THIS
— IS —
STATISTICS

RANKED SET SAMPLING

65 Years Improving the Accuracy in Data Gathering

Edited by
Carlos N. Bouza-Herrera
Amer Ibrahim Falah Al-Omari

Ranked Set Sampling

Ranked Set Sampling

65 Years Improving the Accuracy in Data Gathering

Edited by

Carlos N. Bouza-Herrera
*Faculty of Mathematics and Computation,
University of Havana, Havana, Cuba*

Amer Ibrahim Falah Al-Omari
*Department of Mathematics,
Faculty of Science, Al al-Bayt University,
Mafraq, Jordan*

ELSEVIER

ACADEMIC PRESS
An imprint of Elsevier

Academic Press is an imprint of Elsevier
125 London Wall, London EC2Y 5AS, United Kingdom
525 B Street, Suite 1650, San Diego, CA 92101, United States
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom

Notices
Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

ISBN: 978-0-12-815044-3

Publisher: Candice Janco
Acquisition Editor: Graham Nisbet
Editorial Project Manager: Susan Ikeda
Production Project Manager: Swapna Srinivasan
Cover Designer: Mark Rogers

Cover designed by Ibrahim Amer Al-Omari

Typeset by MPS Limited, Chennai, India

List of Contributors

Dana Majed Rizi Ahmad
Department of Statistics, Yarmouk University, Irbid, Jordan

Sira Allende-Alonso
Faculty of Mathematics and Computation, University of Havana, Havana, Cuba

Amjad D. Al-Nasser
Department of Statistics, Faculty of Science, Yarmouk University, Irbid, Jordan

Amer Ibrahim Falah Al-Omari
Department of Mathematics, Faculty of Science, Al al-Bayt University, Mafraq, Jordan

Mohammad Fraiwan Al-Saleh
Department of Statistics, Yarmouk University, Irbid, Jordan

Saeid Amiri
Department of Natural and Applied Sciences, University of Wisconsin-Green Bay, Green Bay, WI, United States

Antonio Arcos
Department of Statistics and Operational Research, University of Granada, Granada, Spain

Muhammad Aslam
Department of Statistics, Faculty of Science King Abdul Aziz University, Jeddah, Saudi Arabia

Dinesh S. Bhoj
Department of Mathematical Sciences, Rutgers University, Camden, NJ, United States

Vaishnavi Bollaboina
Department of Mathematics Texas A&M University-Kingsville, Kingsville, TX, United States

Carlos N. Bouza-Herrera
Faculty of Mathematics and Computation, University of Havana, Havana, Cuba

Beatriz Cobo
Department of Statistics and Operational Research, University of Granada, Granada, Spain

Jose F. García
DACEA, Universidad Juárez Autónoma de Tabasco, Villahermosa, Tabasco, Mexico

Abdul Haq
Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan

Konul Bayramoglu Kavlak
Department of Actuarial Sciences, Hacettepe University, Ankara, Turkey

Debashis Kushary
Department of Mathematical Sciences, Rutgers University, Camden, NJ, United States

Mahdi Mahdizadeh
Department of Statistics, Hakim Sabzevari University, Sabzevari, Iran

Vishal Mehta
Department of Mathematics, Jaypee University of Information Technology, Waknaghat, Himachal Pradesh, India

Prabhakar Mishra
Department of Statistics, Banaras Hindu University, Varanasi, Uttar Pradesh, India

Reza Modarres
Department of Statistics, The George Washington University, Washington, DC, United States

Omer Ozturk
Department of Statistics, The Ohio State University, Columbus, OH, United States

Kumar Manikanta Pampana
Department of Mathematics, Texas A&M University-Kingsville, Kingsville, TX, United States

María del Mar Rueda
Department of Statistics and Operational Research, University of Granada, Granada, Spain

Veronica I. Salinas
Department of Mathematics, Texas A&M University-Kingsville, Kingsville, TX, United States

Stephen A. Sedory
Department of Mathematics, Texas A&M University-Kingsville, Kingsville, TX, United States

Rajesh Singh
Department of Statistics, Banaras Hindu University, Varanasi, Uttar Pradesh, India

Sarjinder Singh
Department of Mathematics, Texas A&M University-Kingsville, Kingsville, TX, United States

Gajendra K. Vishwakarma
Department of Applied Mathematics, Indian Institute of Technology (ISM), Dhanbad, Jharkhand, India

Ehsan Zamanzade
Department of Statistics, University of Isfahan, Isfahan, Iran

Sayed Mohammed Zeeshan
Department of Applied Mathematics, Indian Institute of Technology (ISM), Dhanbad, Jharkhand, India

Ruiqiang Zong
Department of Mathematics, Texas A&M University-Kingsville, Kingsville, TX, United States

Preface

Ranked set sampling (RSS) gives a new approach to dealing with sample selection. It was proposed in the seminal paper of McIntyre (1952. A method for unbiased selective sampling using ranked sets. Australian Journal of Agricultural Research 3, 385−390). His experience in agricultural application provoked a challenge to the usual simple random sampling (SRS) design introducing a previous ordering of the units. The practical studies suggested that it produces more accurate estimators of the mean. This proposal was taken into account by other practitioners dealing with agricultural studies. They also obtained better results using RSS. The mathematical validity of the claim was sustained by the work of Takahasi and Wakimoto (1968. On unbiased estimates of the population mean based on the sample stratified by means of ordering. Annals of the Institute of Statistical Mathematics 20, 1−31).

That fact also remained unnoticed by the majority of the statistical community but some interesting results were developed for establishing the mathematical reasons sustaining having better results when using RSS.

Nowadays, the results obtained by RSS still seem to be somewhat "magical" to some colleagues and they are doubtful of the accuracy of the reported improvements due to using RSS. They may be simply explained. Ranking changes the working with "pure" random variables to dealing with order statistics (OS). OS have nice properties coming from the basics of statistical inferences. This supports the individual variances of observations (now OSs) being smaller than the variance of the random variables. Doubts arose in discussions, because in practice the variable of interest is not possible to rank. The fact that ranking a correlated and known variable allows ranking the units at a low cost, providing "adequate" ranking, was proved. The original ranking in McIntyre's experiences was made on the basis of "eye estimation" of pasture availability.

Once a series of theoretical facts was established mathematically, RSS obtained attention and different statistical problems started to be revisited. Not only is estimation better, but testing of hypotheses using RSS samples appears to be more powerful.

The number of contributions in RSS is large. Nowadays it is established as a tool for increasing precision and/or diminishing sampling costs.

This book is concerned not only with the celebration of the first 65 years of having RSS as a sampling alternative model, but also present new results in the context of estimation and testing in finite population sampling. The authors are well known in the area. Having a look at the references or the web permits corroborating their role in conforming the body of important and usable models in survey sampling using RSS. Most of the papers illustrate their use and some of them come from real-life applications.

The description is ordered as they appear in the book.

Amiri-Modarre's chapter, about the bootstrap test of ranked set sampling with different rank sizes, considers testing and confidence intervals estimation when RSS is used and bootstrap tests are applied. Studies were developed for illustrating the accurateness of value's derived using the proposed bootstrap methods.

Simultaneous estimation of means of two sensitive variables using RSS is the contribution of Pampana, Sedory, and Singh. They extended the previous results of Ahmed, Sedory, and Singh (2017. Simultaneous estimation of means of two sensitive quantitative variables. Communications in Statistics: Theory and Methods, Online available) and Bouza (2009. Ranked set sampling and randomized response procedure for estimating the mean of a sensitive quantitative character. Metrika 70, 267−277) in the case of two sensitive variables.

Calibration is the theme of the chapter by Salinas, Sedory, and Singh. They consider the estimation of the population mean under the existence of a known auxiliary variable and a new calibrated estimator of the population mean is proposed for RSS.

The chapter by Bouza, García, Vishwakarma, and Zeeshan deals with the analysis of the estimation of the variance of a sensitive variable, when it is applied to a randomized response procedure and the sample is selected using RSS. The performance of the proposal is evaluated through a study of persons infected with HIV/AIDS.

Bouza, Herrera, Singh, and Mishra developed the chapter on ranked set sampling estimation of the population mean when the information on an attribute is available concerning the development of a review in the theme.

The chapter about studying the quality of environmental variables using a randomized response procedure for the estimation of a proportion through ranked set sampling, by Allende, Alonso, Bouza, and Herrera, is concerned with the performance of RSS in the study of the quality of the environment by ranking using measurements of the contaminants in the air and the water.

Extensions of some "randomized response procedures related with Gupta−Thornton method: the use of order statistics" is a contribution of Bouza and Herrera where new scrambling procedures are developed and the results studied in terms of the variance of the involved estimators.

Vishwakarma, Zeeshan, and Bouza present the chapter on ratio and product type exponential estimators for population mean using ranked set sampling. They suggest an improved form of the exponential ratio and product estimators using RSS. The behavior of the suggested estimators is evaluated by developing a simulation study.

Haq presents a chapter on modified partially ordered judgment subset sampling schemes, where modified partially ordered judgment subset sampling schemes are proposed for estimating the population mean. Extensive Monte Carlo simulations and a case study using a real data set illustrate the performance of this proposal.

Estimation of the distribution function using a modification of RSS, called moving extreme ranked set sampling, is the theme of estimation of the distribution function using moving extreme ranked set sampling, this chapter is by Al-Saleh and Ahmad.

The chapter on improved ratio-cum-product estimators of the population mean is authored by Al-Omari. He considers the problem of estimating the population mean using extreme RSS, where different ratio-cum-product estimators of the population mean are suggested, assuming that some information of the auxiliary variable is known.

Kushary reviews issues related to RSS with unequal samples for estimating the population mean and proposes a new median ranked set sampling.

Al-Nasser and Aslam present the chapter on development of a new control chart based on ranked repetitive sampling. They propose a control chart for the quality characteristic under the normal distribution. The performance is evaluated using the average run length over the existing control chart. The application of a proposed control chart is given through simulation and a real example.

The chapter on statistical inference using stratified ranked set samples from finite populations by Ozturk and Kavlak develops statistical inference of the population mean and total using stratified RSS. Inference is constructed under both randomized design and super population models. The empirical evidence is used for evaluating the performance of the proposed estimators and is applied to apple production data in a finite population setting.

Construction of strata boundaries for ranked set sampling is the contribution of Zong, Sedory, and Singh. They address the problem of constructing strata boundaries in stratified ranked set sampling.

Bollaboina, Sedory, and Singh have contributed the chapter on the forced quantitative randomized response model using ranked set sampling. They consider the problem of estimating the mean of a sensitive variable by combining the ideas of Bouza (2009. Ranked set sampling and randomized response procedure for estimating the mean of a sensitive quantitative character. Metrika 70, 267−277) on the use of ranked set sampling and those of Chaudhuri and Stenger (1992. Sampling Survey. Marcel Dekker, New York) on the use of a forced quantitative response.

The contribution of Mehta is a new Morgenstern type bivariate exponential distribution with known coefficient of variation by ranked set sampling. The chapter introduces a new Morgenstern type bivariate exponential distribution, when coefficients of variation are known, using RSS. To demonstrate the relative performance of various estimators considered in this chapter, an empirical study is carried out. Another contribution is on shrinkage estimation of scale parameters toward an interval of Morgenstern type bivariate uniform distribution using ranked set sampling. The chapter deals with the problem of estimating the scale parameter of Morgenstern type bivariate uniform distribution, based on the observations made on the units of RSS. Some improved classes of shrinkage estimators are proposed in the form of intervals. Numerical illustrations are also given.

Nonparametric estimation in RSS is discussed in the chapter by Ehsan Zamanzade. The author discusses the problem of nonparametric estimation of the population mean and entropy, based on RSS selection of units. The chapter describes some estimators and evaluates their performance using Monte Carlo simulation.

The contributors have done a worthy work and we expect that this book will receive a warm welcome from statisticians. We thank the referees who anonymously helped develop this work with the revisions of the chapters.

And last but not least, we appreciate the collaboration of the staff of Elsevier, headed by Susan Ikeda as Editorial Project Manager, which allowed us to arrive at the final version of this book.

Carlos N. Bouza-Herrera and Amer Ibrahim Falah Al-Omari

CHAPTER

1 STUDYING THE QUALITY OF ENVIRONMENT VARIABLES USING A RANDOMIZED RESPONSE PROCEDURE FOR THE ESTIMATION OF A PROPORTION THROUGH RANKED SET SAMPLING

Sira Allende-Alonso and Carlos N. Bouza-Herrera

Faculty of Mathematics and Computation, University of Havana, Havana, Cuba

1.1 INTRODUCTION

Commonly it is required to obtain information on sensitive attributes and a sample is selected for interviewing a sample of persons. Collecting trustworthy responses on sensitive issues through direct questioning in personal interviews using various techniques is not often successful because they do not protect the respondents' privacy. Therefore in practice the data collected on sensitive features are affected by the existence of respondent bias.

Randomized response models are used to decrease both nonresponses and answer bias and to provide privacy protection to the respondents.

Warner (1965) proposed the randomized response (RR) method as a means of avoiding response bias. The initial model looked for the estimation of the proportion of persons with the stigma. The model used a randomized trial. The seminal paper of Warner (1965) has 50 years of since created and still different contributions are being generated. The models are generally based on the selection of a sample using simple random sampling with replacement.

Consider a population U of size N with two strata U_A and U_{A^*}. Therefore to conduct an inquiry is a serious issue. To belong to U_A is stigmatizing. Hence the respondents will tend to use random response (RR). It provides the opportunity of reducing response biases due to dishonest answers to sensitive questioning. Therefore this technique protects the privacy of the respondent by granting that his belonging to a stigmatized group cannot be detected. The interest of the inquiry is to estimate the proportion of individuals carrying a stigma, identified with belonging to A. If $|A|$ denotes the number of units with the stigma and we are interested in estimating the probability $\theta(A) = |U_A|/|U| = N_A/N$.

The RR technique has been successfully applied in many areas and different modifications and extensions to this method have been proposed in the literature on sampling. It is still receiving

Ranked Set Sampling. DOI: https://doi.org/10.1016/B978-0-12-815044-3.00001-0

attention from the researchers, see for example Gupta et al. (2002), Ryu et al. (2005), and Saha (2006).

A challenging sampling design is ranked set sampling (RSS). It was suggested by McIntyre (1952) and appears as a more efficient than simple random sampling with replacement (SRSWR). Takahashi and Wakimoto (1968) and Dell-Clutter (1972) gave a mathematical support to RSS and the list of new results is growing rapidly. See Patil (2002) for a review on this theme.

Chen et al. (2008) suggested a randomized response model for ordered categorical variables. They used an ordinal logistic regression for ranking. We present these results in Section 1.2. Considering that a sensitive variable is evaluated, we consider the use of RR for collecting the information. We develop an extension of the RSS estimator of Chen et al. (2008) using Warner's model. The proposal is presented in Section 1.3. The derived variance of the proposed estimator is larger than the variance of Chen's proposal. Considering that a sensitive question is evaluated we suspect that its use will reduce answer biases. Section 1.4 develops a study using real-life data. The experiments sustained our suspicion. The answers to the direct question of the interviewed produced estimations more different than the real one. The proposed estimator was closer. These facts support the recommendation of using it to obtain a gain in accuracy with respect to the usual simple random sampling with replacement model.

1.2 RANKING ORDERED CATEGORICAL VARIABLES

The proposal of Chen et al. (2008) for ordered categorical variables allows the use of RSS. They used a set of explanatory variables $\mathbf{Z} = (Z_1,\ldots,Z_K)$ for fitting a logistic regression. Take the variable of interest X_j in an item where

$$X_j = i \text{ if item } j \text{ is classified in the class } C(i)$$

Hence the probability distribution function is the multinomial $M(1,P_1,\ldots P_q)$, $P_i = Prob\{X = i\}$, $i = 1,\ldots,q$. Initially a random sample is selected and in each sample item are measured $\mathbf{Z}$ and $\mathbf{X}^*$. The ordinal logistic regression (ORL) is fitted to the data using a statistical package. Considering

$$c_i = P(\text{classifying an item in a category1to i}) = \sum\nolimits_{t\leq i} P_t, i = 1,\ldots,q$$

The logit function is $logit(c_i) = log\left(\frac{c_i}{1-c_i}\right) = L_i$. Using the collected data the fitted logit model is the proportional odds model

$$L_i = \alpha_i + \boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{z}, \text{i} = 1,\ldots,q$$

The model's probability of classifying a particular item r in the ith category is denoted π_{ri} and its cumulative probability by c_{ri}. The model fitted produces the corresponding estimates $\hat{\pi}_{ri}$ and $\hat{c}_{ri}$.

The procedure proposed by Chen et al. (2008) considers the selection of a random sample of size m using SRSWR. The class of the rth judgmental order statistic for X is denoted by $X_{(r)}$. The ranking is made as follows.

Chen et al. (2008) ranking procedure for ordinal variables:

Step 1 Use the fitted model and compute $\left\{\hat{\pi}_{ri,}\hat{c}_{ri}\right\}, i = 1,..,q, r = 1,..,m$

Step 2 Classify item r in the category h such that $\hat{\pi}_{hi,} = Max\left\{\hat{\pi}_{ri,} i = 1,..,q\right\}, r = 1,\ldots,m$.

Step 3 $Rank(r) > Rank\ (r^*)$ if r is assigned to $C(i)$ and r^* to $C(j)$ being $j < i$.
Step 4 An item in $C(i)$ is ranked using the computed $\hat{c}_{ri}'s$: $Rank(r) > Rank(r^*)$ if $\hat{c}_{ri} < \hat{c}_{r^*i}$.

The procedure is repeated n_r times for each $X_{(r)}$—*class*, $t = 1,..,m$. For the experiment j the item with rank j is interviewed. The RSS sample sets is

$$\begin{array}{ccc}
X_{(1)1}\cdots & X_{(1)t}\cdots & X_{(1)n_1} \\
\vdots & \vdots & \vdots \\
X_{(r)1}\cdots & X_{(r)t} & X_{(r)n_t} \\
\vdots & \vdots & \vdots \\
X_{(m)1}\cdots & X_{(m)t} & X_{(m)n_m}
\end{array}$$

The n_t*'s* are not necessarily equal. The use of an equal number of experiments yields a balanced RSS sampling design; in another case it is unbalanced.

The rth row is a sample from the stratum defined by the rth order statistic. The probability of mass function is $p_{(r)i}$, $i = 1,..,q$.

Let us consider the particular case in which the interviewed persons are questioned to declare belonging to a certain group A. The response can be modeled as

$$I\left[X_{(r)j}\right] = \begin{cases} 1 & \text{if a YES is the answer} \\ 0 & \text{otherwise} \end{cases}$$

We are interested in estimating $\theta(A)$, the proportion of persons belonging to A in the population. $\theta(A)$ may be estimated using the RSS proposed by Chen et al. (2008) by

$$p_c = \frac{\sum_{r=1}^{m} \frac{1}{n_r} \sum_{t=1}^{n_r} I\left[X_{(r)j}\right]}{m}$$

Now we have $p_{(r)A} = \mu_{(r)}$ and $m\mu = mP(A) = \mu_{(1)} + \ldots + \mu_{(m)}$. Hence

$$E(p_c) = \frac{\sum_{r=1}^{m} \frac{1}{n_r} \sum_{t=1}^{n_r} p_{(r)A}}{m} = \theta(A)$$

It has been derived; see that the variance of the statistics of order r is

$$\sigma_{(r)}^2 = \sigma^2 - \left(\mu_{(r)} - \mu\right)^2$$

Therefore we may consider that

$$V\left(I\left[X_{(r)j}\right]\right) = p_{(r)A}\left(1 - p_{(r)A}\right) = \theta(A)(1 - \theta(A)) - \left(p_{(r)A} - \theta(A)\right)^2$$

and, as result, taking $\vartheta = \sum_{t=1}^{n_r} \frac{1}{n_r}$

$$V(p_c) = \sum_{r=1}^{m} \frac{p_{(r)A}\left(1 - p_{(r)A}\right)}{m^2} \sum_{t=1}^{n_r} \frac{1}{n_r} = \frac{\theta(A)(1 - \theta(A))}{m} - \vartheta \sum_{r=1}^{m} \frac{\left(p_{(r)A} - \theta(A)\right)^2}{m^2}$$

The second sum is positive and represents the gain in accuracy due to the use of the proposal of Chen et al. (2008) with respect to the use of SRSWR.

The optimal choice of the sample sizes is given by the expression:

$$n_{r(opt)} = n\frac{\sqrt{p_{(r)A}\left(1-p_{(r)A}\right)}}{\sum_{r=1}^{m}\frac{\sqrt{p_{(r)A}(1-p_{(r)A})}}{m}}, n = \sum_{r=1}^{m} n_r$$

It establishes that the order statistics with larger standard deviation should have larger samples sizes. That is, the order statistics with smaller gains in accuracy measured by $(p_{(r)A}-\theta(A))^2$.

We will consider the case in which A is a sensitive group and evaluate the behavior of this sampling design when a randomized response mechanism is introduced for obtaining the responses.

1.3 A RANDOMIZED RESPONSE STRATEGY

The probability of carrying a stigma $\theta(A)$ is the parameter to be estimated. The usual approach is to ask a selected individual if he/she belongs to A (to carry the stigma). Warner (1965) proposed providing a random mechanism to the interviewed who develops an experiment that selects between the statements:

1. I belong to A, with probability $p \neq 0.5$ and
2. I do not belong to A, with probability $1 - p$. The evaluated variable is Y = 1 if the response is "YES", 0 otherwise

The individual does not reveal which statement is evaluated. The random sample permits evaluating the number of "Yes" answers

$$.n_Y = \sum_{I=I^n} Y_i$$

Commonly, each respondent in the sample is asked to select a card from a deck after shuffling. The deck has a proportion p of cards with statement 1. After deselection the respondent answers "Yes" or "No," without revealing the selected statement. This technique is known as the related question method. Warner (1965) derived that

$$p_W = \frac{\frac{n_y}{n}}{2p-1} + \frac{p-1}{2p-1}$$

is the maximum likelihood estimator of θ(A). It is unbiased and its existence is supported by the use of ≠ 0.5. Its variance is

$$V(p_w) = \frac{\theta(A)(1-\theta(A))}{n} + \frac{p(1-p)}{n(2p-1)^2}$$

The second term in the above expression is the increase in the variance due to the introduction of the randomized mechanism.

Let us consider the use of this RR model when RSS is used.

After conforming the RSS sample using Chen et al.'s (2008) procedure the interviewer uses the RR mechanism for selecting the statement to be evaluated. The response obtained will be again

$$I\left[X_{(r)j}\right] = \begin{cases} 1 & \text{if a YES is the answer} \\ 0 & \text{otherwise} \end{cases}$$

but

$$Prob\left(I\left[X_{(r)j}\right] = 1\right) = pp_{(r)A} + (1-p)(1-p_{(r)A})$$

Now the estimator of the probability of carrying the stigma for the sample of the class of the *r*th order statistics is

$$\hat{p}_{W(r)A} = \frac{\sum_{t=1}^{n_r} I\left[X_{(r)j}\right]}{n_r(2p-1)} - \frac{1-p}{n_r(2p-1)}$$

A naïve estimator based on the RSS sample is

$$p_{cW} = \frac{\sum_{r=1}^{m} \hat{p}_{W(r)A}}{m} = \sum_{r=1}^{m} \left(\frac{\sum_{t=1}^{n_r} I\left[X_{(r)j}\right]}{mn_r(2p-1)} - \frac{1-p}{n_r(2p-1)} \right)$$

Its unbiasedness follows from the fact that, for any $r = 1, \ldots, m$,

$$E(\hat{p}_{W(r)A}) = \frac{n_r\left[pp_{(r)A} + (1-p)(1-p_{(r)A})\right]}{n_r(2p-1)} - \frac{1-p}{n_r(2p-1)} = p_{(r)A}$$

The variance of the estimator is readily obtained as

$$V(\hat{p}_{cW}) = \frac{\sum_{r=1}^{m} V\left(\hat{p}_{W(r)A}\right)}{m^2} = \sum_{r=1}^{m} \frac{p_{W(r)A}(1-p_{W(r)A})}{mn_r} + \vartheta \frac{p(1-p)}{m(2p-1)^2} - \vartheta \sum_{r=1}^{m} \frac{\left(p_{W(r)A} - \theta(A)\right)^2}{m^2}$$

The second term represents an increment in the variance due to the use of the randomization procedure. In practice the nonsampling error produced by providing incorrect answers, for avoiding being stigmatized, is present when direct questions are asked.

We performed a large study to evaluate the behavior of the proposal when managers are interviewed for establishing the quality of the protection of the environment by their enterprises.

1.4 EVALUATION OF THE PERFORMANCE OF $\hat{p}_{cW}$

To test the model proposed we interviewed the directors of different enterprises that produce highly contaminated garbage. They were asked to report if they send contaminated garbage to municipal sites. They gave a report. Afterwards they were provided with a set of cards where 60% of the cards fixed the selection of the sensitive question, $p = 0.60$

The enterprise contaminates the environment

The cards were shuffled by the interviewed for reporting “yes” or “no.”

The characterization of leaching of elements from solid waste compost was made by evaluating grab samples. We consider that it provided the real result. That is, a “Yes” or “No” was produced by analyzing the grab. The grab samples were prepared from multiple grab samples using coning and quartering methods. The compost was collected from composting facilities which were screened to reduce the particles mechanically six times separated in a trammel and passed through

Table 1.1 Logistic Regression Models Used for Estimating the Proportion of Contaminating Enterprises

Model	Explanatory Variables
WQ: main metallic contaminators in the river	Percentage of lead, chrome, and nickel
AQ: main contaminators of the quality of the air	Percentage of sulfuric acid and carbon dioxide
GQ: main metallic contaminators in the river and main contaminators of the quality of the air	Percentage of lead is a test of the level of contamination of "metal" present in the water, chrome, nickel, sulfuric acid, and carbon dioxide

The population census was performed. The collected population data were sampled. Three sampling fractions were used $f = 0.05$, 0.10, *and* 0.20. *The evaluation of the behavior of the estimators was made by selecting* 1000 *samples using each sample fraction.*

Table 1.2 Average of *1000* Proportion Estimates for $m = 2$, $n_r = 10$

Model	Aliment Factories		Metallurgical Factories		Textile Factories		Chemical Factories	
True Proportion	**0.87**		**0.78**		**0.90**		**0.85**	
Model	$\hat{p}_c$	$\hat{p}_{cW}$	$\hat{p}_c$	$\hat{p}_{cW}$	$\hat{p}_c$	$\hat{p}_{cW}$	$\hat{p}_c$	$\hat{p}_{cW}$
WQ	0.74	0.89	0.65	0.72	0.45	0.92	0.55	0.86
AQ	0.75	0.87	0.51	0.73	0.68	0.90	0.67	0.82
GQ	0.76	0.84	0.62	0.71	0.77	0.89	0.3	0.79

a fine. The type of grab came from aliment, metallurgical, textile, and chemical factories. The grab sample procedure is described in Tissdel and Breslin (1995).

We considered three different sets of variables for fitting the logistic regression. The measurement of contamination in the air and the rivers, of the basin used for sending the residuals of the industries, produced the explanatory variables. The reports of the closest monitoring station were used for measuring them in a large research conducted for detecting the highly contaminating enterprises. Table 1.1 gives a description. An inspection to the enterprises established whether they were contaminating the environment. The inquiry took place a year after the auditing performed. The objective was to check if they changed their status. Presumably the managers would avoid declaring their incompetence to solve the problems detected previously.

Table 1.2 presents the average of the proportions computed with the two estimators for an overall sample size $n = 20$ with $m = 2$ and constant value of n_r*'s*. It is clear that the managers cheated. The direct responses produced an underestimation of the true proportion. The use of the RR allows obtaining a more accurate estimation.

Table 1.3 presents the average of the proportions computed with the two estimators with $m = 4$, $n_r = 4$. Comparison of them leads to a similar conclusion. Note that it seems to be better to use RR, which allows to obtain a closer estimation.

Table 1.3 Average of *1000* Proportion Estimates for $m = 4$, $n_r = 4$

Model	**Aliment Factories**		**Metallurgical Factories**		**Textile Factories**		**Chemical Factories**	
True Proportion	**0.87**		**0.78**		**0.90**		**0.85**	
Model	$\hat{p}_c$	$\hat{p}_{cW}$	$\hat{p}_c$	$\hat{p}_{cW}$	$\hat{p}_c$	$\hat{p}_{cW}$	$\hat{p}_c$	$\hat{p}_{cW}$
WQ	0.75	0.86	0.62	0.70	0.41	0.92	0.57	0.81
AQ	0.74	0.37	0.55	0.71	0.65	0.93	0.69	0.81
GQ	0.72	0.88	0.64	0.70	0.74	0.91	0.62	0.76

Table 1.4 Computed ε_u. For $u = c, cW$ and for $m = 2$, $n_r = 10$

Model	**Aliment Factories**		**Metallurgical Factories**		**Textile Factories**		**Chemical Factories**	
Model	ε_c	ε_{cW}	ε_c	ε_{cW}	ε_c	ε_{cW}	ε_c	ε_{cW}
WQ	1.87	0.81	1.85	0.89	1.90	0.91	1.87	0.81
AQ	1.92	0.80	1.91	0.88	1.96	0.92	1.92	0.80
GQ	1.91	0.75	1.91	0.91	1.93	0.92	1.91	0.75

Table 1.5 Computed ε_u. For $u = c, cW$ and for $m = 4$, $n_r = 5$

Model	**Aliment Factories**		**Metallurgical Factories**		**Textile Factories**		**Chemical Factories**	
Model	ε_c	ε_{cW}	ε_c	ε_{cW}	ε_c	ε_{cW}	ε_c	ε_{cW}
WQ	1.91	0.87	1.92	0.89	1.93	0.90	1.91	0.87
AQ	1.91	0.88	1.93	0.89	1.93	0.91	1.91	0.88
GQ	1.92	0.87	1.93	0.90	1.94	0.93	1.92	0.87

Note that the estimator based on the randomized response procedure performs better for smaller values of m.

The accuracy of the estimators was analyzed by computing:

$$\varepsilon_u = \sum\nolimits_{h=1}^{1000} \frac{\left|\hat{p}_u - \theta(A)\right|_u}{1000\theta(A)}, u = c, cW$$

The results are given in Tables 1.4 and 1.5. The direct question is considerably more inaccurate than the randomized one.

REFERENCES

Chen, H., Stasny, E.A., Wolfe, D.A., 2008. Ranked set sampling for ordered categorical variables. Can. J. Stat. 36, 179–191.

Dell, G.P., Clutter, J.L., 1972. Ranked set sampling theory with order statistics background. Biometrics 28, 545–553.

Gupta, S., Gupta, B., Singh, S., 2002. Estimation of sensitivity level of personal interview survey questions. J. Stat. Plan. Inference 100, 239–247.

McIntyre, G.A., 1952. A method for unbiased selective sampling using ranked sets. Aust. J. Agric. Res. 3, 385–390.

Patil, G.P., 2002. Ranked set sampling. In: El-Shaarawi, A.H., Pieegoshed, W.W. (Eds.), Encyclopedia of Enviromentrics, vol. 3. Wiley, Chichester, pp. 1684–1690.

Ryu, J.B., Kim, J.-M., Heo, T.-Y., Park, C.G., 2005. On stratified randomized response sampling. Model Assist. Stat. Appl. 1, 31–36.

Saha, A., 2006. A generalized two-stage randomized response procedure in complex sample surveys. Aust. N. Z. J. Stat. 48, 429–443.

Takahashi, K., Wakimoto, K., 1968. On unbiased estimates of the population mean based on sample stratified by means of ordering. Ann. Inst. Stat. Math. 20, 1–31.

Tissdel, S.E., Breslin, V.T., 1995. Characterization of leaching of element from municipal solid waste compost. J. Environ. Qual. 24, 827–833.

Warner, S.L., 1965. Randomized response: a survey technique for eliminating evasive answer bias. J. Am. Stat. Assoc. 60, 63–69.

CHAPTER

DEVELOPMENT OF A NEW CONTROL CHART BASED ON RANKED REPETITIVE SAMPLING

2

Amjad D. Al-Nasser[1] **and Muhammad Aslam**[2]

[1]*Department of Statistics, Faculty of Science, Yarmouk University, Irbid, Jordan*

[2]*Department of Statistics, Faculty of Science King Abdul Aziz University, Jeddah, Saudi Arabia*

2.1 INTRODUCTION

Statistical control charts are tools for understanding variation of a product; they are considered to be one of the most important statistical tools that can be used for monitoring a product and then help in maintaining the quality of a product based on a given specification criterion. In general, control charts can be divided into two different types, control charts for attributes and control charts for variables, depending on the product quality characteristics. The original idea of control charts was introduced by Shewhart (1924) to improve the quality of telephone transmission; by suggesting a control chart that consists of three components, the chart fences which are also known as control chart limits, namely; upper control limit (UCL) and lower control limit (LCL), in addition to the center line (CL). The main idea of Shewhart charts is to monitor the process mean; then, if the process mean is stabile and located between the chart limits, the process will be considered under control. However, it will be out of control if the value of the process mean deviated from the chart limits in a specific number of process standard deviations (i.e., say k). For example, in normal product populations, if k is equal to 2 then only 5% of the product is expected to exceed the control chart limits (Fig. 2.1).

Assuming we are sampling from a normal distribution with mean μ and standard deviation σ, and let $\{X_{ij}{:}i=1,2,\ldots,m\}j=1,2,\ldots,r$ be r independent simple random samples (SRS) each of size m are selected from this population; then the sample mean $\overline{X}_j=\frac{1}{m}\sum_{i=1}^{m}x_{ij};j=1,2,..,r$ is distributed normally, with mean μ with standard deviation $\sigma/\sqrt{m}$. Then, the Shewhart control charts limits will be:

$$\begin{cases} UCL=\mu+Z_{1-\frac{\alpha}{2}}\sigma/\sqrt{m} \\ CL=\mu \\ LCL=\mu-Z_{1-\frac{\alpha}{2}}\sigma/\sqrt{m} \end{cases}$$

where $Z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})^{th}$ percentile from the standard normal distribution; and $(1-\alpha)$ is the probability that any sample mean will be between the UCL and LCL. Usually, a 3σ rule is implemented in these limits and we replace $Z_{1-\frac{\alpha}{2}}$ with 3. For normal distributions, the 3σ limits are equivalent to 0.001 probability limits; which means 99.7% of the sample means will fall within the control limits (Montgomery, 2009).

Ranked Set Sampling. DOI: https://doi.org/10.1016/B978-0-12-815044-3.00002-2

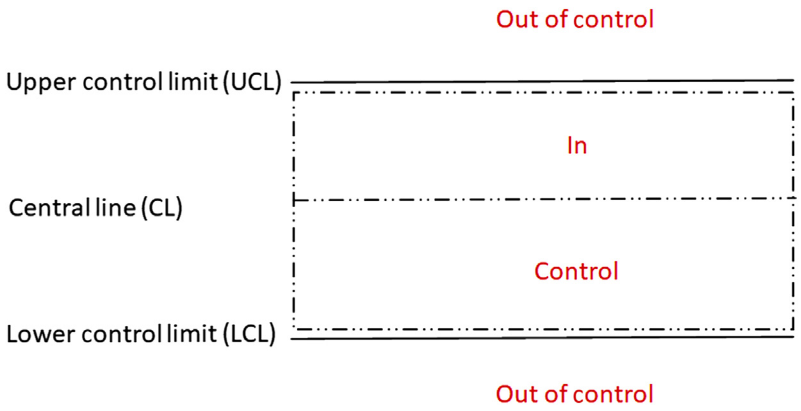


FIGURE 2.1

Shewhart control chart.

Moreover, if the process mean μ and standard deviation σ are unknown, then an unbiased estimator will be used in the limits and the control charts are estimated as:

$$\begin{cases} UCL = \overline{\overline{X}} + 3\hat{\sigma}_{\overline{X}} \\ CL = \overline{\overline{X}} \\ LCL = \overline{\overline{X}} - 3\hat{\sigma}_{\overline{X}} \end{cases}$$

where the unbiased estimators of mean μ and σ are:

$$\overline{\overline{X}} = \frac{1}{r}\sum_{j=1}^{r} \overline{X}_j$$

and

$$\hat{\sigma}_{\overline{X}} = \frac{\Gamma\left(\frac{m-1}{2}\right)}{r\sqrt{m}\left(\frac{2}{m-1}\right)^{2}\Gamma(m-1)} \sum_{j=1}^{r} \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}\left(X_{ij} - \overline{X}_j\right)^2}$$

One of the most important indicators of the control chart is the run length (RL), which is the sample number when a data point is out of the control chart limits. The average RL (ARL) is a key indicator used to evaluate the performance of a control chart and represents the expected number of samples until a control chart has one point of the control limits. There are two types of ARL:

- In-control ARL (ARL0) is the expected number of samples until a control chart signals, under the condition that the actual process is truly in control; noting that, the ARL is a geometric random variable with probability of success equal to α which represents also type I error and is equivalent to "Pr (signal/ in-control process)." Therefore the ARL for the Shewhart control chart is the expected value of a geometric experiment and equal to $\text{ARL}_0 = \frac{1}{\alpha}$.

The following table illustrates the possible sequences leading to an "out of control" signal:

Run Length	Probability
1	α
2	$\alpha(1-\alpha)$
3	$\alpha(1-\alpha)^2$
:	:
R	$\alpha(1-\alpha)^{r-1}$

then $\text{ARL}_0 = \sum_{j=1}^{\infty} \text{RL} \cdot \text{Probability} = \sum_{j=1}^{\infty} j\alpha(1-\alpha)^{j-1} = \frac{1}{\alpha}$

- Out-of-control ARL (ARL1) is the expected number of samples until a control chart signals, under the condition that the actual process is in fact out-of-control; then the Pr (signal/out-of-control process) $= 1 - \beta$; therefore ARL1 $= \frac{1}{1-\beta}$.

Most statisticians consider ARL0 = 370 to be the desired value for ARL0 as it achieves a balance between α and β. Shewhart control charts have weaknesses in detecting a small shift in the process. Therefore extensive researches are found to improve the performance of the Shewhart control chart (Sindhumol et al., 2016; Amiri et al., 2014; Franco et al., 2014; Chan et al., 2003; Kumar et al., 2017; Prajapati and Singh, 2016; Woodall, 2000).

One of the most important techniques used to improve the performance of the control chart is the sampling scheme that is used in selecting the item in a given process. Several sampling techniques were used to improve the performances of Shewhart control charts. Al-Nasser and Al-Rawwash (2007) developed a Shewhart control chart based on ranked data, the main idea proposed in their work is of using ranked set sampling (RSS) schemes. Al-Omari and Al-Nasser (2011) used a robust extreme ranked set sampling scheme in developing a new control chart limit for the mean. Al-Omari et al. (2016) used double acceptance sampling for time truncated life tests based on transmuted new Weibull—Pareto distribution. Al-Nasser et al. (2013) suggested using folded ranked set sampling in developing the control chart. Shafqat et al. (2017) discussed the attribute control charts for several distributions.

Resampling or the repetitive sampling scheme is an interesting scheme that could be implemented to improve the performances of the control chart. Repetitive sampling is similar to the sequential sampling scheme, which required multiple control chart limits. Moreover, control charts based on multiple control limits are of interest of many researchers as they are more robust than the classical Shewhart chart. These charts depend on a resampling criterion to accept or reject a product under investigation. Repetitive sampling control charts were originally proposed by Sherman (1965), who suggested using this idea for developing an attributes acceptance sampling plan. Balamurali and Jun (2006) used repetitive sampling to develop more efficient acceptance sampling plans. The idea of repetitive sampling is different from the sequential or basically the double-sampling approach. The double-sampling scheme has four parameters, while repetitive sampling has only two parameters.

Therefore in using repetitive sampling the control chart is divided more precisely into different subregions using two pairs of control limits (inner and outer limits) as shown in Fig. 2.2, instead of one pair of limits as it is in the novel Shewhart control chart.

