Scientia Iranica E (2021) 28(1), 465–476

Sharif University of Technology
Scientia Iranica
Transactions E: Industrial Engineering
http://scientiairanica.sharif.edu

On the use of ranked set sampling for estimating super-population total: Gamma population model

S. Ahmed* **and J. Shabbir**

Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan.

Received 11 May 2018; received in revised form 13 May 2019; accepted 17 June 2019

KEYWORDS
Superpopulation;
Proportional relationship;
Finite populations;
Ranked set sampling;
Prediction.

Abstract. The application of superpopulation models to estimate population parameters is an advantageous practice when recognizing the relationship between the study variable and one or more auxiliary variables is a simple matter. This paper aims to estimate the finite population total under Ranked Set Sampling Without Replacement (RSSWOR) employing the model relationship, especially Gamma Population Model (GPM), between the study and auxiliary variables. Behavior of the proposed estimator, in terms of relative efficiency, is studied in the case of a constant γ through Monte Carle experiment. The simulation study demonstrates the superiority of the proposed estimator to existing estimators under the same model. The sampling procedure, in particular, facilitates collecting data from a continuous production process.

1. Introduction

In the survey sampling literature, much attention has been given to the design-based approach, which assumes that the values of units in the population of interest are fixed constants. However, in many real-life situations, population values are generated as a result of the realization of a set of stochastic variables with specified means and variances only, i.e., higher order moments often remain unknown. Such populations are called superpopulations and the statistical models for them are called superpopulation models. Superpopulation models facilitate sample selection, constructing the estimators for population quantities of interest and enhancing the precision of estimates. Superpopulation model uses the relationship between the study variable and the auxiliary variable(s) for sample values to predict the population values of the non-sampled units, assuming that the selected sample is non-informative. In the agriculture field, an estimate of the average or total production of a certain crop can be obtained using the relationship between the production and the amount of fertilizer used or area under production. Similar examples can be found in other fields of research, especially in business, economics, and social and medical surveys. In the framework of model-based inference, Fuller [1] attempted to estimate the finite population mean or total. Royall et al. [2,3] obtained optimal model-unbiased estimators for the population mean or total using Least Square (LS) estimation methods and the well-known Gauss Markov theorem using the regression population model. Discussion on the model-based approach can be found in [4–11]. Royall [12] applied the linear LS prediction approach to two-stage sampling. Hansen et al. [13] and Rao [14] demonstrated the poor performance of the model-based approach, especially in large samples under non-self-weighting designs, even for small departure from the model. Brewer and Gregoire [15] attempted to compare the model-based approach with the model-assisted approach. For a recent comparison between the model-

*. *Corresponding author.*
E-mail address: sahmed@stat.qau.edu.pk (S. Ahmed)

doi: 10.24200/sci.2019.50976.1946

based approach and the designed-based approach, see [16]. An updated review of the model-based approach can be found in [17]. Cheruiyot et al. [18] applied this approach to predict the total and average numbers of peoples with HIV/AIDS living in Nakuru Central district in Kenya. Current work in this area can be found in [19–24].

In the same era, many experts of survey sampling have worked on efficient methods of data collection. Among them, Ranked Set Sampling (RSS) technique is a good alternative in terms of Relative Efficiency (RE) to Simple Random Sampling (SRS) for obtaining experimental data that are truly representative of the population under investigation. This is true across all of the sciences including agricultural, biological, environmental, engineering, physical, medical, and social sciences. Because in RSS, measurements are likely to be more regularly spaced than SRS. The RSS procedure facilitates stratification of the entire population at the sampling stage, i.e., we randomly select samples from the subpopulations of small, medium, and large units without constructing the subpopulations (strata) in advance. RSS method, proposed originally by McIntyre [25] to estimate mean pasture yields, has recently been modified by many authors to estimate the population parameters. Dell and Clutter [26] showed that the sample mean was an unbiased estimator for the population mean under RSS for both perfect as well as imperfect ranking. To take advantage of the negative correlation between the observations, Patil et al. [27] extended the idea of RSS for finite population assuming sampling without replacement. Muttlak [28] suggested Median Ranked Set Sampling (MRSS) for the estimation of finite population mean. [29] used Multi-Stage Ranked Set Sampling (MSRSS) to improve the efficiency of an estimator of the population mean for certain values of the sample size. Although Multi-Stage Ranked Set Sampling (MSRSS) leads to improved estimators than what is possible to obtain in RSS, this sampling scheme requires a large number of population units to be ranked before actual quantifications. Mahdizadeh and Zamanzade [30] developed a new variation in MRSS called Multi-Stage Paired Ranked Set Sampling (MSPRSS) to reduce ranking burden in MRSS and use it for estimation of bodyfat. Many other authors have worked on estimation of parameters in RSS (see [31–34] among others). RSS has been applied, after modifications, for estimation of different population parameters such as mean, median, distribution function, etc. Moreover, Haq et al. [35] proposed a mixture of SRS and RSS for estimation of population mean and median. Salehi and Jafari [36] worked on the estimation of stress-strength reliability with the help of record values obtained through the RSS. Ahmed and Shabbir [37] suggested the extreme-cum-median RSS for estimation of population mean by sub-sampling non-respondents. Similarly, Priya and Thomas [38] developed a method for estimation of common location and scale parameters using suitable RSS schemes. Mahdizadeh and Zamanzade [39] worked on reliability estimation in Multi-Stage Ranked Set Sampling (MSRSS) and [38] developed tests of perfect rankings applied with binary data. Recently, Dümbgen and Zamanzade [40] worked on estimation of cumulative distribution function in RSS.

Predicting the nature of the behavior of some future observations using the information contained in sample and the previous knowledge about the parameter involved in the density is an important problem in statistical data analysis such as estimation and inference, etc. The method is called Bayesian prediction. It has many applications in quality control and reliability engineering and biological sciences. One might construct a desirable confidence limit for the future observations. A wide range of literature pieces are available regarding predictive inference for future observations. Some of the related works are cited as [18,41–45].

Chambers and Clark [46] discussed model-based estimation in detail under the application of different population models. The current paper discusses Gamma Population Model (GPM) for estimation of finite population mean in SRS in Section 2. Ranked Set Sampling Without Replacement (RSSWOR) is employed under the model-based approach to estimation of different superpopulation total in Section 3. A comparison between the proposed estimators and existing ones is made using Monte Carle (MC) experiment in Section 4. Section 5 concludes the paper.

