$		r = 3, $m' = 3$		r =3, $m' = 4$	
	MSE_1	MSE_2	MSE_1	MSE_2	MSE_1	MSE_2
4	227.4776	89.4379	193.0703	90.0924	204.7227	91.5629
5	55.4979	69.9861	51.0138	69.3528	55.0475	70.0885
6	25.8019	56.7316	26.5967	54.9603	26.0757	54.2575
8	9.2579	37.5100	9.0912	37.3246	9.2049	36.5891

Finally, the relative efficiency of the new estimator $\bar{y}_{ppsrss}$ to $\hat{\mu}^{r}_{TSCRSS}$ at different sample sizes in the two stages of the sampling designs are obtained by:

$$RE = \frac{MSE\left(\hat{\mu}^{r}_{TSCRSS}\right)}{MSE\left(\bar{y}_{ppsrss}\right)} \tag{17}$$

The computed relative efficiencies are presented in Table 7.

Table 7. Relative efficiencies of the new estimator $\bar{y}_{ppsrss}$ to $\hat{\mu}^{r}_{TSCRSS}$

Number of selected FSU's	r = 3 $m' = 2$	r = 3 $m' = 3$	r = 3 $m' = 4$
4	0.393	0.467	0.447
5	1.261	1.359	1.273
6	2.199	2.067	2.080
8	4.052	4.105	3.9749

From the results in Table 7, the relative efficiency of the new estimator compared to Nematollahi et al. (2008) estimator shows that the new estimator is more efficient for different sizes of first stage units sample except, in the case when n = 4. This suggests that as the sample size in the first stage of sampling increases, the relative efficiency of the new estimator keeps improving. For instance, when n = 5 in the first stage and m = 5 in the second stage, the relative efficiency improves from 0.39 to 1.26. Similarly, when n = 8 and m = 12, it changes from 2.08 when n = 6 and m = 12 to 3.97. However, it is

important to note that the preferred performance of the new estimator to the Nematollahi et al. (2008) estimator may have been because the population in question has significantly varied sizes in the first stage units. If a situation where somewhat equality in sizes of the FSU's is encountered, this preference may not be guaranteed.

5. Conclusion

A new two-stage sampling design has been developed where probability proportional to size sampling with replacement (PPSWR) is used in the first stage and ranked set sampling is used in the second stage. The empirical comparative study carried out revealed that our new sampling design is more efficient as it produced better estimator for estimating the population mean than similar design built with simple random sampling in the first stage and ranked set sampling in the second stage units under the condition of significant variation in the sizes of the first stage units.

Acknowledgements

The authors wish to thank the two anonymous reviewers and the managing editor for their comments and criticisms which have greatly improved the quality of this paper.

References

Al-Omari A. I., Bouza C. N., (2014). Review of Ranked Set Sampling: Modifications and Applications. *Revista Investigacion Operacional*, 35(3), pp. 215–240.

Chen Z., Bai Z., Sinha B. K., (2003). *Ranked Set Sampling: Theory and Applications.* Springer, New York.

Damon, V., (2018). Advantages and disadvantages of multistage sampling. https://classroom.synonym.com/advantages-disadvantages-multistage-sampling-8544049.html

Dell, T. R., Clutter, J. I., (1972). Ranked set sampling theory with order statistics background. *Biometrics*, 28, pp. 545–553.

Dilip, N. C., (2015). Two-stage sampling design for estimation of total fertility rate: with an illustration for slum dweller married woman. *Electronic Journal of Applied statistical Analysis,* 8(1), pp. 112–121.

Fears, T. R., Gail, M. H., (2000). Analysis of a two-stage case-control study with cluster sampling of controls: Application to Nonmelanoma skin cancer. *Biometrics,* 56(1), pp. 190–198.

Galway, L. P., Bell, N., Shatari, S. AE. Al., Hagopian, A., Burnham, G., Flaxman, A., Weiss, W. M., Rajaratnam, J. and Takaro, T. K., (2012). A two-stage cluster sampling method using gridded population data, a GIS and Google Earth TM imagery in population-based mortality survey in Iraq. *International Journal of Health Geographics,* 11(12), pp. 1–9.

Halls, L. S., Dell, T. R., (1966). Trial of ranked set sampling for forage yields. *Forest Science,* 12(1), pp. 22–26.

Hansen, M. H., Hurwitz, W. N., (1943). On the theory of sampling from a finite population. *Annals of Mathematical Statistics,* 14, pp. 333–362.

Horney, J. J., Dickinson, M., Hsai, J., Williams, A. and Zotti, M., (2010). Two-stage cluster sampling with referral: Improving the efficiency of estimating unmet needs among pregnant and postpartum women after flooding in Northwest Georgia. *Remote Sensing of Environment,* 113(6), pp. 1236–1249.

Innocenti, F., Candel, M. J. J. M., Tan, F. E. S. and van Breukelen, G. J. P., (2019). Relative efficiencies of two-stage sampling schemes for mean estimation in multilevel populations when cluster size is informative. *Statistics in Medicine,* 38(10), pp. 1817–1834.

Innocenti, F., Candel, M. J. J. M., Tan, F. E. S. and van Breukelen, G. J. P., (2021). Optimal two-stage sampling for mean estimation in multilevel populations when cluster size is informative. *Statistical Methods in Medical Research,* 30(2), pp. 357–375.

McIntyre, G. A., (1952). A method of unbiased selective sampling, using ranked sets. *Australian Journal of Agricultural Research,* 3, pp. 385–390.

Nematollahi, N., Salehi, M. M. and Saba, A. R., (2008). Two-stage cluster sampling with ranked set sampling in the secondary sampling frame. *Communications in Statistics–Theory and methods,* 37(15), pp. 2404–2415.

Okafor, F. C., (2002). *Sample Survey Theory with Application.* Afro-Orbis Publications Ltd. Nsukka.

Ozturk, O., (2019). Two-stage cluster samples with ranked set sampling designs. *Annals of the Institute of Statistical Mathematics,* 71, pp. 63–91.

Patil, G. P., (2002). Ranked set sampling. *Encyclopedia of Environmetrics,* 3, pp. 1684–1690.

Phillips, A. E., Boily, M. C., Lowndes, C. M., Garnett, G. P., Gurav, K., Ramesh, B. M., Anthony, J., Watts, R., Moses, S. and Alary, M., (2008). Sexual identity and its contribution to MSM risk behaviour in Bangaluru (Bangalore) India: The results of a two-stage cluster sampling survey. *Journal of LGBT Health Research,* 4, pp. 111–126.

Shahbaz, M. Q., Hanif, M., (2010). *Some developments in unequal probability sampling: selection procedures and estimators*. Lap Lambert Academic Publishing, GmbH & Co. KG, Deutschland.

Stehman, S. V., Wichham, J. D., Fattorini, L., Wade, T. D., Baffetta, F. and Smith, J. H., (2009). Estimating accuracy of land-cover composition from two-stage cluster sampling. *Remote Sensing of Environment,* 113(6), pp. 1236–1249.

Stokes, S. L., (1980). Estimation of variance using judgment ordered ranked set samples. *Biometrics*, 36, pp. 35–42.

Takahasi, K., Wakimoto, K., (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics*, 21, pp. 249–255.