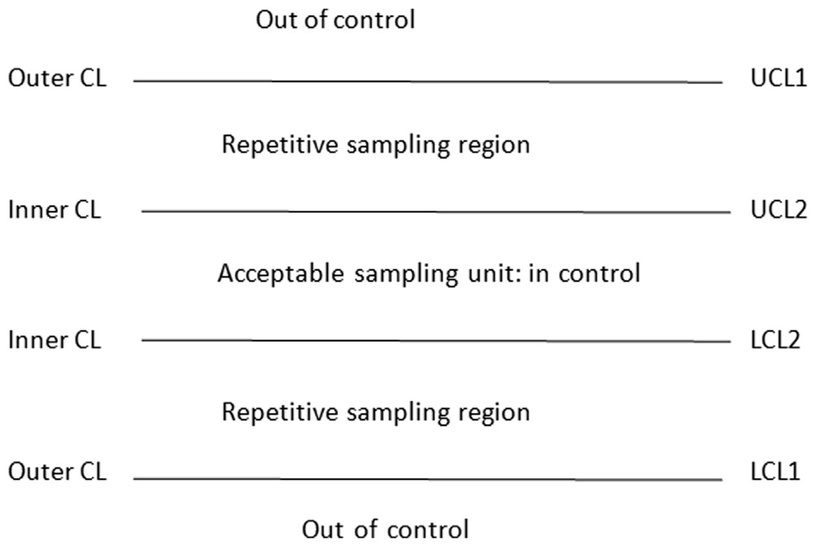

Out of control
Outer CL
UCL1
Repetitive sampling region
Inner CL
UCL2
Acceptable sampling unit: in control
Inner CL
LCL2
Repetitive sampling region
Outer CL
LCL1
Out of control

FIGURE 2.2

Repetitive sampling control chart limits.

When using repetitive sampling, the process is declared out of control using the same rule as the Shewhart control chart, however, it is declared to be in control only if the process mean hardly deviates from the center of the chart, and it should be located within the inner control limits. If the process mean is located between the inner and outer control limits, then a geometric sampling procedure should be applied by keep inspecting repetitively new samples until we observe a process mean within the inner limits.

In using the repetitive sampling control charts, the calculations of control limits depend on two limits (inner and outer limits), multipliers, e.g., k_1 and k_2 ($k_2 < k_1$). In the case that ARL_0 is around 370 the value of k_1 is close to 3. Recently, Aslam et al. (2014a,b) proposed a t-control chart using repetitive sampling and Ahmad et al. (2014) designed an X-bar control chart based on the process capability index using repetitive sampling and proved its efficiency. Azam et al. (2015) designed a hybrid EWMA chart using repetitive sampling for normal distribution. Lee et al. (2015) proposed a control chart using an auxiliary variable and repetitive sampling to detect the process mean. Aslam et al. (2014a,b) designed some attribute and variable control charts using repetitive sampling for monitoring the process mean. Other published work can be found in Ahmad et al. (2014), Aslam et al. (2015), and Aslam et al. (2013).

All of the suggested control charts used the idea of drawing a simple random sample from a given population. The sampling scheme is very important, in the literature many researches have shown that the precision of the sampling units using ranked set sampling is much better than using SRS (McIntyre, 1952; Chen et al., 2004; Al-Nasser, 2007).

It is noted that a lot of work is available on repetitive sampling plans use an ordinary single sampling plan. By exploring the literature and to the best of the authors' knowledge, there is no work available on the design of a repetitive sampling plan using rank set sampling. Therefore in this chapter, we will introduce the design of repetitive sampling plan for the rank set sampling by assuming that the variable of interest follows the normal distribution. In the next section we discuss the repetitive sampling control charts based on SRS. In Section 2.3 is an overview of the ranked set sampling scheme. Section 2.4 discusses the control chart for the sample mean based on the idea of a repetitive sampling scheme. The performance and a comparative study will be given in Section 2.5 and the chapter ends with some concluding remarks in Section 2.6.

2.2 SHEWHART CONTROL CHART UNDER REPETITIVE SAMPLING

Suppose that the quality characteristics follow a probability density function $f(x)$ that has a distribution $F(x)$ with mean μ and standard deviation σ. Also, when the process is under control assume that the target mean is μ_0. Then, the repetitive control chart for the sample mean $\overline{X}$ has the following steps:

Step 1: Draw a SRS of size n.
Step 2: Calculate the sample mean $\overline{X}$
Step 3: Declare the following decision about the entire process:

$$\begin{cases} \text{Out of Control,} & \text{if; } \overline{X} > \text{UCL}_1 \text{ or } \overline{X} < \text{LCL}_1 \\ \quad\text{In Control,} & \text{if; } \text{LCL}_2 < \overline{X} < \text{UCL}_2 \\ \text{Otherwise;} & \text{Re Sample} \end{cases}$$

Where the outer control chart limits are given by:

$$\text{UCL}_1 = \mu_0 + k_1 \frac{\sigma}{\sqrt{n}}$$

$$\text{LCL}_1 = \mu_0 - k_1 \frac{\sigma}{\sqrt{n}}$$

Similarly, the inner control chart limits are given by:

$$\text{UCL}_2 = \mu_0 + k_2 \frac{\sigma}{\sqrt{n}}$$

$$\text{LCL}_2 = \mu_0 - k_2 \frac{\sigma}{\sqrt{n}}$$

Then the probability that the process is declared as in control is:

$$P_{\text{in}} = \frac{P\left(\text{LCL}_2 < \overline{X} < \text{UCL}_2 | \mu = \mu_0\right)}{1 - P_{\text{rep}}}$$

where the probability that repetitive sampling is needed can be obtained by:

$$P_{\text{rep}} = P\left(\text{UCL}_2 < \overline{X} < \text{UCL}_1\right) + P(\text{LCL}_1 < \overline{X} < \text{LCL}_2)$$

Hence, the in-control average run length (ARL) is given by:

$$\text{ARL}_0 = \frac{1}{1 - P_{\text{rep}}}$$

Suppose now that the process mean has shifted from m to $m + \delta\sigma$. Then, the probability that the process is declared as out of control is obtained by:

$$P_{\text{in}}^{*} = \frac{P\left(\text{LCL}_2 < \overline{X} < \text{UCL}_2 | \mu = \mu_0 + \delta\sigma\right)}{1 - P_{\text{rep}}^{*}}$$

Similarly, the ARL for an out-of-control process will be

$$\text{ARL}_1 = \frac{1}{1 - P_{\text{rep}}^{*}}$$

Moreover, the control limits will be obtained when the process is under control by using a nonlinear programming system where the objective function is the average sample number (ASN) $(\text{ASN} = \frac{n}{1 - P_{\text{rep}}})$:

*Minimize **ASN***

Subject to:

1. $\text{ARL}_0 \geq r_0$
2. $k_1 > k_2$

After obtaining the control chart limit's coefficients k_1 and k_2, then we will use them to find out the ARL of the process. Now, if we are sampling from a normal distribution, then

$$P_{\text{in}} = \frac{2\Phi(k_2) - 1}{1 - 2(\Phi(k_1) - \Phi(k_2))}$$

$$P_{\text{in}}^{*} = \frac{\Phi\left(k_2 - \delta\sqrt{n}\right) + \Phi\left(k_2 + \delta\sqrt{n}\right) - 1}{\left(\Phi\left(k_2 + \delta\sqrt{n}\right) - \Phi\left(k_1 + \delta\sqrt{n}\right)\right) - \left(\Phi\left(k_1 - \delta\sqrt{n}\right) - \Phi\left(k_2 - \delta\sqrt{n}\right)\right)}$$

Which can be used to compute the ARL of the process for normal distribution.

2.3 RANKED SET SAMPLING SCHEME

Ranked set sampling (RSS) is a visual sampling scheme that has been proposed by McIntyre (1952). The samples obtained by this scheme depend on drawing several simple random samples, and each sample is ranked using a free cost method or based on an auxiliary variable that relates to the variable of interest for actual measurement. The steps in the ranked set sampling scheme can be described as follows:

Step 1: Randomly select m^2 sample units from the population;
Step 2: Allocate the m^2 selected units as randomly as possible into m sets, each of size m;
Step 3: Without yet knowing any values for the variable of interest, rank the units within each set based on personal judgment or with measurements of a covariate that is correlated with the variable of interest;
Step 4: Choose a sample for actual analysis by including the smallest ranked unit in the first set, then the second smallest ranked unit in the second set, continuing in this fashion until the largest ranked unit is selected in the last set.

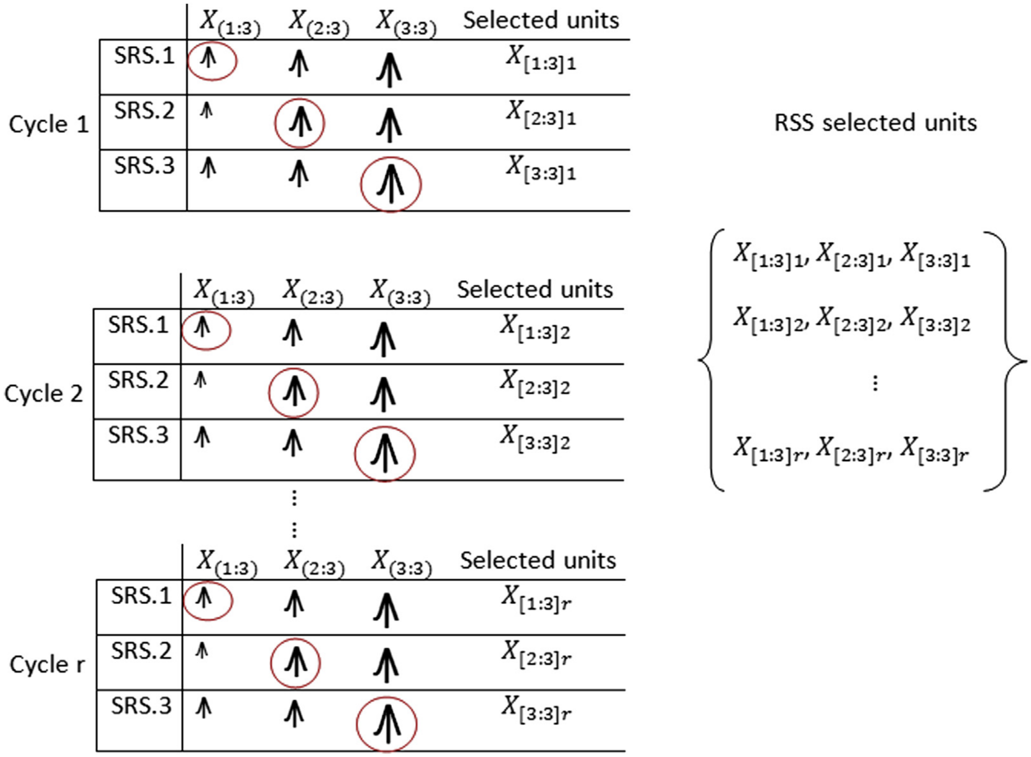


FIGURE 2.3

RSS scheme.

To explain more for this method, assuming that three sample sets are randomly selected to collect three RSS, the procedure is repeated r times. This can be visualized as shown in Fig. 2.3.

The selected observations are an RSS of size m denoted by $X_{[i:m]}$ $i = 1, 2, \ldots, m$, which represents the i^{th} ordered statistic obtained from the ith SRS of size m, and it is denoted by the ith judgment order statistics. It can be noted that the selected elements are independent-order statistics but not identically distributed. Also, note that we actually need m^2 observations selected via SRS to obtain m RSS units which means that we have to, unfortunately, discard $m(m-1)/2$ observations. In practice, the sample size m is kept small to ease the visual ranking, RSS literature suggested that $m = 3$, 4, 5, or 6. Therefore if a sample of larger size is needed, then the entire cycle may be repeated several times; say r times, to produce an RSS sample of size $n = rm$. Then the element of the desired sample will be in the form:

$$\{X_{[i:m]j},\ i = 1, 2, \ldots, m,\ \ j = 1, 2, \ldots, r\}$$

where $X_{[i:m]j}$ is the ith judgment order statistics in the jth cycle, which is the ith order statistics of the ith random sample of size m in the jth cycle. It should be noted that all of $X_{[i:m]j}$'s are mutually independent, in addition, the $X_{[i:m]j}$ are identically distributed for all i.

Let μ and σ^2 be the population mean and variance for variable X, respectively. Then the unbiased estimator of the population mean under RSS is defined as:

$$\overline{X}_{\mathrm{RSS}} = \frac{1}{rm}\sum_{j=1}^{r}\sum_{i=1}^{m} X_{[i:m]j}$$

which is more efficient than the usual sample mean $\overline{X}$ under SRS when both estimators are constructed on the basis of the same number n of actual measurements (McIntyre, 1952; Takahasi and Wakimoto, 1968). The variance of $\overline{X}_{\mathrm{RSS}}$ is given by:

$$\begin{aligned} Var(\overline{X}_{\mathrm{RSS}}) &= \frac{1}{rm^2}\sum_{j=1}^{r}\sum_{i=1}^{m} Var(X_{[i:m]j}) \\ &= \frac{1}{rm}\left(\sigma_X^2 - \frac{1}{m}\sum_{i=1}^{m}\left(E(X_{[i:m]i}) - \mu\right)^2\right) \end{aligned}$$

where $E\left(X_{[i:m]i}\right)$ is the expected value of the ith order statistics of a sample of size m:

$$E\left(X_{[i:m]i}\right) = \int_{-\infty}^{\infty} xf\left(X_{[i:m]}\right)dx$$

where

$$f\left(X_{[i:m]}\right) = m\binom{m-1}{i-1}F(x)^{i-1}(1-F(x))^{m-i}f(x)$$

Noting that the relative efficiency (RE) of estimating the population mean using novel RSS with respect to the traditional estimator by SRS is defined as follows:

$$RE(\overline{X}_{\mathrm{RSS}}, \overline{X}_{\mathrm{SRS}}) = \frac{\sigma^2/n}{\mathrm{Var}(\overline{X}_{\mathrm{RSS}})}$$

Takahasi and Wakimoto (1968) concluded that the RE for all continuous distributions is between 1 and $(m+1)/2$ with equal allocation and by using the same number of quantifications, where the maximum value holds for the standard uniform distribution. However, unequal allocation can actually increase the performance of RSS above and beyond that achievable with standard equal allocations. Actually the RE with unequal allocation will be between 0 and m.

2.3.1 SHEWHART CONTROL CHARTS UNDER THE RSS SCHEME

As mentioned earlier, the quality control charts are determined via the lower and upper control limits as well as the central limit term. The estimates of the three parts are necessary when the population mean and variance are unknown. This leads us to present new set of estimates of (μ,σ^2) using RSS so that we may construct the quality control charts. Salazar and Sinha (1997) proposed the following:

$$\begin{aligned} \mathrm{LCL} &= \mu - 3\sigma_{\overline{X}_{\mathrm{RSS}}} \\ \mathrm{CL} &= \mu \\ \mathrm{UCL} &= \mu + 3\sigma_{\overline{X}_{\mathrm{RSS}}} \end{aligned}$$

$\sigma_{\overline{X}_{\mathrm{RSS}}} = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n} E\left(X_{[i:n]i} - E\left(X_{[i:n]i}\right)\right)^2}$ is the standard deviation obtained via RSS (Chen et al., 2004). Muttlak and Al-Sabah (2003) proposed an estimator for $\sigma_{\overline{X}_{\mathrm{RSS}}}$:

$$\hat{\sigma}_{\overline{X}_{\mathrm{RSS}}} = \sqrt{\frac{1}{m}\left(\hat{\sigma}^2_{\mathrm{RSS}} - \frac{1}{m}\sum_{i=1}^{m}\left(\overline{X}_{[i]} - \overline{X}_{\mathrm{RSS}}\right)^2\right)}$$

where

$$\hat{\sigma}^2_{\mathrm{RSS}} = \frac{1}{rm-1}\sum_{i=1}^{m}\sum_{j=1}^{r}(\overline{X}_{[i:m]j} - \overline{X}_{\mathrm{RSS}})^2 \text{ and } \overline{X}_{[i]} = \frac{1}{r}\sum_{j=1}^{r} X_{[i:m]j}$$

2.4 SHEWHART CONTROL CHART UNDER RANKED REPETITIVE SAMPLING

We propose a Shewhart ranked control chart using repetitive sampling. Under a repetitive sampling scheme, there are two types of limits, outer (LCL1 and UCL1) and inner (LCL2 and UCL2) control chart limits. Then, the ranked repetitive control chart for the sample mean has the following steps:

Step 1: Draw an RSS of size n;
Step 2: Calculate the sample mean $\overline{X}_{\mathrm{RSS}}$;
Step 3: Declare the following decision about the entire process:

$$\begin{cases} \text{Out of Control}, \quad \text{if; } \overline{X}_{\mathrm{RSS}} > UCL_1 \text{ or } \overline{X}_{\mathrm{RSS}} < LCL_1 \\ \text{In Control}, \quad \text{if; } \mathrm{LCL}_2 < \overline{X}_{\mathrm{RSS}} < \mathrm{UCL}_2 \\ \text{Otherwise; Re Sample} \end{cases}$$

Where the outer control chart limits are given by:

$$\mathrm{UCL}_1 = \mu_{\mathrm{rss0}} + k_1\sigma_{\overline{X}_{\mathrm{RSS}}}$$

$$\mathrm{LCL}_1 = \mu_{\mathrm{rss0}} - k_1\sigma_{\overline{X}_{\mathrm{RSS}}}$$

Similarly, the inner control chart limits are given by:

$$\mathrm{UCL}_2 = \mu_{\mathrm{rss0}} + k_2\sigma_{\overline{X}_{\mathrm{RSS}}}$$

$$\mathrm{LCL}_2 = \mu_{\mathrm{rss0}} - k_2\sigma_{\overline{X}_{\mathrm{RSS}}}$$

Then the probability that the process is declared as in control is:

$$P_{\mathrm{in\_RSS}} = \frac{P\left(\mathrm{LCL}_2 < \overline{X}_{\mathrm{RSS}} < \mathrm{UCL}_2 | \mu_{\mathrm{rss}} = \mu_{\mathrm{rss0}}\right)}{1 - P_{\mathrm{rep}}}$$

where the probability that repetitive sampling is needed can be obtained by:

$$P_{\mathrm{rep\_RSS}} = P\left(\mathrm{UCL}_2 < \overline{X}_{\mathrm{RSS}} < \mathrm{UCL}_1\right) + P(\mathrm{LCL}_1 < \overline{X}_{\mathrm{RSS}} < \mathrm{LCL}_2)$$

Hence, the in-control average run length (ARL) is given by:

$$\mathrm{ARL}_{\mathrm{rss0}} = \frac{1}{1 - P_{\mathrm{rep\_RSS}}}$$

Suppose now that the process mean has shifted from μ_{rss0} to $\mu_{\mathrm{rss0}} + \delta\sigma$. Then, the probability that the process is declared as out of control is obtained by:

$$P^*_{\mathrm{in\_rss}} = \frac{P\left(\mathrm{LCL}_2 < \overline{X}_{\mathrm{RSS}} < \mathrm{UCL}_2 | \mu_{\mathrm{rss}} = \mu_{\mathrm{rss0}} + \delta\sigma_{\mathrm{rss}}\right)}{1 - P^*_{\mathrm{rep\_rss}}}$$

Similarly, the ARL for out-of-control process will be

$$\mathrm{ARL}_{1\_\mathrm{rss}} = \frac{1}{1 - P^*_{\mathrm{rep\_rss}}}$$

In general, the steps of the ranked repetitive sampling control chart can be summarized as follows:

Step 1: Using the assumption that the control chart is under control, specify the value of ARL_0;
Step 2: Find the value of the control charts multipliers k_1 and k_2 ($k_1 > k_2$) by minimizing ASN0 given that ARL0 is more than or equals the target;
Step 3: Find the value of ARL when the process is out of control.

2.5 PERFORMANCES OF THE PROPOSED CONTROL CHART

Monte Carlo simulation experiments were used to study the performance of the ranked control charts under the following assumptions:

Step 1: Setting up the control chart components: Sample mean and sample variance
- Generate 1,000,000 ranked set sampling each of size $m = 3$, 4, 5, and 6 from the standard normal distribution
- Calculate the mean and the variance for each subgroup
- Compute the grand mean and grand variance from the 1000000 subgroups;

Step 2: Setting up control limits multipliers
- Chose initial values of the $\mathrm{ARL}_{0\_\mathrm{rss}} = 350$ and 400
- Select the initial values of k_1 and k_2
- Using the generating samples from step 1 and an optimization problem to minimize the ASS0_rss find the optimal values of k_1 and k_2
- Compute the control chart limits (LCL1, UCL1) and (LCL2, UCL2);

Step 3: Compute the ARL_0 and ARL_1
- Follow the procedure of the proposed control chart and check if the process is declared as in-control, out-of-control, or resampling

- Compute the number of subgroups so far as the in-control run length say (R). Then the $\text{ARL}_{0\_\text{rss}} = R/1{,}000{,}000$
- Compute $\text{ARL}_{1\_\text{rss}}$ as $\delta = 0.1, 0.2, \ldots, 3.0$.

The results of this Monte Carlo experiment are given in Tables 2.1 and 2.2. The simulation results indicated that for the same values of m, k_1, and k_2, we note a decreasing trend in average run length as δ changes from 0.0 to 2.9.

Table 2.1 ASN and ARL When r_o is 350

	$n = 3$ $k_1 = 2.99; k_2 = 2.471$		$n = 4$ $k_1 = 2.98; k_2 = 2.245$		$n = 5$ $k_1 = 3.03; k_2 = 2.303$		$n = 6$ $k_1 = 3.001; k_2 = 1.875$	
δ	ASN	ARL	ASN	ARL	ASN	ARL	ASN	ARL
0	34	350.432	79	350.877	90	350.643	305	350.222
0.1	32	286.533	53	298.508	57	309.598	178	313.480
0.2	24	204.499	26	210.971	41	223.214	146	229.885
0.3	13	153.610	28	170.940	25	142.046	74	146.628
0.4	9	110.375	16	129.199	20	94.697	47	92.593
0.5	9	71.582	11	86.430	15	61.087	49	63.131
0.6	3	50.429	12	59.453	9	40.420	29	44.524
0.7	4	35.448	8	44.543	11	26.532	19	30.321
0.8	3	26.399	6	32.798	5	17.973	10	20.833
0.9	3	18.702	4	24.073	6	12.606	12	15.237
1	3	14.085	4	18.083	7	9.112	9	11.011
1.1	3	10.350	4	13.770	5	6.971	6	8.275
1.2	3	8.131	4	10.733	5	5.295	6	6.332
1.3	3	6.304	4	8.478	5	4.159	6	4.969
1.4	4	5.073	4	6.748	5	3.350	6	3.993
1.5	3	4.091	4	5.444	5	2.746	6	3.263
1.6	4	3.412	4	4.575	5	2.306	6	2.728
1.7	3	2.877	4	3.779	5	1.980	6	2.324
1.8	3	2.443	4	3.225	5	1.742	6	1.994
1.9	3	2.145	4	2.764	5	1.555	6	1.766
2	3	1.893	4	2.407	5	1.412	6	1.589
2.1	3	1.700	4	2.137	5	1.303	6	1.450
2.2	3	1.546	4	1.902	5	1.224	6	1.338
2.3	3	1.420	4	1.720	5	1.159	6	1.254
2.4	3	1.326	4	1.576	5	1.112	6	1.190
2.5	3	1.253	4	1.458	5	1.080	6	1.139
2.6	3	1.190	4	1.363	5	1.054	6	1.100
2.7	3	1.145	4	1.287	5	1.036	6	1.073
2.8	3	1.108	4	1.227	5	1.024	6	1.050
2.9	3	1.078	4	1.177	5	1.015	6	1.036

Table 2.2 ASN and ARL When r_o is 400

	$n = 3$ $k_1 = 3.005; k_2 = 2.595$		$n = 4$ $k_1 = 3.009; k_2 = 2.40$		$n = 5$ $k_1 = 3.02; k_2 = 2.92$		$n = 6$ $k_1 = 3.10; k_2 = 2.13$	
δ	ASN	ARL	ASN	ARL	ASN	ARL	ASN	ARL
0	59	400.00	66	400.01	87	400.01	316	400.02
0.1	37	250.00	45	277.77	65	344.828	220	346.783
0.2	29	227.273	37	200.001	31	285.714	131	263.158
0.3	22	175.439	25	192.307	30	169.492	98	196.078
0.4	11	138.889	8	135.135	17	163.934	55	102.041
0.5	13	80.000	12	90.090	10	95.238	40	80.645
0.6	10	51.546	7	59.523	7	65.360	32	40.984
0.7	8	39.526	9	49.261	7	55.556	14	34.130
0.8	5	27.855	5	34.246	5	34.722	11	21.368
0.9	3	19.724	4	25.316	5	27.322	8	14.085
1	3	13.986	4	19.723	6	19.685	8	10.142
1.1	3	10.627	4	14.347	5	15.060	6	7.686
1.2	3	8.117	4	11.481	5	10.965	6	5.794
1.3	3	6.618	4	8.703	5	9.033	7	4.686
1.4	3	5.152	4	6.747	5	6.998	6	3.670
1.5	3	4.225	4	5.803	5	5.828	6	2.998
1.6	3	3.516	4	4.683	5	4.744	6	2.449
1.7	3	2.847	4	3.930	5	4.005	6	2.110
1.8	3	2.421	4	3.290	5	3.356	6	1.800
1.9	3	2.125	4	2.819	5	2.867	6	1.608
2	3	1.925	4	2.513	5	2.562	6	1.458
2.1	3	1.679	4	2.171	5	2.210	6	1.342
2.2	3	1.532	4	1.969	5	1.969	6	1.245
2.3	3	1.437	4	1.740	5	1.755	6	1.185
2.4	3	1.339	4	1.609	5	1.602	6	1.136
2.5	3	1.259	4	1.493	5	1.492	6	1.093
2.6	3	1.197	4	1.387	5	1.384	6	1.063
2.7	3	1.147	4	1.313	5	1.305	6	1.042
2.8	3	1.107	4	1.232	5	1.244	6	1.030
2.9	3	1.081	4	1.196	5	1.197	6	1.018

2.5.1 COMPARATIVE STUDY: MONTE CARLO EXPERIMENT 2

In this section we use a simulation study to illustrate the quality control mechanism via different sampling approaches. Three sampling schemes are considered for computing the ARL of the Shewhart control chart in this experiment: the Shewhart control chart based on simple random sampling (SRS); Shewhart control chart based on ranked set sampling (RSS); and Shewhart control chart based on repetitive sampling (Rep-RSS). The simulation study is conducted under the normality assumption with mean μ_0 and variance σ_0^2 assuming the ranking is perfect. Note that under the SRS procedure, the ARL of the $\overline{X}$ chart will be 370. Therefore we set ARL0_rss equal to 370 to find the optimal multiplier values for the proposed control limits. The ARL is computed for $\delta = 0.0$, 0.1, 0.2, 0.3, 0.4, 0.5, 1.0, 1.8, 2.6, and 3.4 under the same Monte Carlo step as given in Experiment 1. The results are given in Tables 2.3–2.6

Table 2.3 ARL Using Different Methods When $m = 3$; $r_0 = 370$

Shift	SRS	RSS	Rep-RSS $k_1 = 3.03$, $k_2 = 2.09$
0.0	369.68	340.56	370.37
0.1	351.27	321.85	318.58
0.2	305.25	254.77	217.39
0.3	254.71	185.19	163.93
0.4	202.47	128.51	108.67
0.5	153.23	128.50	86.95
1.0	43.31	18.89	14.99
1.8	8.67	3.27	2.96
2.6	2.91	1.38	1.24
3.4	1.52	1.04	1.00

Table 2.4 ARL Using Different Methods When $m = 4$; $r_0 = 370$

Shift	SRS	RSS	Rep-RSS $k_1 = 3.05$, $k_2 = 2.12$
0.0	369.41	349.04	370.37
0.1	341.71	312.30	303.03
0.2	312.98	229.41	227.27
0.3	256.01	166.75	117.64
0.4	200.79	115.94	81.96
0.5	156.12	76.70	62.11
1.0	45.32	14.15	11.57
1.8	8.88	2.48	2.24
2.6	2.95	1.19	1.16
3.4	1.52	1.01	1.01

Table 2.5 ARL Using Different Methods When $m = 5$; $r_0 = 370$

Shift	SRS	RSS	Rep-RSS $k_1 = 3.06$, $k_2 = 2.51$
0.0	369.41	356.76	370.37
0.1	341.71	301.93	292.61
0.2	312.98	225.83	185.18
0.3	256.01	152.46	136.98
0.4	200.79	98.42	69.93
0.5	156.12	65.33	51.81
1.0	45.32	11.05	9.38
1.8	8.88	2.01	1.84
2.6	2.95	1.10	1.07
3.4	1.52	1.00	1.00

Table 2.6 ARL Using Different Methods When $m = 6$; $r_0 = 370$

Shift	SRS	RSS	Rep-RSS $k_1 = 3.07$, $k_2 = 2.73$
0.0	370.52	346.14	370.37
0.1	349.41	300.84	285.71
0.2	309.02	218.77	212.55
0.3	248.04	137.12	107.52
0.4	198.89	87.00	62.50
0.5	154.94	55.95	40.81
1.0	45.57	9.00	7.73
1.8	7.69	1.71	1.58
2.6	2.93	1.05	1.04
3.4	1.52	1.00	1.00

2.6 CONCLUDING REMARKS

A new ranked Shewhart control chart based on repetitive sampling is proposed in this chapter. The average run length properties are analyzed, and the ARL tables are provided for various parameters. The proposed control charts provide smaller values of ARL1 as compared to the existing control charts based on SRS, RSS, and Rep-RSS when ARL0 remains the same for all charts. It may be concluded that the proposed control charts perform better than the traditional control charts in terms of the ARL. It may be an interesting future work to design other ranked control charts under repetitive sampling.

REFERENCES

Ahmad, L., Aslam, M., Jun, C.-H., 2014. Designing of X-bar control charts based on process capability index using repetitive sampling. Trans. Inst. Meas. Control. 36 (3), 367–374.

Al-Nasser, A.D., 2007. L Ranked set sampling: a generalization procedure for robust visual sampling. Commun. Stat.: Simul. Comput. 36 (1), 33–43.

Al-Nasser, A.D., Al-Rawwash, M., 2007. A control chart based on ranked data. J. Appl. Sci. 7 (14), 1936–1941.

Al-Nasser, A.D., Al-Omari, A., Al-Rawwash, M., 2013. Monitoring the process mean based on quality control charts using folded ranked set sampling. Pak. J. Stat. Oper. Res. IX (1), 79–91.

Al-Omari, A., Al-Nasser, A.D., 2011. Statistical quality control limits for the sample mean chart using robust extreme ranked set sampling. Econ. Qual. Control. 26 (1), 73–89.

Al-Omari, A., Al-Nasser, A., Gogah, F., 2016. Double acceptance sampling plan for time truncated life tests based on transmuted new Weibull-Pareto distribution. Electron. J. Appl. Stat. Anal. 9 (3), 520–529.

Amiri, F., Noghondarian, K., Safaei, A.S., 2014. Evaluating the performance of variable scheme X-bar control chart: a Taguchi loss approach. Int. J. Prod. Res. 1–11.

Aslam, M., Azam, M., Jun, C.-H., 2013. A mixed repetitive sampling plan based on process capability index. Appl. Math. Model. 37 (2013), 10027–10035.

Aslam, M., Khan, N., Azam, M., Jun, C.-H., 2014a. Designing of a new monitoring T-chart using repetitive sampling. Inf. Sci. 269, 210–216.

Aslam, M., Azam, M., Jun, C.-H., 2014b. New attributes and variables control charts under repetitive sampling. Ind. Eng. Manag. Syst. 13 (1), 101–106.

Aslam, M., Khan, N., Jun, C.-H., 2015. A new S 2 control chart using repetitive sampling. J. Appl. Stat. 42 (11), 2485–2496.

Azam, M., Aslam, M., Jun, C.-H., 2015. Designing of a hybrid exponentially weighted moving average control chart using repetitive sampling. Int. J. Adv. Manuf. Technol. 77 (9–12), 1927–1933.

Balamurali, S., Jun, C.H., 2006. Repetitive group sampling procedure for variables inspection. J. Appl. Stat. 33 (3), 327–338.

Chan, L.Y., Lai, C.D., Xie, M., Goh, T.N., 2003. A two-stage decision procedure for monitoring processes with low fraction nonconforming. Eur. J. Oper. Res. 150 (2), 420–436.

Chen, Z., Bai, Z.D., Sinha, B.K., 2004. Ranked Set Sampling: Theory and Applications. Springer, New York.

Franco, B.C., Celano, G., Castagliola, P., Costa, A.F.B., 2014. Economic design of Shewhart control charts for monitoring auto correlated data with skip sampling strategies. Int. J. Prod. Econ. 151, 121–130.

Kumar, N., Chakraborti, S., Rakitzis, A.C., 2017. Improved Shewhart-type charts for monitoring times between events. J. Qual. Technol. 49 (3), 278–296.

Lee, H., Aslam, M., Shakeel, Q., Lee, W., Jun, C.-H., 2015. A control chart using an auxiliary variable and repetitive sampling for monitoring process mean. J. Stat. Comput. Simul. 85 (16), 3289–3296.

McIntyre, G.A., 1952. A method for unbiased selective sampling, using ranked sets. Aust. J. Agric. Res. 3, 385–390.

Montgomery, D.C., 2009. Introduction to Statistical Quality Control, 6th ed. John Wiley & Sons, Inc, New York.

Muttlak, H.A., Al-Sabah, W., 2003. Statistical quality control based on pair and selected ranked set sampling. Pak. J. Stat. 19 (1), 107–128.

Prajapati, D.R., Singh, S., 2016. Determination of level of correlation for products of pharmaceutical industry by using modified X-bar chart. Int. J. Qual. Reliab. Manag. 33 (6), 724–746.

Salazar, R.D., Sinha, A.K., 1997. Control chart $\bar{X}$ based on ranked set sampling. Comunicacion Tecica, No. 1-97-09 (PE/CIMAT).

Shafqat, A., Hussain, J., Al-Nasser, A.D., Aslam, M., 2017. Attribute control chart for some popular distributions. Commun. Stat: Theory Methods. 47 (8), 1978–1988.

Sherman, R.E., 1965. Design and evaluation of a repetitive group sampling plan. Technometrics 7 (1), 11–21.

Shewhart, W.A., 1924. Some applications of statistical methods to the analysis of physical and engineering data. Bell Techn. J. 3, 43–87.

Sindhumol, M., Srinivasan, M., Gallo, M., 2016. A robust dispersion control chart based on modified trimmed standard deviation. Electron. J. Appl. Stat. Anal. 9 (1), 111–121.

Takahasi, K., Wakimoto, K., 1968. On unbiased estimates of the population mean based on the sample stratified by means of ordering. Ann. Inst. Stat. Math. 20, 1–31.

Woodall, W.H., 2000. Controversies and contradictions in statistical process control. J. Qual. Technol. 32, 341–378.

CHAPTER

IMPROVED RATIO-CUM-PRODUCT ESTIMATORS OF THE POPULATION MEAN 3

Amer Ibrahim Falah Al-Omari

Department of Mathematics, Faculty of Science, Al al-Bayt University, Mafraq, Jordan

3.1 INTRODUCTION

Assume that the random variables X and Y have a joint probability density function (PDF) $f(x,y)$, and a joint cumulative distribution function (CDF) $F(x,y)$, with population means μ_X,μ_Y and population variances σ_X^2,σ_Y^2, of X and Y, respectively, and let ρ be the correlation coefficient between X and Y. Let $C_X = \frac{\sigma_X}{\mu_X}$ and $C_Y = \frac{\sigma_Y}{\mu_Y}$ be the population coefficients of variations of X and Y, respectively. Let (X_1, Y_1), (X_2, Y_2), …, (X_m, Y_m) be a bivariate simple random sample of size m from $f(x,y)$, and $\overline{X}_{\text{SRS}} = \frac{1}{m}\sum_{i=1}^{m} X_i$ be the sample mean of the auxiliary variable X with $\text{Var}(\overline{X}_{\text{SRS}}) = \frac{\sigma_X^2}{m}$ and $\overline{Y}_{\text{SRS}} = \frac{1}{m}\sum_{i=1}^{m} Y_i$ be the sample mean of the study variable Y with $\text{Var}(\overline{Y}_{\text{SRS}}) = \frac{\sigma_Y^2}{m}$. The usual simple random sampling (SRS) ratio estimator of the population mean μ_Y of the study variable Y is defined as

$$\hat{\mu}_Y^{\text{SRS}} = \mu_X \left(\frac{\overline{Y}_{\text{SRS}}}{\overline{X}_{\text{SRS}}}\right), \tag{3.1}$$

provided that the mean of X is known. Since this estimator is biased of the population mean, then the mean square error (MSE) of $\hat{\mu}_Y^{SRS}$ is given by

$$\text{MSE}\left(\hat{\mu}_Y^{\text{SRS}}\right) \cong \frac{1-f}{m}\left(\sigma_Y^2 + R^2\sigma_X^2 - 3R^2\sigma_X^2\rho\frac{C_Y}{C_X}\right), \tag{3.2}$$

where $f = \frac{m}{M}$, M is the population size, m is the sample size, R is the population ratio defined as $R = \frac{\mu_Y}{\mu_X}$, $\rho = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$, and $\sigma_{XY} = Cov(X,Y)$ is the covariance between X and Y, for more details see Cochran (1977). Al-Omari et al. (2009) suggested new ratio estimators of the population mean of the variable of interest Y using simple random sampling and ranked set sampling methods (RSS). Their SRS estimator is given by

$$\hat{\mu}_{Y-A}^{\text{SRS1}} = \left(\mu_X + q_1\right)\frac{\overline{Y}_{\text{SRS}}}{\overline{X}_{\text{SRS}} + q_1} \quad \text{and} \quad \hat{\mu}_{Y-A}^{\text{SRS3}} = \left(\mu_X + q_3\right)\frac{\overline{Y}_{\text{SRS}}}{\overline{X}_{\text{SRS}} + q_3}, \tag{3.3}$$

Ranked Set Sampling. DOI: https://doi.org/10.1016/B978-0-12-815044-3.00003-4

where q_1 and q_3 are the first and third quartiles of the auxiliary variable X, respectively, with MSE defined as

$$\mathrm{MSE}\left(\hat{\mu}_{Y-A}^{\mathrm{SRS}j}\right) \cong \frac{1}{m}\left(\frac{\mu_Y}{\mu_X+q_j}\right)\left[\left(\frac{\mu_Y}{\mu_X+q_j}\right)\sigma_X^2+\sigma_Y^2-2\sigma_X\sigma_Y\rho\right]\ j=1,3 \tag{3.4}$$

Singh and Tailor (2003) proposed another ratio estimator of the population mean using the SRS method given by

$$\hat{\mu}_{Y-ST}^{\mathrm{SRS}}=\overline{Y}_{\mathrm{SRS}}\left(\frac{\mu_X+\rho}{\overline{X}_{\mathrm{SRS}}+\rho}\right), \tag{3.5}$$

with MSE

$$\mathrm{MSE}\left(\hat{\mu}_{Y-ST}^{\mathrm{SRS}}\right)=\frac{1-f}{m}\mu_Y^2\left[C_Y^2+\frac{\mu_X}{\mu_X+\rho}C_X^2\left(\frac{\mu_X}{\mu_X+\rho}-2\rho\frac{C_Y}{C_X}\right)\right], \tag{3.6}$$

and bias given by

$$\mathrm{Bias}\left(\hat{\mu}_{Y-ST}^{\mathrm{SRS}}\right)=\frac{1-f}{m}\mu_Y C_X^2\frac{\mu_X}{\mu_X+\rho}\left(\frac{\mu_X}{\mu_X+\rho}-\rho\frac{C_Y}{C_X}\right), \tag{3.7}$$

where $C_Y^2=\frac{\sigma_Y^2}{\mu_Y^2}$, $C_X^2=\frac{\sigma_X^2}{\mu_X^2}$, $\rho=\frac{\sigma_{XY}}{\sigma_X S_Y}$, $\sigma_X^2=(M-1)^{-1}\sum_{i=1}^{M}\left(X_i-\mu_X\right)^2$, $\sigma_Y^2=(M-1)^{-1}\sum_{i=1}^{M}\left(Y_i-\mu_Y\right)^2$ and $\sigma_{XY}^2=(M-1)^{-1}\sum_{i=1}^{M}\left(X_i-\mu_X\right)\left(Y_i-\mu_Y\right)$.