2. Model-based estimation under SRS

Let Y and X denote the study and auxiliary variables, respectively, for the corresponding units in population $U = \{U_i; i = 1, 2, ..., N\}$. Let U be comprised of two mutually exclusive sets s (set of sampled elements) and $\bar{s}$ (set of non-sampled elements) having n and $(N - n)$ elements, respectively. We assume the following three population models:

1. $y_i = \mu + \epsilon_i$ (Homogenous Population Model, HPM)
2. $y_i = \beta x_i + \epsilon_i x_i^{\gamma}$ (Gamma Population Model, GPM)
3. $y_i = \alpha + \beta x_i + \epsilon_i$ (Linear Population Model, LPM)

for $i = 1, 2, ..., N$,

where y_i, x_i, and ϵ_i are the ith population values corresponding to the study variable Y, auxiliary variable X, and the random error term ϵ, respectively. The random error term ϵ_i is iid with zero mean and constant variance. Further, α and β are unknown constants to be estimated using sample data. Here, γ is the rate parameter as Y varies with this rate; it may also

be unknown, but is chosen in advance using expert judgment or pilot surveys with cross validation. Many studies on prediction under LPM are available in the model-based estimation literature.

This study first briefly discusses the estimation of population total under HPM and GPM.

2.1. Homogenous Population Model (HPM)

Under HPM, we have the relationship $y_i = \mu + \epsilon_i$, which assumes that there is no auxiliary variable at design stage or/and estimation stage. We can express the population total as:

$$T_y = \sum_{i\in s} y_i + \sum_{i\in \bar{s}} y_i. \tag{1}$$

The notations $\sum_{i\in s}$ and $\sum_{i\in \bar{s}}$ show that the summation is applied over the samples s and $\bar{s}$, respectively. A Best Linear Unbiased Predictor (BLUP) for T_y suggested by Chambers and Clark [46] is as follows:

$$t_y = \sum_{i\in s} y_i + E(t_{y\bar{s}}|y_i, i \in s) = n\bar{y}_s + (N-n)\bar{y}_s$$

$$= N\bar{y}_s, \tag{2}$$

where $t_{y\bar{s}} = \sum_{i\in \bar{s}} y_i$. The prediction variance of t_y, is given by:

$$Var(t_y - T_y) = \sigma^2\big(N-n\big)\big(\frac{N}{n}\big), \tag{3}$$

where $\sigma^2 = \frac{1}{N}\sum_{i\in U}(y_i - \mu)^2$. Proof of Eq. (3) can be found in [46].

2.2. Gamma Population Model (GPM)

When population under study is heterogeneous, the estimator given in Eq. (2) may not work well. One possible way to overcome this deficiency is stratification; however, in some occasions, it is difficult to stratify the population according to certain stratification variable(s), e.g., stratifying units in the production process may cause destruction of units. In such a situation, the best way to handle the problem of heterogeneity is to search for an auxiliary variable that has some correlation with the study variable. GPM deals with such problems by controlling variance in the study variate Y when there is a proportional relationship between the study variable and some auxiliary variable whose values for all population units are available in advance. Another condition that must hold in such a model is that the marginal distribution of sampled and non-sampled values of Y for a given value of the auxiliary variable should be the same. In other words, by conditioning on X, we obtain a non-informative sample [46]. Under GPM, we have a relationship $y_i = \beta x_i + \epsilon_i x_i^{\gamma}$ between Y and X. A BLUP for T_y is given by:

$$t_{yg} = t_{ys} + E\big(t_{y\bar{s}}|y_i, i \in s; x_i, i \in U\big) = t_{y_s} + bt_{x\bar{s}}, \tag{4}$$

where $b = \sum_{i\in s} c_i y_i$ and $c_i = \frac{x_i^{1-2\gamma}}{\sum_{i\in s} x_i^{2-2\gamma}}$ for $i = 1, 2, ..., n$. The conditional expectation of t_{yg} for the given sample information is:

$$E(t_{yg}|x_i, i \in s) = \beta x_i = \mu \text{ (say)}. \tag{5}$$

This reveals that for fixed values of X, t_y is unbiased conditioning on values of X with conditional variance:

$$Var(t_{yg}|x_i, i \in U) = Var(t_{ys}) + t_{x\bar{s}}^2 Var\left(\sum_{i\in s} c_i y_i\right)$$

$$= \sigma^2 \sum_{i\in s}\big(1 + \lambda^2 x_i^{2-4\gamma}\big), \tag{6}$$

where $\lambda = \frac{t_{y\bar{s}}}{\sum_{i\in s} x_i^{2-2\gamma}}$. The variance goes down when larger values of X are selected in the sample. Comparing Eqs. (6) with (3), we see that $Var(t_{yg}|x_i, i \in U) < Var(t_y)$ if:

$$n + \lambda \sum_{i\in s} x_i^{1-2\gamma} < \big(N-n\big)\left(\frac{N}{n}\right). \tag{7}$$

The unbiasedness and efficiency properties are computed with respect to the model, although the total estimator with gamma population under the design-based approach is biased.

3. Model-based estimation under RSS

To obtain a more accurate dataset, [25] proposed the RSS assuming that ranking small sets of units was economical, while taking actual measurement from a large sample was costly. This section provides the application of a RSS scheme to the model-based approach upon making some modifications and discussion on estimation of population total in RSS assuming HPM and GPM. Consider a finite population U generated from a superpopulation with mean $\mu(i)$ and variance $\sigma^2_{(i)}$ for the ith ordered random variable $y_{(i)}$ for $i \in U$. For any given underlying superpopulation model:

1. Take sub-populations of size N_j for $j = 1, 2, ..., t$ from a superpopulation such that $N = \sum_{j=1}^{t} N_j$, in which t is the number of cycles or time frame. It is also assumed that every subpopulation is large enough to select m^2 units from them, i.e., $N_j > m^2$. The concept of so-called sub-populations is defined just for taking larger sets to ensure that sampling is without replacement. For a valid statistical inference, this division must be random and independent of the survey variable;

2. Select m^2 units from each sub-population, i.e., units produced at the same time on the same day can be taken as sub-population in the production process;

3. Divide each m^2 unit in m sets, each with size m, and rank each set within itself according to some ranking mechanism;

4. Select the ith ranked unit from the ith set for $i = 1, 2, 3, ..., m$, and $j = 1, 2, .., t$. In this way, a RSSWOR of size tm is obtained. An illustration of RSSWOR scheme is provided in Figure 1.

Figure 1 explains our sampling scheme assuming that a finite population of size N is coming from a large superpopulation with specified mean and variance generated through the stochastic process. Top stream of Figure 1 shows the continuous population. From the finite population of size N units, we consider t different cycles with sizes $N_1, N_2, ..., N_t$ randomly, leaving $N - \sum_{j=1}^{t} N_j$ as non-sampled. For example, in the production process (for quality control), one might consider the units produced in 20 days as a finite population; then, we take $t = 8$ randomly selected days as cycles. In this way, we are left with t so called sub-populations. From each sub-population, we then select m^2 units for ranking, leaving $(N_j - m^2)$ units from each sub-population as non-sampled. Finally, the RSS is applied for selecting m units from each cycle by returning the remaining $m^2 - m$ non-sampled units. The total non-sampled units are obtained through three different stages, as shown in Figure 1.

$$\begin{aligned}\text{Non-sampled} &= \text{Non-sampled at Stage-1} \\ &\quad + \text{Non-sampled at Stage-2} \\ &\quad + \text{Non-sampled at Stage-3} \\ &= N - \sum_{j=1}^{t} N_j + \sum_{j=1}^{t} (N_j - m^2) \\ &\quad + \sum_{j=1}^{t} (m^2 - m) = N - tm.\end{aligned}$$

Let s be a set of tm units selected using the above mechanism and $\bar{s}$ a set of units that are not in s. A ranked set sample, s, can be defined as:

$$s = \{y_{1(1)1}, ..., y_{m(m)1}, ...y_{1(1)2}, ..., y_{m(m)2}, ... \\ y_{1(1)t}, ...y_{m(m)t}\}.$$

3.1. RSS under HPM

To determine the ith population value of the study variable Y, we have $y_{(i)} = \mu_{(i)} + \epsilon_{(i)}$ for $i \in U$, where $\epsilon_{(i)}$ for all $i \in U$ is i.i.d with zero mean and variance $\sigma^2_{(i)}$. Hence $E(y_{(i)}) = 0$, $Var(y_{(i)}) = \sigma^2_{(i)}$, and $Cov(y_{(i)}, y_{(j)}) = 0$ for $i \neq j$, when $y_{(i)}$ and $y_{(j)}$ are taken from different ranked sets. The condition of zero mean for error term hold and only some variables, other

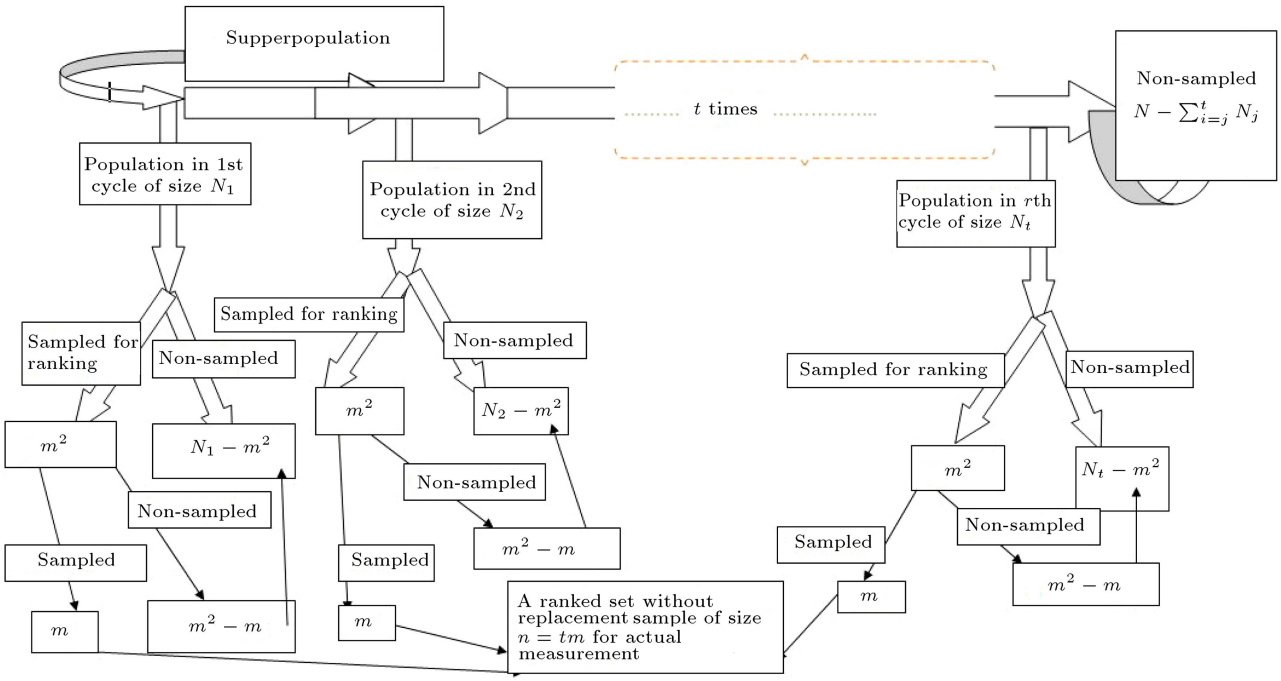

Figure 1. Ranked Set Sampling Without Replacement (RSSWOR) layout for set size m and number of cycles t.