Kadilar and Cingi (2004), based on SRS, suggested the following ratio estimator of the population mean

$$\hat{\mu}_{Y-KG}^{\mathrm{SRS}}=\frac{\overline{Y}^{\mathrm{SRS}}+b\left(\mu_X-\overline{X}^{\mathrm{SRS}}\right)}{\overline{X}^{\mathrm{SRS}}+\rho}\left(\mu_X+\rho\right), \tag{3.8}$$

where $b=\frac{S_{XY}}{S_X^2}$, with MSE given by

$$MSE\left(\hat{\mu}_{Y-KG}^{\mathrm{SRS}}\right)\cong\frac{1-f}{m}\left[R^2\sigma_X^2+\sigma_Y^2\left(1-\rho^2\right)\right]. \tag{3.9}$$

Also, for more about ratio and product method of estimation, see Jemain et al. (2007, 2008) and Haq and Shabbir (2010, 2013).

3.2 SAMPLING METHODS

In this section, we will define the sampling methods which are used throughout the work, namely; ranked set sampling and extreme ranked set sampling as well as the commonly used simple random sampling method.

3.2.1 RANKED SET SAMPLING

The RSS method can be described as follows:

Step 1: Select m random samples each of size m bivariate units from the population of interest.
Step 2: Rank the units within each set with respect to the variable of interest by visual inspection or any cost-free method.

Step 3: From the first set of m units, the smallest ranked unit X is selected together with the associated Y, and from the second set of m units the second smallest ranked unit X is selected together with the associated Y. The procedure is continued until from the mth set of m units the largest ranked unit X is selected with the associated Y.

The procedure can be repeated n times to increase the sample size to nm RSS bivariate units.

In this chapter, we assume that the ranking is performed on the variable X for estimating the population mean of the study variable Y. However, the whole process can be repeated while the ranking can be formed on the variable Y. Let $\left(X_{i(1)}, Y_{i[1]}\right)$, $\left(X_{i(2)}, Y_{i[2]}\right)$, $\ldots$, $\left(X_{i(m)}, Y_{i[m]}\right)$ be the order statistics of $X_{i1}, X_{i2}, \ldots, X_{im}$ and the judgment order of Y_{i1}, Y_{i2}, $\ldots$, Y_{im}, $(i = 1, 2, \ldots, m)$. Then the RSS units are $\left(X_{1(1)}, Y_{1[1]}\right)$, $\left(X_{2(2)}, Y_{2[2]}\right)$, $\ldots$, $\left(X_{m(m)}, Y_{m[m]}\right)$, where ( ) and [ ] indicate that the ranking of X is perfect and the ranking of Y has errors.

McIntyre (1952) proposed that the sample mean based on RSS as an estimator of the population mean defined as

$$\overline{Y}^{\text{RSS}} = \frac{1}{m}\sum_{i=1}^{m} Y_{i[i]} \tag{3.10}$$

Takahasi and Wakimoto (1968) provided the necessary mathematical theory of RSS and showed that

$$f(y) = \frac{1}{m}\sum_{i=1}^{m} f_{i[i]}(y), \mu_Y = \frac{1}{m}\sum_{i=1}^{m} \mu_{Y[i]}, \sigma_Y^2 = \frac{1}{m}\sum_{i=1}^{m} \sigma_{Y[i]}^2 - \frac{1}{m}\sum_{i=1}^{m} \left(\mu_{Y[i]} - \mu_Y\right)^2 ,$$

where $f_{i[i]}(y)$, $\mu_{Y[i]} = \int_{-\infty}^{\infty} y f_{i[i]}(y)dy$, and $\sigma_{Y[i]}^2 = \int_{-\infty}^{\infty} \left(y - \mu_{Y[i]}\right)^2 f_{i[i]}(y)dy$, respectively are the probability density function, mean, and the variance of the ith order statistics.

3.2.2 EXTREME RANKED SET SAMPLING

The extreme ranked set sampling (ERSS) method, as suggested by Samawi et al. (1996), can be described as follows:

Step 1: Select m random samples, each of size m units, from the target population and rank the units within each sample with respect to a variable of interest by visual inspection or any other cost-free method.

Step 2: For actual measurement, if the sample size m is even, from the first $\frac{m}{2}$ sets select the smallest ranked unit and from the other $\frac{m}{2}$ sets select the largest ranked unit. If the sample size is odd, from the first $\frac{m-1}{2}$ sets select the lowest ranked unit and from the other $\frac{m-1}{2}$ sets select the largest ranked unit, and from the remaining set the median is selected. The procedure can be repeated n times if needed to obtain a sample of size nm units.

If m is even, then the measured ERSSE units are $\left(X_{1(1)}, Y_{1[1]}\right)$, $\left(X_{2(1)}, Y_{2[1]}\right)$, $\ldots$, $\left(X_{\frac{m}{2}(1)}, Y_{\frac{m}{2}[1]}\right)$, $\left(X_{\frac{m+2}{2}(m)}, Y_{\frac{m+2}{2}[m]}\right)$, $\left(X_{\frac{m+4}{2}(m)}, Y_{\frac{m+4}{2}[m]}\right)$, $\ldots$, $\left(X_{m(m)}, Y_{m[m]}\right)$, where

$$\overline{X}^{\text{ERSSE}} = \frac{1}{m}\left(\sum_{i=1}^{\frac{m}{2}} X_{i(1)} + \sum_{i=\frac{m+2}{2}}^{m} X_{i(m)}\right) \quad \text{and} \quad \overline{Y}^{\text{ERSSE}} = \frac{1}{m}\left(\sum_{i=1}^{\frac{m}{2}} Y_{i[1]} + \sum_{i=\frac{m+2}{2}}^{m} Y_{i[m]}\right),$$

with respective variances

$$\sigma^2_{\overline{X}^{\text{ERSSE}}} = \frac{1}{2m}\left(\sigma^2_{X(1)} + \sigma^2_{X(m)}\right) \quad \text{and} \quad \sigma^2_{\overline{Y}^{\text{ERSSE}}} = \frac{1}{2m}\left(\sigma^2_{Y[1]} + \sigma^2_{Y[m]}\right) \tag{3.11}$$

If m is odd, then the measured ERSSO units are $\left(X_{1(1)}, Y_{1[1]}\right)$, $\left(X_{2(1)}, Y_{2[1]}\right)$, …, $\left(X_{\frac{m-1}{2}(1)}, Y_{\frac{m-1}{2}[1]}\right)$, $\left(X_{\frac{m+1}{2}\left(\frac{m+1}{2}\right)}, Y_{\frac{m+1}{2}\left[\frac{m+1}{2}\right]}\right)$, $\left(X_{\frac{m+3}{2}(m)}, Y_{\frac{m+3}{2}[m]}\right)$, …, $\left(X_{m(m)}, Y_{m[m]}\right)$, where

$$\overline{X}^{\text{ERSSO}} = \frac{1}{m}\left(\sum_{i=1}^{\frac{m-1}{2}} X_{i(1)} + X_{\frac{m+1}{2}\left(\frac{m+1}{2}\right)} + \sum_{i=\frac{m+3}{2}}^{m} X_{i(m)}\right),$$

with variance

$$\sigma^2_{\overline{X}^{\text{ERSSO}}} = \frac{1}{m^2}\left[\frac{m-1}{2}\left(\sigma^2_{X(1)} + \sigma^2_{X(m)}\right) + \sigma^2_{X\left(\frac{m+1}{2}\right)}\right], \tag{3.12}$$

and

$$\overline{Y}^{\text{ERSSO}} = \frac{1}{m}\left(\sum_{i=1}^{\frac{m}{2}} Y_{i[1]} + Y_{\frac{m+1}{2}\left[\frac{m+1}{2}\right]} + \sum_{i=\frac{m+3}{2}}^{m} Y_{i[m]}\right),$$

with variance

$$\sigma^2_{\overline{Y}^{\text{ERSSO}}} = \frac{1}{m^2}\left[\frac{m-1}{2}\left(\sigma^2_{Y[1]} + \sigma^2_{Y[m]}\right) + \sigma^2_{Y\left[\frac{m+1}{2}\right]}\right] \tag{3.13}$$

3.3 THE SUGGESTED ESTIMATORS

In this section, we will introduce the suggested estimators of the population mean of the study variable Y using SRS and ERSS schemes.

3.3.1 THE FIRST SUGGESTED ESTIMATOR

$$\hat{\mu}^{\text{ERSS}}_{Y-\mathbb{C}} = \left(\overline{Y}^{\text{ERSS}} + \mathbb{C}\right)\left(\delta\frac{\overline{X}^{\text{ERSS}} + \mathbb{C}}{\mu_X + \mathbb{C}} + (1-\delta)\frac{\mu_X + \mathbb{C}}{\overline{X}^{\text{ERSS}} + \mathbb{C}}\right) \tag{3.14}$$

where $\mathbb{C}$ can be considered as the coefficient of variation, coefficient of kurtosis, median, correlation coefficient, coefficient of skewness of the auxiliary variable X or any auxiliary information of X.

Using Taylor series expansion to the first order of approximation, this estimator can be written as

$$\hat{\mu}^{\text{ERSS}}_{Y-\mathbb{C}} \cong \overline{Y}^{\text{ERSS}} + (1-2\delta)\frac{\mu_Y + \mathbb{C}}{\mu_X + \mathbb{C}}\left(\overline{X}^{\text{ERSS}} - \mu_X\right) \tag{3.15}$$

Theorem 1: *To the first degree of approximation of the estimator* $\hat{\mu}^{ERSS}_{Y-\mathbb{C}}$, *we have*

1. *The estimator is approximately unbiased.*
2. *If m is even, the MSE of* $\hat{\mu}^{ERSS}_{Y-\mathbb{C}}$ *is*

$$MSE\left(\hat{\mu}^{ERSSE}_{Y-\mathbb{C}}\right) \cong \frac{1}{m}\left\{\frac{1}{2}\left(\sigma^2_{Y[1]}+\sigma^2_{Y[m]}\right)+\frac{(1-2\delta)\left(\mu_Y+\mathbb{C}\right)}{\mu_X+\mathbb{C}}\left[\begin{array}{l}\frac{1}{2}\frac{(1-2\delta)\left(\mu_Y+\mathbb{C}\right)}{\mu_X+\mathbb{C}}\left(\sigma^2_{X(1)}+\sigma^2_{X(m)}\right)\\ +2\left(\sigma_{XY}-\frac{1}{m}\sum_{i=1}^{m}H_{XY(i)}\right)\end{array}\right]\right\}, \tag{3.16}$$

and if m is odd, the MSE is

$$MSE\left(\hat{\mu}^{ERSSO}_{Y-\mathbb{C}}\right) \cong \frac{1}{m^2}\left\{\begin{array}{l}\frac{m-1}{2}\left(\sigma^2_{Y[1]}+\sigma^2_{Y[m]}\right)+\sigma^2_{Y\left[\frac{m+1}{2}\right]}\\ +\frac{(1-2\delta)\left(\mu_Y+\mathbb{C}\right)}{\mu_X+\mathbb{C}}\left\{\begin{array}{l}\frac{(1-2\delta)\left(\mu_Y+\mathbb{C}\right)}{\mu_X+\mathbb{C}}\left[\frac{m-1}{2}\left(\sigma^2_{X(1)}+\sigma^2_{X(m)}\right)+\sigma^2_{X\left(\frac{m+1}{2}\right)}\right]\\ +2\left(m\sigma_{XY}-\sum_{i=1}^{m}H_{XY(i)}\right)\end{array}\right\}\end{array}\right\}, \tag{3.17}$$

where $H_{XY(i)}=\left(\mu_{X(i)}-\mu_X\right)\left(\mu_{Y[i]}-\mu_Y\right)$.

Proof:

1. *The first part of the theorem can be proved by taking the expectation of* Eq. (3.12) *as*

$$E\left(\hat{\mu}^{ERSS}_{Y-\mathbb{C}}\right) \cong E\left[\overline{Y}^{ERSS}+(1-2\delta)\frac{\mu_Y+\mathbb{C}}{\mu_X+\mathbb{C}}\left(\overline{X}^{ERSS}-\mu_X\right)\right] \cong \mu_Y$$

2. *To find the MSE of the estimator* $\hat{\mu}^{ERSS}_{Y-\mathbb{C}}$, *from* Eq. (3.12) *we have*

$$\begin{aligned}\left(\hat{\mu}^{ERSS}_{Y-\mathbb{C}}-\mu_Y\right)^2 &= \left(\overline{Y}^{ERSS}-\mu_Y\right)^2+\left(\mu_Y+\mathbb{C}\right)^2\left(\frac{1-2\delta}{\mu_X+\mathbb{C}}\right)^2\left(\overline{X}^{ERSS}-\mu_X\right)^2\\ &\quad +2\left(\overline{Y}^{ERSS}-\mu_Y\right)\left[\left(\mu_Y+\mathbb{C}\right)\left(\frac{1-2\delta}{\mu_X+\mathbb{C}}\right)\left(\overline{X}^{ERSS}-\mu_X\right)\right]\end{aligned}$$

Taking the expectation of both sides yields

$$\begin{aligned}E\left(\hat{\mu}^{ERSS}_{Y-\mathbb{C}}-\mu_Y\right)^2 &= E\left(\overline{Y}^{ERSS}-\mu_Y\right)^2+\left(\mu_Y+\mathbb{C}\right)^2\left(\frac{1-2\delta}{\mu_X+\mathbb{C}}\right)^2E\left(\overline{X}^{ERSS}-\mu_X\right)^2\\ &\quad +2\left(\mu_Y+\mathbb{C}\right)\left(\frac{1-2\delta}{\mu_X+\mathbb{C}}\right)E\left[\left(\overline{Y}^{ERSS}-\mu_Y\right)\left(\overline{X}^{ERSS}-\mu_X\right)\right]\end{aligned}$$

Hence,

$$MSE\left(\hat{\mu}^{ERSS}_{Y-\mathbb{C}}\right) \cong Var\left(\overline{Y}^{ERSS}\right)+\left(\mu_Y+\mathbb{C}\right)^2\left(\frac{1-2\delta}{\mu_X+\mathbb{C}}\right)^2Var\left(\overline{X}^{ERSS}\right)+2\left(\mu_Y+\mathbb{C}\right)\left(\frac{1-2\delta}{\mu_X+\mathbb{C}}\right)Cov\left(\overline{Y}^{ERSS},\overline{X}^{ERSS}\right)$$

Now, if the sample size is even, the MSE of $\hat{\mu}_Y^{ERSS}$ is given by

$$MSE\left(\hat{\mu}_Y^{ERSSE}\right) \cong \tfrac{1}{m}\left\{\tfrac{1}{2}\left(\sigma^2_{Y[1]}+\sigma^2_{Y[m]}\right)\right.$$

$$\left.+\frac{(1-2\delta)\left(\mu_Y+\mathbb{C}\right)}{\mu_X+\mathbb{C}}\left[\frac{1}{2}\frac{(1-2\delta)\left(\mu_Y+\mathbb{C}\right)}{\mu_X+\mathbb{C}}\left(\sigma^2_{X(1)}+\sigma^2_{X(m)}\right)+2\left(\sigma_{XY}-\frac{1}{m}\sum_{i=1}^{m}H_{XY(i)}\right)\right]\right\}$$

and if the sample size is odd, the MSE is given as

$$MSE\left(\hat{\mu}_Y^{ERSSO}\right) \cong \frac{1}{m^2}\left\{\frac{m-1}{2}\left(\sigma^2_{Y[1]}+\sigma^2_{Y[m]}\right)\right.$$

$$\left.+\sigma^2_{Y\left[\frac{m+1}{2}\right]}+\frac{(1-2\delta)\left(\mu_Y+\mathbb{C}\right)}{\mu_X+\mathbb{C}}\left\{\begin{matrix}\dfrac{(1-2\delta)\left(\mu_Y+\mathbb{C}\right)}{\mu_X+\mathbb{C}}\left[\dfrac{m-1}{2}\left(\sigma^2_{X(1)}+\sigma^2_{X(m)}\right)+\sigma^2_{X\left(\frac{m+1}{2}\right)}\right]\\ +2\left(m\sigma_{XY}-\displaystyle\sum_{i=1}^{m}H_{XY(i)}\right)\end{matrix}\right\}\right\}$$

3.3.2 THE SECOND SUGGESTED ESTIMATOR

$$\hat{\mu}_{Y-\mathbb{C}}^{\mathrm{SRS}}=\left(\overline{Y}^{\mathrm{SRS}}+\mathbb{C}\right)\left(\delta\frac{\overline{X}^{\mathrm{SRS}}+\mathbb{C}}{\mu_X+\mathbb{C}}+(1-\delta)\frac{\mu_X+\mathbb{C}}{\overline{X}^{\mathrm{SRS}}+\mathbb{C}}\right), \tag{3.18}$$

Using Taylor series expansion to the first order of approximation, this estimator can be written as

$$\hat{\mu}_{Y-\mathbb{C}}^{\mathrm{SRS}}\cong\overline{Y}^{\mathrm{SRS}}+(1-2\delta)\frac{\mu_Y+\mathbb{C}}{\mu_X+\mathbb{C}}\left(\overline{X}^{\mathrm{SRS}}-\mu_X\right). \tag{3.19}$$

Theorem 2: *To the first degree of approximation of the estimator $\hat{\mu}_{Y-\mathbb{C}}^{SRS}$, we have*

1. *The estimator is approximately an unbiased estimator of the population mean.*
2. $MSE\left(\hat{\mu}_{Y-\mathbb{C}}^{SRS}\right)\cong\frac{\sigma_Y^2}{m}+\left(\mu_Y+\mathbb{C}\right)^2\left(\frac{1-2\delta}{\mu_X+\mathbb{C}}\right)^2\frac{\sigma_X^2}{m}+2\rho\sigma_Y\sigma_X m\left(\mu_Y+\mathbb{C}\right)\left(\frac{1-2\delta}{\mu_X+\mathbb{C}}\right).$ (3.20)

Proof: *The proof of (1) is directly and the proof of (2) can be obtained as above in Theorem 1 using*

$$MSE\left(\hat{\mu}_{Y-\mathbb{C}}^{SRS}\right)\cong Var\left(\overline{Y}^{SRS}\right)+\left(\mu_Y+\mathbb{C}\right)^2\left(\frac{1-2\delta}{\mu_X+\mathbb{C}}\right)^2 Var\left(\overline{X}^{SRS}\right)+2\left(\mu_Y+\mathbb{C}\right)\left(\frac{1-2\delta}{\mu_X+\mathbb{C}}\right)Cov\left(\overline{Y}^{SRS},\overline{X}^{SRS}\right)$$

3.3.3 THE THIRD SUGGESTED ESTIMATOR

Singh and Espejo (2003) suggested a ratio-product estimator of a population mean using simple random sampling as

$$\hat{\mu}_{Y-W}^{\mathrm{SRS}} = \overline{Y}_{\mathrm{SRS}}\left(w\frac{\overline{X}^{\mathrm{SRS}}}{\mu_X} + (1-w)\frac{\mu_X}{\overline{X}^{\mathrm{SRS}}}\right), \tag{3.21}$$

with bias given by

$$B\left(\hat{\mu}_{Y-W}^{\mathrm{SRS}}\right) = \frac{1-f}{m}\mu_Y C_X^2\left[\rho\frac{C_Y}{C_X} + w\left(1-\rho\frac{C_Y}{C_X}\right)\right],$$

and the associated MSE is

$$\mathrm{MSE}\left(\hat{\mu}_{Y-W}^{\mathrm{SRS}}\right) = \frac{1-f}{m}\mu_Y^2\left\{C_Y^2 + C_X^2[1-2w]\left[1-2w+2\rho\frac{C_Y}{C_X}\right]\right\}, \tag{3.22}$$

where the optimal value of w, which minimizes the MSE in Eq. (3.22), is $w_{\mathrm{Opt}} = \frac{1+\rho C_Y/C_X}{2}$.

Motivated by Singh and Espejo (2003), we propose a new ratio-cum-product type estimator of the population mean using ERSS technique as

$$\hat{\mu}_{Y-W}^{\mathrm{ERSS}} = \overline{Y}^{\mathrm{ERSS}}\left(w\frac{\overline{X}^{\mathrm{ERSS}}}{\mu_X} + (1-w)\frac{\mu_X}{\overline{X}^{\mathrm{ERSS}}}\right), \tag{3.23}$$

which can be written to the first degree of Taylor series approximation as

$$\hat{\mu}_{Y-W}^{\mathrm{ERSS}} = 2\mu_Y - \overline{Y}^{\mathrm{ERSS}} + \frac{w\mu_Y-(1-w)}{\mu_X}\left(\overline{X}^{\mathrm{ERSS}} - \mu_X\right). \tag{3.24}$$

Theorem 3: *To the first degree of approximation of the estimator $\hat{\mu}_{Y-W}^{ERSS}$, we have*

1. *The estimator is approximately an unbiased estimator of the population mean.*
2. *The MSE of $\hat{\mu}_{Y-W}^{ERSSE}$ and $\hat{\mu}_{Y-W}^{ERSSO}$, respectively, are given by*

$$\begin{aligned} MSE\left(\hat{\mu}_{Y-W}^{ERSSE}\right) \cong{}& \frac{1}{2m}\left(\sigma_{Y[1]}^2 + \sigma_{Y[m]}^2\right) + \frac{1}{2m}\left[\frac{w\mu_Y-(1-w)}{\mu_X}\right]^2\left(\sigma_{X(1)}^2 + \sigma_{X(m)}^2\right) \\ &- 2\frac{w\mu_Y-(1-w)}{\mu_X}\left(\sigma_{XY} - \frac{1}{m}\sum_{i=1}^{m} H_{XY(i)}\right) \end{aligned} \tag{3.25}$$

$$\begin{aligned} MSE\left(\hat{\mu}_{Y-W}^{ERSSO}\right) \cong{}& \frac{1}{m^2}\left[\frac{m-1}{2}\left(\sigma_{Y[1]}^2 + \sigma_{Y[m]}^2\right) + \sigma_{Y\left[\frac{m+1}{2}\right]}^2\right] + \frac{1}{m^2}\left[\frac{w\mu_Y-(1-w)}{\mu_X}\right]^2\left[\frac{m-1}{2}\left(\sigma_{X(1)}^2 + \sigma_{X(m)}^2\right) + \sigma_{X\left(\frac{m+1}{2}\right)}^2\right] \\ &- 2\frac{w\mu_Y-(1-w)}{\mu_X}\left(\sigma_{XY} - \frac{1}{m}\sum_{i=1}^{m} H_{XY(i)}\right) \end{aligned} \tag{3.26}$$

Proof: *The proof of (1) is directly and the proof of (2) can be obtained by using*

$$
\begin{aligned}
MSE\left(\hat{\mu}_{Y-W}^{ERSS}\right) &\cong \mathrm{Var}\left(\overline{Y}^{ERSS}\right) - 2\frac{w\mu_Y-(1-w)}{\mu_X}E\left[\left(\overline{X}^{ERSS}-\mu_X\right)\left(\overline{Y}^{ERSS}-\mu_Y\right)\right] + \left[\frac{w\mu_Y-(1-w)}{\mu_X}\right]^2 Var\left(\overline{X}^{ERSS}\right)\\
&\cong Var\left(\overline{Y}^{ERSS}\right) - 2\frac{w\mu_Y-(1-w)}{\mu_X}Cov\left(\overline{X}^{ERSS},\overline{Y}^{ERSS}\right) + \left[\frac{w\mu_Y-(1-w)}{\mu_X}\right]^2 Var\left(\overline{X}^{ERSS}\right)
\end{aligned}
$$

Hence, the results can be obtained by substituting the expressions of $Var\left(\overline{Y}^{ERSS}\right)$ *and* $Var\left(\overline{X}^{ERSS}\right)$ *for odd and even sample sizes, respectively.*

3.4 SIMULATION STUDY

The suggested ratio-cum-product estimators of the population mean are compared within themselves based on simulation study for sample sizes $m = 3, \ldots, 10$ with $\rho = \pm 0.99, \pm 0.90, \pm 0.70, \pm 0.50, \pm 0.30, \pm 0.10$ and $\mathbb{C} = \rho$. The samples are generated from the bivariate normal distribution for $\mu_X = 7$, $\mu_Y = 5$, and $\sigma_X^2 = \sigma_Y^2 = 1$.

The efficiency of $\hat{\mu}_{Y-\mathbb{C}}^{\mathrm{ERSSE}}$ with respect to $\hat{\mu}_{Y-\mathbb{C}}^{\mathrm{SRS}}$ is defined as

$$
\begin{aligned}
\mathrm{Eff}\left(\hat{\mu}_{Y-\mathbb{C}}^{\mathrm{ERSSE}},\ \hat{\mu}_{Y-\mathbb{C}}^{\mathrm{SRS}}\right) &= \frac{\mathrm{MSE}\left(\hat{\mu}_{Y-\mathbb{C}}^{\mathrm{SRS}}\right)}{\mathrm{MSE}\left(\hat{\mu}_{Y-\mathbb{C}}^{\mathrm{ERSSE}}\right)}\\
&\cong \frac{2\left[\left(\mu_X+\mathbb{C}\right)\left(\sigma_Y^2+2\rho m\sigma_Y\sigma_X\right)+(1-2\delta)\left(\mu_Y+\mathbb{C}\right)\sigma_X^2\right]}{\dfrac{\left(\sigma^2_{Y[1]}+\sigma^2_{Y[m]}\right)\left(\mu_X+\mathbb{C}\right)^2}{(1-2\delta)\left(\mu_Y+\mathbb{C}\right)} + \left[\begin{array}{l}\dfrac{(1-2\delta)\left(\mu_Y+\mathbb{C}\right)}{\mu_X+\mathbb{C}}\left(\sigma^2_{X(1)}+\sigma^2_{X(m)}\right)\\ +4\left(\sigma_{XY}-\dfrac{1}{m}\displaystyle\sum_{i=1}^{m}H_{XY(i)}\right)\end{array}\right]}
\end{aligned}
\tag{3.27}
$$

And the efficiency of $\hat{\mu}_{Y-\mathbb{C}}^{\mathrm{ERSSO}}$ with respect to $\hat{\mu}_{Y-\mathbb{C}}^{\mathrm{SRS}}$ is defined as

$$
\begin{aligned}
\mathrm{Eff}\left(\hat{\mu}_{Y-\mathbb{C}}^{\mathrm{ERSSO}},\hat{\mu}_{Y-\mathbb{C}}^{\mathrm{SRS}}\right) &= \frac{\mathrm{MSE}\left(\hat{\mu}_{Y-\mathbb{C}}^{\mathrm{SRS}}\right)}{\mathrm{MSE}\left(\hat{\mu}_{Y-\mathbb{C}}^{\mathrm{ERSSO}}\right)}\\
&\cong \frac{m\left[\dfrac{\mu_X+\mathbb{C}}{(1-2\delta)\left(\mu_Y+\mathbb{C}\right)}\sigma_Y^2+\dfrac{(1-2\delta)\left(\mu_Y+\mathbb{C}\right)}{\mu_X+\mathbb{C}}\sigma_X^2+2\rho\sigma_Y\sigma_X m\right]}{\begin{array}{l}\dfrac{\mu_X+\mathbb{C}}{(1-2\delta)\left(\mu_Y+\mathbb{C}\right)}\left(\dfrac{m-1}{2}\left(\sigma^2_{Y[1]}+\sigma^2_{Y[m]}\right)+\sigma^2_{Y\left[\frac{m+1}{2}\right]}\right)\\ +\left\{\begin{array}{l}\dfrac{(1-2\delta)\left(\mu_Y+\mathbb{C}\right)}{\mu_X+\mathbb{C}}\left[\dfrac{m-1}{2}\left(\sigma^2_{X(1)}+\sigma^2_{X(m)}\right)+\sigma^2_{X\left(\frac{m+1}{2}\right)}\right]\\ +2\left(m\sigma_{XY}-\displaystyle\sum_{i=1}^{m}H_{XY(i)}\right)\end{array}\right\}\end{array}}
\end{aligned}
\tag{3.28}
$$

The efficiency of $\hat{\mu}_{Y-W}^{\text{ERSSE}}$ and $\hat{\mu}_{Y-W}^{\text{ERSSO}}$ with respect to $\hat{\mu}_{Y-W}^{\text{SRS}}$ is defined as

$$\begin{aligned}\operatorname{Eff}\left(\hat{\mu}_{Y-W}^{\text{ERSSE}}, \hat{\mu}_{Y-W}^{\text{SRS}}\right) &= \frac{\operatorname{MSE}\left(\hat{\mu}_{Y-W}^{\text{SRS}}\right)}{\operatorname{MSE}\left(\hat{\mu}_{Y-W}^{\text{ERSSE}}\right)} \\ &\cong \frac{2(1-f)\mu_Y^2\left\{C_Y^2 + C_X^2[1-2w]\left[1-2w+2\rho\frac{C_Y}{C_X}\right]\right\}}{\begin{array}{c}\left(\sigma_{Y[1]}^2+\sigma_{Y[m]}^2\right)+\left[\frac{w\mu_Y-(1-w)}{\mu_X}\right]^2\left(\sigma_{X(1)}^2+\sigma_{X(m)}^2\right) \\ -4m\frac{w\mu_Y-(1-w)}{\mu_X}\left(\sigma_{XY}-\frac{1}{m}\sum_{i=1}^{m}H_{XY(i)}\right)\end{array}}\end{aligned} \tag{3.29}$$

$$\begin{aligned}\operatorname{Eff}\left(\hat{\mu}_{Y-W}^{\text{ERSSO}}, \hat{\mu}_{Y-W}^{\text{SRS}}\right) &= \frac{\operatorname{MSE}\left(\hat{\mu}_{Y-W}^{\text{SRS}}\right)}{\operatorname{MSE}\left(\hat{\mu}_{Y-W}^{\text{ERSSO}}\right)} \\ &\cong \frac{m(1-f)\mu_Y^2\left\{C_Y^2 + C_X^2[1-2w]\left[1-2w+2\rho\frac{C_Y}{C_X}\right]\right\}}{\begin{array}{c}\frac{m-1}{2}\left(\sigma_{Y[1]}^2+\sigma_{Y[m]}^2\right)+\sigma^2_{Y\left[\frac{m+1}{2}\right]}+\left[\frac{w\mu_Y-(1-w)}{\mu_X}\right]^2 \\ \times\left[\frac{m-1}{2}\left(\sigma_{X(1)}^2+\sigma_{X(m)}^2\right)+\sigma^2_{X\left(\frac{m+1}{2}\right)}\right]-2m^2\frac{w\mu_Y-(1-w)}{\mu_X}\left(\sigma_{XY}-\frac{1}{m}\sum_{i=1}^{m}H_{XY(i)}\right)\end{array}}\end{aligned} \tag{3.30}$$

The results of the simulation are presented in Tables 3.1–3.4 for all cases considered in this study.

Remarks:

1. $\hat{\mu}_{Y-\mathbb{C}}^{\text{ERSSE}}$ is more efficient than $\hat{\mu}_{Y-\mathbb{C}}^{\text{SRS}}$ if $\operatorname{Eff}\left(\hat{\mu}_{Y-\mathbb{C}}^{\text{ERSSE}}, \hat{\mu}_{Y-\mathbb{C}}^{\text{SRS}}\right) > 1$.
2. $\hat{\mu}_{Y-\mathbb{C}}^{\text{ERSSO}}$ is more efficient than $\hat{\mu}_{Y-\mathbb{C}}^{\text{SRS}}$ if $\operatorname{Eff}\left(\hat{\mu}_{Y-\mathbb{C}}^{\text{ERSSO}}, \hat{\mu}_{Y-\mathbb{C}}^{\text{SRS}}\right) > 1$.
3. $\hat{\mu}_{Y-W}^{\text{ERSSE}}$ is more efficient than $\hat{\mu}_{Y-W}^{\text{SRS}}$ if $\operatorname{Eff}\left(\hat{\mu}_{Y-W}^{\text{ERSSE}}, \hat{\mu}_{Y-W}^{\text{SRS}}\right) > 1$.
4. $\hat{\mu}_{Y-W}^{\text{ERSSO}}$ is more efficient than $\hat{\mu}_{Y-W}^{\text{SRS}}$ if $\operatorname{Eff}\left(\hat{\mu}_{Y-W}^{\text{ERSSO}}, \hat{\mu}_{Y-W}^{\text{SRS}}\right) > 1$.

We observe from Tables 3.1–3.4 that:

- The ratio-cum-product estimator $\hat{\mu}_{Y-C}^{\text{ERSS}}$ performs better than $\hat{\mu}_{Y-C}^{\text{SRS}}$ for all values of the correlation coefficient and sample sizes. The same thing can be concluded for $\hat{\mu}_{Y-W}^{\text{ERSS}}$ as compared to $\hat{\mu}_{Y-W}^{\text{SRS}}$.
- Without loss of generality, the efficiency of $\hat{\mu}_{Y-W}^{\text{ERSS}}$ with respect to $\hat{\mu}_{Y-W}^{\text{SRS}}$ increases in the sample size for fixed value of the correlation coefficient, especially for $\rho = -0.99, -0.90, -0.70$.
- The efficiency of the suggested estimators $\hat{\mu}_{Y-W}^{\text{ERSS}}$ is increasing as the sample size increasing for most cases in Tables 3.3 and 3.4.
- The bias values of all suggested estimators approaches zero for all cases considered in this study.

Table 3.1 The Efficiency of $\hat{\mu}_{Y-\mathbb{C}}^{ERSS}$ With Respect to $\hat{\mu}_{Y-\mathbb{C}}^{SRS}$ for $m = 3, \ldots, 10$ With $\rho = 0.99,\ 0.90$, 0.70, 0.50, 0.30, 0.10 for $\mu_X = 7$ and $\mu_Y = 5$

	$\hat{\mu}_{Y-\mathbb{C}}^{\mathbf{SRS}}$	$\hat{\mu}_{Y-\mathbb{C}}^{\mathbf{ERSS}}$	$B\left(\hat{\mu}_{Y-\mathbb{C}}^{\mathbf{SRS}}\right)$	$B\left(\hat{\mu}_{Y-\mathbb{C}}^{\mathbf{ERSS}}\right)$	$V\left(\hat{\mu}_{Y-\mathbb{C}}^{\mathbf{SRS}}\right)$	$V\left(\hat{\mu}_{Y-\mathbb{C}}^{\mathbf{ERSS}}\right)$	**Eff**
m		$\rho = 0.99$					
3	5.860076	5.665692	2.860076	2.665692	6.327324	3.216656	1.405402
4	5.711220	5.639494	2.711220	2.639494	4.497059	2.288005	1.280157
5	5.716215	5.586388	2.716215	2.586388	3.754843	1.522393	1.355692
6	5.670257	5.575278	2.670257	2.575278	3.029844	1.282059	1.283796
7	5.640729	5.557779	2.640729	2.557779	2.651196	0.948352	1.284899
8	5.642768	5.542131	2.642768	2.542131	2.299761	0.851347	1.269383
9	5.620538	5.537345	2.620538	2.537345	2.022440	0.695794	1.246112
10	5.607014	5.530943	2.607014	2.530943	1.765726	0.621081	1.218521
		$\rho = 0.90$					
3	5.831177	5.647893	2.831177	2.647893	5.580202	3.002137	1.357747
4	5.689738	5.625215	2.689738	2.625215	4.088180	2.148276	1.252526
5	5.696826	5.582238	2.696826	2.582238	3.307520	1.461004	1.301568
6	5.654074	5.567641	2.654074	2.567641	2.673719	1.234905	1.241469
7	5.627512	5.552080	2.627512	2.552080	2.338993	0.928616	1.242025
8	5.630264	5.536653	2.630264	2.536653	2.030088	0.831683	1.231491
9	5.609221	5.534355	2.609221	2.534355	1.786739	0.695267	1.207433
10	5.597238	5.530898	2.597238	2.530898	1.559441	0.616440	1.182743
		$\rho = 0.70$					
3	5.778805	5.634193	2.778805	2.634193	5.803809	3.194170	1.334785
4	5.663591	5.600695	2.663591	2.600695	2.945515	1.893382	1.159782
5	5.664042	5.566405	2.664042	2.566405	2.468752	1.370683	1.202178
6	5.626662	5.558594	2.626662	2.558594	2.006020	1.147417	1.157471
7	5.605011	5.546434	2.605011	2.546434	1.751260	0.895291	1.156882
8	5.609087	5.532567	2.609087	2.532567	1.523950	0.798513	1.155132
9	5.589609	5.526360	2.589609	2.526360	1.343871	0.692496	1.137803
10	5.580454	5.527079	2.580454	2.527079	1.171936	0.611124	1.119108
		$\rho = 0.50$					

3	5.782869	5.610228	2.782869	2.610228	5.291162	2.380849	1.417808
4	5.652202	5.585049	2.652202	2.585049	2.248846	1.716541	1.105251
5	5.645564	5.551397	2.645564	2.551397	1.856226	1.278483	1.137020
6	5.610764	5.546816	2.610764	2.546816	1.516704	1.076820	1.101770
7	5.592014	5.541929	2.592014	2.541929	1.315125	0.888374	1.093048
8	5.596935	5.529544	2.596935	2.529544	1.149641	0.778516	1.099846
9	5.577442	5.521496	2.577442	2.521496	1.015393	0.692639	1.086237
10	5.570446	5.528277	2.570446	2.528277	0.884619	0.602522	1.071069
***ρ* = 0.30**							
3	5.767762	5.614585	2.767762	2.614585	12.51099	2.211185	2.229574
4	5.658593	5.581372	2.658593	2.581372	1.959375	1.592738	1.093417
5	5.641453	5.552919	2.641453	2.552919	1.486897	1.228390	1.092746
6	5.606446	5.542686	2.606446	2.542686	1.215599	1.046016	1.066286
7	5.588538	5.535083	2.588538	2.535083	1.035881	0.864945	1.061005
8	5.593790	5.528652	2.59379	2.528652	0.911906	0.766370	1.066924
9	5.572733	5.521059	2.572733	2.521059	0.804242	0.689594	1.053633
10	5.567226	5.524619	2.567226	2.524619	0.700309	0.595986	1.046096
***ρ* = 0.10**							
3	5.756092	5.622937	2.756092	2.622937	50.55562	2.188349	6.412739
4	5.640029	5.597807	2.640029	2.597807	45.76353	1.550466	6.354121
5	5.651727	5.563243	2.651727	2.563243	1.380651	1.218637	1.080045
6	5.613796	5.545454	2.613796	2.545454	1.115020	1.019538	1.059752
7	5.594591	5.538331	2.594591	2.538331	0.919467	0.868542	1.046460
8	5.599655	5.526860	2.599655	2.526860	0.816460	0.760321	1.060085
9	5.575500	5.521781	2.575500	2.521781	0.713775	0.690839	1.042092
10	5.570813	5.526857	2.570813	2.526857	0.622306	0.601462	1.035056

Table 3.2 The Efficiency of $\hat{\mu}^{ERSS}_{Y-\mathbb{C}}$ With Respect to $\hat{\mu}^{SRS}_{Y-\mathbb{C}}$ for $m = 3,\ldots,10$ With $\rho = -0.99, -0.90, -0.70, -0.50, -0.30, -0.10$ for $\mu_X = 7$ and $\mu_Y = 5$

	$\hat{\mu}^{\mathrm{SRS}}_{Y-\mathbb{C}}$	$\hat{\mu}^{\mathrm{ERSS}}_{Y-\mathbb{C}}$	$B\left(\hat{\mu}^{\mathrm{SRS}}_{Y-\mathbb{C}}\right)$	$B\left(\hat{\mu}^{\mathrm{ERSS}}_{Y-\mathbb{C}}\right)$	$V\left(\hat{\mu}^{\mathrm{SRS}}_{Y-\mathbb{C}}\right)$	$V\left(\hat{\mu}^{\mathrm{ERSS}}_{Y-\mathbb{C}}\right)$	Eff
m		$\rho = -0.99$					
3	4.261237	4.072545	− 0.73876	− 0.92745	1.362038	0.682503	1.236689
4	4.225636	4.055253	− 0.77436	− 0.94475	0.971821	0.473467	1.150400
5	4.197943	4.045286	− 0.80206	− 0.95471	0.784817	0.317509	1.162022
6	4.183963	4.033338	− 0.81604	− 0.96666	0.651751	0.257177	1.105786
7	4.178395	4.027180	− 0.82161	− 0.97282	0.538343	0.198738	1.059611
8	4.184747	4.024501	− 0.81525	− 0.97550	0.487940	0.174904	1.023146
9	4.167585	4.024336	− 0.83242	− 0.97566	0.422390	0.143096	1.018530
10	4.168594	4.022187	− 0.83141	− 0.97781	0.381368	0.109491	1.006580
		$\rho = -0.90$					
3	4.289380	4.153400	− 0.71062	− 0.8466	1.184162	0.624515	1.259382
4	4.258207	4.138356	− 0.74179	− 0.86164	0.847502	0.436898	1.185216
5	4.233749	4.129966	− 0.76625	− 0.87003	0.684779	0.298240	1.205383
6	4.221397	4.119352	− 0.77860	− 0.88065	0.569186	0.241805	1.155367
7	4.216908	4.114773	− 0.78309	− 0.88523	0.470516	0.190382	1.112669
8	4.223308	4.112375	− 0.77669	− 0.88762	0.426407	0.166779	1.078563
9	4.207604	4.111922	− 0.79240	− 0.88808	0.369550	0.138518	1.075757
10	4.208877	4.110709	− 0.79112	− 0.88929	0.333386	0.124888	1.047542
		$\rho = -0.70$					
3	4.374327	4.336182	− 0.62567	− 0.66382	0.850403	0.519920	1.292842
4	4.351122	4.326305	− 0.64888	− 0.67370	0.611744	0.365798	1.260014
5	4.333398	4.323276	− 0.66660	− 0.67672	0.495386	0.260410	1.308170
6	4.324262	4.313616	− 0.67574	− 0.68638	0.412845	0.212798	1.271297
7	4.321563	4.309888	− 0.67844	− 0.69011	0.341618	0.172636	1.235793
8	4.328062	4.309165	− 0.67194	− 0.69084	0.309609	0.152062	1.209426
9	4.315297	4.308127	− 0.68470	− 0.69187	0.269069	0.129935	1.212388
10	4.316936	4.306374	− 0.68306	− 0.69363	0.242211	0.116650	1.185725
		$\rho = -0.50$					

3	4.493676	4.523603	− 0.50632	− 0.47640	0.598763	0.432934	1.295868
4	4.476022	4.518050	− 0.52398	− 0.48195	0.432340	0.314507	1.292820
5	4.464449	4.516154	− 0.53555	− 0.48385	0.351204	0.232666	1.366873
6	4.457812	4.509214	− 0.54219	− 0.49079	0.293720	0.191054	1.360624
7	4.456252	4.506790	− 0.54375	− 0.49321	0.243155	0.159306	1.338469
8	4.462810	4.507031	− 0.53719	− 0.49297	0.220437	0.140613	1.326819
9	4.452543	4.504319	− 0.54746	− 0.49568	0.192246	0.120476	1.343499
10	4.454318	4.505250	− 0.54568	− 0.49475	0.172407	0.107230	1.335700
		$\rho = -0.30$					
3	4.651730	4.714919	− 0.34827	− 0.28508	0.428698	0.374810	1.205903
4	4.637355	4.715161	− 0.36265	− 0.28484	0.310153	0.275254	1.239283
5	4.631381	4.711000	− 0.36862	− 0.28900	0.253184	0.212427	1.314638
6	4.626527	4.704871	− 0.37347	− 0.29513	0.212421	0.177098	1.331965
7	4.625525	4.705667	− 0.37448	− 0.29433	0.176219	0.147987	1.348789
8	4.632111	4.704215	− 0.36789	− 0.29579	0.159592	0.131665	1.345792
9	4.623905	4.702919	− 0.37610	− 0.29708	0.139843	0.115616	1.379731
10	4.625616	4.703871	− 0.37438	− 0.29613	0.124576	0.102139	1.394601
		$\rho = -0.10$					
3	4.852841	4.909030	− 0.14716	− 0.09097	0.342708	0.340355	1.045129
4	4.839578	4.912014	− 0.16042	− 0.08799	0.247764	0.254762	1.041892
5	4.838692	4.905691	− 0.16131	− 0.09431	0.203779	0.197508	1.113353
6	4.834924	4.902314	− 0.16508	− 0.09769	0.170774	0.167812	1.116546
7	4.833942	4.902154	− 0.16606	− 0.09785	0.142890	0.142678	1.119622
8	4.840546	4.899476	− 0.15945	− 0.10052	0.128645	0.126918	1.124417
9	4.833970	4.900888	− 0.16603	− 0.09911	0.113358	0.113174	1.145752
10	4.835435	4.902309	− 0.16457	− 0.09769	0.099980	0.099909	1.160885

Table 3.3 The Efficiency of $\hat{\mu}_{Y-W}^{ERSS}$ With Respect to $\hat{\mu}_{Y-W}^{SRS}$ for $m = 3, \ldots, 10$ with $\rho = 0.99$, 0.90, 0.70, 0.50, 0.30, 0.10 for $\mu_X = 7$ and $\mu_Y = 5$

	$\hat{\mu}_{Y-W}^{\mathrm{SRS}}$	$\hat{\mu}_{Y-W}^{\mathrm{ERSS}}$	$B\left(\hat{\mu}_{Y-W}^{\mathrm{SRS}}\right)$	$B\left(\hat{\mu}_{Y-W}^{\mathrm{ERSS}}\right)$	$V\left(\hat{\mu}_{Y-W}^{\mathrm{SRS}}\right)$	$V\left(\hat{\mu}_{Y-W}^{\mathrm{ERSS}}\right)$	**Eff**
m		$\rho = 0.99$					
3	5.064542	5.023006	0.064542	0.023006	1.344709	0.696868	1.934156
4	5.024869	5.026544	0.024869	0.026544	0.966415	0.424715	2.273129
5	5.039456	5.010530	0.039456	0.010530	0.808842	0.285859	2.833861
6	5.028388	5.012584	0.028388	0.012584	0.673635	0.207290	3.251126
7	5.029887	5.006276	0.029887	0.006276	0.558039	0.154454	3.617835
8	5.018673	5.004735	0.018673	0.004735	0.492149	0.125657	3.918674
9	5.026638	5.005307	0.026638	0.005307	0.432769	0.101960	4.250274
10	5.017080	5.002376	0.017080	0.002376	0.399693	0.084087	4.756497
		$\rho = 0.90$					
3	5.053566	5.019704	0.053566	1.97E-02	1.168882	0.632648	1.851001
4	5.018317	5.022626	0.018317	2.26E-02	0.840046	0.394705	2.126379
5	5.033236	5.009191	0.033236	9.19E-03	0.702113	0.271617	2.588195
6	5.025186	5.011302	0.025186	1.13E-02	0.566357	0.203106	2.789844
7	5.017951	5.007660	0.017951	7.66E-03	0.499444	0.155984	3.202763
8	5.024302	5.000972	0.024302	9.72E-04	0.432548	0.124620	3.475639
9	5.019415	5.000088	0.019415	8.84E-05	0.381656	0.103259	3.699745
10	5.017534	4.998483	0.017534	-1.52E-03	0.333697	0.085883	3.888983
		$\rho = 0.70$					
3	5.036578	5.013590	0.036578	0.013590	0.841263	0.514557	1.636941
4	5.009103	5.018824	0.009103	0.018824	0.604553	0.338193	1.785971
5	5.022311	5.005556	0.022311	0.005556	0.505071	0.245996	2.054932
6	5.016380	5.008073	0.016380	0.008073	0.408398	0.191313	2.135388
7	5.011105	5.006781	0.011105	0.006781	0.359281	0.150788	2.382789
8	5.017498	4.997649	0.017498	-0.00235	0.311767	0.125316	2.490188
9	5.013195	4.998965	0.013195	-0.00104	0.275313	0.106153	2.595157
10	5.012198	4.998532	0.012198	-0.00147	0.240381	0.090830	2.648076
		$\rho = 0.50$					

3	5.025174	5.008099	0.025174	0.008099	0.59504	0.426069	1.397855
4	5.003785	5.016725	0.003785	0.016725	0.427618	0.294637	1.450009
5	5.014516	5.000456	0.014516	0.000456	0.357031	0.221846	1.610311
6	5.010058	5.010139	0.010058	0.010139	0.290181	0.180357	1.608568
7	5.006417	5.002605	0.006417	0.002605	0.253893	0.147404	1.722631
8	5.012899	4.997767	0.012899	-0.00223	0.221233	0.125188	1.768472
9	5.008673	5.000406	0.008673	0.000406	0.195614	0.109881	1.780909
10	5.008476	4.999094	0.008476	-0.00091	0.170487	0.096080	1.775154
ρ = 0.30							
3	5.019325	5.006501	0.019325	6.50E-03	0.430747	0.367577	1.172737
4	5.002324	5.011326	0.002324	1.13E-02	0.309795	0.266298	1.162802
5	5.009886	5.003053	0.009886	3.05E-03	0.258169	0.207576	1.244145
6	5.006273	5.006430	0.006273	6.43E-03	0.211744	0.171513	1.234496
7	5.003906	5.001069	0.003906	1.07E-03	0.183291	0.142399	1.287257
8	5.010480	5.001502	0.010480	1.50E-03	0.160988	0.125434	1.284301
9	5.005880	4.997816	0.005880	-2.18E-03	0.142509	0.109651	1.299913
10	5.006371	5.000023	0.006371	2.25E-05	0.124112	0.098722	1.257592
ρ = 0.10							
3	5.019020	5.004568	0.019020	4.57E-03	0.349116	0.335439	1.041787
4	5.004694	5.010469	0.004694	1.05E-02	0.251658	0.249393	1.008725
5	5.008439	5.005679	0.008439	5.68E-03	0.208690	0.202758	1.029443
6	5.005055	5.003029	0.005055	3.03E-03	0.173256	0.169076	1.024818
7	5.003578	5.003690	0.003578	3.69E-03	0.147485	0.142915	1.031969
8	5.010230	4.998800	0.010230	-1.20E-03	0.131139	0.124894	1.050833
9	5.004832	4.999874	0.004832	-1.26E-04	0.116018	0.110586	1.049330
10	5.005885	4.999915	0.005885	-8.46E-05	0.101383	0.098591	1.028668

Table 3.4 The Efficiency of $\hat{\mu}_{Y-W}^{ERSS}$ With Respect to $\hat{\mu}_{Y-W}^{SRS}$ for $m = 3, \ldots, 10$ With $\rho = -0.99$, -0.90, -0.70, -0.50, -0.30, -0.10 for $\mu_X = 7$ and $\mu_Y = 5$

	$\hat{\mu}_{Y-W}^{\mathrm{SRS}}$	$\hat{\mu}_{Y-W}^{\mathrm{ERSS}}$	$B\left(\hat{\mu}_{Y-W}^{\mathrm{SRS}}\right)$	$B\left(\hat{\mu}_{Y-W}^{\mathrm{ERSS}}\right)$	$V\left(\hat{\mu}_{Y-W}^{\mathrm{SRS}}\right)$	$V\left(\hat{\mu}_{Y-W}^{\mathrm{ERSS}}\right)$	Eff
m		$\rho = -0.99$					
3	5.114716	5.054574	0.114716	0.054574	1.407696	0.720631	1.963567
4	5.083137	5.037643	0.083137	0.037643	1.011697	0.431141	2.354849
5	5.056966	5.026043	0.056966	0.026043	0.819953	0.294125	2.792366
6	5.044314	5.014090	0.044314	0.014090	0.681727	0.208720	3.272524
7	5.040036	5.011940	0.040036	0.011940	0.564998	0.159215	3.555529
8	5.047179	5.009249	0.047179	0.009249	0.511658	0.125837	4.080950
9	5.030384	5.008221	0.030384	0.008221	0.443865	0.102986	4.316081
10	5.031947	5.007732	0.031947	0.007732	0.400697	0.083324	4.817687
		$\rho = -0.90$					
3	5.100567	5.047233	0.100567	0.047233	1.220611	0.652837	1.878773
4	5.072716	5.033009	0.072716	0.033009	0.878882	0.398956	2.210169
5	5.049334	5.022204	0.049334	0.022204	0.712290	0.278430	2.562442
6	5.038059	5.012769	0.038059	0.012769	0.592516	0.201121	2.950877
7	5.034629	5.009781	0.034629	0.009781	0.491289	0.156718	3.140585
8	5.041612	5.008053	0.041612	0.008053	0.444823	0.125824	3.547221
9	5.026208	5.006442	0.026208	0.006442	0.386184	0.104558	3.698582
10	5.027924	5.005783	0.027924	0.005783	0.348400	0.086260	4.046441
		$\rho = -0.70$					
3	5.060366	5.032169	0.060366	0.032169	0.865062	0.523030	1.657630
4	5.052285	5.025263	0.052285	0.025263	0.640934	0.347427	1.849272
5	5.043939	5.008513	0.043939	0.008513	0.504409	0.248009	2.041026
6	5.036839	5.009845	0.036839	0.009845	0.417407	0.189365	2.210276
7	5.024193	5.010397	0.024193	0.010397	0.361208	0.153748	2.351504
8	5.023876	5.005146	0.023876	0.005146	0.312096	0.124601	2.508801
9	5.020344	5.006023	0.020344	0.006023	0.276069	0.107627	2.568041
10	5.024908	5.003080	0.024908	0.003080	0.251790	0.093096	2.711020
		$\rho = -0.50$					

3	5.048287	5.028532	0.048287	0.028532	0.608419	0.437421	1.393659
4	5.035241	5.018845	0.035241	0.018845	0.445889	0.299598	1.490667
5	5.033372	5.008671	0.033372	0.008671	0.364473	0.222659	1.641360
6	5.014972	5.009753	0.014972	0.009753	0.294769	0.183795	1.604181
7	5.024169	5.005677	0.024169	0.005677	0.251048	0.144800	1.737406
8	5.017391	5.002627	0.017391	0.002627	0.218011	0.125026	1.746050
9	5.016349	4.998879	0.016349	-0.00112	0.198993	0.108786	1.831652
10	5.015232	5.000920	0.015232	0.000920	0.177622	0.096384	1.845238
$\rho = -\mathbf{0.30}$							
3	5.023274	5.020530	0.023274	0.020530	0.429828	0.372432	1.154261
4	5.021981	5.007382	0.021981	0.007382	0.324528	0.266393	1.219792
5	5.018594	5.008728	0.018594	0.008728	0.263492	0.211454	1.247282
6	5.012574	5.002197	0.012574	0.002197	0.214442	0.172163	1.246457
7	5.006930	5.004702	0.006930	0.004702	0.181800	0.145061	1.253402
8	5.015895	5.000963	0.015895	0.000963	0.160197	0.124147	1.292412
9	5.011022	5.003629	0.011022	0.003629	0.142788	0.112370	1.271623
10	5.010192	4.999806	0.010192	-0.00019	0.127686	0.097507	1.310574
$\rho = -\mathbf{0.10}$							
3	5.021547	5.010563	0.021547	0.010563	0.346315	0.340744	1.017378
4	5.020529	5.005282	0.020529	0.005282	0.262093	0.248310	1.057084
5	5.012725	4.999262	0.012725	-0.00074	0.209457	0.202046	1.037478
6	5.009023	5.002840	0.009023	0.002840	0.178853	0.166397	1.075290
7	5.007271	5.001898	0.007271	0.001898	0.150249	0.143952	1.044080
8	5.008075	5.002471	0.008075	0.002471	0.129278	0.126399	1.023243
9	5.008477	4.999239	0.008477	-0.00076	0.114987	0.110568	1.040613
10	5.004838	5.002679	0.004838	0.002679	0.104981	0.099579	1.054405

3.5 CONCLUSION

In this chapter, various ratio-cum-product estimators of the population mean of the study variable are suggested using SRS and ERSS methods based on information on a single concomitant variable. Expressions of the mean squared errors of the proposed estimators are derived. Based on theoretical and simulation comparisons, it is noted that the suggested estimators using ERSS are always better than their competitors using SRS for all cases considered in this study.

REFERENCES

Al-Omari, A.I., Jemain, A.A., Ibrahim, K., 2009. New ratio estimators of the mean using simple random sampling and ranked set sampling methods. Rev. Investig. Oper. 30 (2), 97–108.

Cochran, W.G., 1977. Sampling Techniques, 3rd ed. Wiley and Sons, New York.

Haq, A., Shabbir, J., 2010. A family of ratio estimators for population mean in extreme ranked set sampling using two auxiliary variables. Sort 34 (1), 45–64.

Haq, A., Shabbir, J., 2013. Improved family of ratio estimators in simple and stratified random sampling. Commun. Stat.: Theory Methods 42 (5), 782–799.

Jemain, A.A., Al-Omari, A.I., Ibrahim, K., 2007. Multistage extreme ranked set samples for estimating the population mean. J. Stat. Theory Appl. 6 (4), 456–471.

Jemain, A.A., Al-Omari, A.I., Ibrahim, K., 2008. Modified ratio estimator for the population mean using double median ranked set sampling. Pak. J. Stat. 24 (3), 217–226.

Kadilar, C., Cingi, H., 2004. Ratio estimators in simple random sampling. Appl. Math. Comput. 151, 893–902.

McIntyre, G.A., 1952. A method for unbiased selective sampling using ranked sets. Aust. J. Agric. Res. 3, 385–390.

Samawi, H.M., Ahmed, M.S., Abu-Dayyeh, W., 1996. Estimating the population mean using extreme ranked set sampling. Biom. J. 38, 577–586.

Singh, H.P., Espejo, M.R., 2003. On linear regression and ratio-product estimation of a finite population mean. Statistician 52 (1), 59–67.

Singh, H.P., Tailor, R., 2003. Use of known correlation coefficient in estimating the finite population mean. Stat. Transit. 6 (4), 555–560.

Takahasi, K., Wakimoto, K., 1968. On unbiased estimates of the population mean based on the sample stratified by means of ordering. Ann. Inst. Stat. Math. 20, 1–31.

CHAPTER

4 ESTIMATION OF THE DISTRIBUTION FUNCTION USING MOVING EXTREME RANKED SET SAMPLING (MERSS)

Mohammad Fraiwan Al-Saleh, and Dana Majed Rizi Ahmad
Department of Statistics, Yarmouk University, Irbid, Jordan

4.1 INTRODUCTION

Statistics is the science of collecting, organizing, analyzing, and making inference about a population using information in a sample taken from it. There are several sampling techniques that can be used to choose a suitable sample. *Simple random sampling* (SRS) is the basic sampling technique. Using this method, a sample of size n is selected from a population of size N, such that all groups of n elements in the population are equally likely to be selected. SRS is used when each subgroup within the population needs to be represented in the chosen sample, the population is divided into nonoverlapping groups; each group is called a stratum, a random sample is taken from each stratum. The ideal situation occurs when strata are very similar within and very different among. In *cluster random sampling*, the population consists of groups of elements called clusters; a cluster is preferred to be as heterogeneous as possible. We choose a random sample of clusters. The ideal situation occurs when clusters are very similar among and very different within. *Systematic random sampling*: In this method, a starting point is chosen from the first k elements in the frame, and then every kth element thereafter is included in the sample; usually, $k = [N/n]$. (For more information about these techniques see "Elementary Survey Sampling" by Scheaffer et al., 1986.)

McIntyre (1952) suggested a new sampling technique, which was called *ranked set sampling* (RSS), to estimate more effectively yields of pastures. The technique can be executed as follows:

1. Draw randomly m sets of size m each from the population of interest;
2. Rank the units within each set by judgment, with respect to the variable of interest from smallest to largest. It is assumed here that each element can be ranked by eyes or by a relatively cheap method;
3. From the ith set, take for actual quantification the element ranked by *judgment* as the ith order statistic, $i = 1, 2, \ldots, m$.

   Steps 1–3 give an RSS of size m.
4. The above procedure can be repeated, if necessary, r times to get an RSS of size $n = mr$.

Ranked Set Sampling. DOI: https://doi.org/10.1016/B978-0-12-815044-3.00004-6

Let $\{Y^j_{(i:m)},\ i=1,\ldots,\ m,\ j=1,\ldots,r\}$ be the set of RSS elements, where *under perfect judgment ranking*, $Y^j_{(i:m)}$ is the ith order statistic for a random sample of size m at the jth cycle. Note that for each i, $Y^1_{(i:m)}, Y^2_{(i:m)},\ldots, Y^r_{(i:m)}$ are independent and identically distributed, $f_{(i:m)}$, while for each j, $Y^j_{(1:m)}, Y^j_{(2:m)},\ldots, Y^j_{(m:m)}$, are only independent. $f_{(i:m)}$ is the pdf of the ith order statistic of a random sample of size m.

RSS is applicable whenever a ranking mechanism can be found such that the ranking of sampling units is carried out easily and sampling is much cheaper than the measurement of the variable of interest. McIntyre (1952) mentioned, without mathematical proof, that the mean of quantified elements is an unbiased estimator of the population mean regardless of any error in judgment ranking. With perfect ranking for typical unimodal distributions, the mean of such a sample is nearly $(m+1)/2$ times as efficient as the mean of an SRS of the same size; this upper bound is achieved when the underlying distribution is the uniform.

Takahasi and Wakimoto (1968) established the theory of RSS. Let the population density function be $f(x)$ with mean μ and variance σ^2. Let the $(i{:}m)th$ order statistic from the population have the density function $f_{(i:m)}(x)$ with mean $\mu_{(i:m)}$, and variance $\sigma^2_{(i:m)}$, then from Takahasi and Wakimoto (1968) the basic identity is:

$$f(x)=\frac{1}{m}\sum_{i=1}^{m} f_{(i:m)}(x).$$

Based on this identity, they showed that

$$\mu=\frac{1}{m}\sum_{i=1}^{m}\mu_{(i:m)}\ \&\ \sigma^2=\frac{1}{m}\sum_{i=1}^{m}\sigma^2_{(i:m)}+\frac{1}{m}\sum_{i=1}^{m}(\mu_{(i:m)}-\mu)^2.$$

Let $\hat{\mu}_{\text{RSS}}$ be the RSS mean and $\hat{\mu}_{\text{SRS}}$ be the mean of an SRS of the same size. Takahasi and Wakimoto (1968) compared the performance of the estimators using the efficiency of $\hat{\mu}_{\text{RSS}}$ with respect to $\hat{\mu}_{\text{SRS}}$:

$$\text{Eff}(\hat{\mu}_{\text{RSS}};\hat{\mu}_{\text{SRS}})=\frac{\text{Var}(\hat{\mu}_{\text{SRS}})}{\text{Var}(\hat{\mu}_{\text{RSS}})}.$$

They showed that

$$1\le \text{Eff}(\hat{\mu}_{\text{RSS}};\hat{\mu}_{\text{SRS}})\le\frac{m+1}{2}.$$

The lower bound is attained if and only if the underlying distribution is degenerate, while the upper bound is attained if and only if the underlying distribution is rectangular (uniform).

Stokes and Sager (1988) used RSS to estimate the distribution function $F(t)$. They showed that the empirical distribution function-based RSS, $\hat{F}_{\text{RSS}}(t)$, is unbiased for $F(t)$ and more efficient than the empirical distribution function based on an SRS of size n. The empirical distribution function using RSS is given by:

$$\hat{F}_{\text{RSS}}(t)=\frac{1}{rm}\sum_{j=1}^{r}\sum_{i=1}^{m} I(X^{(j)}_{(i:m)}\le t),$$

where, $X^{(j)}_{(i:m)}$ is the ith order statistics for a random sample of size m at the jth cycle.

For more work on the RSS technique see Kaur et al. (1995) “Ranked Set Sampling: An Annotated Bibliography,” Al-Saleh and Al-Kadiri (2000), Al-Saleh & Al-Omari (2002), Abu-Dayyeh et al. (2002), Al-Saleh and Zheng (2002), Al-Saleh and Samawi (2000), Ozturk and Wolfe (2000), Ozturk and Wolfe (2000), (2001), Ozturk (2002), Al-Saleh and Ababneh (2015), Zheng and Al-Saleh (2002), Al-Saleh and Darabseh (2017).

Moving extreme ranked set sampling (MERSS) is a variation of RSS introduced by Al-Odat and Al-Saleh (2001). The MERSS technique can be described as follows:

1. Select m SRSs of size $1, 2, \ldots, m$, respectively;
2. Order the elements by tudgment, without actual measurement of the characteristic of interest;
3. Measure accurately the maximum ordered observation from the first set, and the maximum ordered observation from the second set. The process continues in this way until the maximum ordered observation from the last mth sample is measured;
4. Steps 1−3 may be repeated if necessary on m samples of size $1, 2, \ldots, m$, respectively, but here the minimum ordered observation is measured instead of the maximum ordered observations.
5. The entire cycle can be repeated, if necessary, many times to obtain a sample of larger size.

Al-Odat and Al-Saleh (2001) investigated this method nonparametrically and concluded that the estimator of the population mean is more efficient than that of SRS in the case of symmetric populations. Al-Saleh and Al-Hadhrami (2003a,b) studied the method in more detail. They insisted that MERSS allows for an increase in set size without introducing too much ranking error. They concluded that the MLE of the mean of the exponential distribution based on MERSS is more efficient than the MLE based on SRS. Also, the information contained in MERSS, measured by Fisher information number, is always greater than that of the SRS with the same size.

Al-Saleh and Al-Ananbeh (2005) considered the estimation of correlation coefficient in the bivariate normal distribution based on a modification of the MERSS using a concomitant random variable. Al-Saleh and Samawi (2010) considered the estimation of the odds using MERSS. Samawi and Al-Saleh (2013) considered the estimation of odds ratio using MERSS.

Al-Saleh and Ababneh (2015) considered testing for perfect ranking in MERSS. Hanandeh (2011) considered the estimation of the parameters of Downton’s bivariate exponential distribution using MERSS. Al-Saleh and Naamneh (2016) studied the performance of “The Five-Number Summary” obtained using different sampling techniques.

There are other techniques for utilizing some available variables that are easy to measure and that have a strong relation with the main variable. One popular technique is the ratio estimation technique. In this method, it is assumed that the obtained sample is $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$; X is the variable of interest and Y is the auxiliary variable. It is assumed that $E(Y) = \mu_y$ is known, and $E(X) = \mu_x$ is estimated from $\hat{\mu}_x = \overline{XY}\mu_x$. This ratio estimator can be compared to other estimators of μ_x. For recent work on ratio estimation, see Subzar et al. (2016, 2017a, 2017b, 2017c) and Sharma et al. (2016).

In this chapter, the use of the MERSS technique to estimate the cumulative distribution function $F(x)$ is investigated. The suggested estimators are compared with the corresponding estimators based on RSS and SRS.

4.1.1 ESTIMATION OF DISTRIBUTION FUNCTION USING METHOD OF MOMENTS

Let $X_1, X_2, \ldots, X_n, n = mr$ be a simple random sample (SRS) of size n, with common absolutely continuous unknown cdf $F(t)$. For a given t, the well-known SRS estimator of $F(t)$ is the empirical distribution function given by:

$$\hat{F}_{\mathrm{SRS}}(t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le t),$$

where, $Y_i = I(X_i \le t) = \begin{cases} 1 & \mathit{if}\ X_i \le t \\ 0 & \mathit{if}\ X_i > t. \end{cases}$.

Clearly, $\hat{F}_{\mathrm{SRS}}(t)$ is the method of moments estimator (MME) and the maximum likelihood estimator (MLE) of $F(t)$. It a suitable estimator when there is no available information about F. Now,

$$E(I(X_i \le t)) = p(X_i \le t) = F(t),\ Var(I(X_i \le t)) = F(t)(1 - F(t)).$$

Thus$E(\hat{F}_{\mathrm{SRS}}(t)) = F(t)$, i.e., $\hat{F}_{\mathrm{SRS}}(t)$ is an unbiased estimator of $F(t)$ Also,

$$\mathrm{Var}(\hat{F}_{\mathrm{SRS}}(t)) = \frac{F(t)(1 - F(t))}{n}.$$

Let $X^j_{(1:m)}, X^j_{(2:m)}, \ldots, X^j_{(m:m)}$, for $j = 1, \ldots, r$, be an RSS of size $n = rm$ (set size m and r cycles) from $F(t)$. Stokes and Sager (1988) suggested the following estimator for $F(t)$:

$$\hat{F}_{\mathrm{RSS}}(t) = \frac{1}{rm}\sum_{j=1}^{r}\sum_{i=1}^{m} I(X^{(j)}_{(i:m)} \le t).$$

They showed that $\hat{F}_{\mathrm{RSS}}(t)$ *is* an unbiased estimator for $F(t)$ and is more efficient than $\hat{F}_{\mathrm{SRS}}(t)$. The variance of $\hat{F}_{\mathrm{RSS}}(t)$ is given by

$$\mathrm{Var}(\hat{F}_{\mathrm{RSS}}(t)) = \frac{1}{r^2m^2}\sum_{j=1}^{r}\sum_{i=1}^{m} F_{(i:m)}(t)(1 - F_{(i:m)}(t)) = \frac{\sum_{i=1}^{m} F_{(i:m)}(t)(1 - F_{(i:m)}(t))}{rm^2}$$

The efficiency, for $m = 2, 3, 4, 5$, of $\hat{F}_{\mathrm{RSS}}(t)$ relative to $\hat{F}_{\mathrm{SRS}}(t)$ is given in Table 4.1.

Based on Table 4.1, it can be seen that $\hat{F}_{\mathrm{RSS}}(t)$ is more efficient than $\hat{F}_{\mathrm{SRS}}(t)$. The efficiency is increasing in m for fixed $F(t)$. For very large or very small $F(t)$ (with small m),

Table 4.1 $\boldsymbol{Eff(\hat{F}_{RSS}(t); \hat{F}_{SRS}(t))}$ **for Some Values of** $\boldsymbol{m}$ **and** $\boldsymbol{F(t)}$

$F(t)$ m	0.05	0.2	0.4	0.5	0.7	0.9	0.95
2	1.05	1.19	1.31	1.33	1.26	1.10	1.05
3	1.10	1.37	1.56	1.60	1.50	1.20	1.10
4	1.15	1.41	1.79	1.83	1.71	1.29	1.15
5	1.20	1.68	2.00	2.03	1.73	1.41	1.20

though the efficiency >1, the improvement is not very significant. For fixed m, the efficiency is increasing in $F(t)$ for $F(t) \leq 0.5$ and decreasing for $F(t) > 0.5$. The best values are for $F(t) = 0.5$.

Let $\{X_{(i:i)}^{(j)}{:}i = 1, \ldots, m, j = 1, \ldots, r\}$ be an MERSS based on a distribution function, $F(t)$. Note that: For fixed i, $X_{(i:i)}^{(1)}, \ldots, X_{(i:i)}^{(r)}$ are *iid* with common distribution $F^i(t)$; while for fixed j, $X_{(1:1)}^{(j)}, \ldots, X_{(m:m)}^{(j)}$ are only independent.

Let $I(X_{(i:i)}^{(j)} \leq t) = Y_i^{(j)}, i = 1, 2, \ldots, m, j = 1, 2, \ldots, r$, then $Y_i^{(j)}$ is $Ber(1, F^i(t))$.

A modified MME (MMME) can be obtained using binomial theorem:

$$\sum_{k=0}^{m} \binom{m}{k} F^{m-k}(t) = (1 + F(t))^m. \text{ (Binomial Theorem)}$$

Thus

$$F(t) = \sqrt[m]{\sum_{k=0}^{m} \binom{m}{k} F^{m-k}(t) - 1}.$$

Therefore the MMME is

$$\hat{F}_{\text{MMME}}(t) = \sqrt[m]{\sum_{k=0}^{m} \binom{m}{k} \overline{Y}_{m-k}} - 1. \tag{4.1}$$

Now,

$$E(\hat{F}_{\text{MMME}}(t)) = \sum_{y_m=0}^{r} \sum_{y_{m-1}=0}^{r} \cdots \sum_{y_2=0}^{r} \sum_{y_1=0}^{r} \left(\sqrt[m]{\sum_{k=0}^{m} \binom{m}{k} \overline{Y}_{m-k}} - 1 \right) \left(\prod_{i=1}^{m} \binom{r}{y_i} (F^i(t))^{y_i} (1 - F^i(t))^{r-y_i} \right)$$

The efficiency of $\hat{F}_{\text{MMME}}(t)$ w.r.t. $\hat{F}_{\text{RSS}}(t)$ is

$$\text{Eff}(\hat{F}_{\text{MMME}}(t), \hat{F}_{\text{SRS}}(t)) = \frac{\text{MSE}(\hat{F}_{\text{SRS}}(t))}{\text{MSE}(\hat{F}_{\text{MMME}}(t))},$$

where

$$\text{MSE}(\hat{F}_{\text{MMME}}(t)) = E\,(F(t) - \hat{F}_{\text{MMME}}(t))^2$$

$\hat{F}_{\text{MMME}}(t)$ and $\hat{F}_{\text{SRS}}(t)$ are compared, for $m = 3, 4, 5$. The numerical results, obtained using a scientific workplace package, are given in Tables 4.2–4.4.

Table 4.2 $\boldsymbol{Eff(\hat{F}_{MMME}(t), \hat{F}_{SRS}(t)), r = 3}$

$F(t)$ m	0.20	0.40	0.50	0.70	0.75	0.90	0.99
3	0.58	0.71	0.82	1.07	1.13	1.31	1.46
4	0.48	0.62	0.71	1.06	1.17	1.49	1.67
5	0.41	0.52	0.65	1.03	1.17	1.59	1.92

Table 4.3 $Eff(\hat{F}_{MMME}(t), \hat{F}_{SRS}(t))$, $r = 5$

m \ $F(t)$	0.20	0.40	0.50	0.70	0.75	0.90	0.99
3	0.57	0.75	0.86	1.11	1.19	1.38	1.49
4	0.47	0.66	0.81	1.13	1.26	1.55	1.72
5	0.48	0.72	0.91	1.38	1.22	2.03	1.91

Table 4.4 $Eff(\hat{F}_{MMME}(t), \hat{F}_{SRS}(t))$, $r = 10$

m \ $F(t)$	0.20	0.40	0.50	0.70	0.75	0.90	0.99
3	0.55	0.77	0.89	1.16	1.23	1.45	1.58
4	0.45	0.68	0.86	1.19	1.30	1.59	1.76
5	0.38	0.63	0.79	1.22	1.29	1.71	1.93

Based on these tables we conclude that:

- For fixed $F(t)$ and m, the efficiency is increasing in r for large and moderate values of $F(t)$ and is decreasing for small values of $F(t)$.
- For fixed values of r and m, the efficiency is increasing in $F(t)$.
- For fixed values of $F(t)$ and r, the efficiency is increasing in m for large values of $F(t)$ and decreasing for small and moderate values of $F(t)$.

In general, it can be seen that for large values of $F(t)$ the efficiency tends to be larger than 1 and is increasing in m. But for small to moderate values of $F(t)$, the efficiency is less than 1.

4.1.2 ESTIMATION OF DISTRIBUTION FUNCTION USING MAXIMUM LIKELIHOOD ESTIMATOR

In this subsection, we consider the estimation of $F(t)$ using the maximum likelihood estimator (MLE) based on all elements of MERSS. Let

$$Y_i = \sum_{j=1}^{r} Y_i^{(j)}, \ i = 1, 2, \ldots, m$$

Then Y_i is $bin(r, F^i(t))$. The likelihood function is

$$L(F(t)|y_i) = \prod_{i=1}^{m} \binom{r}{y_i} (F^i(t))^{y_i} (1 - F^i(t))^{r - y_i}.$$

Therefore

$$\begin{aligned} L* &= (\ln(L(F(t)|y_i)) = \sum_{i=1}^{m} \ln[F^{iy_i}(t)(1-F^i(t))^{r-y_i}] \\ &= \sum_{i=1}^{m}[iy_i \ln F(t) + (r-y_i)\ln(1-F^i(t))]. \\ \frac{\partial L*}{\partial F(t)} &= \sum_{i=1}^{m} \frac{iy_i}{F(t)} - \frac{i(r-y_i)F^{i-1}(t)}{(1-F^i(t))} = \sum_{i=1}^{m} \frac{iy_i - irF^i(t)}{F(t)(1-F^i(t))}.(*) \end{aligned}$$

Note that:

$\frac{\partial L^*}{\partial F(t)}$ is positive when $F(t) \to 0$ and negative when $F(t) \to 1$. Therefore $\frac{\partial L^*}{\partial F(t)} = 0$ has a root and the root maximizes L. Thus the root of

$$\sum_{i=1}^{m} \frac{iy_i - irF^i(t)}{F(t)(1-F^i(t))} = 0$$

is the MLE of $F(t)$.

Note: If $Y_i = r$ for all i then $\hat{F}_{\text{MLE}}(t) = 1$, while, $Y_i = 0$ for all i then $\hat{F}_{\text{MLE}}(t) = 0$.

Special cases:

- For $m = 1$, setting equation (*) to zero, we get

$$\hat{F}_{\text{MLE1}}(t) = \frac{Y_1}{r} = \overline{Y}_1.$$

- For $m = 2$, setting equation (*) to zero, we get

$$\frac{y_1 + 2y_2}{F(t)} = \frac{r - y_1}{1 - F(t)} + \frac{2(r - y_2)}{1 - F^2(t)}$$

$$\begin{aligned} &\Rightarrow (y_1 + 2y_2)(1 - F^2(t)) = (r - y_1)F(t)(1 + F(t)) + 2(r - y_2)F^2(t) \\ &\Rightarrow y_1 + 2y_2 - (y_1 + 2y_2)F^2(t) = (r - y_1)F(t) + (r - y_1)F^2(t) + 2rF^2(t) - 2y_2F^2(t) \\ &\Rightarrow y_1 + 2y_2 = (3r)F^2(t) + (r - y_1)F(t) \\ &\Rightarrow (3r)F^2(t) + (r - y_1)F(t) - (y_1 + 2y_2) = 0. \end{aligned}$$

Thus

$$\hat{F}_{\text{MLE2}}(t) = \frac{y_1 - r + \sqrt{r^2 + 10ry_1 + 24ry_2 + y_1^2}}{6r}.$$

Numerical evaluation is needed for large m. The bias of $\hat{F}_{MLE2}(t)$ and its efficiency w.r.t. $\hat{F}_{\text{SRS}}(t)$ when $m = 2$ for $r = 3.5$ and 10, were obtained using a scientific workplace package and are given in Table 4.5.

It can be seen that the efficiency of $\hat{F}_{\text{MLE2}}(t)$ w.r.t. $\hat{F}_{\text{SRS}}(t)$ is increasing in $F(t)$ for fixed m. In general, it can be seen that for large and moderate values of $F(t)$ the efficiency tends to be larger than 1 and increasing in r. For small values of $F(t)$ the efficiency is less than 1 and is decreasing in r. $\hat{F}_{\text{MLE2}}(t)$ is a negatively biased estimator.