than study variables, are ranked. Hence the ranking process is considered judgmental rather than perfect ranking. In this respect, consider a predictor for the total population given in Eq. (1):

$$t_{y(rss)} = t_{y(rss)s} + t_{y\bar{s}}, \tag{8}$$

where $t_{y(rss)s} = \sum_{j=1}^{t} \sum_{i\in s} y_{i(i)j}$ and $t_{y\bar{s}} = \sum_{i\in U} y_i - \sum_{j=1}^{t} \sum_{i\in s} y_{i(i)j}$. The problem is to predict $t_{y\bar{s}}$ using the information at hand such that (i) $E(t_{y(rss)} - T_y) = 0$, the prediction error, and (ii) $E(t_{y(rss)} - T_y)^2$, the squared prediction error, are minimum. $t_{y(rss)}$ can be expressed as a linear combination of the ranked data as follows:

$$t_{y(rss)} = \sum_{i\in s} w_{(i)} y_{i(i)}. \tag{9}$$

To simplify the computation, we take $t = 1$, i.e., only one cycle is performed.

$$\begin{aligned} t_{y(rss)} - T_y &= \sum_{i\in s} w_{(i)} y_{i(i)} + \sum_{i\in s} y_{i(i)} - \sum_{i\in s} y_{i(i)} - T_y \\ &= \sum_{i\in s} \big(w_{(i)} - 1\big) y_{i(i)} - t_{y\bar{s}}, \end{aligned} \tag{10}$$

where $(w_{(i)} - 1) = u_{(i)}$ (say) is the prediction weight of the ith non-sampled unit. Taking expectation of Eq. (10), we have:

$$E(t_{y(rss)} - T_y) = \sum_{i\in s} u_{(i)} \mu_{(i)} - (N - m)\mu. \tag{11}$$

Therefore, $t_{y(rss)}$ will be unbiased when $\sum_{i\in s} u_{(i)} \mu_{(i)} = (N - m)\mu$. Similarly, variance of $t_{y(rss)} - T_y$ can be found as follows:

$$\begin{aligned} Var(t_{y(rss)} - T_y) &= Var\left(\sum_{i\in s} u_{(i)} y_{i(i)} - t_{y\bar{s}}\right) \\ &= Var\left(\sum_{i\in s} u_{(i)} y_{i(i)}\right) + Var(t_{y\bar{s}}) \end{aligned}$$

$$Var(t_{y(rss)} - T_y) = \sum_{i\in s} u_{(i)}^2 \sigma_{(i)}^2 + (N - m)\sigma^2. \tag{12}$$

Given that the sampled and non-sampled values are independent, the covariance term on the right-hand side of Eq. (12) is zero. The value of u_i which provides unbiased estimate of $t_{y(rss)}$ is $u_{(i)} = \frac{N-m}{m}$.

Moreover, the second term in variance expression is $(N - m)\sigma^2$ as there is no-ranking on non-sampled data. Inserting the value of $u_{(i)}$ in variance expression, we get:

$$\begin{aligned} Var(t_{y(rss)} - T_y) &= \sum_{i\in s} \left(\frac{N-m}{m}\right)^2 \sigma_{(i)}^2 + (N-m)\sigma^2 \\ &= \frac{N}{m}(N - m)\sigma^2 - \left(\frac{N-m}{m}\right)^2 \sum_{i\in s} \delta_{(i)}^2 \\ &= Var(t_y - T_y) - \left(\frac{N-m}{m}\right)^2 \sum_{i\in s} \delta_{(i)}^2, \end{aligned} \tag{13}$$

where $\delta_{(i)} = (\mu_{(i)} - \mu)$ and $(m\sigma^2 - \sum_{i\in s} \delta_{(i)}^2) = \sum_{i\in s} \sigma_{(i)}^2$. From Eqs. (3) and (13), it is clear that $t_{y(rss)}$ is always more efficient than t_y.

3.2. RSS under GPM

Under GPM, the ith population value of the study variable Y is expressed as $y_{(i)} = x_{[i]}\beta + x_{[i]}^{\gamma}\epsilon_{(i)}$ for $i \in U$, where $E(y_{(i)}) = x_{[i]}\beta$, $Var(y_{(i)}) = x_{[i]}^{2\gamma}\sigma_{(i)}^2$ and $Cov(y_{(i)}, y_{(j)}) = 0$ for $i \neq j$, when $y_{(i)}$ and $y_{(j)}$ are taken from different ranked sets. It is also assumed that ranking is applied to the study variable itself (based on personal judgment or some other mechanism). The best predictor for $t_{y\bar{s}}$ is $E\big(t_{y\bar{s}}|y_{(i)} i \in s, x_{[i]}, i \in U\big)$, see [46] for detail.

$$\begin{aligned} t_{y(rss)g} &= t_{y(rss)s} + E\big(t_{y\bar{s}}|y_{(i)} i \in s, x_{(i)}, i \in U\big), \\ t_{y(rss)g} &= t_{y(rss)s} + \sum_{i\in\bar{s}} x_{[i]}\beta. \end{aligned} \tag{14}$$

In Eq. (14) β is assumed unknown. A Best Linear Unbiased Predictor (BLUP) b for β is obtained by minimizing the following sum of squared error for sample data with respect to b:

$$\sum_{i\in s} e_{i(i)}^2 = \sum_{i\in s} x_{i[i]}^{-2\gamma}\big(y_{i(i)} - x_{i[i]}b\big), \tag{15}$$

which is given by $b = \sum_{i\in s} q_{(i)} y_{i(i)}$, where $q_{(i)} = \frac{x_{i[i]}^{1-2\gamma}}{\sum_{i\in s} x_{i[i]}^{2-2\gamma}}$ and the resulting estimator is:

$$t_{y(rss)g} = t_{y(rss)s} + \sum_{i\in\bar{s}} x_{[i]}b.$$

Inserting the value of b to the above and simplifyin the previous relations, we get:

$$t_{y(rss)g} = \sum_{i\in s} \left(1 + \Phi x_{i[i]}^{1-2\gamma}\right) y_{(i)}, \tag{16}$$

where $\Phi = \frac{t_{x\bar{s}}}{\sum_{i\in s} x_{i[i]}^{2-2\gamma}}$. It is now clear that $t_{y(rss)g}$ is unbiased with respect to variance, given by:

$$Var(t_{y(rss)g}|x_{[i]}, i \in U) = Var\left(\sum_{i\in s} \Phi^*_{[i]} y_{(i)}\right)$$

$$= \sum_{i\in s} \Phi^{*2}_{[i]} \sigma^2_{(i)} = \sigma^2 \sum_{i\in s} \Phi^{*2}_{[i]} - \sum_{i\in s} \Phi^{*2}_{[i]} \delta^2_{(i)}, \quad (17)$$

where $\Phi^*_{[i]} = 1 + \Phi x^{1-2\gamma}_{i[i]}$. We can also express Eq. (17) as:

$$Var(t_{y(rss)g}|x_{[i]}, i \in U) = Var(t_{yg}|x_{[i]}, i \in U)$$

$$- \sum_{i\in s} \Phi^{*2}_{[i]} \delta^2_{(i)},$$

where $\delta_{(i)} = \mu_{(i)} - \mu$. This provides that $t_{y(rss)g}$ is more efficient than its counterpart in SRS.

In this section, GPM is considered a general population model for situations where the values of the study variable generated from the stochastic process are proportional to the corresponding values of the auxiliary variable. Further the variation in Y depends on the value of X^γ, where γ is the rate parameter that controls how much the variation in Y depends on X. Chambers and Clark [46] suggested choosing the value of gamma out of 0 and 1. Ratio population model is a particular case of the GPM for $\gamma = \frac{1}{2}$. We can derive BLUP in RSS for the ratio population model by inserting $\gamma = \frac{1}{2}$. In practical data sets the value of γ can be guessed using the scatter plot or through the value of correlation coefficient between X and Y. Similarly, by setting $\gamma = 0$ and adding the intercept term the GPM is reduced to LPM. In the subsequent section, real world data are used to check the efficiency of the proposed estimators for determining the population total.

4. MC study

To make a comparison between the models in terms of efficiency, MC experiment was employed by generating hypothetical data on variable X and obtaining Y using the relationship $Y = \rho^2 X + X^\gamma e$ for $\gamma = 0.3, 0.5, 0.8$, where e is an i.i.d error term, normally distributed with zero mean and variance σ^2 with $\rho = 0.7$. The data on X is generated from gamma distribution assuming different combinations of parameters a and b. Figure 2 provides different shapes of gamma distribution for the given combinations of parameters. A RSSWOR procedure is obtained by using the steps given in Section 3. The estimators for the sample total under the ranked set sampling with replacement for HPM and GPM models are obtained. For efficiency comparison, we also obtain an SRSWOR of size $n = tm$. Repeat the sampling process 10,000 times to obtain bias and variances of the proposed estimators. The Absolute Biases (ABs) of the total estimators are obtained from

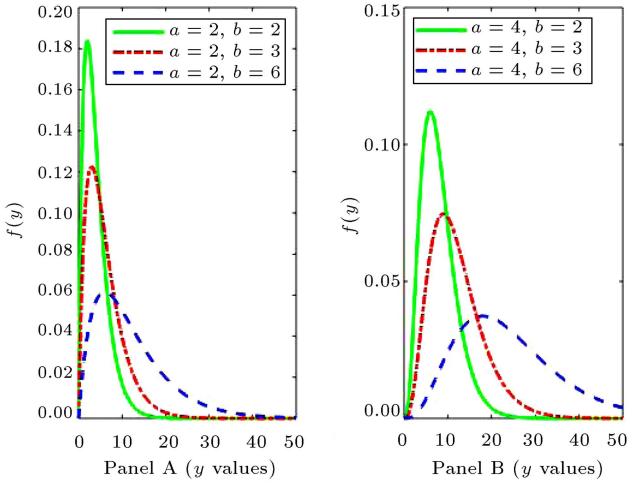

Figure 2. Effect of distribution parameters on efficiency.

the designed-based point of view as the unbiasedness is conditioned on X. The RE of the suggested estimators, is given by:

$$RE_r = \frac{Var(t_y)}{MSE(t_{yg})}, \quad RE_{rss} = \frac{Var(t_y)}{MSE(t_{y(rss)})},$$

and:

$$RE_{R.rss} = \frac{Var(t_{y(rss)})}{MSE(t_{y(rss)g})}.$$

Tables 1–3 provide the RE and AB of the proposed estimators. Different sections of Tables 1–3 are constructed for gamma distribution $G(a, b)$ for different combinations of a and b.

The results can be interpreted in the following ways:

- It is clear that the RE rates of $t_{y(rss)g}$ and t_{yg} both are high when $\gamma = 1/2$ as compared to the RE for other choices of gamma. It is suggested that the proposed estimator be used in case of a proportional relationship between the two variables with $\gamma = 1/2$;
- The RE of the estimator depends on the shape of the population from which X is generated. If the ratio a/b increases then the performance of the proportional model increases more than that of the HPM;
- According to different sections of Tables 1–3, i.e., $G(2, 6)$ and $G(4, 2)$ have the lowest and highest efficiencies, respectively, with respect to their competitors with other combinations. In other words, it can be inferred that the relative performance of GPM model is higher for skewed populations than that for the HPMs;
- In case of fat tail distribution, the predictors under GPM had the worst performance than their counterparts under HPM for both SRSWOR and RSSWOR;

Table 1. Absolute bias and relative efficiency of proposed estimators for $\gamma = 0.3$.