For $m = 3$, setting equation (*) to zero, we get

$$\frac{y_1 + 2y_2 + 3y_3}{F(t)} = \frac{(r - y_1)}{1 - F(t)} + \frac{2(r - y_2)F(t)}{1 - F^2(t)} + \frac{3(r - y_3)F^2(t)}{1 - F^3(t)}$$

Table 4.5 The Bias of $\hat{F}_{MLE2}(t)$ and the Efficiency of $\hat{F}_{MLE2}(t)$ w.r.t. $\hat{F}_{SRS}(t)$ for $m = 2$

	Bias			Efficiency		
$F(t)$	$r = 3$	$r = 5$	$r = 10$	$r = 3$	$r = 5$	$r = 10$
0.05	−0.0120	−0.0088	−0.0053	0.91	0.79	0.69
0.20	−0.0296	-0.0191	−0.0098	0.86	0.82	0.81
0.40	−0.0287	−0.0165	−0.0079	0.94	0.98	1.02
0.50	−0.0241	−0.0136	−0.0064	1.01	1.07	1.12
0.70	−0.0138	−0.0077	−0.0037	1.17	1.23	1.28
0.75	−0.0113	−0.0063	−0.0030	1.21	1.27	1.32
0.90	−0.0043	−0.0024	−0.0012	1.31	1.37	1.41
0.99	−0.0004	−0.0002	−0.0001	1.37	1.42	1.46

Table 4.6 The Bias of $\hat{F}_{MLE3}(t)$ and the Efficiency of $\hat{F}_{MLE3}(t)$ w.r.t. $\hat{F}_{SRS}(t)$ for $m = 3$

	Bias			Efficiency		
$F(t)$	$r = 3$	$r = 5$	$r = 10$	$r = 3$	$r = 5$	$r = 10$
0.05	−0.0141	−0.0096	−0.0061	0.71	0.59	0.50
0.20	−0.0399	−0.0126	−0.0136	0.73	0.64	0.62
0.40	−0.0370	−0.0199	−0.0096	0.80	0.87	0.93
0.50	−0.0332	−0.0174	−0.0030	0.92	1.02	1.12
0.70	−0.0133	−0.0099	−0.0035	1.32	1.38	1.57
0.75	−0.0103	−0.0066	−0.0010	1.41	1.52	1.58
0.90	−0.0038	−0.0029	−0.0014	1.71	1.77	1.81
0.99	−0.0015	−0.0009	−0.0007	1.78	1.97	2.06

$$\Rightarrow y_1 + 2y_2 + 3y_3 = (r - 2y_1 - 2y_2 - 3y_3)F(t) + (4r - 2y_1 - 2y_2 - 3y_3)F^2(t) + (7r - y_1)F^3(t) + (6r)F^4(t)$$

$$\begin{aligned}\Rightarrow (6r)F^4(t) + (7r - y_1)F^3(t) + (4r - 2y_1 - 2y_2)F^2(t)\\ + (r - 2y_1 - 2y_2 - 3y_3)F(t) - y_1 - 2y_2 - 3y_3 = 0\end{aligned}$$

This equation can be solved numerically. We used Minitab Programming to obtain numerical values for the bias and efficiency.

Table 4.6 gives the bias of $\hat{F}_{\text{MLE3}}(t)$ and its efficiency w.r.t. $\hat{F}_{\text{SRS}}(t)$ when $m = 3$ for $r = 3$, 5, and 10.

Table 4.6 shows that the efficiency of $\hat{F}_{\text{MLE3}}(t)$ w.r.t. $\hat{F}_{\text{SRS}}(t)$ is increasing in $F(t)$.

- For fixed values of $F(t)$, the efficiency is increasing in r for large and moderate values of $F(t)$ and decreasing for small values of $F(t)$. $\hat{F}_{\text{MLE3}}(t)$ is more efficient than $\hat{F}_{\text{SRS}}(t)$ for large and moderate values of $F(t)$.

- For fixed r, $\hat{F}_{\mathrm{MLE3}}(t)$ is more efficient than $\hat{F}_{\mathrm{MLE2}}(t)$ when $F(t) \geq 0.7$, while $\hat{F}_{\mathrm{MLE2}}(t)$ is more efficient then $\hat{F}_{\mathrm{MLE3}}(t)$ for $F(t) < 0.7$.
- $\hat{F}_{\mathrm{MLE3}}(t)$ is a negatively biased estimator.

4.1.3 FISHER INFORMATION NUMBER IN MERSS ABOUT *F*(*T*)

For SRS, the Fisher information number about $F(t)$ is

$$I_{\mathrm{SRS}} = \frac{mr}{F(t)(1 - F(t))}.$$

The Fisher information number about $F(t)$ in MERSS, I_{MERSS} is obtained from the following second derivative:

$$\begin{aligned}
\frac{\partial^2 L^*}{\partial F^2(t)} &= \sum_{i=1}^{m} \frac{[(-i^2 rF^{i-1}(t))F(t)(1 - F^i(t))] - [(iy_i - irF^i(t))(-iF^i(t) + (1 - F^i(t))]}{[F(t)(1 - F^i(t))]^2} \\
&= \sum_{i=1}^{m} \frac{-i^2 rF^i(t) + i^2 rF^{2i}(t) + i^2 y_i F^i(t) - i^2 rF^{2i}(t) - iy_i + iy_i F^i(t) + irF^i(t) - irF^{2i}(t)}{[F(t)(1 - F^i(t))]^2}
\end{aligned}$$

The Fisher information number is:

$$\begin{aligned}
&-E\left[\sum_{i=1}^{m} \frac{-i^2 rF^i(t) + i^2 Y_i F^i(t) - iY_i + iY_i F^i(t) + irF^i(t) - irF^{2i}(t)}{[F(t)(1 - F^i(t))]^2}\right] \\
&= \sum_{i=1}^{m} \frac{i^2 rF^i(t) - i^2 rF^{2i} + irF^i(t) - irF^{2i}(t) - irF^i(t) + irF^{2i}(t)}{[F(t)(1 - F^i(t))]^2} \\
&= \sum_{i=1}^{m} \frac{i^2 rF^i(t)(1 - F^i(t))}{F^2(t)(1 - F^i(t))^2}.
\end{aligned}$$

Thus the MERSS Fisher information number about $F(t)$ is

$$I_{\mathrm{MERSS}}(m, r) = \sum_{i=1}^{m} \frac{i^2 rF^{i-2}(t)}{(1 - F^i(t))},$$

and the corresponding one using SRS is

$$I_{\mathrm{SRS}} = I_{\mathrm{MERSS}}(1, mr) = \frac{mr}{F(t)(1 - F(t))}.$$

Table 4.7 gives the Fisher relative efficiency of $F_{\mathrm{MERSS}}(t)$ w.r.t. $F_{\mathrm{SRS}}(t)$ when $m = 2, \ldots, 5$.

From Table 4.7 we can see that MERSS has more information about $F(t)$ than SRS for large and moderate values of $F(t)$. For fixed values of $F(t)$, the Fisher relative efficiency is increasing for large and moderate values of $F(t)$ and decreasing for small values of $F(t)$. For fixed m, the Fisher relative efficiency is increasing in $F(t)$.

4.2 MERSS BASED ON MINIMA

In this section, we use MERSS with minima instead of maxima to estimate $F(t)$. Let the elements of MERSS based on minima be

Table 4.7 Fisher Relative Efficiency = $\frac{I_{\text{MERSS}}}{I_{\text{SRS}}}$

F(*t*) *m*	0.05	0.20	0.40	0.50	0.70	0.75	0.90	0.99
2	0.60	0.83	1.18	1.16	1.32	1.35	1.44	1.49
3	0.40	0.65	1.02	1.21	1.55	1.63	1.86	1.99
4	0.30	0.51	0.92	1.17	1.88	1.84	2.24	2.47
5	0.24	0.42	0.82	1.10	1.79	1.99	2.60	2.96

$$\{X^{(j)}_{(1:i)}: i = 1, 2, \ldots, m, j = 1, 2, \ldots, r\}$$

Let $Y_i = \sum_{j=1}^{r} I(X^{(j)}_{(1:i)} \leq t)$. Then$Y_i \sim bin(r, (1-(1-F(t))^i)$.

Consider the following likelihood function:

$$L(F(t)|y_i) = \prod_{i=1}^{m} \binom{r}{y_i} [1-(1-F(t))^i]^{y_i} [1-(1-(1-F(t))^i]^{r-y_i}.$$

Then $L^* = \ln(L(F(t)|y_i) = \sum_{i=1}^{m} [y_i \ln(1-(1-F(t))^i + i(r-y_i)\ln(1-F(t))]$

$$\begin{aligned}\frac{\partial L^*}{\partial F(t)} &= \sum_{i=1}^{m} \left[\frac{iy_i(1-F(t))^{i-1}}{(1-(1-F(t))^i)} - \frac{i(r-y_i)}{(1-F(t))} \right] \\ &= \sum_{i=1}^{m} \left[\frac{iy_i(1-F(t))^i - i(r-y_i) + ir(1-F(t))^i - iy_i(1-F(t))^i}{(1-F(t))(1-(1-F(t))^i)} \right. \\ &= \sum_{i=1}^{m} \frac{ir(1-F(t))^i - i(r-y_i)}{(1-F(t))(1-(1-F(t))^i)}.\end{aligned}$$

Also

$$\frac{\partial L^*}{\partial F(t)} = \sum_{i=1}^{m} \frac{ir(1-F(t))^i - i(r-y_i)}{(1-(1-F(t))^i)}.$$

Let $1 - F(t) = F^*(t)$ and $r - y_i = y_i{}^*$, then $Y_i^* \sim bin(r, 1 - F^{*i}(t))$. Thus

$$\frac{\partial L^*}{\partial F(t)} = \sum_{i=1}^{m} \frac{irF^{*i}(t) - iy_i^*}{(1-F^{*i}(t))}.$$

Special cases:

- For $m = 1$, setting equation $\frac{\partial L^*}{\partial F(t)}$ to zero, we get

$$\hat{F}^*(t) = \frac{y_1^*}{r} \Rightarrow 1 - \hat{F}(t) = \frac{r - y_1}{r} = 1 - \frac{y_1}{r}$$
$$\therefore \hat{F}_{\text{MLE}^*1}(t) = \frac{Y_1}{r}.$$

- For $m = 2$, we get
- $\frac{y_1^* + 2y_2^*}{F^*(t)} = \frac{y_1}{1-F*(t)} + \frac{2y_2^* F*(t)}{1-F^{2*}(t)}$,

Table 4.8 The Bias Values of $\hat{F}_{MLE^*2}(t)$ and the Efficiency of $\hat{F}_{MLE^*2}(t)$ w.r.t. $\hat{F}_{SRS}(t)$ for $m = 2$

	Bias			Efficiency		
$F(t)$	$r = 3$	$r = 5$	$r = 10$	$r = 3$	$r = 5$	$r = 10$
0.05	−0.0004	−0.0019	−0.0008	1.30	1.39	1.40
0.20	−0.0113	−0.0074	−0.0021	1.21	1.29	1.37
0.40	−0.0178	−0.0105	−0.0057	1.08	1.14	1.20
0.50	−0.0241	−0.0136	−0.0064	1.01	1.07	1.12
0.70	−0.0317	−0.0171	−0.0104	0.90	0.91	0.85
0.75	−0.0313	−0.0182	−0.0092	0.86	0.85	0.86
0.90	−0.0022	−0.0115	−0.0085	0.87	0.76	0.73
0.99	−0.0025	−0.0024	−0.0017	0.91	0.84	0.71

Thus

$$\hat{F}_{\text{MLE}^*2}(t) = \frac{y_1^* - r + \sqrt{r^2 + 10ry_1^* + 24ry_2^* + y_1^{2*}}}{6r}$$

Table 4.8 gives the bias of $\hat{F}_{\text{MLE}^*2}(t)$ and its efficiency w.r.t. $\hat{F}_{\text{SRS}}(t)$ when $m = 2$ for $r = 3$, 5, and 10. It can be seen that the efficiency of $\hat{F}_{\text{MLE}^*2}(t)$ w.r.t. $\hat{F}_{\text{SRS}}(t)$ is increasing in r for small and moderate values of $F(t)$ (efficiency >1) and decreasing for large values of $F(t)$ (efficiency < 1). $\hat{F}_{\text{MLE}^*2}(t)$ is a negatively biased estimator.

- For $m = 3$, we get

$$\frac{y_1^* + 2y_2^* + 3y_3^*}{F*(t)} = \frac{y_1}{1 - F^*(t)} + \frac{2y_2F*(t)}{1 - F^{2*}(t)} + \frac{3y_3F^{2*}(t)}{1 - F^{3*}(t)}$$
$$\Rightarrow (6r)F^{4*}(t) + (7r - y_1^*)F^{3*}(t) + (4r - 2y_1^* - 2y_2^*)F^{2*}(t)$$
$$+ (r - 2y_1^* - 2y_2^* - 3y_3^*)F^*(t) - y_1^* - 2y_2^* - 3y_3^* = 0$$

This equation can be solved numerically. Table 4.9 gives the bias of $\hat{F}_{\text{MLE}^*2}(t)$ and its efficiency w.r.t. $\hat{F}_{\text{SRS}}(t)$ when $m = 3$ for $r = 3$, 5, and 10. It can be seen that the efficiency of $\hat{F}_{\text{MLE}^*3}(t)$ w.r.t. $\hat{F}_{\text{SRS}}(t)$ is increasing in r for small and moderate values of $F(t)$ (efficiency >1) and decreasing for large values of $F(t)$ (efficiency < 1). $\hat{F}_{\text{MLE}^*3}(t)$ is more efficient than $\hat{F}_{\text{MLE}^*2}$ when $F(t) \leq 0.4$ for fixed r and m. However, $\hat{F}_{\text{SRS}}(t)$ is more efficient when $F(t) > 0.4$.

4.3 ESTIMATION OF *F*(*X*) USING MOVING EXTREME RSS BASED ON MINIMA AND MAXIMA

It is clear from previous sections that the estimators of F using MERSS based on maxima and MERSS based on minima are not always better than the corresponding estimators based on SRS. In this section, we compare the estimator of the distribution function based on MERSS with *minima* and *maxima* to the corresponding estimator based on SRS. The MLE of $F(t)$ based on MERSS with

Table 4.9 The Bias of $\hat{F}_{MLE^*3}(t)$ and Efficiency of $\hat{F}_{MLE^*3}(t)$ w.r.t. $\hat{F}_{SRS}(t)$ for $m = 3$

	Bias			**Efficiency**		
***F*(*t*)**	***r* = 3**	***r* = 5**	***r* = 10**	***r* = 3**	***r* = 5**	***r* = 10**
0.05	−0.00215	−0.00157	−0.00052	1.79	1.85	1.91
0.20	−0.00955	−0.00454	−0.00206	1.48	1.59	1.66
0.40	−0.02159	−0.01285	−0.00618	1.10	1.29	1.29
0.50	−0.0332	−0.0174	−0.0030	0.92	1.02	1.12
0.70	−0.04142	−0.02489	−0.01254	0.71	0.72	0.77
0.75	−0.04141	−0.02590	−0.01288	0.68	0.67	0.69
0.90	−0.02421	−0.01877	−0.00967	0.67	0.58	0.54
0.99	−0.00200	−0.00164	−0.00057	0.71	0.59	0.48

minima and maxima is derived. Its efficiency with respect to the estimator based on SRS is obtained.

4.3.1 MERSS BASED ON BOTH MINIMA AND MAXIMA

Let $Y_i \sim bin\,(r,\ F^i(t))$ and $Y_i^* \sim bin\,(r, (1-(1-F(t))^i))$, $i = 1, \ldots, m$. The likelihood function based on Y_i and Y_i^* is

$$L(F|y_i, y_i^*) = \prod_{i=1}^m \binom{r}{y_i} (F^i(t))^{y_i} (1-F^i(t))^{r-y_i} \binom{r}{y_i^*} (1-(1-F(t))^i)^{y_i^*} (1-(1-(1-F(t))^i)^{r-y_i^*}$$

$$L^* = \ln(L(F|y_i, y_i^*) = \sum_{i=1}^m [iy_i \ln F(t) + (r-y_i)\ln(1-F^i(t)) + y_i^* \ln(1-(1-F(t))^i) + i(r-y_i^*)\ln(1-F(t))]$$

$$\frac{\partial L^*}{\partial F(t)} = \sum_{i=1}^m (\frac{iy_i}{F(t)} - \frac{i(r-y_i)F^{i-1}(t)}{1-F^i(t)} + \frac{iy_i^*(1-F(t))^{i-1}}{(1-(1-F(t))^i)} - \frac{i(r-y_i^*)}{(1-F(t))})(**)$$

$$= \sum_{i=1}^m (\frac{iy_i - irF^i(t)}{F(t)(1-F^i(t))} + \frac{ir(1-F(t))^i - i(r-y_i^*)}{(1-F(t))(1-(1-F(t))^i)}).(**)$$

Special cases:

- For $m = 1$, setting equation (**) to zero, we get

$$\hat{F}(t) = \frac{y_1 + y_1^*}{2r}.$$

- For $m = 2$, setting equation (**) to zero, we get

$$\frac{y_1}{F(t)} - \frac{(r-y_1)}{1-F(t)} + \frac{y_1^*}{F(t)} - \frac{(r-y_1^*)}{1-F(t)} + \frac{2y_2}{F(t)} - \frac{2(r-y_2)F}{1-F^2(t)} + \frac{2y_2^*(1-F)}{(1-(1-F(t))^2)} - \frac{2(r-y_2^*)}{1-F(t)} = 0$$

$$\Rightarrow \frac{y_1 + y_1^* + 2y_2}{F(t)} - \left[\frac{(r-y_1) + (r-y_1^*) + 2(r-y_2^*)}{1-F(t)}\right] - \frac{2(r-y_2)F}{1-F^2(t)} + \frac{2y_2^* - 2y_2^*}{(1-(1-F(t))^2)} = 0,$$

Table 4.10 Bias of $\hat{F}_{\mathrm{MLE}^*4}(t)$ and Efficiency of $\hat{F}_{MLE*4}(t)$ w.r.t. $\hat{F}_{SRS}(t)$ with $m = 2$

	Bias			**Efficiency**		
***F*(*t*)**	***r* = 3**	***r* = 5**	***r* = 10**	***r* = 3**	***r* = 5**	***r* = 10**
0.05	0.02101	0.02554	0.02625	1.13	0.99	0.83
0.20	0.04758	0.05013	0.05318	1.70	1.56	1.20
0.40	0.04359	0.04220	0.04358	2.31	2.25	1.88
0.50	0.03354	0.03440	0.03599	2.56	2.52	2.20
0.70	0.01655	0.01685	0.01706	2.69	2.58	2.57
0.75	0.01467	0.01156	0.01168	2.58	2.61	2.56
0.90	0.00501	0.00332	0.00269	2.38	2.32	2.39
0.99	0.00036	0.00031	0.00055	2.22	2.24	2.22

Thus

$$(6r)F^3(t) + (2y_2^* - 12r)F^2(t) - (2y_2 + 4r)F(t) + 2y_1 + 2y_1^* + 4y_2 + 2y_2^* = 0$$

This equation is solved numerically. We used Minitab programming to solve this equation. Table 4.10 gives the bias of $\hat{F}_{\mathrm{MLE}^*4}(t)$ and its efficiency w.r.t. $\hat{F}_{\mathrm{SRS}}(t)$ when $m = 2$ for $r = 3$, 5, and 10. It can be seen that the efficiency of $\hat{F}_{MLE*4}(t)$ w.r.t. $\hat{F}_{\mathrm{SRS}}(t)$ is increasing in r for all values of $F(t)$. For very small values of $F(t)$, the efficiency could be smaller than 1. $\hat{F}_{\mathrm{MLE}^*4}(t)$ is a positively biased estimator. The bias tends to increase in r for small to moderate values of $F(t)$, but decrease in r for large values of $F(t)$.

For $m = 3$, setting Equation (**) to zero, we have

$$\Rightarrow \frac{y_1}{F(t)} - \frac{(r-y_1)}{1-F(t))} + \frac{y_1^*}{F(t)} - \frac{(r-y_1^*)}{1-F(t)} + \frac{2y_2}{F(t)} - \frac{2(r-y_2)F(t)}{1-F^2(t)} + \frac{2y_2^*(1-F(t))}{1-(1-F(t))^2}$$
$$- \frac{2(r-y_2^*)}{1-F(t)} + \frac{3y_3}{F(t)} - \frac{3(r-y_3)F^2(t)}{1-F^3(t)} + \frac{3y_3^*(1-F(t))^2}{1-(1-F(t))^3} - \frac{3(r-y_3^*)}{1-F(t)} = 0$$

$$\begin{aligned}&(12r)F^7(t) - (y_1 + y_2^* + 41r)F^6(t) + (3y_1 + 3y_1^* - 2y_2 + 2y_2^* + 29r)F^5(t)\\&- (y_1 + y_1^* - 8y_2 + 2y_2^* + 3y_3 + 3y_3^* - 26r)F^4(t) - (3y_1 + 3y_1^* + 10y_2 + 2y_2^* - 12y_3 + 5r)F^3(t)\\&- (y_1 + y_1^* - 4y_2 - 2y_2^* + 12y_3 - 6y_3^* + 33r)F^2(t) + (3y_1 + 3y_1^* - 6y_2 + 6y_2^* - 9y_3 + 9y_3^* - 42r)F(t)\\&+ 6y_1 + 6y_1^* + 12y_2 + 6y_2^* + 18y_3 + 6y_3^* = 0\end{aligned}$$

This equation is solved numerically. We used Minitab programming to solve this equation. Table 4.11 gives the bias of $\hat{F}_{\mathrm{MLE}^*5}(t)$ and its efficiency w.r.t. $\hat{F}_{\mathrm{SRS}}(t)$ when $m = 3$ for $r = 3$, 5, and 10. The results given in Table 4.11 show that $\hat{F}_{\mathrm{MLE}^*5}(t)$ is more efficient than $\hat{F}_{\mathrm{SRS}}(t)$ with different values of r, the efficiency >1 for all values of $F(t)$. The values of the bias are very close to zero.

Table 4.11 Bias of $\hat{F}_{MLE*5}(t)$ and Efficiency of $\hat{F}_{MLE*5}(t)$ w.r.t. $\hat{F}_{SRS}(t)$ with $m = 3$

	Bias			**Efficiency**		
$F(t)$	$r = 3$	$r = 5$	$r = 10$	$r = 3$	$r = 5$	$r = 10$
0.05	0.00067	−0.00028	0.00085	1.78	2.43	2.29
0.20	0.00176	0.00163	0.00005	2.34	2.34	2.37
0.40	0.00095	−0.0004	0.00012	2.41	2.39	2.40
0.50	0.0001	−0.00120	−0.00109	2.40	2.42	2.37
0.70	−0.00238	0.00013	0.00020	2.37	2.39	2.40
0.75	−0.00043	0.00032	0.00035	2.42	2.37	2.40
0.90	−0.00162	0.00023	−0.00002	2.33	2.34	2.36
0.99	−0.00153	−0.00075	−0.00046	2.20	2.44	2.44

4.4 CONCLUDING REMARKS AND SUGGESTED FUTURE WORK

The estimation of distribution function based on RSS and some of its modifications have been considered in this chapter. MERSS, with maxima, minima, and with both were investigated. Method of moments estimation and maximum likelihood estimation were used. It turned out that some of these estimators can be more efficient than the corresponding counterparts using SRS for some of the range of $F(t)$. It is important to mention here that MERSS can be easily executed with less chance of ranking error than RSS. When both minima and maxima are used in MERSS, the MLE is more efficient than the estimator based on SRS. Taking into account the lower amount of effort needed to obtain MERSS compared to that needed to obtain RSS, we recommend the use of this estimator.

In this work, it is assumed that the obtained sample is accurate, in the sense that there is no error in ranking. This rarely happens; some ranking errors may occur in obtaining RSS and MERSS. Therefore it is of interest to see the performance of RSS and MERSS, when there is some error in ranking. Also, when the variable of interest has a strong relation with another easier variable, then we can use the other variable (concomitant variable) to rank the values of the variable of interest. In this case ratio estimation can be used to make an inference about the variable of interest. This topic is a future research work topic.

ACKNOWLEDGMENTS

The authors are very thankful to the editor and the referee for their comments. One of the comments about ratio estimation is very useful and could be a topic for future work.

REFERENCES

Abu-Dayyeh, W.A., Samawi, H.M., Bani-Hani, L.A., 2002. On distribution estimation using double ranked set samples with application. J. Mod. Appl. Stat. Methods 1, 443−451.

Al-Odat, M., Al-Saleh, M.F., 2001. A variation of ranked set sampling. J. Appl. Stat. Sci. 10, 137–146.

Al-Saleh, M.F., Ababneh, A., 2015. Test for accuracy in ranking in moving extreme ranked set sampling. Int. J. Comput. Theor. Stat. 2, 67–77.

Al-Saleh, M.F., Al-Ananbeh, A.M., 2005. Estimation the correlation coefficient in a bivariate normal distribution using moving extreme ranked set sampling with a concomitant variable. J. Korean Stat. Soc. 34, 125–140.

Al-Saleh, M.F., Al-Hadhrami, S.A., 2003a. Parametric estimation for the location parameter for symmetric distributions using moving extreme ranked set sampling with application to trees data. Environmetrics 14, 651–664.

Al-Saleh, M.F., Al-Hadhrami, S.A., 2003b. Estimation of the mean of the exponential distribution using moving extreme ranked set sampling. Stat. Papers 44, 367–382.

Al-Saleh, M.F., Al-Kadiri, M., 2000. Double ranked set sampling. Stat. Probab. Lett. 48, 205–212.

Al-Saleh, M.F., Al-Omari, A.I., 2002. Multistage ranked set sampling. J. Stat. Plan. Inference 102, 273–286.

Al-Saleh, M.F., Darabseh, M., 2017. Inference on the skew normal distribution using ranked set sampling. Int. J. Comput. Theor. Stat. 4, 65–76.

Al-Saleh, M.F., Naamneh, A., 2016. Properties of the elements of SRS, RSS, MERSS. J. Appl. Stat. Sci. 22, 75–85.

Al-Saleh, M.F., Samawi, H.M., 2000. On the efficiency of Monte Carlo methods using steady state ranked simulated samples. Commun. Stat. Simul. Comput. 29, 941–954.

Al-Saleh, M.F., Samawi, H., 2010. On estimating the odds using moving extreme ranked set sampling. Stat. Methodol. 7, 133–140.

Al-Saleh, M.F., Zheng, G., 2002. Estimation of bivariate characteristics using ranked set sampling. Austr. N. Z. J. Stat. 44, 221–232.

Hanandeh, A., 2011. Estimation of the parameters of Downton's bivariate exponential distribution using moving extreme ranked set sampling. Department of Statistics, Yarmouk University, Jordan.

Kaur, A., Patil, G., Sinha, A.K., Taillie, C., 1995. Ranked set sampling an annotated bibliography. Environ. Ecol. Stat. 2, 25–54.

McIntyre, G., 1952. A method for unbiased selective sampling using ranked set sampling. Aust. J. Agric. Res. 3, 385–390.

Ozturk, O., Wolfe, D.A., 2000a. An improved ranked set two-sample Mann-Whitney-Wilcoxon test. Can. J. Stat. 13, 57–76.

Ozturk, O., Wolfe, D.A., 2000b. Alternative ranked set sampling protocol for the sign test. Stat. Probab. Lett. 47, 15–23.

Ozturk, O., Wolfe, D.A., 2001. A new ranked set sampling protocol for the signed rank test. J. Stat. Plann. Inference 96, 351–370.

Ozturk, O., 2002. Ranked set sample inference under a symmetry restriction. J. Stat. Plann. Inference 102, 317–336.

Scheaffer, R.L., Mendenhall, W., Ott, L., 1986. Elementary Survey Sampling. *Duxbury Press*, Boston USA.

Sharma, P., Bouza, C., Verma, H., Sghin, R., Sautto, J., 2016. A Generalized class of estimators for the finite population mean when the study variable is qualitative in nature. Rev. Investig. Oper. 37 (2), 163–172.

Stokes, S.L., Sager, T.W., 1988. Characterization of a ranked-set sample with application to estimating distribution functions. J. Am. Stat. Assoc. 83, 374–381.

Subzar, M., Maqbool, S., Raja, T., Shabber, M., 2017a. A new ratio estimators for estimation of population mean using conventional location parameters. World Appl. Sci. J. 35, 377–384.

Subzar, M., Maqbool, S., Raja, T., Shabber, M., Lone, B., 2017b. A class of improved ratio estimators for population mean using conventional location parameters. Int. J. Mod. Math. Sci. 15, 187–205.

Subzar, M., Maqbool, S., Raja, T., Mir, S., Jeelani, M., Bhat, M., 2017c. Improved family of ratio type estimators for estimating population mean using conventional and non conventional location parameters. Rev. Investig. Oper. 38 (5), 499–513.

Takahasi, K., Wakimoto, K., 1968. On unbiased estimates of the population mean based on the sample stratified by means of ordering. Ann. Inst. Stat. Math. 20, 1–31.

Zheng, G., Al-Saleh, M.F., 2002. Modified maximum likelihood estimators based on RSS. Ann. Inst. Stat. Math. 54, 641–658.

FURTHER READING

Ababneh, A., 2011. *Test for perfect ranking in moving extreme ranked set sampling*. Master Thesis. Department of Statistics, Yarmouk Univ., Jordan.

Ahmad, Dana, 2012. Estimation of the Distribution Function Using Some Variations of Ranked Set Sampling. Master Thesis. Department of Statistics, Yarmouk University, Jordan.

Evans, M.J., 1967. *Application of Ranked Set Sampling to Regeneration Surveys in Areas Direct-Seeded to Longleaf Pine*. Masters Thesis. School of Forestry and Wildlife Management, Louisiana State University, Louisiana.

Ghosh, K., Tiwari, R., 2008. Estimating the distribution function using k-tuple ranked set samples. J. Stat. Plan. Inference 138, 929–949.

Halls, L., Dell, T., 1966. Trials of ranked set sampling for forage yields. Forest Sci. 12, 22–26.

Kim, D.H., Kim, D.W., Kim, H.G., 2005. On the estimation of the distribution function using extreme median ranked set sampling. J. Korean Data Anal. Soc. 2, 429–439.

Na'amnih, A., 2011. *The Performance of "The Five-Number Summary" Obtained Using Different Sampling Techniques*. Master Thesis. Department of Statistics, Yarmouk University, Jordan.

Samawi, H., Al-Saleh, M.F., 2013. Valid estimation of odds ratio using two types of moving extreme ranked set sampling. J. Korean Stat. Soc. 42, 17–24.

Sinha, B.K., Sengpta, S., Mukhuti, S., 2006. Unbiased estimation of the distribution function of the exponential population using order statistics with application ranked set sampling. Commun. Stat.: Theory Methods 35, 1655–1670.

Sunzar, M., Raja, A., Maqbool, S., Nazir, N., 2016. New alternative to ratio estimator of population mean. Int. J. Agric. Stat. Sci. 1, 221–225.

CHAPTER

STATISTICAL INFERENCE OF RANKED SET SAMPLING VIA RESAMPLING METHODS

5

Saeid Amiri[1] **and Reza Modarres**[2]

[1] *Department of Natural and Applied Sciences, University of Wisconsin-Green Bay, Green Bay, WI, United States*

[2] *Department of Statistics, The George Washington University, Washington, DC, United States*

5.1 INTRODUCTION

Rank-based sampling (RSS) provides powerful inference alternatives to simple random sampling (SRS) and often leads to large improvements in the precision of estimators. Several variants of RSS are designed to further improve the performance of RSS. Theoretical underpinnings must be developed for such RSS designs. However, these results are often nontrivial due to many factors, including unknown parent distributions, small sample sizes, and nonidentical order statistics that form the cornerstones of any RSS design. These difficulties make bootstrap methods more attractive for RSS. Bootstrap is a well-known resampling method that provides accurate inference for SRS. The bootstrap method has also been explored in the different RSS contexts: confidence interval estimation (Hui et al., 2005), resampling algorithms (Modarres et al., 2006), one sample test (Amiri et al., 2014), confidence bands for the CDF (Frey, 2014), empirical likelihood (Amiri et al., 2016), censored RSS (Mahdizadeh and Strzalkowska-Kominiak, 2017), and tests of perfect ranking (Amiri et al., 2017).

In this work, we consider the parametric statistical inference of one sample and two samples. RSS is concerned with small sample sizes and distribution-free methods such as sign, sign ranked, and Mann−Whitney tests have been investigated by Bohn and Wolfe (1992) and Ozturk and Wolfe (2000a, 2000b). However, proposed nonparametric RSS tests might be sensitive to the ranking procedure. Fligner and MacEachern (2006) consider the center of the observations to eliminate the impact of ranking. The sensitivity of these tests to the ranking procedure occurs due to the use of the distribution function of the rth order statistic to test the mean. To overcome the sensitivity, we explore $H_0 : \mu_x = \mu_0$ and its two-sample variant using the t test statistics and show that the bootstrap methods provide more accurate inference. We will also consider RSS with different ranks sizes.

The reminder of the chapter is organized as follows. Section 5.2 provides an overview of the data structure of an RSS, and then formally defines the test statistics for one and two samples. Section 5.3 is devoted to the bootstrap methods. It gives a number of theoretical results that allow us to use the bootstrap methods. In Section 5.4, we compare the proposed methods using Monte Carlo simulation. The simulations show that a hybrid method, based on the average of the p-values of pivotal and nonpivotal bootstrap tests, outperforms the competing tests.

Ranked Set Sampling. DOI: https://doi.org/10.1016/B978-0-12-815044-3.00005-8

5.2 STATISTICAL INFERENCE FOR RSS

Suppose a total number of n units are to be measured from the underlying population on the variable of interest. Let n sets of units, each of size k, be randomly chosen from the population using a simple random sampling (SRS) technique. The units of each set are ranked by any means other than actual quantification of the variable. Finally, one unit in each ordered set with a prespecified rank is measured on the variable. Let m_r be the number of measurements on units with rank r, $r = 1, \ldots, k$, such that $n = \sum_{r=1}^{k} m_r$. Let $X_{(r)j}$ denote the measurement on the jth measured unit with rank r. This results in a URSS of size n from the underlying population as $X_{(r)j}$; $r = 1, \ldots, k, j = 1, \ldots, m_r$. When $m_r = m$, $r = 1, \ldots, k$, URSS reduces to the balanced RSS. It is worth mentioning that, in ranked set sampling designs, $X_{(1)j}, \ldots, X_{(k)j}$ are independent order statistics (as they are obtained from independent sets) and each $X_{(r)j}$ provides information about a different stratum of the population. One can represent the structure of a URSS as follows:

$$\mathscr{X}_r = \left\{X_{(r)1}, X_{(r)2}, \ldots, X_{(r)m_r}\right\} \overset{i.i.d.}{\sim} F_{(r)}, \quad r = 1, \ldots, k_1,$$

where $F_{(r)}$ is the distribution function (df) of the rth order statistic. The second sample can be generated using the same procedure. We assume the second sample is generated using k_2 which can be different from $k = k_1$

$$\mathscr{Y}_r = \{Y_{(r)1}, Y_{(r)2}, \ldots, Y_{(r)m_r}\} \overset{i.i.d.}{\sim} G_{(r)}, \quad r = 1, \ldots, k_2.$$

It is of interest to test $H_0 : F(x) \overset{d}{=} G(x - \Delta)$. Specifically, we are concerned with the null hypothesis $H_0 : \mu_x = \mu_y + \Delta$ versus $H_0 : \mu_x \neq \mu_y + \Delta$. Two sample tests are commonly used to determine whether the samples come from the same unknown distribution. In our setting, we assume X and Y are collected with different ranks sizes. Therefore, even under the same parent distributions, the variance of the estimator would not be the same.

The following proposition can be used to establish the asymptotic normality of statistic under the null hypothesis.

Proposition 1: *Let F denote the cdf of a member of the family with $\int x^2 dF(x) < \infty$ and $\hat{F}_{(r)}$ is the empirical distribution function (edf) of the rth row. If $\vartheta_i = (\overline{X}_{(i)} - \mu_{(i)})$, then $(\vartheta_1, \ldots, \vartheta_k)$ converges in distribution to a multivariate normal distribution with mean vector zero and covariance matrix diag$(\sigma^2_{(1)}/m_1, \ldots, \sigma^2_{(k)}/m_k)$ where $\sigma^2_{(i)} = \int (x - \mu_{(i)})^2 dF_{(i)}(x)$ and $\mu_{(i)} = \int x dF_{(i)}(x)$.*

Proposition 1 suggests the following statistic for testing H_0: $\mu = \mu_0$,

$$Z = \frac{1}{k}\sum\nolimits_{r=1}^{k} \overline{X}_{(r)} - \mu_0 \hat{\sigma} \overset{d}{\to} N(0, 1),$$

where $\hat{\sigma}^2$ is the plug-in estimator for the $V\left(\frac{1}{k}\sum_{r=1}^{k} \overline{X}_{(r)}\right)$,

$$\hat{\sigma}^2 = \frac{1}{k^2}\sum_{r=1}^{k} \frac{\hat{\sigma}^2_{(r)}}{m_r},$$

and $\sigma^2_{(r)}$ is the estimate of $V(\overline{X}_{(r)})$. Using the central limit theorem, one obtains a confidence interval where

$$P\left(\mu \in \left(\overline{X} + t_{\alpha/2,n-1}\frac{\sigma}{\sqrt{n}}, \overline{X} + t_{1-\alpha/2,n-1}\frac{\sigma}{\sqrt{n}}\right)\right) \approx 1 - \alpha.$$

One needs $\sigma^2_{(r)}$ to estimate the variance of the mean. Hence it is necessary to have $m_r \geq 2$. The estimate of the variance for small sample sizes would be very inaccurate, suggesting that a pivotal statistic might be unreliable. We show in Section 5.4 that parametric statistics are very conservative. Bootstrap provides a nonparametric alternative to estimate the variance. The bootstrap method can be used to obtain the sampling distribution of the statistic of interest and allows for estimation of the standard error of any well-defined functional. Hence, bootstrap enables us to draw inferences when the exact or the asymptotic distribution of the statistic of interest is unavailable. A procedure of generating resamples to calculate the variance is discussed in Section 5.3.

Proposition 1 can be used to obtain a test statistic for two samples $\mathscr{X}_1, \ldots, \mathscr{X}_{k_1}$ and $\mathscr{Y}_1, \ldots, \mathscr{Y}_{k_2}$. One can show that

$$T(\mathscr{X}, \mathscr{Y}) = \left(\frac{1}{k_1}\sum\nolimits_{r=1}^{k_1} \overline{X}_{(r)} - \frac{1}{k_2}\sum\nolimits_{r=1}^{k_2} \overline{Y}_{(r)}\right) - (\mu_1 - \mu_2)\hat{\sigma} \xrightarrow{d} N(0,1),$$

where

$$\hat{\sigma}^2 = \frac{1}{k_1^2}\sum_{r_1=1}^{k_1} \frac{\hat{\sigma}^2_{(r_1)}}{m_{r_1}} + \frac{1}{k_2^2}\sum_{r_2=1}^{k_2} \frac{\hat{\sigma}^2_{(r_2)}}{m_{r_2}}.$$

We can consider the parametric statistical inference for the skewed distribution: let $X_1, \ldots, X_n$ be i.i.d. random variable with the mean μ and finite variance σ^2. Since the characteristic function of S_n converges to $e^{-t^2/2}$, the characteristic function of the standard normal, $\sqrt{n}S_n = \sqrt{n}(\mu - \mu)/\sigma$, is asymptotically normally distributed with zero mean and unit variance. To take the sample skewness into account, the following proposition obtains the Edgeworth expansion of $\sqrt{n}S_n$.

Proposition 2: *If $E\left(Y_i^6\right) < \infty$ and Cramer's condition holds, the asymptotic distribution function of $\sqrt{n}S_n$ is*

$$P(\sqrt{n}S_n \leq x) = \Phi(x) + \frac{1}{\sqrt{n}}\gamma(ax^2 + b)\phi(x) + O(n^{-1}),$$

where a and b are known constants, γ is an estimable constant, and Φ and ϕ denote the standard normal distribution and density functions, respectively.

Hall (1992) suggested two functions,

$$S_1(t) = t + a\hat{\gamma}t^2 + \frac{1}{3}a^2\hat{\gamma}^2t^3 + n^{-1}b\hat{\gamma},$$
$$S_2(t) = (2an^{-\frac{1}{2}}\hat{\gamma})^{-1}\left\{\exp\left(2an^{-\frac{1}{2}}\hat{\gamma}t\right) - 1\right\} + n^{-1}b\hat{\gamma},$$

where $a = 1/3$ and $b = 1/6$. Zhou and Dinh (2005) suggested

$$S_3(t) = t + t^2 + \frac{1}{3}t^3 + n^{-1}b\hat{\gamma}.$$

Using $S_i(t)$, for $i = 1, 2, 3$, one can construct new confidence intervals for μ as

$$(\hat{\mu} - S_i(n^{-1/2}t_{1-\alpha/2,n-1})\hat{\sigma}, \hat{\mu} - S_i(n^{-1/2}t_{\alpha/2,n-1})\hat{\sigma}),$$

where $t_{1-\alpha/2,n-1}$ is the $1-\alpha/2$ quartile of the t distribution. However, use of the sample skewness in the asymptotic distribution makes the inference less reliable, especially for the parametric methods. For example, the asymptotic distribution of test for the coefficient of variation depends on the skewness. This parameter makes the inference for coefficient of variation inaccurate, see Amiri (2016). It is of interest to study this problem using a fully nonparametric approach via the bootstrap.