		Bias and RE for $\gamma = 0.3$				
r	m	RE_r	RE_{rss}	$RE_{R.rss}$	AB.srs	AB.rss
				$G(2,2)$		
5	2	2.7477	1.0162	2.4882	60.4370	74.3182
	5	8.4049	1.0416	7.8941	13.3828	28.1902
	8	14.2844	1.0674	12.5380	3.6897	15.7625
10	2	6.4891	1.0371	6.0193	16.6093	37.0989
	5	18.7730	1.0968	15.9923	14.8950	12.1114
	8	28.1957	1.1624	25.2679	40.2373	8.4270
				$G(2,3)$		
5	2	1.9541	1.0162	1.7817	39.3351	49.5385
	5	5.9852	1.0416	5.5614	9.2946	19.1030
	8	10.0640	1.0673	8.7772	1.7740	10.4089
10	2	4.5732	1.0371	4.2278	10.4029	24.9263
	5	13.3438	1.0967	11.1148	9.1333	8.4639
	8	20.4853	1.1623	17.5920	25.9294	6.0248
				$G(2,6)$		
5	2	0.7002	1.0162	0.6532	18.2332	24.7587
	5	2.1652	1.0416	1.9770	5.2064	10.0158
	8	3.5829	1.0665	3.1030	0.1417	5.0553
10	2	1.6207	1.0370	1.5014	4.1965	12.7538
	5	4.7780	1.0962	3.8859	3.3716	4.8163
	8	7.6737	1.1615	6.1247	11.6216	3.6227
				$G(4,2)$		
5	2	6.6740	1.0162	6.5199	70.2027	71.4624
	5	18.6256	1.0414	17.5572	3.7536	26.4852
	8	30.0509	1.0678	28.2431	32.0967	20.8870
10	2	14.6985	1.0371	13.7490	14.6898	36.9738
	5	37.2922	1.0969	34.4925	50.2069	13.9068
	8	52.2661	1.1625	54.6377	96.6559	8.0505
				$G(4,3)$		
5	2	5.4981	1.0162	5.3625	46.4492	47.4090
	5	15.4862	1.0413	14.5785	2.2278	16.6957
	8	24.8446	1.0677	23.5442	22.2571	13.2721
10	2	12.2151	1.0371	11.3978	9.7364	24.3884
	5	31.2106	1.0969	28.5025	33.9727	8.7211
	8	45.2124	1.1624	44.7444	64.8684	4.4050
				$G(4,6)$		
5	2	2.8007	1.0162	2.7367	22.6957	23.3556
	5	8.0255	1.0412	7.5142	0.7021	6.9062
	8	12.8020	1.0676	12.1435	12.4175	5.6571
10	2	6.3329	1.0371	5.8753	4.7830	11.8029
	5	16.5931	1.0968	14.5284	17.7384	3.5353
	8	26.1513	1.1619	22.3871	33.0809	0.7596

Table 2. Absolute bias and relative efficiency of proposed estimators for $\gamma = 0.5$.

		Bias and RE for $\gamma = 0.5$				
r	m	RE_r	RE_{rss}	$RE_{R.rss}$	B.srs	B.rss
				$G(2,2)$		
5	2	4.1635	1.0162	3.8755	26.3349	38.4283
	5	11.7600	1.0416	10.8376	0.2269	14.6806
	8	19.0647	1.0674	16.6432	12.6135	7.0302
10	2	9.0123	1.0371	8.3368	0.3577	19.2577
	5	24.6040	1.0968	20.8750	22.0938	5.2963
	8	34.8054	1.1624	32.5903	45.1165	3.7419
				$G(2,3)$		
5	2	2.4424	1.0162	2.2918	16.6708	25.5938
	5	6.9412	1.0416	6.3435	0.4148	9.9228
	8	11.2630	1.0673	9.7110	8.0589	4.2787
10	2	5.2569	1.0371	4.8709	0.7021	12.7774
	5	14.6075	1.0967	12.1113	14.1533	3.5794
	8	21.8860	1.1623	19.0112	29.5143	2.5279
				$G(2,6)$		
5	2	0.6876	1.0162	0.6540	7.0068	12.7592
	5	1.9717	1.0416	1.7872	0.6028	5.1650
	8	3.2106	1.0665	2.7341	3.5042	1.5272
10	2	1.4733	1.0370	1.3731	1.7620	6.2971
	5	4.1606	1.0962	3.3897	6.2128	1.8626
	8	6.6925	1.1615	5.3228	13.9122	1.3139
				$G(4,2)$		
5	2	11.6353	1.0162	11.3209	30.6179	36.0369
	5	30.9743	1.0414	29.1134	10.1908	12.5094
	8	47.6155	1.0678	46.2887	39.1973	11.9721
10	2	24.7812	1.0371	23.0511	3.4527	20.0803
	5	57.9638	1.0969	55.5549	56.0029	7.7684
	8	72.2816	1.1625	85.5164	98.7919	4.2145
				$G(4,3)$		
5	2	8.2487	1.0162	8.0154	20.4934	24.2274
	5	22.1472	1.0413	20.7583	6.4045	7.8630
	8	34.1948	1.0677	33.1922	26.5797	8.0502
10	2	17.6948	1.0371	16.4005	1.6694	13.8710
	5	42.7483	1.0969	39.5368	37.2152	5.1830
	8	57.7674	1.1624	60.4255	65.7651	2.3580
				$G(4,6)$		
5	2	3.2104	1.0162	3.1190	10.3689	12.4179
	5	8.7104	1.0412	8.1194	2.6182	3.2166
	8	13.6275	1.0676	13.0141	13.9622	4.1283
10	2	6.9433	1.0371	6.4016	0.1139	7.6617
	5	17.7594	1.0968	15.3926	18.4274	2.5975
	8	27.7148	1.1619	23.3248	32.7382	0.5016

Table 3. Absolute bias and relative efficiency of proposed estimators for $\gamma = 0.8$.

		Bias and RE for $\gamma = 0.8$				
r	m	RE_r	RE_{rss}	$RE_{R.rss}$	B.srs	B.rss
				$G(2,2)$		
5	2	2.9772	1.0162	2.9319	0.7295	11.7680
	5	7.8813	1.0416	7.3193	6.4027	6.8744
	8	13.0466	1.0674	11.2503	17.6501	2.4392
10	2	6.1334	1.0371	5.7045	9.8893	6.0968
	5	16.3307	1.0968	13.8569	25.1742	3.3341
	8	24.7321	1.1624	22.0765	46.4994	2.8797
				$G(2,3)$		
5	2	1.3124	1.0162	1.2942	0.7727	9.2188
	5	3.4763	1.0416	3.2281	2.4282	5.9425
	8	5.8088	1.0673	4.9801	10.6486	1.9791
10	2	2.7028	1.0371	2.5220	6.4683	4.7646
	5	7.3178	1.0967	6.1311	15.0885	3.1778
	8	11.7379	1.1623	9.8098	29.6435	2.7129
				$G(2,6)$		
5	2	0.2952	1.0162	0.2914	0.8158	6.6696
	5	0.7815	1.0416	0.7259	1.5462	5.0106
	8	1.3146	1.0665	1.1239	3.6471	1.5191
10	2	0.6073	1.0370	0.5688	3.0473	3.4325
	5	1.6602	1.0962	1.3828	5.0029	3.0215
	8	2.7646	1.1615	2.2174	12.7877	2.5461
				$G(4,2)$		
5	2	19.0692	1.0162	18.5281	7.1291	4.0326
	5	48.3596	1.0414	45.5339	24.8539	2.3440
	8	70.4115	1.0678	72.8449	48.5345	1.7598
10	2	38.7544	1.0371	36.2560	21.4485	4.0871
	5	83.1682	1.0969	85.3838	63.7828	0.5404
	8	90.1236	1.1625	127.4162	103.5377	2.0718
				$G(4,3)$		
5	2	9.4384	1.0162	9.1597	5.7765	2.1612
	5	24.1982	1.0413	22.6224	16.4848	2.9508
	8	36.5374	1.0677	36.3521	33.6606	0.9746
10	2	19.2451	1.0371	17.9593	14.0669	3.0742
	5	45.4154	1.0969	42.6479	42.8306	0.9760
	8	59.2750	1.1624	64.0248	69.4480	2.4260
				$G(4,6)$		
5	2	2.5469	1.0162	2.4673	4.4240	0.2897
	5	6.5779	1.0412	6.1133	8.1157	3.5577
	8	10.2535	1.0676	9.8319	18.7867	0.1893
10	2	5.1960	1.0371	4.8419	6.6852	2.0612
	5	13.2630	1.0967	11.5543	21.8784	1.4116
	8	20.9691	1.1619	17.4367	35.3583	2.7801