5.3 BOOTSTRAP METHOD

Bootstrap resampling is a well-known statistical method to conduct statistical inference. Bootstrap mimics the underlying distribution of the observations by resampling from the URSS sample. Several papers have explored the application of bootstrap in RSS. URSS bootstrap was considered in Amiri et al. (2014). The idea of URSS bootstrap is to obtain a sample of size n_0 from each stratum in order to transform the URSS to an RSS dataset. The RSS dataset is then resampled to provide inference. Amiri et al. (2017) consider more general resampling techniques that obtain resamples from the entire dataset instead the resampling each stratum. The procedure is described below.

Algorithm:

1. Select a row randomly and select an observation, continue until k observations have been collected (obviously any row can appear more than once).

$$\text{Order them as } X^{\Diamond}_{(1)} \leq \ldots \leq X^{\Diamond}_{(k)} \text{ and retain } X^{*}_{(r)1} = X^{\Diamond}_{(r)}.$$

2. Perform steps 1–2 m_r times and collect $X^*_{(r)1}, \ldots, X^*_{(r)m_r}$.
3. Perform step 3 for $r = 1, \ldots, k$.
4. Repeat steps 1–4, B times to obtain the bootstrap samples.

Using step 1 of the algorithm,

$$\{X^{\Diamond}_1, \ldots, X^{\Diamond}_k\} \sim \hat{F}_n(t) = \frac{1}{k}\sum_{r=1}^{k}\frac{1}{m_r}\sum_{j=1}^{m_r} I(X_{(r)j} \leq t),$$

and using steps 2 and 3,

$$\mathbf{X}^*_r = \{\boldsymbol{X}^*_{(r)1}, \boldsymbol{X}^*_{(r)2}, \ldots, \boldsymbol{X}^*_{(r)m_r}\} \sim \hat{F}_{(r)}(.), \tag{5.1}$$

where $\hat{F}_{(r)}(t) = \frac{1}{m_r}\sum_{j=1}^{m_r} I(X_{(r)j} \leq t)$. Let

$$\hat{\boldsymbol{F}}^*_{(r)}(\boldsymbol{t}) = \frac{1}{\boldsymbol{m_r}}\sum_{j=1}^{\boldsymbol{m_r}} \boldsymbol{I}\left(\boldsymbol{X}^*_{(r)j} \leq \boldsymbol{t}\right), \tag{5.2}$$

$$\hat{\boldsymbol{F}}^*_{\boldsymbol{n}}(\boldsymbol{t}) = \frac{1}{\boldsymbol{k}}\sum_{r=1}^{\boldsymbol{k}}\frac{1}{\boldsymbol{m_r}}\sum_{j=1}^{\boldsymbol{m_r}} \boldsymbol{I}\left(\boldsymbol{X}^*_{(r)j} \leq \boldsymbol{t}\right). \tag{5.3}$$

Amiri et al. (2017) proved the following propositions for the proposed bootstrap algorithm. These properties are essential to draw inference using the resamples.

Proposition 3: *Let $F_{(r)}(t)$ denote the cdf of the rth row of a member of the family with the continuous density function, and $\hat{F}^*_{(r)}$ denote the edf of the rth row given in (Eq. (5.2)), it follows that*

$$\hat{F}^*_{(r)}(t) \overset{a.s.}{\to} F_{(r)}(t).$$

Proposition 4: *Let $F(t)$ denote the cdf of a member of the family with the continuous density function, and suppose $\mathscr{X}^*_1, \ldots, \mathscr{X}^*_k$ are samples obtained using the proposed bootstrap algorithm, it follows that*

$$\sup_{t\in\mathbb{R}} |\hat{F}^*_n(t) - F(t)| = 0.$$

Proposition 4 shows a desirable property for the bootstrap method that can be used to draw statistical inference. The direct application of bootstrap is in the estimation of variance. Suppose $\mathscr{X}^*_1, \ldots, \mathscr{X}^*_k$ and we are interested in $V(\theta(F_{(1)}, \ldots, F_{(K)})) = \frac{1}{k^2}\sum\limits_{r=1}^{k}\frac{\sigma^2_{(r)}}{m_r}$. The plug-in estimation is $V(\theta(\hat{F}_{(1)}, \ldots, \hat{F}_{(K)})) = \hat{\sigma}^2 = \frac{1}{k^2}\sum\limits_{r=1}^{k}\frac{\hat{\sigma}^2_{(r)}}{m_r}$ where $\hat{F}_{(r)}$ is the edf on the $r-$th stratum. Clearly, the plug-in estimate does not work for $m_r = 1$. However, one can use the proposed bootstrap to estimate the variance. Generate the resamples using the proposed algorithm and compute $\theta(\hat{F}^*_{(1)}, \ldots, \hat{F}^*_{(k)}) = \overline{X}^* = \frac{1}{\mathrm{k}}\sum_{r=1}^{k}\overline{X}^*_{(r)}$, and repeat the procedure B times to obtain $\overline{X}^*_b, b = 1, \ldots, B$. The most important property of the bootstrap lies in the conditional independence, given the original sample. Hence, we view bootstrap resample as iid random samples and compute the sample mean and the sample variance with,

$$\overline{\overline{X}}^* = \frac{1}{B}\sum_{b=1}^{B}\overline{X}^*_b,$$

$$\hat{V}^*\left(\theta\left(\hat{F}_{(1)}, \ldots, \hat{F}_{(K)}\right)\right) = \frac{1}{B}\sum_{b=1}^{B}(\overline{X}^*_b - \overline{\overline{X}}^*)^2.$$

The confidence interval can be found using the bootstrap estimate of variance as,

$$\frac{1}{k}\sum_{r=1}^{k}\overline{X}_{(r)} \pm t_{\alpha/2,n-1}\sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\overline{X}^*_b - -\overline{\overline{X}}^*\right)^2}.$$

The nonparametric confidence interval can be obtained using the percentile confidence interval

$$(\overline{X}^*_{\alpha/2}, \overline{X}^*_{1-\alpha/2}),$$

where $\overline{X}^*_{\alpha/2}$ is the $\alpha/2$ percentile of bootstrap resample mean.

5.4 NUMERICAL STUDY

This section is devoted to assessing the accuracy and comparisons of the proposed test statistics for finite sample sizes. We study the type I error rate and the statistical power. The proposed tests are based on the same simulated data in order to provide a meaningful comparison. The resampling is carried out using $B = 800$ resamples. In order to make a comparative evaluation of the testing procedures, we seek certain desirable features, such as robustness, power, and small sample test validity in terms of observed type I error rates. In the following, the significance and the power of the proposed tests are studied for different sample sizes.

To compare two group means: $H_0 : \mu_x = \mu_y + \delta$ vs. $H_0 : \mu_x \neq \mu_y + \delta$, the appropriate test statistic is

$$\boldsymbol{T}_0(\mathscr{X},\mathscr{Y}) = \left(\frac{1}{\boldsymbol{k}_1}\sum\nolimits_{\boldsymbol{r}=1}^{\boldsymbol{k}_1}\overline{\boldsymbol{X}}_{(\boldsymbol{r})} - \frac{1}{\boldsymbol{k}_2}\sum\nolimits_{\boldsymbol{r}=1}^{\boldsymbol{k}_2}\overline{\boldsymbol{Y}}_{(\boldsymbol{r})}\right) - \boldsymbol{\delta}\hat{\boldsymbol{\sigma}}, \tag{5.4}$$

where, $\hat{\sigma}^2 = \frac{1}{k_1^2}\sum_{r_1=1}^{k_1}\frac{\hat{\sigma}^2_{(r_1)}}{m_{r_1}} + \frac{1}{k_2^2}\sum_{r_2=1}^{k_2}\frac{\hat{\sigma}^2_{(r_2)}}{m_{r_2}}$. Under the null hypothesis, $T_0(\mathscr{X},\mathscr{Y}) \sim t_{n_1+n_2-2}$. We refer to this test as the parametric test and denote it with PT.

The bootstrap test, referred to as BT, is constructed as follows. Calculate the statistic given in (Eq. (5.4)), take the resamples according to the algorithm described in Section 5.3, and calculate the following statistic,

$$\boldsymbol{T}^*(\mathscr{X}^*,\mathscr{Y}^*,\mathscr{X},\mathscr{Y}) = \frac{\left(\frac{1}{k_1}\sum\nolimits_{\boldsymbol{r}=1}^{k_1}\overline{\boldsymbol{X}}^*_{(\boldsymbol{r})} - \frac{1}{k_2}\sum\nolimits_{\boldsymbol{r}=1}^{k_2}\overline{\boldsymbol{Y}}^*_{(\boldsymbol{r})}\right) - \left(\frac{1}{k_1}\sum\nolimits_{\boldsymbol{r}=1}^{k_1}\overline{\boldsymbol{X}}_{(\boldsymbol{r})} - \frac{1}{k_2}\sum\nolimits_{\boldsymbol{r}=1}^{k_2}\overline{\boldsymbol{Y}}_{(\boldsymbol{r})}\right)}{\hat{\boldsymbol{\sigma}}^*}, \tag{5.5}$$

where σ^* is the estimate of variance using the bootstrap samples. Generate B resamples and calculate the test statistics,

$$T_1^*(\mathscr{X}^*,\mathscr{Y}^*,\mathscr{X},\mathscr{Y}),\ldots,T_B^*(\mathscr{X}^*,\mathscr{Y}^*,\mathscr{X},\mathscr{Y}).$$

The approximate p-value can be estimated with

$$p^* = \frac{\#T_b^*(\mathscr{X}^*,\mathscr{Y}^*,\mathscr{X},\mathscr{Y}) \le T_0(\mathscr{X},\mathscr{Y})}{B},$$

$$p - \text{value} = \min\{p^*, 1-p^*\}.$$

Since the RSS often uses small sample sizes and the plug-in estimate of variance is not very accurate, one may consider a third approach and use nonpivotal test statistics to calculate the p-value. That is,

$$\begin{aligned} T_0(\mathscr{X},\mathscr{Y}) &= \left(\frac{1}{k_1}\sum_{r=1}^{k_1}\overline{X}_{(r)} - \frac{1}{k_2}\sum_{r=1}^{k_2}\overline{Y}_{(r)}\right) - \delta, \\ T^*(\mathscr{X}^*,\mathscr{Y}^*,\mathscr{X},\mathscr{Y}) &= \left(\frac{1}{k_1}\sum_{r=1}^{k_1}\overline{X}^*_{(r)} - \frac{1}{k_2}\sum_{r=1}^{k_2}\overline{Y}^*_{(r)}\right) - \left(\frac{1}{k_1}\sum_{r=1}^{k_1}\overline{X}_{(r)} - \frac{1}{k_2}\sum_{r=1}^{k_2}\overline{Y}_{(r)}\right). \end{aligned} \tag{5.6}$$

This bootstrap test using the nonpivotal statistic is denoted as BNT.

We compare the following test statistics:

1. PT: Parametric two-sample t-test (Eq. (5.4));
2. BT: Bootstrap test (Eq. (5.5));
3. BNT: Nonpivotal bootstrap test (Eq. (5.6));
4. BHT : Hybrid test of BT and BNT.

Table 5.1 includes the simulation of the 10th percentile of p-value for the proposed methods with different sample sizes $(n_X, n_Y) = (k_1 m, k_2 m)$, and the following underlying distributions:

1. $X \overset{d}{=} Y \sim N(0,1)$,
2. $X \overset{d}{=} Y \sim exp(1.5)$,
3. $X \overset{d}{=} Y \sim logistic(1,1)$,
4. $X \overset{d}{=} Y \sim Gamma(1,2)$.

Table 5.1 Observed α-levels of the Proposed Tests at $\alpha = 0.1$

			m					***m***			
dist.	k_1, k_2	**test**	**3**	**4**	**5**	**6**	**dist.**	**3**	**4**	**5**	**6**
$X \overset{d}{=} Y \sim N(0,1)$	(3,3)	PT	0.315	0.226	0.188	0.169	$X \overset{d}{=} Y \sim exp(1.5)$	0.317	0.231	0.191	0.165
		BT	0.133	0.119	0.115	0.111		0.144	0.127	0.116	0.108
		BNT	0.080	0.093	0.099	0.101		0.076	0.099	0.101	0.102
		BHT	0.095	0.100	0.104	0.104		0.098	0.105	0.104	0.103
	(3,4)	PT	0.310	0.226	0.194	0.174		0.316	0.237	0.186	0.178
		BT	0.117	0.117	0.115	0.11		0.127	0.116	0.108	0.117
		BNT	0.077	0.095	0.097	0.103		0.088	0.100	0.100	0.111
		BHT	0.087	0.101	0.104	0.104		0.095	0.104	0.101	0.111
	(3,5)	PT	0.316	0.236	0.196	0.184		0.333	0.243	0.196	0.177
		BT	0.132	0.119	0.110	0.116		0.132	0.117	0.113	0.112
		BNT	0.090	0.104	0.094	0.103		0.088	0.102	0.110	0.107
		BHT	0.100	0.106	0.098	0.108		0.094	0.105	0.111	0.108
	(4,4)	PT	0.312	0.225	0.189	0.166		0.316	0.217	0.196	0.182
		BT	0.114	0.110	0.106	0.106		0.125	0.105	0.107	0.115
		BNT	0.092	0.095	0.100	0.101		0.089	0.097	0.103	0.111
		BHT	0.094	0.098	0.103	0.100		0.098	0.096	0.105	0.113
	(4,5)	PT	0.316	0.218	0.197	0.176		0.307	0.230	0.198	0.161
		BT	0.112	0.107	0.109	0.105		0.118	0.112	0.110	0.097
		BNT	0.092	0.092	0.098	0.102		0.091	0.101	0.103	0.094
		BHT	0.095	0.094	0.102	0.104		0.097	0.102	0.104	0.093
$X \overset{d}{=} Y \sim logistic(1,1)$	(3,3)	PT	0.331	0.244	0.204	0.178	$X \overset{d}{=} Y \sim Gamma(1,2)$	0.338	0.231	0.202	0.185
		BT	0.136	0.124	0.127	0.109		0.133	0.118	0.121	0.114
		BNT	0.072	0.090	0.111	0.112		0.082	0.090	0.109	0.112
		BHT	0.077	0.099	0.116	0.106		0.086	0.098	0.111	0.109
	(3,4)	PT	0.338	0.240	0.204	0.185		0.337	0.251	0.196	0.179
		BT	0.136	0.122	0.123	0.116		0.138	0.126	0.111	0.116
		BNT	0.091	0.103	0.112	0.114		0.093	0.105	0.105	0.113
		BHT	0.093	0.105	0.116	0.112		0.101	0.111	0.106	0.110

(*Continued*)

Table 5.1 Observed α-levels of the Proposed Tests at $\alpha = 0.1$ *Continued*

			m					***m***			
dist.	k_1, k_2	**test**	**3**	**4**	**5**	**6**	**dist.**	**3**	**4**	**5**	**6**
	(3,5)	PT	0.353	0.244	0.209	0.191		0.333	0.256	0.203	0.175
		BT	0.145	0.124	0.119	0.121		0.131	0.132	0.120	0.113
		BNT	0.101	0.100	0.107	0.117		0.093	0.110	0.106	0.106
		BHT	0.105	0.103	0.109	0.114		0.095	0.117	0.110	0.105
	(4,4)	PT	0.323	0.245	0.184	0.180		0.331	0.233	0.197	0.182
		BT	0.121	0.121	0.102	0.113		0.121	0.112	0.105	0.111
		BNT	0.088	0.104	0.099	0.118		0.095	0.101	0.105	0.114
		BHT	0.090	0.107	0.098	0.114		0.092	0.102	0.103	0.108
	(4,5)	PT	0.337	0.238	0.202	0.189		0.339	0.242	0.197	0.178
		BT	0.126	0.115	0.115	0.111		0.124	0.117	0.111	0.102
		BNT	0.099	0.100	0.113	0.115		0.093	0.102	0.109	0.108
		BHT	0.096	0.101	0.110	0.111		0.091	0.102	0.107	0.104

Since X and Y are generated from the same distributions, it is expected that an accurate test maintains the nominal level. In the frequentist approach, the appealing property of the p-value is its (asymptotic) uniformity on $Unif(0,1)$ under the null hypothesis. When a test statistic is conservative (or liberal), the actual type I error of the test will be small (large) compared with the nominal level. For a conservative (or liberal) test, the power values can be misleading. It is easy to see that a conservative p-value hardly rejects an incorrect null hypothesis and a liberal test easily rejects a correct null hypothesis too often and both lead to incorrect inferences.

Clearly the PT leads to an overly conservative test, i.e., fails to reject the null hypothesis when it should. However, this problem tends to diminish with an increase in sample size. Here, BT and BNP have better performances, and are closer to the actual p-value. It is noteworthy that for $m = 3$, $(k_1, k_2) = (3,3), (3,4)$ which have very small sample sizes $((n_X, n_Y) = (9,9), (9,12))$, BT and BNT are conservative and liberal, respectively. It is of interest to explore the average of the p-values. We refer to this hybrid test as the BHT method. Clearly BHT has better performance.

To compare the statistical power, we consider

1. $X \sim N(0,1), Y \sim N(0.5,1)$,
2. $X \sim exp(1), Y \sim exp(1.5)$,
3. $X \sim logistic(0,1), Y \sim logistic(1,1)$,
4. $X \sim Gamma(1,1), Y \sim Gamma(1,2)$.

Since X and Y are generated from different distributions with different parameters, a powerful test should reject the null hypothesis with high probability. The result is presented in Table 5.2. Since the PT is conservative for small sample sizes, we expect a large value for the power. However, this power value is overly optimistic and not accurate. Among the bootstrap tests, BHT has better power than BNT. BHT has less power than PT, keeping in mind that PT performs conservatively for small sample sizes.

5.5 CONCLUSIONS

A considerable amount of research has been conducted in the past few decades to advance the theoretical foundation of RSS and present its applications. RSS draws on additional information from inexpensive and easily obtained sources to collect a more representative sample. In this work, we review the statistical tests of means under one and two samples. RSS is often applied with small sample sizes. Presenting nonparametric methods and exploring the performance of test statistics are essential in obtaining a better understanding of their behavior. In our empirical study, we mainly considered small samples and compared the performance of proposed tests using Monte Carlo investigations under different distributions. We proposed a hybrid method, based on the average of the p-values of pivotal and nonpivotal bootstrap tests and demonstrate its better performance. The hybrid method provides a more accurate inference for small sample sizes and enables one to maintain the nominal level with comparable power.

Table 5.2 The Empirical Power of the Proposed Tests

			m					m			
dist.	k_1, k_2	**test**	**3**	**4**	**5**	**6**	**dist.**	**3**	**4**	**5**	**6**
$X \sim N(0,1), Y \sim N(0.5,1)$	(3,3)	PT	0.552	0.553	0.605	0.666	$X \sim exp(1), Y \sim exp(1.5)$	0.590	0.604	0.652	0.703
		BT	0.342	0.422	0.502	0.583		0.395	0.481	0.559	0.631
		BNT	0.224	0.359	0.457	0.553		0.267	0.403	0.510	0.602
		BHT	0.276	0.392	0.485	0.571		0.322	0.444	0.537	0.619
	(3,4)	PT	0.607	0.626	0.670	0.724		0.640	0.678	0.733	0.789
		BT	0.391	0.476	0.568	0.653		0.439	0.556	0.645	0.722
		BNT	0.269	0.423	0.522	0.617		0.309	0.488	0.596	0.683
		BHT	0.325	0.452	0.547	0.639		0.369	0.524	0.623	0.706
	(3,5)	PT	0.628	0.657	0.722	0.771		0.676	0.714	0.773	0.819
		BT	0.427	0.517	0.613	0.691		0.479	0.598	0.687	0.759
		BNT	0.314	0.455	0.576	0.664		0.353	0.521	0.635	0.714
		BHT	0.367	0.489	0.596	0.677		0.407	0.562	0.662	0.739
	(4,4)	PT	0.670	0.702	0.755	0.822		0.702	0.749	0.808	0.859
		BT	0.445	0.563	0.655	0.747		0.491	0.624	0.721	0.803
		BNT	0.356	0.514	0.630	0.729		0.389	0.567	0.679	0.773
		BHT	0.399	0.546	0.647	0.741		0.441	0.596	0.703	0.790
	(4,5)	PT	0.708	0.740	0.817	0.859		0.754	0.793	0.845	0.893
		BT	0.491	0.599	0.725	0.794		0.562	0.685	0.771	0.846
		BNT	0.407	0.569	0.702	0.776		0.459	0.628	0.735	0.823
		BHT	0.449	0.581	0.716	0.788		0.514	0.661	0.756	0.835
$X \sim logistic(0,1)$, $Y \sim logistic(1,1)$	(3,3)	PT	0.481	0.440	0.454	0.476	$X \sim Gamma(1,1)$, $Y \sim Gamma(1,2)$	0.640	0.695	0.752	0.808
		BT	0.257	0.302	0.347	0.380		0.437	0.563	0.654	0.736
		BNT	0.142	0.233	0.314	0.377		0.245	0.452	0.617	0.734
		BHT	0.175	0.263	0.327	0.381		0.320	0.510	0.635	0.737
	(3,4)	PT	0.529	0.500	0.535	0.562		0.668	0.720	0.791	0.844
		BT	0.323	0.365	0.438	0.488		0.450	0.581	0.684	0.773
		BNT	0.211	0.302	0.398	0.454		0.241	0.468	0.666	0.777
		BHT	0.252	0.337	0.421	0.470		0.321	0.531	0.676	0.781

	(3,5)	PT	0.580	0.567	0.598	0.631		0.706	0.740	0.813	0.861
		BT	0.382	0.452	0.508	0.560		0.476	0.597	0.714	0.799
		BNT	0.254	0.379	0.448	0.506		0.259	0.479	0.689	0.798
		BHT	0.305	0.412	0.475	0.531		0.337	0.540	0.708	0.800
	(4,4)	PT	0.539	0.548	0.581	0.624		0.780	0.819	0.880	0.927
		BT	0.321	0.394	0.456	0.525		0.595	0.721	0.817	0.887
		BNT	0.216	0.329	0.440	0.521		0.403	0.634	0.788	0.869
		BHT	0.251	0.362	0.449	0.524		0.488	0.686	0.805	0.879
	(4,5)	PT	0.587	0.598	0.654	0.700		0.798	0.855	0.909	0.946
		BT	0.372	0.456	0.553	0.615		0.611	0.752	0.851	0.911
		BNT	0.265	0.397	0.517	0.593		0.410	0.667	0.835	0.905
		BHT	0.306	0.426	0.539	0.606		0.509	0.716	0.847	0.912

REFERENCES

Amiri, S., 2016. Revisiting inference of coefficient of variation: nuisances parameters. Stat 5 (1), 234–241.

Amiri, S., Jozani, M.J., Modarres, R., 2014. Resampling unbalanced ranked set samples with applications in testing hypothesis about the population mean. J. Agric., Biol., Environ. Stat. 19 (1), 1–17.

Amiri, S., Jozani, M.J., Modarres, R., 2016. Exponentially tilted empirical distribution function for ranked set samples. J. Korean Stat. Soc. 45 (2), 176–187.

Amiri, S., Modarres, R., Zwanzig, S., 2017. Tests of perfect judgment ranking using pseudo samples. Comput. Stat. 32 (4), 1309–1322.

Bohn, L.L., Wolfe, D.A., 1992. Nonparametric two-sample procedures for ranked-set samples data. J. Am. Stat. Assoc. 87 (418), 552–561.

Fligner, M.A., MacEachern, S.N., 2006. Nonparametric two-sample methods for ranked set sample data. J. Am. Stat. Assoc. 101, 1107–1118.

Frey, J., 2014. Bootstrap confidence bands for the CDF using ranked-set sampling. J. the Korean Stat. Soc. 43 (3), 453–461.

Hall, P., 1992. On the removal of skewness by transformation.. J. R. Stat. Soc.. Ser. B 221–228.

Hui, T.P., Modarres, R., Zheng, G., 2005. Bootstrap confidence interval estimation of mean via ranked set sampling linear regression. J. Stat. Comput. Simul. 75 (7), 543–553.

Mahdizadeh, M., Strzalkowska-Kominiak, E., 2017. Resampling based inference for a distribution function using censored ranked set samples. Comput. Stat. 1–24.

Modarres, R., Hui, T.P., Zheng, G., 2006. Resampling methods for ranked set samples. Comput. Stat. Data Anal. 51 (2), 1039–1050.

Ozturk, O., Wolfe, D.A., 2000a. Alternative ranked set sampling protocols for the sign test. Stat. Probab. Lett. 47 (1), 15–23.

Ozturk, O., Wolfe, D.A., 2000b. An improved ranked set two-sample mann-whitney-wilcoxon test. Can. J. Stat. 28 (1), 123–135.

Zhou, X.H., Dinh, P., 2005. Nonparametric confidence intervals for the one-and two-sample problems. Biostatistics 6 (2), 187–200.

FURTHER READING

Hall, P., 2013. The Bootstrap and Edgeworth Expansion. Springer Science & Business Media, New York.

CHAPTER

6

EXTENSIONS OF SOME RANDOMIZED RESPONSE PROCEDURES RELATED WITH GUPTA-THORNTON METHOD: THE USE OF ORDER STATISTICS

Carlos N. Bouza-Herrera

Faculty of Mathematics and Computation, University of Havana, Havana, Cuba

6.1 INTRODUCTION

We will consider that the interest is in estimating the mean of a sensitive variable Y. Some persons in the population carry a stigma and tend to give an incorrect value of Y or to refuse to give a report. The seminal work of Warner (1965a,b) opened a way to deal with this problem by using the so-called technique of randomized response (RR). The use of RR provides the opportunity of reducing response biases, as well as nonresponses, due to dishonest answers when questioning on Y. This technique protects the privacy of the respondent by ensuring that his belonging to a stigmatized group cannot be detected by the sampler.

Greenberg et al. (1971) extended the theory of RR to the quantitive case. Different extensions of RR have been introduced since then. A usual approach for estimating the mean of a quantitative sensitive variable Y is scrambling the responses using some auxiliary variables. Celebrating the 50th anniversary of the publication of Warner's paper, Chaudhuri et al. (2016) edited a set of recent research results on this theme. Some important particular new models are due to Gupta and Thornton (2002), Hussain and Shabbir (2011), Singh and Chen (2009), and Tarray and Singh (2015).

In this chapter, we introduce the use of order statistics (OS) as an alternative to some scrambling procedures reported in the literature in Section 6.2. A detailed discussion on them may be obtained, for example, in Bouza and Singh (2009), Chaudhuri and Mukherjee (1988), and Gupta and Thornton (2002).

Section 6.3 is concerned with the development of their counterparts, which use OS of the distribution of the auxiliary variable X.

A comparison of the estimators is developed by comparing their variances. An important result is that the use of scrambling using the OS of X provides, in general, an improvement in the accuracy.

Ranked Set Sampling. DOI: https://doi.org/10.1016/B978-0-12-815044-3.00006-X

6.2 THE CONSIDERED SCRAMBLING PROCEDURES

In Chaudhuri and Mukherjee (1988) a simple scrambling procedure can be seen. Take the sensitive variable Y and a variable X with a known distribution with $E(X) = \mu_X$ and $V(X) = \sigma_X^2$. The ith respondent performs an experiment and obtains a value of X. Then he/she reports

$$S_i = Y_i + X_i$$

Its expectation is $E(S_i) = \mu_Y + \mu_X$ and its variance $V(S_i) = V(Y_i) + V(X_i)$. We may compute the sample mean of

$$\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i - \mu_X = \frac{1}{n}\sum_{i=1}^{n} Y_i + X_i - \mu_X$$

It is unbiased, as

$$E\left(\overline{Z}\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) + E(X_i) - \mu_X = \mu_Y$$

Its variance is

$$V(\overline{Z}) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) + V(X_i) = \frac{\sigma_Y^2 + \sigma_X^2}{n}$$

A variation is that each respondent selects randomly a value from $U^* = \{U_1, \ldots, U_k\}$ with probability π_t. U^* is determined previously by the sampler and the sample he/she makes a selection U_t. As we know U^* and π_t

$$\mu_U = \sum_{t=1}^{k} U_t \pi_t, \sigma_U^2 = \sum_{t=1}^{k} \left(U_t - \mu_U\right)^2 \pi_t,$$

It seems that the respondents should think that an extra protection is given to his possible stigmatization if the report is

$$S_{Ui} = Y_i + U_i X_i$$

Under this scrambling procedure the expectation of the report is

$$E(S_{Ui}) = E(Y_i) + E(U_i)E(X_i) = \mu_Y + \mu_U \mu_X.$$

We are able to compute

$$\overline{Z}_U = \frac{1}{n}\sum_{i=1}^{n} Z_{Ui} = \frac{1}{n}\sum_{i=1}^{n} S_{Ui} - \mu_U \mu_X$$

As its expectation is

$$E(\overline{Z}_U) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) + E(U_i)E(X_{i(i)}) = \mu_Y + \mu_U\left(\frac{1}{n}\sum_{i=1}^{n} \mu_{X_{(i)}}\right) - \mu_U \mu_X = \mu_Y$$

it is an unbiased estimator of μ_Y. The variance of this estimator is:

$$V(\overline{Z}_U) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} V(Z_{Ui}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\sigma_Y^2 + \sigma_U^2\sigma_X^2}{n} = \frac{\sigma_Y^2 + \sigma_U^2\sigma_X^2}{n}$$

Note that

$$V\left(\overline{Z}\right) - \mathrm{V}\left(\overline{Z}_U\right) = \frac{(1-\sigma_U^2)\sigma_X^2}{n}$$

Therefore, if the sampler determines U^* in such a way that $\sigma_U^2 > 0$ is preferred estimating μ_Y employing $\overline{Z}_U$.

Gupta and Thornton (2002) proposed generating a random Bernoulli variable A with parameter α and obtaining as response

$$S_{iG} = \begin{cases} Y_i \textit{ if } A = 1 \\ Y_i + X_i \textit{ if } A = 0 \end{cases}$$

That is, the report is modeled by

$$S_{iG} = AY_i + (1-A)(Y_i + X_i)$$

Let us analyze its expected value.

$$E(\overline{S}_G) = \frac{1}{n}\sum_{i=1}^{n} E(S_{iG}) = \alpha\mu_Y + (1-\alpha)\left(\mu_Y + \mu_X\right) = \mu_Y + (1-\alpha)\mu_X$$

Take the transformed variable $Z_{iG} = S_{iG} - (1-\alpha)\mu_X$, its expectation is

$$E(Z_{iG}) = E(S_{iG}) - (1-\alpha)\mu_X = \mu_Y$$

Clearly, for estimating μ_Y unbiasedly a good decision is taking its sample mean

$$\overline{Z}_G = \frac{1}{n}\sum_{i=1}^{n} Z_{iG}$$

The sampling errors of the sample means of $\overline{S}_G$ and $\overline{Z}_G$coincide :

$$V(\overline{S}_G) = V(\overline{Z}_G) = \frac{\sigma_Y^2}{n} + \frac{(1-\alpha)\sigma_X^2 + \alpha(1-\alpha)\mu_X^2}{n}$$

Comparing the accuracy of $\overline{Z}$ with that of $\overline{Z}_G$ we have that

$$V\left(\overline{Z}\right) \le V\left(\overline{Z}_G\right) \text{ if } 1 - \frac{\sigma_X^2}{\mu_X^2} = 1 - CV(X)^2 \ge \alpha$$

These results allow the sampler to design the preference of one of the methods with respect to a convenient distribution function of X. For example, if is used the distribution described below

$$f(x) = \begin{cases} \frac{1}{9} \textit{ if } x \in [3, 12] \\ 0 \textit{ otherwise} \end{cases}$$

By preferring the proposal of Gupta and Thornton (2002) as, in this case $\frac{\sigma_X^2}{\mu_X^2} = \frac{\frac{81}{12}}{\frac{225}{4}} \cong 0.122$, is enough using $\alpha > 0, 9$.

Consider the difference $V(\overline{Z}_U) - V\left(\overline{Z}_G\right)$. It is equal to

$$V(\overline{Z}_U) - V\left(\overline{Z}_G\right) = \frac{(\alpha + \sigma_U^2 - 1)\sigma_X^2}{n} - \frac{\alpha(1-\alpha)\mu_X^2}{n}$$

We have the second-degree equation inequality $\alpha^2\mu_X^2 + \alpha\left(\sigma_X^2 - \mu_X^2\right) + (\ \alpha + \sigma_U^2 - 1)\sigma_X^2 < 0$. Its solution provides an adequate value of α, once the sampler fixes $f\,(x)$, if he/she decides to use the Gupta and Thornton (2002) scrambling method.

6.3 USING ORDER STATISTICS (OS) FOR SCRAMBLING

We propose using order statistics (OS) instead of values of an auxiliary variable for scrambling. Consider that the respondent selected in the ith draw is provided with a mechanism for generating, using SRSWR, a sequence of positive independent random variables $X_1, \ldots, X_k, X_j \in X^*$. The interviewee, included in the ith drawn, obtains a sequence, ranks it, and determines $X_{i(1)}, \ldots, X_{i(k)}, \text{where} X_{i(t)} < X_{i(h)}, \textit{if}\ \ t < h$. The report is made as follows:

$$S_{(i)} = Y_i + X_{i(i)}$$

We have that the expectation of the report is

$$E(S_{(i)}) = E(Y_i) + E\left(X_{i(i)}\right) = \mu_Y + \mu_{X_{(i)}}$$

We may compute from the response

$$Z_{(i)} = S_{(i)} - \mu_X$$

Under the described model, we have that:

$$V(S_{(i)}) = V(Y_i) + V\left(X_{i(i)}\right) = \sigma_Y^2 + \sigma_{X_{(i)}}^2$$

We select, from the population, a simple random sample with replacement of size n, and take the sample mean:

$$\overline{Z}_{os} = \frac{1}{n}\sum_{i=1}^{n} Z_{(i)} = \frac{1}{n}\sum_{i=1}^{n} Y_i + X_{i(i)} - \mu_X$$

We derive its unbiasedness because, see Chen et al. (2004), $\frac{1}{n}\sum_{i=1}^{n} \mu_{X_{(i)}} = \mu_X$. Therefore

$$E\left(\overline{Z}_{os}\right) = \frac{1}{n}\sum_{i=1}^{n} \mu_Y + \mu_{X_{(i)}} - \mu_X = \mu_Y$$

The random mechanism used sustains that the OS are mutually independent and they are also independent of Y. Taking into account these facts, the variance is given by

$$V(\overline{Z}_{os}) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} V(Z_{(i)}) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} V(Y_i) + V(X_{i(i)}) = \frac{\sigma_Y^2}{n} + \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \sigma_{X_{(i)}}^2 =$$

Denoting $\mu_{X_{(i)}} - \mu_X = \Delta_{X_{(i)}}$ we have, see Bouza and Singh (2009) and Chen et al. (2004) for example, that

$$\sigma^2_{X_{(i)}} = \sigma^2_X - \left(\mu_{X_{(i)}} - \mu_X\right)^2$$

Then, the variance of the estimated mean is

$$V(\overline{Z}_{os}) = \frac{\sigma^2_Y + \sigma^2_X}{n} - \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \Delta^2_{X_{(i)}}$$

Let us consider again that each respondent selects also randomly a value from $U^* = \{U_1, \ldots, U_k\}$ with probability π_t. U* is determined previously by the sampler and provides a device for performing the random selection of U_t. We know the expectation and variance of U_t. They are

$$\mu_U = \sum_{t=1}^{k} U_t \pi_t, \sigma^2_U = \sum_{t=1}^{k} (U_t - \mu_U)^2 \pi_t,$$

It will be more reliable for the respondents to report the scrambled variable

$$S_{U(i)} = Y_i + U_i X_{i(i)}$$

U_i is the selection made by respondent I from U^*. The expectation of the report is

$$E(S_{U(i)}) = E(Y_i) + E(U_i)E(X_{i(i)}) = \mu_Y + \mu_U \mu_{X_{(i)}}.$$

We are able to compute

$$\overline{Z}_{(U)} = \frac{1}{n}\sum_{i=1}^{n} Z_{U(i)} = \frac{1}{n}\sum_{i=1}^{n} S_{U(i)} - \mu_U \mu_{\mathrm{X}}$$

Its expectation is

$$E\left(\overline{Z}_{(U)}\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) + E(U_i)E(X_{i(i)}) = \mu_Y + \mu_U\left(\frac{1}{n}\sum_{i=1}^{n} \mu_{X_{(i)}}\right) - \mu_U \mu_X = \mu_Y$$

Hence, it is an unbiased estimators of μ_Y. The variance of this estimator is derived as follows

$$V(\overline{Z}_{(U)}) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} V(Z_{U(i)}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\sigma^2_Y + \sigma^2_U V(X_{i(i)})}{n} = \frac{\sigma^2_Y + \sigma^2_U \sum_{i=1}^{n} \sigma^2_{X_{(i)}}}{n} = \frac{\sigma^2_Y + \sigma^2_U \sigma^2_X}{n} - \sigma^2_U \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \Delta^2_{X_{(i)}}$$

Comparing the variances of $\overline{Z}_{os}$ with the above expression, we have the preference for $\overline{Z}_U$ whenever

$$V\left(\overline{Z}_{os}\right) - V(\overline{Z}_{(U)}) = (1 - \sigma^2_U)\left(\frac{\sigma^2_X}{n} - \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \Delta^2_{X_{(i)}}\right) > 0$$

We know that $\frac{\sigma^2_X}{n} - \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \Delta^2_{X_{(i)}} > 0$, therefore, this relationship holds unless $1 \leq \sigma^2_U$.