- It can also be noticed that RE i.e. RE_R (ratio), RE_{rss} (RSS), and $RE_{R.rss}$ (ratio estimator in RSS) are all increasing functions of the set size (m) and the number of cycles (t);
- The last two columns of Tables 1–3 provide ABs of the total estimators under gamma population in SRSWOR and RSSWOR. AB of the total estimator decreases with increase in set sizes m and number of cycles r in the RSSWOR scheme;
- ABs are relatively smaller in case of $\gamma = 1/2$ in the ratio population model.

5. Concluding remarks

A new version of RSS for obtaining a sample without replacement under the Gamma Population Model (GPM) (general form of proportional population model) was introduced. Figure 1 shows an image of the Ranked Set Sampling Without Replacement (RSSWOR) which assumes that the finite population derives from an infinite superpopulation in the stochastic process with finite mean and variance. It was also assumed that a population could be generated from different points, i.e., cycles, and the m sets taken from one cycle were totally different from the m set in other cycles. After selecting a sample using RSSWOR, the model relationship between the study variable and the auxiliary variable was used to predict the non-sampled values while obtaining a point predictor for the population total. The mathematical expressions and Monte-Carlo experiment both supported the superiority of the predictor under RSSWR to the total predictor under SRSWOR for GPM and Homogenous Population Model (HPM). Hence, the proposed predictors may perform well for the process controls to construct control charts given that in such situations, there are highly dimensional data in terms of the number of observations. They are applicable to social surveys conducted on social media in which one deals with a large population with unending size.

Acknowledgments

The authors are very grateful to the referees for their valuable comments that significantly improved this paper.

References

1. Fuller, W.A. "Simple estimators for the mean of skewed populations", Technical Report, Iowa State University, Dept. of Statistics (1970).
2. Royall, R. "An old approach to finite population sampling theory", *Journal of American Statistical Association*, **63**, pp. 1269–1279 (1969).
3. Royall, R.M. and Cumberland, W.G. "The finite-population linear regression estimator and estimators of its variance an empirical study", *Journal of the American Statistical Association*, **76**(376), pp. 924–930 (1981).
4. Godambe, V.P. "A unified theory of sampling from finite populations", *Journal of the Royal Statistical Society: Series B (Methodological)*, **17**(2), pp. 269–278 (1955).
5. Godambe, V. and Joshi, V. "Admissibility and Bayes estimation in sampling finite populations", *The Annals of Mathematical Statistics*, **36**(6), pp. 1707–1722 (1965).
6. Basu, D., *An Essay on the Logical Foundations of Survey Sampling Part i, in Foundations of Statistical Inference*, eds. Godambe and Sprott, Holt, Rinehart and Winston of Canada, Toronto, pp. 203–233 (1971).
7. Smith, T.M.F. "The foundations of survey sampling: a review", *Journal of the Royal Statistical Society: Series A (General)*, **139**(2), pp. 183–195 (1976).
8. Särndal, C.E., Thomsen, I., Hoem, J.M., Lindley, D.V., Barndorff-Nielsen, O., and Dalenius, T. "Design-based and model-based inference in survey sampling [with discussion and reply]", *Scandinavian Journal of Statistics*, pp. 27–52 (1978).
9. Smith, T.M.F. "On the validity of inferences from non-random samples", *Journal of the Royal Statistical Society: Series A (General)*, **146**(4), pp. 394–403 (1983).
10. Royall, R.M. "The model based (prediction) approach to finite population sampling theory", *Lecture Notes-Monograph Series*, **17**, pp. 225–240 (1992).
11. Sarndal, C.E., Swensson, B., and Wretman, J., *Model Assisted Survey Sampling*, Springer Science & Business Media (2003).
12. Royall, R. "The linear least-squares prediction approach to two-stage sampling", *Journal of American Statistical Association*, **71**, pp. 657–664 (1976).
13. Hansen, M.H., Madow, W.G., and Tepping, B.J. "An evaluation of model-dependent and probability-sampling inferences in sample surveys", *Journal of the American Statistical Association*, **78**(384), pp. 776–793 (1983).
14. Rao, J. "Development in sample survey theory", *The Canadian Journal of Statistics*, **25**, pp. 1–21 (1996).
15. Brewer, K.R., *Combined Survey Sampling Inference: Weighing Basu's Elephants*, Oxford University Press (2002).

16. Brewer, K. and Gregoire, T.G. "Introduction to survey sampling", *Handbook of Statistics*, **29**, pp. 9–37 (2009).

17. Valliant, R. "Model-based prediction of finite population totals", *Sample Surveys: Inference and Analysis*, **29B**, pp. 23–31 (2009).

18. Cheruiyot, R., Cheruiyot, T., and Jepchumba, L. "Estimation of population total using model-based approach: A case of hiv/aids in nakuru central district, kenya", *International Journal of Scientific and Technology Research*, **3**(11), pp. 171–175 (2014).

19. Podlaski, R. and Roesch, F.A. "Modelling diameter distributions of two-cohort forest stands with various proportions of dominant species: a two-component mixture model approach", *Mathematical Biosciences*, **249**, pp. 60–74 (2014).

20. Bohning, D. "Ratio plot and ratio regression with applications to social and medical sciences", *Statistical Science*, **31**(2), pp. 205–218 (2016).

21. Ogundimu, E.O., Altman, D.G., and Collins, G.S. "Adequate sample size for developing prediction models is not simply related to events per variable", *Journal of Clinical Epidemiology*, **76**, pp. 175–182 (2016).

22. Kumar, S., Sisodia, B.V.S., Singh, D., and Basak, P. "Calibration approach based estimation of finite population total in survey sampling under super population model when study variable and auxiliary variable are inversely related", *Journal of Reliability and Statistical Studies*, **10**(2), pp. 83–93 (2017).

23. Lovasi, G.S., Fink, D.S., Mooney, S.J., and Link, B.G. "Model-based and design-based inference goals frame how to account for neighborhood clustering in studies of health in overlapping context types", *SSM-Population Health*, **3**, pp. 600–608 (2017).

24. Li, J. "Assessing the accuracy of predictive models for numerical data: Not r nor r2, why not? then what?", *PloS One*, **12**(8), e0183250 (2017).

25. McIntyre, G. "A method for unbiased selective sampling using ranked sets", *Crop and Pasture Science*, **3**, pp. 385–390 (1952).

26. Dell, T. and Clutter, J. "Ranked set sampling theory with order statistics background", *Biometrica*, **28**, pp. 545–555 (1972).

27. Patil, G., Sinha, A., and Taillie, C. "Finite population corrections for ranked set sampling", *Annals of the Institute of Statistical Mathematics*, **47**(4), pp. 621–636 (1995).

28. Muttlak, H. "Median ranked set sampling", *Journal of Applied Statistical Sciences*, **6**(4), pp. 577–586 (1997).

29. Al-Saleh, M.F. and Al-Omari, A.I. "Multistage ranked set sampling", *Journal of Statistical Planning and Inference*, **102**(2), pp. 273–286 (2002).

30. Mahdizadeh, M. and Zamanzade, E. "Efficient body fat estimation using multistage pair ranked set sampling", *Statistical Methods in Medical Research*, SAGE Publications Sage UK: London, England (2018).

31. Samawi, H.M. and Muttlak, H.A. "Estimation of ratio using rank set sampling", *Biometrical Journal*, **38**, pp. 753–764 (1996).

32. Ohyama, T.D.J. and Yanagawa, T. "Estimating population characteristics by incorporating prior values in stratified random sampling/ranked set sampling", *Journal of Statistical Planning and Inference*, **138**, pp. 4021–4032 (1999).

33. Bouza, C. "Ranked set subsampling the non-response strata for estimating the difference of means", *Biometrical Journal*, **1**, pp. 203–243 (2002).

34. Al-Omari, A. and Jaber, K. "Percentile double ranked set sampling", *Journal of Mathematics and Statistics*, **44**, pp. 903–915 (2008).

35. Haq, A., Brown, J., Moltchanova, E., and Al-Omari, A.I. "Mixed ranked set sampling design", *Journal of Applied Statistics*, **41**(10), pp. 2141–2156 (2014).

36. Salehi, M. and Jafari, A. "Estimation of stress-strength reliability using record ranked ret sampling scheme from the exponential distribution", *Filomat*, **29**(5), pp. 1149–1162 (2015).

37. Ahmed, S. and Shabbir, J. "Extreme-cum-median ranked set sampling", *Brazilian Journal of Probability and Statistics*, **33**(1), pp. 24–38 (2019).

38. Priya, R. and Thomas, P.Y. "An application of ranked set sampling when observations from several distributions are to be included in the sample", *Communications in Statistics-Theory and Methods*, **45**(23), pp. 7040–7052 (2016).

39. Mahdizadeh, M. and Zamanzade, E. "Efficient body fat estimation using multistage pair ranked set sampling", *Statistical Methods in Medical Research*, **28**(1), pp. 223–234 (2019).

40. Dümbgen, L. and Zamanzade, E. "Inference on a distribution function from ranked set samples" *Annals of the Institute of Statistical Mathematics*, **72**(1), pp. 157–185 (2020).

41. Aitchison, J. and Dunsmore, I.R., *Statistical Prediction Analysis*, Cambridge, MA: Cambridge University Press (1975).

42. Bain, L.J., *Statistical Analysis of Reliability and Life Testing Model*, New York, NY: Marcel Dekker (1978).

43. Sinha, S.K. "On the prediction limits for Rayleigh life distribution", *Calcutta Statistical Association Bulletin*, **39**, pp. 105–109 (1990).

44. Raqab, M.Z. "Modified maximum likelihood predictors of future order statistics from normal samples", *Computational Statistics and Data Analysis*, **25**, pp. 91–106 (1997).

45. Raqab, M.Z. and Madi, M.T. "Bayesian prediction of the total time on test using doubly censored Rayleigh data", *Journal of Statistical Computational and Simulation*, **72**, pp. 781–789 (2002).

46. Chambers, R. and Clark, R., *An Introduction to Model-Based Survey Sampling with Applications*, OUP Oxford, **37** (2012).

Biographies

Shakeel Ahmed completed his MPhil in Statistics from the Department of Statistics Quaid-i-Azam University Islamabad and won Vice Chancellor Gold Medal from the University in 2015. He has published 10 papers in internationally reputed journals in the field of survey methodologies and estimation of parameters under new data collection mechanism, especially in the ranked set sampling scheme. He is now a PhD scholar at the Department of Statistics Quaid-i-Azam University Islamabad, Pakistan. The present paper is a part of his PhD research.

Javid Shabbir is working as a Tenured Professor at Statistics at Department of Statistics Quaid-i-Azam University, Islamabad. He completed his PhD in Statistics from Kent University at Canterbury, UK in 1997. He had Post-Doctoral positions at University of Southern Maine, USA in 2003 and University of North Carolina at Greensboro USA in 2005. He has published about 300 article papers in different internationally reputed journals. His area of research includes survey sampling and randomized response techniques. He has supervised many MPhil and PhD students at the department.