Another RR-scrambling method based on OS is derived by using the scrambling procedure of Gupta and Thornton (2002). We suggest scrambling by using the OS obtained by the ith respondent. A random Bernoulli variable A with parameter α is generated by the respondent and is obtained as response

$$S_{iG} = \begin{cases} Y_i \; if \; A = 1 \\ Y_i + X_{i(i)} \; if \; A = 0 \end{cases}$$

That is, the report is modeled by

$$S_{(i)G} = AY_i + (1 - A)(Y_i + X_{i(i)})$$

Let us analyze its expected value. It is

$$E(S_{(i)G} = \alpha\mu_Y + (1-\alpha)\left(\mu_Y + \mu_{X_{(i)}}\right) = \mu_Y + (1-\alpha)\mu_{X_{(i)}}$$

The variance of it is given by

$$V\left(S_{(i)G}\right) = \sigma_Y^2 + (1-\alpha)\sigma_{X_{(i)}}^2 + \alpha(1-\alpha)\mu_{X_{(i)}}^2$$

We may compute

$$Z_{(i)G} = S_{(i)G} - (1-\alpha)\mu_X$$

In addition, derive as an estimator of μ_Y its sample mean

$$\overline{Z}_{(G)} = \frac{1}{n}\sum_{i=1}^{n} Z_{(i)G}$$

We have that

$$E(\overline{Z}_{(G)}) = \mu_Y + \frac{1-\alpha}{n}\left(\sum_{i=1}^{n}\mu_{X_{(i)}} - \mu_X\right) = \mu_Y$$

Its sampling error is given by

$$V(\overline{S}_{(G)}) = V(\overline{Z}_{(G)}) = \frac{\sigma_Y^2}{n} + \frac{(1-\alpha)\sum_{i=1}^{n}\sigma_{X_{(i)}}^2 + \alpha(1-\alpha)\sum_{i=1}^{n}\mu_{X_{(i)}}^2}{n}$$

Note that

$$\sum_{i=1}^{n}\sigma_{X_{(i)}}^2 = \frac{\sigma_X^2}{n} - \left(\frac{1}{n}\right)^2\sum_{i=1}^{n}\Delta_{X_{(i)}}^2$$

Hence

$$V(\overline{Z}_{(G)}) = \frac{\sigma_Y^2}{n} + (1-\alpha)\left(\frac{\sigma_X^2}{n} - \left(\frac{1}{n}\right)^2\sum_{i=1}^{n}\Delta_{X_{(i)}}^2\right) + \frac{\alpha(1-\alpha)\sum_{i=1}^{n}\mu_{X_{(i)}}^2}{n}$$

Comparing $\overline{Z}_{(G)}$ with $\overline{Z}_{os}$ we have that $\overline{Z}_{os}$ is more accurate if is satisfied the relationship

$$\alpha^2\sum_{i=1}^{n}\mu_{X(i)}^2 + \alpha\left(\sigma_X^2 - \frac{1}{n}\sum_{i=1}^{n}\Delta_{X_{(i)}}^2 + \sum_{i=1}^{n}\mu_{X_{(i)}}^2\right) + \frac{1}{n}\sum_{i=1}^{n}\Delta_{X_{(i)}}^2 < 0$$

A comparison with $\overline{Z}_U$ is developed by computing

$$V(\overline{Z}_{(U)}) - V(\overline{Z}_{(G)}) = \frac{(\alpha - 1 + \sigma_U^2)\sigma_X^2}{n} - \left(\alpha - 1 + \sigma_U^2\right)\left(\frac{1}{n}\right)^2\sum_{i=1}^{n}\Delta_{X_{(i)}}^2 - \frac{\alpha(1-\alpha)\sum_{i=1}^{n}\mu_{X_{(i)}}^2}{n} < 0$$

In terms of α this means that

$$\alpha^2 \sum_{i=1}^{n} \mu_{X_{(i)}}^2 + \alpha\left(\sigma_X^2 - \frac{1}{n}\sum_{i=1}^{n} \Delta_{X_{(i)}}^2 + \sum_{i=1}^{n} \mu_{X_{(i)}}^2\right) + \frac{1}{n}\sum_{i=1}^{n} \Delta_{X_{(i)}}^2 < 0$$

Then a sampler with a preference for the Gupta-Thornton procedure is able to tune the value of α once $f(x)$ is fixed by solving a second-degree equation.

Let us consider the effect of using $X_{i(i)}$ instead of X_i. The paired comparisons of the procedures of scrambling using X or the OS yields the following criteria:

1. $V(\overline{Z}) vs V(\overline{Z}_{os})$
   As $V(\overline{Z}) - V(\overline{Z}_{os}) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \Delta_{X_{(i)}}^2 \geq 0$, we should prefer $\overline{Z}_{os}$.
2. $V(\overline{S}_G) vs\ V(\overline{Z}_{(G)})$

$$V(\overline{S}_G) - V(\overline{Z}_{(G)}) = \frac{+\alpha(1-\alpha)\left(\mu_X^2 - \sum_{i=1}^{n} \mu_{X_{(i)}}^2\right)}{n} + (1-\alpha)\left(\left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \Delta_{X_{(i)}}^2\right)$$

   As α is a probability,$(1-\alpha) > 0$, a sufficient condition for preferring $\overline{Z}_{(G)}$ is the positiveness of X because $\mu_X^2 - \sum_{i=1}^{n} \mu_{X_{(i)}}^2 = \sum_{i \neq j} \mu_{X_{(i)}} \mu_{X_{(j)}}$.
3. $V(\overline{S}_U) vs\ V(\overline{Z}_{(U)})$
   The difference of the variance is always positive: $V(\overline{S}_U) - V\left(\overline{Z}_{(U)}\right) = \sigma_U^2 \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \Delta_{X_{(i)}}^2$

6.4 CONCLUSIONS

From the developed paired comparison, we have that the use of scrambling using OS should be preferred in all cases.

The introduction of an additional randomization through a set of values U^* improves the accuracy with respect to the direct use of $Y + X$.

The procedure of Gupta and Thornton (2002) may be preferred to the other scrambling proposed with an adequate selection of α previous a fixation of the distribution of the variable X.

REFERENCES

Bouza, C.N., Singh, L., 2009. Optimal aspects of a model based randomized responses procedure under unequal selection of insensitive variables. Rev. Investig. Oper. 30, 234–243.

Chaudhuri, A., Mukerjee, R., 1988. Randomized Response: Theory and Techniques. Marcel Dekker Inc, New York.

Chaudhuri, A., Christofides, T.C., Rao, C.R., 2016. Data gathering, analysis and protection of privacy through randomized response Techniques: Qualitative and Quantitative Human Traits. Handbook of Statistics, vol. 34. Elsevier, Amsterdam.

Chen, Z., Bai, Z., Sinha, B.K., 2004. Ranked Set Sampling: Theory and Applications. Lectures Notes in Statistics, 176. Springer, N. York.

Greenberg, B.G., Kubler, R.R., Horvitz, D.G., 1971. Applications of RR technique in obtaining quantitative data. J. Am. Stat. Assoc. 66, 243–250.

Gupta, S., Thornton, B., 2002. Circumventing social desirability response bias in personal interview surveys. Am. J. Math. Manag. Sci. 22, 369–383.

Hussain, Z., Shabbir, J., 2011. Improved estimation of mean in randomized response models. Hacet. J. Math. Stat. 40, 91–104.

Singh, S., Chen, C.C., 2009. Utilization of higher order moments of scrambling variables in randomized response sampling. J. Stat. Plan. Inference 139, 3377–3380.

Tarray, T.A., Singh, H., 2015. A general procedure for estimating the mean of a sensitive variable using auxiliary information. Rev. Investig. Oper. 36, 268–279.

Warner, S.L., 1965a. Randomized response: a survey technique for eliminating evasive answer bias. J. Am. Stat. Assoc. 60, 63–69.

Warner, S.L., 1965b. Randomized Response: A Survey Technique Tarray, T.A. and Housila.

FURTHER READING

Liu, P., Gao, G., He, Z., Ruan, Y., Li, X., Yu, M., 2011. Two-stage sampling on additive model for quantitative sensitive question survey and its application. Prog. App. Math. 2, 67–72.

CHAPTER 7

RANKED SET SAMPLING ESTIMATION OF THE POPULATION MEAN WHEN INFORMATION ON AN ATTRIBUTE IS AVAILABLE

Carlos N. Bouza-Herrera[1], Rajesh Singh[2] and Prabhakar Mishra[2]

[1]Faculty of Mathematics and Computation, University of Havana, Havana, Cuba [2]Department of Statistics, Banaras Hindu University, Varanasi, Uttar Pradesh, India

7.1 INTRODUCTION

Consider a variable of interest Y and a concomitant variable X, which are correlated and the coefficient of correlation ρ. The population ratio of the population mean of the two variables is $R = \frac{\mu_y}{\mu_x} = \frac{\overline{Y}}{\overline{X}}$, and its usual estimator is $\hat{R} = \frac{\overline{y}}{\overline{x}}$, $\overline{y}$ and $\overline{x}$ are the sample means.

Textbooks consider that the sample is selected using simple random sampling with replacement (SRSWR). The ratio estimator is biased and it is negligible under certain conditions. The expression of the bias is developed using Taylor series expansion, see classic textbooks, such as Cochran (1977) and Murthy (1967). The approximated variance of $\hat{R}$, considering such development, is

$$Var(\hat{R}) \cong \frac{R^2}{n}\left(V_x^2 + V_y^2 - 2\rho_{xy} V_x V_y\right),$$

where $V_x = \frac{\sigma_x}{\mu_x}$, $V_y = \frac{\sigma_y}{\mu_y}$ and $\rho_{xy} = \sum_{i=1}^{N} (x_i - \mu_x)(\mu_i - \mu_y)/N\sigma_x\sigma_y$, σ_x and σ_y are the standard deviations of the populations of the variables X and Y, respectively.

The available information may be used in different ways and many modified ratio estimators have been developed in recent years. The information on X, as the coefficient of variation, quartiles, median, coefficient of kurtosis, coefficient of skewness, is used for improving the estimation of R. Modified ratio estimators have been proposed by Murthy (1967), Cochran (1977), Kadilar and Cingi (2004), Singh et al. (2008), Al-Omari et al. (2009), and Singh and Solanki (2012).

An alternative to simple random sampling (SRS) is the sample design known as ranked set sampling (RSS). McIntyre (1952) introduced it looking to increase the efficiency of the estimation of the population mean. The method is useful when the variable of interest is very expensive or difficult to measure but it can be easily ranked at a negligible cost. The original form of RSS, conceived by McIntyre (1952), can be described as follows. First, a simple random sample of size k is drawn from the population and the k sampling units are ranked with respect to the variable of interest, say X, without measuring Y. Then the unit with rank 1 is identified and taken for the

Ranked Set Sampling. DOI: https://doi.org/10.1016/B978-0-12-815044-3.00007-1

measurement. The remaining units of the sample are discarded. Next, another simple random sample of size k is drawn and the units of the sample are ranked by judgment, the unit with rank 2 is taken using the measurement of X and the remaining units are discarded. This process is continued until a simple random sample of size k is taken and ranked and the unit with rank k is taken for the measurement of X. This whole process is referred to as a cycle. The cycle then repeats m times and yields a ranked set sample of size $n = mk$. In the recent past a lot of research has been done in RSS by Samawi et al. (1996), Muttlak (1997), Philip and Lam (1997), Muttlak (1998), Al-Saleh and Al-Kadiri (2000), Al-Odat and Al-Saleh (2001), Al-Saleh and Al-Omari (2002), Jozani and Johnson (2011), and Jeelani et al. (2013, 2014a,b,c,d).

Takahasi and Wakimoto (1968) gave mathematical support to their claims. Dell and Clutter (1972) established that even if the ranking is not perfect the proposed estimator is still unbiased. The use of RSS is the theme of a growing number of papers. Patil et al. (2002), Bouza (2005), and Al-Omari and Bouza (2014) gave reviews of the theme as well as a large list of papers.

Different ratio type estimators have been developed for RSS, see for example, Wolfe (2004), Ganeslingam and Ganesh (2006), Ohyama et al. (2008), Al-Omari et al. (2009), Herrera and Al-Omari (2011), Al-Omari (2012), Singh et al. (2014), Jeelani and Bouza (2015), Al-Omari et al. (2016), and Khan and Shabbir (2016).

In the last 65 years the theory of RSS has been extended and is now thoroughly applied. Its popularity is due to the fact that RSS is expected to improve the accuracy of the estimation of the population mean of Y.

Take a finite population $U = \{u_1, \ldots, u_N\}$ and a variable X correlated with the variable of interest Y. It may be used for obtaining an accurate ranking of Y cheaply. Consider that in addition to X each unit is attached to an attribute γ, which is highly correlated to Y in some sense. Denote the information on U by $\vec{Z} = (Z_1, \ldots, Z_N)$, $Z = X, Y, \gamma$. X and Y are real variables and

$$\gamma_i = \begin{cases} 1 \text{ if } u_i \text{ belong to a group } \vartheta \\ \quad 0 \text{ otherwise} \end{cases}$$

X and γ are known in advance by the statistician. This is a common situation. Take for example the study of the response to a treatment of cancer patients. Take X as the size of the tumor, existing in the patient's expedient, and γ as the sex. Measuring the size of the tumor after the treatment, Y, is to be obtained using an expensive method, such as tomography axial computing.

Note that we know in advance the values of the totals $\gamma_T = \sum_{i=1}^{N} \gamma_i$ and $X_T = \sum_{i=1}^{N} X_i$. Therefore we may compute the proportion of units belonging to ϑ, $P = \gamma_T / N$ as well as the population mean of X: $\overline{X} = X_T / N$.

We are interested in estimating the population mean of Y

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

Commonly, a sample s is selected from U using simple random sampling with replacement (SRSWR) and $\overline{Y}$ is estimated using the sample mean

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

The selected sample may be used for estimating the mean of X and P by $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\mathrm{p} = \frac{1}{n}\sum_{i=1}^{n} \gamma_i$. As they are unbiased estimators their mean squared errors (MSE) are their variances. They are, respectively:

$$V(\bar{z}) = \frac{1}{Nn}\sum_{i=1}^{N}\left(Z_i - \overline{Z}\right)^2 = \frac{\sigma_Z^2}{n}, Z = X, Y$$
$$V(p) = \frac{\sigma_\gamma^2}{n} = \frac{P(1-P)}{n}$$

The ratio of the true proportion and the estimation provides information, which may be introduced in the estimation process to improve the accuracy of the estimate. Different authors have used attributes for deriving ratio type estimators of $\overline{Y}$ based on SRSWR. See, for example, Singh et al. (2008).

In this chapter we will extend some results, when RSS is used for selecting a sample and is decided estimating $\overline{Y}$ by means of ratio type estimators, based on an auxiliary attribute γ.

Some exponential ratio type estimators of the finite population mean $\overline{Y}$ are considered. The proposed RSS-estimators perform better under conditions that generally hold in practice.

Section 7.2 is concerned with presenting some ratio type estimators, based on auxiliary information provided by attributes. Section 7.3 is devoted to the development of their RSS counterparts. An auxiliary variable X is used for ranking the units. The proposed estimators are analyzed and approximate expressions of their mean squared errors (MSE) are obtained by developing Taylor series. The expressions of the gains in accuracy of the RSS-estimators are developed and their meanings are discussed. Finally, in Section 7.4, a numerical study is developed using real-life data for illustrating the performance of the proposal. We compare the proposed RSS-estimators with the existing SRSWR-estimators of the population mean in terms of their MSE and a simulation study of the approximation error (AE).

7.2 RATIO TYPE ESTIMATORS IN SRSWR USING γ

In SRSWR, n units out of N units of a population $U = \{u_1, \ldots, u_N\}$ are drawn independently and every possible combination of items, for the given sample size, has an equal chance of being selected.

Ratio estimators are of wide use when looking for increasing the estimation of the precision of the estimates of the population mean. They take advantages from the existence of a correlation between an auxiliary variable and the variable of interest. The basic theory of ratio estimation is presented in standard textbooks, such as Cochran (1977), Murthy (1967), and Hedayat and Sinha (1992). The common framework takes into account first-order Taylor series developments. Commonly the concomitant variable is quantitative but some approaches consider that it is an attribute. Then particular ratio estimators have been developed. We will analyze some of the most popular ones.

The classic ratio estimator is determined by

$$\bar{y}_r = \bar{y}\left(\frac{\overline{X}}{\bar{x}}\right),$$

Its MSE is approximately

$$MSE\left(\bar{y}_r\right) \cong \frac{\sigma_Y^2}{n} + \bar{Y}^2 \frac{\sigma_X^2}{n\bar{X}^2} - 2\frac{\bar{Y}}{\bar{X}}\rho_{y\gamma}\sigma_Y\sigma_X$$

Improving estimation using ratio type estimators is giving a new look to the use of additional information. Some papers on the theme are Sharma et al. (2013) and Kadilar and Cingi (2004).

Among the first proposals of using attributes as an auxiliary variable is the paper by Naik and Gupta (1996). A seminal paper is the contribution of Prasad (1989). More recently, contributions are Shabbir and Gupta (2007) and Yadav and Adewara (2013). They considered the use of SRSWR and proposed as ratio estimator of $\bar{Y}$

$$\bar{y}_1 = \bar{y}r_p, r_p = \frac{P}{p}$$

They obtained that the MSE of $\bar{y}_1$ is given by

$$MSE\left(\bar{y}_1\right) \cong \frac{\sigma_Y^2}{n} + \bar{Y}^2 \frac{\sigma_\gamma^2}{nP^2}\left(1 - 2\rho_{y\gamma}\left(\frac{P\sigma_Y}{\bar{Y}\sigma_\gamma}\right)\right)$$

Note that $\rho_{y\gamma}$ is the point biserial coefficient of correlation. That is $\rho_{y\gamma} = \frac{\sqrt{P(1-P)}\left(\bar{Y}_\vartheta - \bar{Y}_{\bar{\vartheta}}\right)}{\sigma_y}$, where

$$\bar{Y}_\vartheta = \frac{\sum_{u_i \in \vartheta} Y_i}{\sum_{i=1}^N I_\vartheta(i)}, \bar{Y}_{\bar{\vartheta}} = \frac{\sum_{u_i \in \vartheta} Y_i}{N - \sum_{i=1}^N I_\vartheta(i)}, I_\vartheta(i) = \begin{cases} 1 \text{ if } u_i \in \vartheta \\ 0 \text{ otherwise} \end{cases}$$

Jhajj et al. (2006) proposed to work within a general class of estimators. Their proposal was considering the parametric class

$$\zeta_2 = \left\{\bar{y}_2 \middle| \bar{y}_2 = g(\bar{y}, \tau); \tau = \frac{p}{P}\right\}$$

The parametric function $g(a, b)$ should satisfy a set of regularity conditions. One of them is that $g\left(\bar{Y}, 1\right) = \bar{Y}$, for any value of the population mean. The optimal estimator in this class is the linear regression estimator

$$\bar{y}_{2\text{opt}} = \bar{y} + b(P - p), b = \hat{\beta}, \beta = \frac{\text{Cov}(y, p)}{V(p)} = \rho_{y\gamma}\left(\frac{\sigma_y}{\sigma_\gamma}\right).$$

as its MSE equals

$$\text{Min}\left\{\text{MSE}(\bar{y}_2)\right\} \cong \frac{\sigma_y^2}{n}\left(1 - \rho_{y\gamma}^2\right)$$

Singh et al. (2007) considered the ratio type exponential estimators

$$\bar{y}_{3t} = \bar{y}\exp(\tau_t)$$

,

$$\tau_t = \begin{cases} \dfrac{P - p}{P + p} if\, t = 1 \\ \dfrac{p - P}{P + p} if\, t = 2 \end{cases}$$

The MSEs of the estimators obtained considering up to the first order of approximation are:

$$\mathrm{MSE}(\overline{y}_{3t}) = \begin{cases} \dfrac{\sigma_y^2}{n} + \dfrac{\overline{Y}^2\sigma_\gamma^2}{4n} - \dfrac{\overline{Y}\rho_{y\gamma}\sigma_y\sigma_\gamma}{nP} & \mathit{if}\ t = 1 \\ \dfrac{\sigma_y^2}{n} + \dfrac{\overline{Y}^2\sigma_\gamma^2}{nP} + \dfrac{\overline{Y}\rho_{y\gamma}\sigma_y\sigma_\gamma}{nP} & \mathit{if}\ t = 2 \end{cases}$$

Another suggested estimator was developed fixing a constant α :

$$\overline{y}_4 = \overline{y}[\alpha\exp(\tau_1) + (1-\alpha)\exp(\tau_2)]$$

Minimizing MSE ($\mathrm{MSE}(\overline{y}_4)$) is obtained that

$$\mathrm{Min}\left\{\mathrm{MSE}\left(\overline{y}_4\right) = \mathrm{Min}\left\{\mathrm{MSE}(\overline{y}_2)\right\}\right\} \cong \frac{\sigma_y^2}{n}\left(1 - \rho_{y\gamma}^2\right)$$

7.3 RATIO TYPE ESTIMATORS IN RSS USING γ

7.3.1 SOME BASIC ELEMENTS OF RSS

McIntyre (1952) considered ranking with respect to the prediction of the values of the variable of interest Y. Hence he considered as valid the hypothesis of having a perfect ranking of Y. By ordering in terms of the latent values, we have that the measured values of Y are indeed order statistics. Then the density function of the ith order statistic (OS) of a simple random sample (SRS) of size m, $f_{[i]} = f_{(i)}$, should be derived from distribution of Y: F. We have that from the probability density function of the OSs, for any y,

$$f(y) = \frac{1}{m}\sum\nolimits_{i=1}^{m} f_{(i)}(y).$$

This equality has an important role in RSS as it gives rise to deriving its statistical merits.

Perfect ranking with respect to the latent values of Y is consistent. When ranking errors exist, the density function of the ranked statistic with rank i is not $f_{(i)}$,but the corresponding cumulative distribution function $F_{[r]}$, which is:

$$F_{[i]} = \sum_{s=1}^{m} p_{si}F_{(s)}(y)$$

Here p_{si} denotes the probability with which the sth (numerical) order statistic is considered having the rank i.

The RSS procedure involves selecting independently m sets of m units from U. In the first set we evaluate Y in the lowest ranked unit, the remaining units of it are discarded. In the second set of m units, Y is evaluated in the second lowest ranked unit and the remaining units are discarded. The procedure is continued until the mth set is evaluated. This completes one cycle and a ranked set sample $s(1)$ of size m is obtained. The whole process can be repeated k times (cycles) and the ranked set sample of size $n = mk$ is given by the sequence $s(1),\ldots,s(k)$. In practical studies m takes values of 2, 3, or 4.

The theory considers that Y may be ranked with some error. Lynne Stokes (1977) derived the effect of ranking using concomitant variables. This fact affects the behavior of the RSS estimator by reducing the associated gain in accuracy, with respect to the SRSWR-estimator.

Let us denote by $\left(X_{j(i)}, Y_{j[i]}\right)$ the pair of the ith order statistics of X and the associated element Y in the jth cycle. Then the ranked set sampling can be explained as follows

First we select m SRS each of size m as

$$\begin{bmatrix} \{(X_{11}, Y_{11}), (X_{12}, Y_{12}), \ \ldots, \ (X_{1m}, Y_{1m})\} \\ \{(X_{21}, Y_{21}), (X_{22}, Y_{22}), \ \ldots, \ (X_{2m}, Y_{2m})\} \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ \{(X_{m1}, Y_{m1}), (X_{m2}, Y_{m2}), \ldots, (X_{mm}, Y_{mm})\} \end{bmatrix}$$

Rank the units within each set according to the variable X as

$$\begin{bmatrix} \left\{\underline{\left(X^*_{1(1)}, Y^*_{1[1]}\right)}, \left(X_{1(2)}, Y_{1[2]}\right), \ldots, \left(X_{1(m)}, Y_{1[m]}\right)\right\} \\ \left\{\left(X_{2(1)}, Y_{2[1]}\right), \underline{\left(X^*_{2(2)}, Y^*_{2[2]}\right)}, \ldots, \left(X_{2(m)}, Y_{2[m]}\right)\right\} \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ \left\{\left(X_{m(1)}, Y_{m[1]}\right), \left(X_{m(2)}, Y_{m[2]}\right), \ldots, \underline{\left(X^*_{m(m)}, Y^*_{m[m]}\right)}\right\} \end{bmatrix}.$$

Then the measured RSS units are $\left\{\left(X^*_{1(1)}, Y^*_{1[1]}\right), \left(X^*_{2(2)}, Y^*_{2[2]}\right), \ldots, \left(X^*_{m(m)}, Y^*_{m[m]}\right)\right\}$. The process is repeated k times (cycles).

If the error probabilities are the same within each cycle of a balanced RSS, we have that

$$\frac{1}{m}\sum_{i=1}^{m} F_{[i]}(y) = \frac{1}{m}\sum_{i=1}^{m} \ \sum_{s=1}^{m} p_{si} F_{(s)}(y) = F(y)$$

Then, X can be used for the ranking of the sampling units; it is measured on each unit in the selected simple random samples. The units are ranked according to the measured X values. We have induced order statistics $Y_{(i)}$. Let $f(Y|X_{(i)})(y|x)$ denote the conditional density function of Y given $X_{(i)} = x$ and $g_{(i)}(x)$ the marginal density function of $X_{(i)}$. Hence

$$f_{[i]}(y) = \int f_{Y|X_{(i)}}(yx) g_{(i)}(x)dx$$

As a result

$$f(y) = \int \frac{1}{m}\sum_{i=1}^{m} f_{Y|X_{(i)}}(y|x) g_{(i)}(x)dx = \frac{1}{m}\sum_{i=1}^{m} f_{(i)}(y)$$

Let us consider $h(y)$ as a function of y, $\mu_h = E[h(Y)]$, and the existence of $V[h(Y)] = \sigma^2_h$.

Denote $\hat{\mu}_{h,\text{rss}} = \frac{1}{mk}\sum_{i=1}^{m}\sum_{r=1}^{k} h\left(Y_{[i]r}\right)$. The relative efficiency of RSS with respect to SRS for estimating the mean of Y is based on the well-known fundamental theorem of RSS:

Theorem 1: Suppose that the ranking mechanism in RSS is consistent. Then,

i. The estimator $\hat{\mu}_{h,\text{rss}}$ is unbiased.

ii. $V\left[\hat{\mu}_{h,\mathrm{rss}}\right] \leq \frac{\sigma_h^2}{mk}$

iii. If $m \to \infty$ then $\sqrt{mk}\left(\hat{\mu}_{h,rss} - \mu_h\right) \sim N(0, V(\hat{\mu}_{h,\mathrm{rss}}))$.

7.3.2 RATIO TYPE ESTIMATORS

Extensions of ratio estimators, when RSS is used for selecting the samples, is a theme of theoretical and practical interest. The ratio estimation based on RSS usually is more efficient compared with the SRS ratio estimate. The usual SRSWR estimator was extended by Samawi and Muttlak (1996). Some modified ratio estimators have been developed. See, for example, Kadilar et al. (2009), Al-Omari and Gupta (2014), Jeelani et al. (2014a, 2014b), and Jeelani et al. (2017).

Basically, the naïve RSS ratio estimator of the mean is

$$\overline{y}_{r-\mathrm{rss}} = \overline{y}_{\mathrm{rss}}\left(\frac{\overline{X}}{\overline{x}_{\mathrm{rss}}}\right)$$

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

Consider that $g(x_1, .., x_n)$ and $q(y_1, .., y_n)$ are statistics related to the parametric functions represented by

$$t_n = T + \frac{\delta_T}{n} + \frac{\sum_{i=1}^{n} \tau_0(Z_i)}{n^2} + \frac{\sum_{i=1}^{n} \tau_1(Z_i)}{n} + \frac{\sum_{C_2^n} \tau_2(Z_i, Z_j)}{n^2} + \frac{\sum_{C_3^n} \tau_3(Z_i, Z_j, Z_k)}{n^3}$$
$$+ o_P\left(n^{-\frac{3}{2}}\right), t = g, q; T = G, Q, Z = X, Y$$

δ_T is a bias term. We have that $E(\tau_0(Z_i)) = E(\tau_1(Z_i)) = 0$; for the cross terms of second-order $E(\tau_2(Z_i, Z_j)\big|Z_i) = 0$ and for the third-order cross terms $\tau_3\left(Z_i, Z_j, Z_k\big|Z_i, Z_j\right) = 0$.

The corresponding expansion in TS of $E\left(\overline{y}_{r-\mathrm{rss}} - \overline{Y}\right)^2$, using this development, leads to the approximate expression of the MSE:

$$MSE(\overline{y}_{r-\mathrm{rss}}) \cong \frac{\left(\sigma_y^2 - \sum_{i=1}^{m} \frac{\Delta_{y(i)}^2}{m}\right) + R^2\left(\sigma_x^2 - \sum_{i=1}^{m} \frac{\Delta_{x(i)}^2}{m}\right) - 2R\rho_{xy}\left[\sqrt{\left(\sigma_x^2 - \sum_{i=1}^{m} \frac{\Delta_{x(i)}^2}{m}\right)\left(\sigma_y^2 - \sum_{i=1}^{m} \frac{\Delta_{y(i)}^2}{m}\right)}\right]}{n}$$

Then we prefer this estimator to the SRSWR one when

$$\Gamma_{r-\mathrm{rss}} = \sum_{i=1}^{m} \frac{\Delta_{y(i)}^2}{m} + R^2 \sum_{i=1}^{m} \frac{\Delta_{x(i)}^2}{m} + 2R\rho_{xy}\left[\sqrt{\left(\sigma_x^2 - \sum_{i=1}^{m} \frac{\Delta_{x(i)}^2}{m}\right)\left(\sigma_y^2 - \sum_{i=1}^{m} \frac{\Delta_{y(i)}^2}{m}\right)}\right] > 0$$

Let us look at the counterpart of $\overline{y}_1$. It is

$$\overline{y}_{1\mathrm{rss}} = \overline{y}_{\mathrm{rss}} r_p, r_p = \frac{P}{p}$$

In this case, the use of RSS is involved only with $\overline{y}_{\mathrm{rss}}$, as p does not depend on OSs. Now, the MSE of $\overline{y}_{1\mathrm{rss}}$ is given by

$$E(\overline{y}_{1\text{rss}}-\overline{Y})^2=\text{MSE}\left(\overline{y}_{1\text{rss}}\right)\cong V\left(\overline{y}_{\text{rss}}\right)+\overline{Y}^2\frac{V(p)}{P^2}\left(1-2\rho_{y\gamma}\left(P\overline{Y}\sqrt{\frac{V\left(\overline{y}_{\text{rss}}\right)}{V(p)}}\right)\right)$$

$$=\frac{\sigma_y^2-\sum_{i=1}^{m}\frac{\Delta_{y_{(i)}}^2}{m}}{n}+\overline{Y}^2\frac{\sigma_\gamma^2}{nP^2}\left(1-2\rho_{y\gamma}\left(P\overline{Y}\sqrt{\frac{\sigma_y^2-\sum_{i=1}^{m}\frac{\Delta_{y_{(i)}}^2}{m}}{\sigma_\gamma^2}}\right)\right)$$

Let us compare $\text{MSE}\left(\overline{y}_{1\text{rss}}\right)$and $\text{MSE}\left(\overline{y}_1\right)$ by computing $\text{MSE}\left(\overline{y}_1\right)-\text{MSE}\left(\overline{y}_{1\text{rss}}\right)$
This difference is approximately

$$\varGamma_{1-\text{rss}}\cong\frac{\sum_{i=1}^{m}\frac{\Delta_{y_{(i)}}^2}{m}}{n}-2\rho_{y\gamma}\frac{\overline{Y}\sigma_\gamma}{nP}\left(\left(\sqrt{\sigma_y^2}-\sqrt{\sigma_y^2-\sum_{i=1}^{m}\frac{\Delta_{y_{(i)}}^2}{m}}\right)\right)$$

Therefore RSS provides a more accurate estimator if $\Gamma_{1-\text{rss}}>0$. The term in brackets in the second term is positive. Then we have that a sufficient condition for preferring $\overline{y}_{1\text{rss}}$ is that $\rho_{y\gamma}<0$. As $\rho_{y\gamma}$ is a point biserial coefficient of correlation, and it is negative only if $\overline{Y}_\vartheta<\overline{Y}_{\overline{\vartheta}}$. On the other hand, $\Gamma_{1-\text{rss}}\cong 0$ only if, for any $i(=1,\ldots., m)$, $\Delta_{y_{(i)}}^2=0$. This is true iff the ranking is made at random.

The class of estimators $\zeta_2=\left\{\overline{y}_2\middle|\overline{y}_2=g(\overline{y},\tau);\ \tau=\frac{p}{P}\right\}$ has as RSS counterpart

$$\zeta_{2.\text{rss}}=\left\{\overline{y}_{2\text{rss}}\middle|\overline{y}_{2\text{ss}}=g\left(\overline{y}_{\text{rss}},\tau\right);\tau=\frac{p}{P}\right\}$$

We may use the same parametric function $g(a, b)$ in $\zeta_{2.\text{rss}}$ and the optimal estimator is

$$\overline{y}_{2\text{rss,opt}}=\overline{y}_{\text{rss}}+b(P-p), b=\hat{\beta}, \beta=\frac{\text{Cov}(y,p)}{V(p)}=\rho_{y\gamma}\left(\frac{\sigma_y}{\sigma_\gamma}\right)$$

as the minimum MSE equals

$$\text{Min}\left\{\text{MSE}(\overline{y}_{2\text{rss,opt}})\right\}\cong\frac{\sigma_y^2-\sum_{i=1}^{m}\frac{\Delta_{y_{(i)}}^2}{m}}{n}\left(1-\rho_{y\gamma}^2\right)$$

Now the gain in accuracy is

$$\varGamma_{2-\text{rss,opt}}\cong\sum_{i=1}^{m}\frac{\Delta_{y_{(i)}}^2}{mn}\left(1-\rho_{y\gamma}^2\right)$$

This expression is positive unless $\Delta_{y_{(i)}}^2=0$ for any $i=(1,.., m)$, or if the correlation between Y and γ is perfect.

Let us develop the RSS counterparts of $\overline{y}_{3t}$. They are:

$$\overline{y}_{3t-\text{rss}}=\overline{y}_{\text{rss}}\exp(\tau_t),$$

$$\tau_t = \begin{cases} \dfrac{P-p}{P+p}\text{if}t=1 \\ \dfrac{p-P}{P+p}\text{if}t=2 \end{cases}$$

The corresponding approximations to the MSEs of these estimators are easily obtained. They are given by:

$$\text{MSE}(\overline{y}_{3-\text{rss},t}) = \begin{cases} \dfrac{\sigma_y^2}{n} - \displaystyle\sum_{i=1}^{m} \dfrac{\Delta_{y(i)}^2}{nm} + \dfrac{\overline{Y}^2\sigma_\gamma^2}{4n} - \dfrac{\overline{Y}\rho_{y\gamma}\sigma_\gamma\sqrt{\left(\sigma_y^2 - \sum_{i=1}^{m}\dfrac{\Delta_{y(i)}^2}{m}\right)}}{nP}, \text{ if } t=1 \\ \dfrac{\sigma_y^2}{n} - \displaystyle\sum_{i=1}^{m} \dfrac{\Delta_{y(i)}^2}{nm} + \dfrac{\overline{Y}^2\sigma_\gamma^2}{4n} + \dfrac{\overline{Y}\rho_{y\gamma}\sigma_\gamma\sqrt{\left(\sigma_y^2 - \sum_{i=1}^{m}\dfrac{\Delta_{y(i)}^2}{m}\right)}}{nP} \text{ if } t=2 \end{cases}$$

The gain in accuracy differs seriously. Note that

$$\mathit{\Gamma}_{3-\text{rss},1} = \sum_{i=1}^{m} \frac{\Delta_{y(i)}^2}{nm} + \frac{\overline{Y}\rho_{y\gamma}\sigma_\gamma\sqrt{\left(\sigma_y^2 - \sum_{i=1}^{m}\frac{\Delta_{y(i)}^2}{m}\right)}}{nP}$$

This expression is always positive if $\rho_{y\gamma} > 0$, but for $t = 2$ we have that

$$\mathit{\Gamma}_{3-\text{rss},2} = \sum_{i=1}^{m} \frac{\Delta_{y(i)}^2}{nm} - \frac{\overline{Y}\rho_{y\gamma}\sigma_\gamma\sqrt{\left(\sigma_y^2 - \sum_{i=1}^{m}\frac{\Delta_{y(i)}^2}{m}\right)}}{nP}$$

is always positive if $\rho_{y\gamma} < 0$. Therefore the surveyor will prefer one of them after considering the sign of the correlation coefficient.

The analysis of $\overline{y}_4$ yields as RSS-estimator

$$\overline{y}_{4-\text{rss}} = \overline{y}_{\text{rss}}[\alpha\exp(\tau_1) + (1-\alpha)\exp(\tau_2)]$$

Its MSE is minimized as in the case of $\overline{y}_2 - rss$ and the MSEs are equal

$$\text{Min}\left\{\text{MSE}\left(\overline{y}_{4-\text{rss}}\right) = \text{Min}\left\{\text{MSE}(\overline{y}_{2-\text{ss}})\right\}\right\} \cong \left(\frac{\sigma_y^2}{n} - \sum_{i=1}^{m}\frac{\Delta_{y(i)}^2}{nm}\right)\left(1-\rho_{y\gamma}^2\right)$$

Therefore

$$\mathit{\Gamma}_{4-\text{rss}} = \sum_{i=1}^{m}\frac{\Delta_{y(i)}^2}{nm}\left(1-\rho_{y\gamma}^2\right)$$

7.4 A NUMERICAL STUDY OF THE EFFECT OF A VACCINE FOR LUNG CANCER

Many medical institutions are developing top-level research for the evaluation of so-called personalized medicine. The mainstream is to look for adequate vaccines for improving the quality of life of terminal lung cancer patients. The development of such vaccines is on the front line of research and development. Clinical facts have shown that target therapies recurrently do not meet their primary endpoint in open population analysis, but they showed certain benefits in some patients. New treatments must be validated in terms of their behavior in improving a quality of life index, which must be sensitive to changes. Life quality depends on several variables, most of which are categorical.

We obtained data on the evaluation of the success of a new product (vaccine) during 3 years. It was applied to 132 lung cancer patients in a terminal status. Their life expectancy was 3–6 months. A treatment with the new vaccine was applied and the improvement of survival time, Y, was measured.

The ranking variable X was the volume of the tumor when the treatment began. Different categorical variables from their expedients were used as markers. They were:

1. Being a smoker (yes, no);
2. Sex (male, female);
3. Being diabetic (yes, no);
4. Being more than 60-year-old (yes, no);
5. Anemic (yes, no).

Using the population information, the MSEs were computed and the gain in accuracy, diminution of the MSE, was analyzed. We measured the gain in percent. The measures were:

$$\omega_{r-\mathrm{rss}} = \frac{\Gamma_{r-\mathrm{rss}}}{\mathrm{MSE}\left(\overline{y}_r\right)}, \omega_{1-\mathrm{rss}} = \frac{\Gamma_{1-\mathrm{rss}}}{\mathrm{MSE}\left(\overline{y}_1\right)} \omega_{2-\mathrm{rss,opt}} = \frac{\Gamma_{2-\mathrm{rss,opt}}}{\mathrm{Min}\left\{\mathrm{MSE}\left(\overline{y}_2\right)\right\}}$$

$$\omega_{3-\mathrm{rss},t} = \frac{\left(\Gamma_{3-\mathrm{rss},t}\right)}{\mathrm{MSE}\left(\overline{y}_{3,t}\right)}, t = 1, 2 \frac{\Gamma_{4-rss}}{\mathrm{Min}\left\{\mathrm{MSE}\left(\overline{y}_4\right)\right\}}$$

Simulation experiments for evaluating the behavior of the RSS-estimators studied in this paper were conducted. RSS samples of size 40 were selected using the combinations m, k (=2, 20; 4, 10; 5, 8).

One of the aspects of the behavior of the estimators was their accuracy. The difference between the true value of the population mean and the computed estimators was analyzed. One thousand samples were generated randomly and we evaluated

$$D_S = \frac{1}{1000}\sum_{b=1}^{1000}\left|\overline{y}_S - \overline{YY}\right|_b, S = 1-\mathrm{rss}, 2-\mathrm{rssopt}, 3-\mathrm{rss1}, 3-\mathrm{rss2}, 4-\mathrm{rss}$$

The analysis of the behavior of the RSS estimators appear in the tables presented below.

The results evidence that using $\overline{y}_{r-\mathrm{rss}}$ is generally the best alternative. On many occasions, using an attribute seems to be more satisfactory for the clients. Hence, the statistician may opt for evaluating the effect of using an attribute, and fix which properties make one of them preferred. P and $\rho_{y\gamma}$ are involved in the formula of the MSEs, and hence in the gain in accuracy.

The results in Table 7.1 establish that $\bar{y}_{1\mathrm{rss}}$ obtains the largest diminution in MSE, but $\bar{y}_{2\mathrm{rss,opt}}$ and $\bar{y}_{4-rss}$ are more accurate. Note that P may be considered large, while the correlation is negative and notably different from zero. These facts have as an effect preferring $\bar{y}_{3-\mathrm{rss},2}$ to $\bar{y}_{3-\mathrm{rss},1}$.

Table 7.2 gives support to considering again $\bar{y}_{1\mathrm{rss}}$ as having the best behavior in terms of the percent of gain in accuracy. It is the second best in terms of the average difference of the estimates with the population mean. P is not considerably high. $\rho_{y\gamma}$ is negative but closer to zero than in the case of using smoking as an attribute. We may argue that also in the case of not too high a correlation $\bar{y}_{3-\mathrm{rss},2}$ is to be preferred to $\bar{y}_{3-\mathrm{rss},1}$. The estimates produced by $\bar{y}_{2\mathrm{rss,opt}}$ and $\bar{y}_{4-\mathrm{rss}}$ are the closest to the population mean.

Table 7.3 presents the study of the use of being diabetic as an attribute. $\bar{y}_{1\mathrm{rss}}$ has one of the largest improvements of the percent gain in accuracy, but the average of the difference, of the corresponding estimates with $\bar{Y}$, is not notable. P may not be considered as high and $\rho_{y\gamma}$ is almost equal to zero. Also, in this case $\bar{y}_{3-\mathrm{rss},2}$ is to be preferred to $\bar{y}_{3-\mathrm{rss},1}$. Note that $\bar{y}_{2\mathrm{rss,opt}}$ and $\bar{y}_{4-\mathrm{rss}}$ have the second best accuracy in terms of the mean difference of the estimates and the population mean.

Table 7.1 Performance of the RSS-Estimators. $\gamma = 1$ If Being a Smoker, $P = 0.86$, $\rho_{y\gamma} = -0.72$

S	Percent of Gain in Accuracy			Mean Difference		
	$m = 2$	$m = 4$	$m = 5$	$m = 2$	$m = 4$	$m = 5$
r-rss	11.25	9.97	8.42	4.78	4.84	5.04
1-rss	20.03	19.95	19.78	7.51	7.01	6.78
2-rssopt	7.54	7.28	7.27	3.67	3.14	3.02
3-rss1	0.25	0.16	0.15	7.88	7.66	7.09
3-rss2	2.80	2.36	2.35	7.53	7.50	7.48
4-rss	7.54	7.28	7.27	3.67	3.14	3.02

Table 7.2 Performance of the RSS-Estimators. $\gamma = 1$ if Male, $P = 0.69$, $\rho_{y\gamma} = -0.35$

S	Percent of Gain in Accuracy			Mean Difference		
	$m = 2$	$m = 4$	$m = 5$	$m = 2$	$m = 4$	$m = 5$
r-rss	11.25	9.97	8.42	4.78	4.84	5.04
1-rss	11.71	11.63	11.44	7.35	7.26	7.09
2-rssopt	7.54	7.28	7.27	3.67	3.14	3.02
3-rss1	1.48	1.16	1.15	7.80	7.73	7.68
3-rss2	2.66	2.59	2.45	7.44	7.38	7.34
4-rss	7.54	7.28	7.27	3.67	3.14	3.02

Table 7.3 Performance of the RSS-Estimators. $\gamma = 1$ If Diabetic, $P = 0.47$, $\rho_{y\gamma} = -0.08$

S	Percent of Gain in Accuracy			Mean Difference		
	m **= 2**	***m*** **= 4**	***m*** **= 5**	***m*** **= 2**	***m*** **= 4**	***m*** **= 5**
r-rss	11.25	9.97	8.42	4.78	4.84	5.04
1-rss	8.78	8.72	8.64	7.30	7.22	7.05
2-rssopt	7.59	7.50	7.47	3.17	3.08	2.98
3-rss1	1.25	1.14	1.07	7.73	7.70	7.57
3-rss2	4.21	4.17	4.05	7.40	7.35	7.22
4-rss	7.59	7.50	7.47	3.17	3.08	2.98

Table 7.4 Performance of the RSS-Estimators. $\gamma = 1$ If More Than 40 Year Old, $P = 0.79$, $\rho_{y\gamma} = 0.13$

S	Percent of Gain in Accuracy			Mean Difference		
	m **= 2**	***m*** **= 4**	***m*** **= 5**	***m*** **= 2**	***m*** **= 4**	***m*** **= 5**
r-rss	11.25	9.97	8.42	4.78	4.84	5.04
1-rss	3.53	3.42	3.34	7.32	7.28	7.16
2-rssopt	7.54	7.28	7.27	3.67	3.14	3.02
3-rss1	7.02	6.97	6.87	7.73	7.70	7.56
3-rss2	2.15	2.10	2.01	7.40	7.35	7.20
4-rss	7.54	7.28	7.27	3.67	3.14	3.02

Table 7.5 Performance of the RSS-Estimators. $\gamma = 1$ if anemic, $P = 0.92$, $\rho_{y\gamma} = -0.84$

S	Percent of Gain in Accuracy			Mean Difference		
	m **= 2**	***m*** **= 4**	***m*** **= 5**	***m*** **= 2**	***m*** **= 4**	***m*** **= 5**
r-rss	11.25	9.97	8.42	4.78	4.84	5.04
1-rss	6.56	6.52	6.39	7.07	6.97	6.86
2-rssopt	7.50	7.30	7.29	3.55	3.50	3.46
3-rss1	2.12	2.02	1.97	7.75	7.70	7.64
3-rss2	6.00	5.93	5.84	7.50	7.42	7.38
4-rss	7.50	7.30	7.29	3.55	3.50	3.46

Table 7.4 permits to evaluate the effect of having a positive correlation. It is not so high but the performance of $\bar{y}_{3-\mathrm{rss},2}$, $\bar{y}_{3-\mathrm{rss},1}$. $\bar{y}_{2\mathrm{rss,opt}}$ and $\bar{y}_{4-\mathrm{rss}}$ is relatively better, though P is large. $\bar{y}_{2\mathrm{rss,opt}}$ and $\bar{y}_{4-\mathrm{rss}}$ produced estimates very close to $\bar{Y}$.

When having anemia is the attribute (Table 7.5), P is close to 1 and $\rho_{y\gamma}$ may be considered as close to -1. Now the performance of $\bar{y}_{3-\mathrm{rss},2}$ is much better than that exhibited by $\bar{y}_{3-\mathrm{rss},1}$ in terms of the percent of gain in accuracy. Their mean differences are very similar. $\overline{y}_{2\text{rss,opt}}$ and $\overline{y}_{4-\text{rss}}$ are relatively better though P is large. $\overline{y}_{2\text{rss,opt}}$ and $\overline{y}_{4-\text{rss}}$ have the second best behavior in both measures.

ACKNOWLEDGMENTS

Bouza-Herrera is indebted for the support received from the projects iMODA, Métodos Cuantittivos para Estudios Médicos, Medio-ambientales y de Bienestar Human,o and CYTED 516RT0513.

REFERENCES

Al-Saleh, M.F., Al-Kadiri, M.A., 2000. Double-ranked set sampling. Stat. Prob. Lett. 48 (2), 205–212.

Al-Odat, M.T., Al-Saleh, M.F., 2001. A variation of ranked set sampling. J. Appl. Stat. Sci. 10 (2), 137–146.

Al-Saleh, M.F., Al-Omari, A.I., 2002. Multistage ranked set sampling. J. Stat. Plan. Inference 102 (2), 273–286.

Al-Omari, A.I., 2012. Ratio estimation of the population mean using auxiliary information in simple random sampling and median ranked set sampling. Stat. Probab. Lett. 82 (11), 1883–1890.

Al-Omari, A.I., Bouza, C.N., 2014. Review of ranked set sampling: modifications and applications. Rev. Investig. Oper. 3, 215–240.

Al-Omari, A.I., Gupta, S., 2014. Double quartile ranked set sampling for estimating population ratio using auxiliary information. Pak. J. Stat. 30 (4).

Al-Omari, A.I., Jemain, A.A., Kamarulzaman, I., 2009. New ratio estimators of the mean using simple random sampling and ranked set sampling methods. Rev. Investig. Oper. 30, 97–108.

Al-Omari, A.I., Bouza, C.N., Covarrubias, D., Pal, R., 2016. A new estimator of the population mean: an application to bioleaching studies. J. Mod. Appl. Stat. Methods 15 (2), 9.

Bouza, C.N., 2005. Sampling using ranked sets: concepts, results and perspectives. Rev. Investig. Oper. 26 (3), 275–292.

Cochran, W.G., 1977. Sampling Techniques, 3rd ed. Wiley Eastern Limited.

Dell, T.R., Clutter, J.L., 1972. Ranked set sampling theory with order statistics background, Biometrics, 28. pp. 545–555.

Ganeslingam, S., Ganesh, S., 2006. Ranked set sampling versus simple random sampling in the estimation of the mean and the ratio. J. Stat. Manag. Syst. 9 (2), 459–472.

Hedayat, A.S., Sinha, B.K., 1992. Design and Inference in Finite Population Sampling.. Wiley, New York.

Herrera, C.N.B., Al-Omari, A.I., 2011. Ranked set estimation with imputation of the missing observations: the median estimator. Rev. Investig. Oper. 32 (1), 30–37.

Jozani, M.J., Johnson, B.C., 2011. Design based estimation for ranked set sampling in finite populations. Environ. Ecol. Stat. 18 (4), 663–685.

Jeelani, M.I., Maqbool, S., Mir, S.A., Khan, I., Nazir, N., Jeelani, F., 2013. Modified ratio estimators of population mean using linear combination of co-efficient of kurtosis and quartile deviation. Int. J. Mod. Math. Sci. 8 (3), 149–153.

Jeelani, M.I., Mir, S.A., Nazir, N., Jeelani, F., 2014a. Modified ratio estimators using linear combination of co-efficient of skewness and median of auxiliary variable under rank set sampling and simple random sampling. Ind. J. Sci. Technol. 7 (5), 722–727.

Jeelani, M.I., Mir, S.A., Pukhta, M.S., 2014b. A class of modified ratio estimators using linear combination of quartile deviation and median of auxiliary variable under rank set sampling. Univ. J. Appl. Math. 2 (6), 245–249.

Jeelani, M.I., Mir, S.A., Maqbool, S., Khan, I., Singh, K.N., Zaffer, G., et al., 2014c. Role of rank set sampling in improving the estimates of population mean under stratification. Am. J. Math. Stat. 4 (1), 46–49.

Jeelani, M.I., Mir, S.A., Khan, I., Nazir, N., Jeelani, F., 2014d. Non-response problems in ranked set sampling. Pak. J. Stat. 30 (4), 555–562.

Jeelani, M.I., Bouza, C.N., 2015. New ratio method of estimation under ranked set sampling. Rev. Investig. Oper. 36, 151–155.

Jeelani, M.I., Bouza, C.N., Sharma, M., 2017. Modified ratio estimator under rank set sampling. Investig. Oper. 38 (1), 103–107.

Jhajj, H.S., Sharma, M.K., Grover, L.K., 2006. A family of estimators of population mean using information on auxiliary attribute. Pak. J. Stat. 22 (1), 43–50.

Kadilar, C., Cingi, H., 2004. Ratio estimators in simple random sampling. Appl. Math. Comput. 151, 893–902.

Kadilar, C., Unyazici, Y., Cingi, H., 2009. Ratio estimator for the population mean using ranked. Stat. Papers 50, 301–309.

Khan, L., Shabbir, J., 2016. Improved ratio-type estimators of population mean in ranked set sampling using two concomitant variables. Pak. J. Stat. Oper. Res. 12 (3).

McIntyre, G.A., 1952. A method of unbiased selective sampling using ranked sets. Aust. J. Agric. Res. 3, 385–390.

Murthy, M.N., 1967. Sampling Theory and Methods. Statistical Publishing Society, Calcutta, India.

Muttlak, H.A., 1997. Median ranked set sampling. J. Appl. Stat. Sci. 6 (4), 245–255.

Muttlak, H.A., 1998. Median ranked set sampling with concomitant variables and a comparison with ranked set sampling and regression estimators. Environmetrics 9 (3), 255–267.

Naik, V.D., Gupta, P.C., 1996. A note on estimation of mean with known population proportion of an auxiliary character. J. Ind. Soc. Agric. Stat. 48 (2), 151–158.

Ohyama, T., Doi, J.A., Yanagawa, T., 2008. Estimating population characteristics by incorporating prior values in stratified random sampling/ranked set sampling. J. Stat. Plan. Inference 138 (12), 4021–4032.

Prasad, B., 1989. Some improved ratio type estimators of population mean and ratio in finite population sample surveys. Commun. Stat. Theory Methods 18, 379–392.

Philip, L.H., Lam, K., 1997. Regression estimator in ranked set sampling. Biometrics 1070–1080.

Patil, G.P., Surucu, B., Egemen, D., 2002. Ranked set sampling. Encyclopedia of Environmetrics .

Samawi, H.M., Muttlak, H.A., 1996. Estimation of ratio using rank set sampling. Biom. J. 38, 753–764.

Samawi, H.M., Ahmed, M.S., Abu-Dayyeh, W., 1996. Estimating the population mean using extreme ranked set sampling. Biom. J. 38 (5), 577–586.

Shabbir, J., Gupta, S., 2007. On estimating the finite population mean with known population proportion of an auxiliary variable. Pak. J. Stat. 23 (1), 1–9.

Sharma, P., Singh, R., Kim, J.M., 2013. Study of some improved ratio type estimators using information on auxiliary attributes under second order approximation. J. Sci. Res. 57, 138–146.

Singh, R., Cauhan, P., Sawan, N., Smarandache, F., 2007. Auxiliary Information and a Priori Values in Construction of Improved Estimators.. Renaissance High Press, USA.

Singh, R., Chauhan, P., Sawan, N., Smarandache, F., 2008. Ratio estimators in simple random sampling using information on auxiliary attribute. Pak. J. Stat. Oper. Res. 4, 47–53.

Singh, H.P., Solanki, R.S., 2012. Improved estimation of population mean in simple random sampling using information on auxiliary attribute. Appl. Math. Comp. 218, 7798–7812.

Singh, H.P., Tailor, R., Singh, S., 2014. General procedure for estimating the population mean using ranked set sampling. J. Stat. Comput. Simul. 84 (5), 931–945.

Lynne Stokes, S., 1977. Ranked set sampling with concomitant variables. Commun. Stat.: Theory Methods 6 (12), 1207–1211.

Takahasi, K., Wakimoto, K., 1968. On unbiased estimates of the population mean based on the stratified sampling by means of ordering. Ann. Inst. Stat. Math. 20, 1–31.

Wolfe, D.A., 2004. Ranked set sampling: an approach to more efficient data collection. Stat. Sci. 19 (4), 636–643.

Yadav, S.K., Adewara, A.A., 2013. On improved estimation of population mean using qualitative auxiliary information. Mat. Theory Model. 3 (11), 42–50.

FURTHER READING

Bouza, C., 2001. Model assisted ranked survey sampling. Biom. J. 43, 249–259.

CHAPTER

MODIFIED PARTIALLY ORDERED JUDGMENT SUBSET SAMPLING SCHEMES

8

Abdul Haq
Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan

8.1 INTRODUCTION

Cost-effective sampling schemes are of major concern in surveys of natural resources in biology, ecology, environmental management, forestry, etc. One of the most commonly used sampling schemes is simple random sampling (SRS). In environmental, ecological, and biomedical studies, there are situations where taking the actual measurement of sample observations is not only difficult, but also costly, destructive, and time-consuming. However, ranking a small set of sample observations is relatively cheap, easy, and reliable. Ranking of the experimental units may be accomplished through a visual inspection with respect to the study variable or by using any less-expensive method or using ranks of a highly correlated auxiliary variable. For example, if one in interested in estimating the average height of a plant species in a forest, then, a small set of randomly selected plants can be ranked visually with respect to their heights or weights. Likewise, an ecological assessment of the hazardous waste sites involves expensive radiochemical techniques to find the value of the study variable. The hazardous waste sites with different levels of contamination, however, could be ranked by a visual inspection of soil discoloration. In all such situations, McIntyre (1952) proposed a sampling scheme—later called ranked set sampling (RSS)—that could be employed as an efficient alternative to SRS. The RSS scheme incorporates inexpensive auxiliary information related to the study variable as a way of gathering additional information in order to rank the selected sampling units. This use of the auxiliary information at the sampling stage helps in selecting better representative samples from the target population.

Takahasi and Wakimoto (1968) were the first to lay the mathematical foundation of the RSS scheme. They proved that the mean of a ranked set sample is not only an unbiased estimator of the population mean but it is also more precise than the sample mean of a simple random sample. An interesting finding was put forward by Dell and Clutter (1972); they showed that, despite the presence of ranking errors, the mean estimator with RSS is not only unbiased but it also outperforms the mean estimator with SRS. For a brief introduction, bibliography, literature review, applications, and monograph on RSS, readers are referred to Patil (1995), Patil et al. (1999), Wolfe (2012), and Chen (2007), respectively.

Ranked Set Sampling. DOI: https://doi.org/10.1016/B978-0-12-815044-3.00008-3

In the last few decades, there have been many new advancements and variations in the classical RSS scheme. Samawi et al. (1996) and Muttlak (1996, 1997, 2003) introduced extreme RSS (ERSS), paired RSS (PRSS), median RSS (MRSS), and quartile RSS (QRSS) schemes, respectively, for estimating the population mean. The ERSS, MRSS, and QRSS are called unbalanced RSS schemes because these schemes select units on some ranks more frequently than the others. An RSS scheme is called balanced if units on all ranks are selected an equal number of times. The unbalanced RSS schemes may provide efficient mean estimators when sampling from a symmetric population, but efficiency of the mean estimator may depend on the modality (unimodal, bimodal, or multimodal) of the underlying population (cf., Kaur et al., 1997; Ozturk and Wolfe, 2000). For an asymmetric population, however, these mean estimators are not precise and in some cases they may get worse than the mean estimator with SRS. Al-Saleh and Al-Kadiri (2000) introduced double RSS (DRSS) for estimating the population mean. They proved mathematically that the mean estimator with DRSS is always more efficient than the mean estimator with RSS. Al-Naseer (2007) suggested an L RSS (LRSS) scheme for estimating the population mean based on the ideal of L moments. This scheme encompasses RSS, MRSS, and QRSS schemes. A simple modification of LRSS has been suggested by Al-Omari and Raqab (2013), named truncation-based RSS (TBRSS), for estimating the population mean. Both RSS and ERSS schemes are special cases of TBRSS. Haq et al. (2014) suggested mixed RSS for estimating the population mean. The mixed RSS is a suitable mixture of SRS and RSS schemes. For some more recent works on RSS scheme, we refer to Haq et al. (2013, 2015, 2016a,b) and Haq (2017a,b), and the references cited therein.

In practice, when conducting an RSS scheme, the ranker is forced to rank all units from the smallest to the largest without actual measurement, this may not be realistic in certain settings when the ranker lacks in confidence to rank all the selected units accurately. Ozturk (2011) came up with a wonderful idea that, instead of ranking units, it may be possible to rank the tied-ranked units. It is more realistic for a ranker to rank all units in a set by allowing ties among the units when their ranks cannot be identified with full confidence. Following these ideas, Ozturk (2011) suggested a partially ordered judgment subset sampling (POJSS) scheme for estimating the population mean. It was shown that, under perfect ranking and with reasonable assumptions on the partitioning of sets, the mean estimator with POJSS surpasses the mean estimator with RSS.

In this chapter, we extend the work on the POJSS scheme and propose new modified POJSS schemes for efficiently estimating the population mean. Using the ideas of PRSS, LRSS, and DRSS, we propose paired POJSS (PPOJSS), L POJSS (LPOJSS), and ranked POJSS (RPOJSS) schemes. The mathematical properties of the mean estimators under these sampling schemes are derived. It turns out that the proposed schemes with both perfect and imperfect rankings are efficient alternatives to their existing counterparts in terms of providing more precise mean estimators.

The rest of this chapter is outlined as follows: in Section 8.2, some existing RSS schemes are briefly reviewed. The modified POJSS schemes are presented in Section 8.3. Under perfect and imperfect rankings, the mean estimators with the existing and proposed sampling schemes are compared theoretically and numerically in Section 8.4. A real data example is considered in Section 8.5. Section 8.6 summarizes the main findings and concludes the chapter.

8.2 SAMPLING SCHEMES

In this section, some recent and existing RSS schemes are briefly reviewed, along with their mathematical setups when estimating the population mean.

Let $(Y_1, \ldots, Y_n)$ denote a simple random sample of size n drawn from an absolutely continuous distribution having the cumulative distribution function (CDF) $F(y)$ and the probability density function (PDF) $f(y)$, with the mean μ_Y and the variance σ_Y^2. Let $\overline{Y}_{\text{SRS}} = (1/n)\sum_{r=1}^{n} Y_r$ be the sample mean based on a simple random sample of size n. Here, $\overline{Y}_{\text{SRS}}$ is an unbiased estimator of μ_Y, i.e., $E(\overline{Y}_{\text{SRS}}) = \mu_Y$, with variance $\text{Var}(\overline{Y}_{\text{SRS}}) = (1/n)\sigma_Y^2$. Let $(Y_{(1:n)}, \ldots, Y_{(n:n)})$ denote the order statistics corresponding to $(Y_1, \ldots, Y_n)$, where $Y_{(r:n)} = r\text{th} \min\{Y_1, \ldots, Y_n\}$ for $r = 1, \ldots, n$. The CDF and PDF of $Y_{(r:n)}$ $(1 \le r \le n)$ are, respectively, given by

$$F_{(r:n)}(y) = \sum_{i=r}^{n} \binom{n}{i} \{F(y)\}^i \{1 - F(y)\}^{n-i}, \ -\infty < y < \infty,$$

$$f_{(r:n)}(y) = \frac{n!}{(r-1)!(n-r)!} \{F(y)\}^{r-1} \{1 - F(y)\}^{n-r} f(y).$$

The mean and variance of $Y_{(r:n)}$ $(1 \le r \le n)$ are

$$\mu_{Y(r:n)} = \int y f_{(r:n)}(y)\text{d}y \quad \text{and} \quad \sigma^2_{Y(r:n)} = \int (y - \mu_{Y(r:n)})^2 f_{(r:n)}(y)\text{d}y,$$

respectively. Similarly, the covariance between $Y_{(r:n)}$ and $Y_{(s:n)}$ $(1 \le r < s \le n)$ is

$$\sigma_{Y(r,s:n)} = \int\int \left(y_r - \mu_{Y(r:n)}\right)\left(y_s - \mu_{Y(s:n)}\right) f_{(r,s:n)}(y_r, y_s)\text{d}y_r \, \text{d}y_s,$$

where

$$f_{(r,s:n)}(y_r, y_s) = \frac{n!}{(r-1)!(s-r-1)!(n-s)!} \left\{F(y_r)\right\}^{r-1} \left\{F(y_s) - F(y_r)\right\}^{s-r-1}$$
$$\left\{1 - F(y_s)\right\}^{n-s} f(y_r) f(y_s), -\infty < y_r < y_s < \infty,$$

which is the joint PDF of $Y_{(r:n)}$ and $Y_{(s:n)}$. The joint CDF of $Y_{(r:n)}$ and $Y_{(s:n)}$ is

$$F_{(r,s:n)}(y_r, y_s) = \int_{-\infty}^{y_s} \int_{-\infty}^{y_r} f_{(r,s:n)}(y_r, y_s)\text{d}y_r \text{d}y_s.$$

These results will be used in Section 8.3. More details on the order statistics may be seen in David and Nagaraja (2003).

8.2.1 RANKED SET SAMPLING

The RSS scheme is an efficient alternative to the SRS scheme in those sampling situations where a small set of selected units can be ranked visually with respect to the study variable or by using the ranks of an auxiliary variable.

The RSS scheme works as follows: select a simple random sample of size m^2 units from the underlying population. These m^2 units are then partitioned into m sets, each set comprising m units. The ranking of units within each set is accomplished through a visual inspection and/or personal judgment with respect to the study variable or the units can be ranked using any less-expensive method—the ranks of the study variable could be judged using the ranks of a highly correlated auxiliary variable. Then, from the rth set, the rth smallest ranked unit is quantified, for $r = 1, ..., m$. This is one complete cycle of a ranked set sample of size m. The whole procedure can be repeated t times to get t cycles of a ranked set sample of size m with total sample size $n = mt$ units.

Let $(Y_{11j}, \ldots, Y_{1mj}), \ldots, (Y_{m1j}, \ldots, Y_{mmj})$ denote m simple random samples, each of size m, obtained in the jth cycle for $j = 1, \ldots, t$. Apply the RSS scheme on these samples to get a ranked set sample of size m for the jth cycle, denoted by $Y_{r(r:m)j}$, $r = 1, \ldots, m$ for $j = 1, \ldots, t$, where $Y_{r(r:m)j} = r\text{th} \min\{Y_{r1j}, \ldots, Y_{rmj}\}$. It is to be noted that, having fixed r, $Y_{r(r:m)j}$, $j = 1, ..., t$, are independent and identically distributed (IID) random variables, i.e., $Y_{r(r:m)j} \equiv Y_{(r:m)}$, $j = 1, \ldots, t$. Having fixed j, however, $Y_{r(r:m)j}$, $r = 1, \ldots, m$, are independent but not identically distributed (INID) random variables, i.e., $Y_{r(r:m)j} \equiv Y_{(r:m)}$, $r = 1, \ldots, m$. The sample mean and its variance under RSS are

$$\overline{Y}_{\text{RSS}} = \frac{1}{n} \sum_{j=1}^{t} \sum_{r=1}^{m} Y_{r(r:m)j} \quad \text{and} \quad \text{Var}(\overline{Y}_{\text{RSS}}) = \frac{1}{nm} \sum_{r=1}^{m} \sigma^2_{Y(r:m)},$$

respectively. Takahasi and Wakimoto (1968) showed that $\overline{Y}_{\text{RSS}}$ is an unbiased estimator of μ_Y, and it is more precise than $\overline{Y}_{\text{SRS}}$, i.e.,

$$\text{Var}(\overline{Y}_{\text{RSS}}) = \text{Var}(\overline{Y}_{\text{SRS}}) - \frac{1}{nm} \sum_{r=1}^{m} \left(\mu_{Y(r:m)} - \mu_Y\right)^2.$$

8.2.2 PAIRED RANKED SET SAMPLING

The PRSS scheme was first suggested by Muttlak (1996) for estimating the population mean. The PRSS scheme is a cost-efficient alternative to the RSS scheme, i.e., it requires fewer observations than the RSS scheme when selecting a sample from the underlying population—thus it helps in reducing the ranking cost.

The PRSS scheme works as follows: for the even set size m, select $m(m/2)$ units from the underlying population and partition them into $m/2$ sets, each comprising m units. Now rank the units within each set. Then select the rth and $(m - r + 1)$th smallest ranked units from the rth set, for $r = 1, \ldots, m/2$. Similarly, for the odd set size m, select $m(m + 1)/2$ units from the underlying population and partition them into $(m + 1)/2$ sets, each comprising m units. Then select the rth and $(m - r + 1)$th smallest ranked units from the rth set, for $r = 1, \ldots, (m - 1)/2$, and the $\{(m + 1)/2$th smallest ranked unit is selected from the $\{(m + 1)/2$th set. This completes one cycle of a paired ranked set sample of size m. The whole procedure could be repeated t times to get a total sample of size n units.

The sample means under PRSS depending upon even and odd set sizes m are, respectively, given by

$$\overline{Y}^{\text{E}}_{\text{PRSS}} = \frac{1}{n} \sum_{j=1}^{t} \left(\sum_{r=1}^{m/2} Y_{r(r:m)j} + \sum_{r=1}^{m/2} Y_{r(m-r+1:m)j} \right) \text{and}$$

$$\overline{Y}^{\text{O}}_{\text{PRSS}} = \frac{1}{n} \sum_{j=1}^{t} \left(\sum_{r=1}^{(m+1)/2} Y_{r(r:m)j} + \sum_{r=1}^{(m-1)/2} Y_{r(m-r+1:m)j} \right)$$

with variances

$$\mathrm{Var}(\overline{Y}^{\mathrm{E}}_{\mathrm{PRSS}}) = \mathrm{Var}(\overline{Y}_{\mathrm{RSS}}) + \frac{2}{nm}\sum_{r=1}^{m/2}\sigma_{Y(r,m-r+1:m)} \text{ and}$$

$$\mathrm{Var}(\overline{Y}^{\mathrm{O}}_{\mathrm{PRSS}}) = \mathrm{Var}(\overline{Y}_{\mathrm{RSS}}) + \frac{2}{nm}\sum_{r=1}^{(m-1)/2}\sigma_{Y(r,m-r+1:m)}.$$

It is clear that the mean estimator based on RSS is always more precise than the mean estimator based on PRSS—all covariances in the above expressions are always positive. However, the ranking cost associated with PRSS is less than that of RSS [cf., Muttlak, 1996].

8.2.3 L RANKED SET SAMPLING

The LRSS scheme was suggested by Al-Naseer (2007) for estimating the population mean. The LRSS encompasses some existing RSS schemes, and it is an efficient alternative to the RSS scheme when estimating the mean of a symmetric population. Here, we modify LRSS so that ERSS and TBRSS can be made its special cases. Note that the modified LRSS procedure is here referred to as LRSS.

The LRSS scheme works as follows: select the LRSS coefficient, say $k = [\alpha m]$ for $0 \le \alpha < 0.5$, where $[\cdot]$ is the largest integer value less than or equal to $(\cdot)$. Identify m^2 units from the underlying population and partition them into m sets, each comprising m units. Now rank the units within each set. Then select the vth and $(m - v + 1)$th smallest ranked units from the first and last k sets, respectively, where $v \in 1,\ldots,[\mathrm{m}/2]$. Moreover, the rth smallest ranked unit is selected from the rth set, for $r = k + 1,\ldots, m - k$. This completes one cycle of an L ranked set sample of size m. The whole procedure could be repeated t times to get a total sample of size n units. For different choices of k the LRSS reduces to balanced and unbalanced RSS schemes. For example, LRSS is equivalent to RSS and MRSS with $k = 0$ and $k = [(m + 1)/2], v = k + 1$, respectively. Note that, when $v = k + 1$, the above modified LRSS reduces to LRSS suggested by Al-Naseer (2007).

The sample mean and its variance under LRSS are, respectively, given by

$$\overline{Y}_{\mathrm{LRSS}} = \frac{1}{n}\sum_{j=1}^{t}\left(\sum_{r=1}^{k} Y_{r(v:m)j} + \sum_{r=k+1}^{m-k} Y_{r(r:m)j} + \sum_{r=m-k+1}^{m} Y_{r(m-v+1:m)j}\right),$$

$$\mathrm{Var}(\overline{Y}_{\mathrm{LRSS}}) = \frac{1}{nm}\left(\sum_{r=1}^{k}\sigma^2_{Y(v:m)} + \sum_{r=k+1}^{m-k}\sigma^2_{Y(r:m)} + \sum_{r=m-k+1}^{m}\sigma^2_{Y(m-v+1:m)}\right).$$

It is easy to show that $\overline{Y}_{\mathrm{LRSS}}$ is an unbiased estimator of μ_Y and it is more precise than $\overline{Y}_{\mathrm{RSS}}$ when the underlying population is symmetric. For an asymmetric population, however, it is biased and may become less efficient than $\overline{Y}_{\mathrm{SRS}}$ and $\overline{Y}_{\mathrm{RSS}}$ [cf., Al-Naseer, 2007].

8.2.4 DOUBLE-RANKED SET SAMPLING

The DRSS scheme was suggested by Al-Saleh and Al-Kadiri (2000) for estimating the population mean. They showed that DRSS is always more efficient than RSS when estimating the population mean. Moreover, when conducting the DRSS scheme, ranking the units on the second stage is much easier than that on the first stage—this makes DRSS an efficient alternative to RSS.

The DRSS scheme works as follows: identify m^3 units from the underlying population and partition them into m sets, each comprising m^2 units. The RSS scheme is then applied on each set to get m ranked set samples, each of size m units. Again apply the RSS scheme to get a double-ranked set sample of size m. This completes one cycle of a double-ranked set sample of size m. The whole procedure can be repeated t times to get a total sample of size n units.

Let $Y_{r(r:m)j}^{(r)(r:m)}$, $r = 1,\ldots,m$, denote a double-ranked set sample of size m for the jth cycle, $j = 1,\ldots,t$, where $Y_{r(r:m)j}^{(r)(r:m)} =$ rth min of the rth ranked set sample in the rth set $(Y_{1(1:m)j}^{(r)},\ldots,Y_{m(m:m)j}^{(r)})$ in the jth cycle. Al-Saleh and Al-Kadiri (2000) showed that the sample mean with DRSS is not only unbiased, it is also more precise than the sample mean with RSS, i.e.,

$$E(\overline{Y}_{\text{DRSS}}) = \frac{1}{n}\sum_{j=1}^{t}\sum_{r=1}^{m} E(Y_{r(r:m)j}^{r(r:m)}) = \mu_Y \text{ and}$$

$$\text{Var}(\overline{Y}_{\text{DRSS}}) = \text{Var}(\overline{Y}_{\text{RSS}}) - \frac{1}{nm}\sum_{r \neq s}^{m} \sigma_{Y(r,s:m)}^{(r,s:m)},$$

where $\sigma_{Y(r,s:m)}^{(r,s:m)} > 0$ is the covariance between $Y_{r(r:m)j}^{(r)(r:m)}$ and $Y_{s(s:m)j}^{(r)(s:m)}$. More details on DRSS may be seen in Al-Saleh and Al-Kadiri (2000).

8.2.5 PARTIALLY ORDERED JUDGMENT SUBSET SAMPLING

A new sampling scheme has been introduced by Ozturk (2011), in which a ranker is allowed to declare ties among the units within subsets of prefixed sizes. The units within these subsets are partially ranked ordered so that any unit in subset i possesses a smaller rank than any other unit in subset i', where $i < i'$, called partially ordered judgment subsets. A single observation is then quantified from one of these subsets present in a set. This sampling scheme is named POJSS. Ozturk (2011) further imposed some restrictions on the number of units within each subset—comprising the whole set—in order to increase the precision of the mean estimator based on POJSS, i.e., all subsets within a set should comprise equal number of units.

The POJSS scheme works as follows: identify wm^2 units from the underlying population and partition them into m sets, each comprising wm units. The units within each set are further partitioned into m subsets, each comprising w units. These subsets are then partially ranked ordered as mentioned by Ozturk (2011). Select one unit from the rth smallest ranked subset of the rth set, for $r = 1, \ldots, m$. This completes one cycle of a partially ordered judgment subset sample of size m. The whole procedure could be repeated t times to get a total sample of size n units.

Symbolically; under perfect ranking, let $(S_{r(1)j}^{(w)},\ldots,S_{r(m)j}^{(w)})$ denote the rth set that comprises m partially ordered judgment subsets, each of size w, $r = 1, \ldots, m$, in the jth cycle, for $j = 1,\ldots,t$, where $S_{r(r)j}^{(w)} = (Y_{r((r-1)w+1:mw)j},\ldots,Y_{r(rw:mw)j})$ for $w = 1,\ldots,m$, i.e., each subset contains w elements. Let $Y_{r(r:m)j}^{*} =$ Select one element randomly from $S_{r(r)j}^{(w)}$. Having fixed w, this constitutes one complete sample of size m for the jth cycle. Clearly, having fixed j, $Y_{r(r:m)j}^{*}$, $r = 1,\ldots,m$, are INID random variables. However, having fixed r, $Y_{r(r:m)j}^{*}$, $j = 1,\ldots,t$, are IID random variables. For brevity of discussion, let $Y_{r(r:m)j}^{*} \equiv Y_{(r:m)}^{*}$ for $j = 1,\ldots,t$.

The sample mean and its variance under POJSS are, respectively, given by

$$\overline{Y}_{\text{POJSS}} = \frac{1}{n}\sum_{j=1}^{t}\sum_{r=1}^{m} Y^{*}_{r(r:m)j} \quad \text{and} \quad \text{Var}(\overline{Y}_{\text{POJSS}}) = \frac{1}{nm}\sum_{r=1}^{m}\sigma^{*2}_{Y(r:m)},$$

where $\text{Var}(Y^{*}_{r(r:m)j}) = \sigma^{*2}_{Y(r:m)}$. Note that RSS is a special case of POJSS when $w = 1$. Ozturk (2011) showed that POJSS is more precise than RSS when estimating the population mean when $w > 1$. For more details, we refer to Ozturk (2011).

8.3 PROPOSED SAMPLING SCHEMES

In this section, some modified POJSS schemes are proposed for estimating the population mean. We develop unbiased estimators of the population mean under the proposed sampling schemes and study their mathematical properties. Moreover, the unbiased estimators of the variances of these mean estimators are also derived.

8.3.1 PAIRED PARTIALLY ORDERED JUDGMENT SUBSET SAMPLING

As aforementioned, PRSS is a cost-effective alternative to RSS, i.e., it requires less number of ranked units than that using RSS when selecting a sample from the underlying population. On similar lines, we modify POJSS to propose the PPOJSS scheme for estimating the population mean. The PPOJSS is a cost-effective alternative to POJSS.

The PPOJSS scheme works as follows: for an even set size m, identify $wm(m/2)$ units from the underlying population and partition them into $m/2$ sets, each comprising m units. The units within each set are further partitioned into m subsets, each comprising w units. These subsets are then partially ranked ordered. Select one unit from the rth and one from the $(m - r + 1)$th smallest ranked subsets of the rth set, for $r = 1,\ldots, m/2$. Similarly, for an odd set size m, identify $wm(m + 1)/2$ units from the underlying population and partition them into $(m + 1)/2$ sets, each comprising m units. The units within each set are further partitioned into m subsets, each comprising w units. These subsets are then partially ranked ordered. Then select one unit from the rth and one from the $(m - r + 1)$th smallest ranked subsets of the rth set, for $r = 1,\ldots,(m - 1)/2$, and select one unit from the $\{(m + 1)/2\}$th smallest ranked subset of the $\{(m + 1)/2\}$th set. This completes one cycle of a paired partially ordered judgment subset sample of size m. The whole procedure can be repeated t times to get a total sample of size n units.

The sample means under PPOJSS depending upon even and odd set sizes m are, respectively, given by

$$\overline{Y}^{\text{E}}_{\text{PPOJSS}} = \frac{1}{n}\sum_{j=1}^{t}\left(\sum_{r=1}^{m/2} Y^{*}_{r(r:m)j} + \sum_{r=1}^{m/2} Y^{*}_{r(m-r+1:m)j}\right) \text{ and}$$

$$\overline{Y}^{\text{O}}_{\text{PPOJSS}} = \frac{1}{n}\sum_{j=1}^{t}\left(\sum_{r=1}^{(m+1)/2} Y^{*}_{r(r:m)j} + \sum_{r=1}^{(m-1)/2} Y^{*}_{r(m-r+1:m)j}\right)$$

with variances

adksf

$$\mathrm{Var}(\overline{Y}^{\mathrm{E}}_{\mathrm{PPOJSS}}) = \mathrm{Var}(\overline{Y}_{\mathrm{POJSS}}) + \frac{2}{nm}\sum_{r=1}^{m/2} \sigma^*_{Y(r,m-r+1:m)} \text{ and} \tag{8.1}$$

$$\mathrm{Var}(\overline{Y}^{\mathrm{O}}_{\mathrm{PPOJSS}}) = \mathrm{Var}(\overline{Y}_{\mathrm{POJSS}}) + \frac{2}{nm}\sum_{r=1}^{(m-1)/2} \sigma^*_{Y(r,m-r+1:m)}, \tag{8.2}$$

where $\sigma^*_{Y(r,m-r+1:m)} > 0$ is the covariance between $Y^*_{r(r:m)j}$ and $Y^*_{r(m-r+1:m)j}$. Similar to POJSS, the mean estimators with PPOJSS also turn out to be unbiased. Moreover, as expected, these mean estimators can never be more precise than the mean estimator with POJSS. However, the ranking cost of PPOJSS is less than that of POJSS. Thus, it is more economical and practical to employ the PPOJSS scheme when ranking costs are high or constrained by budgets or it may not be possible to use POJSS with full confidence.

The following proposition helps in computing the variances and covariances of the random variables under PPOJSS.

Proposition 1. *Having fixed w,*

i. *the CDF of* $Y^*_{(r:m)}$ $(1 \le r \le m)$ *is*

$$F^*_{(r:m)}(y) = \frac{1}{w}\sum_{i=(r-1)w+1}^{rw} F_{(i:wm)}(y). \tag{8.3}$$

ii. *the joint CDF of* $Y^*_{(r:m)}$ *and* $Y^*_{(s:m)}$ $(1 \le r < s \le m)$ *is*

$$F^*_{(r,s:m)}(y_r, y_s) = \frac{1}{w^2}\sum_{i=(r-1)w+1}^{rw}\;\sum_{j=(s-1)w+1}^{sw} F_{(i,j:mw)}(y_r, y_s). \tag{8.4}$$

The proofs of (i) and (ii) are trivial.

The mean and variance of $Y^*_{(r:m)}$ are, respectively, given by

$$\mu^*_{Y(r:m)} = \int y f^*_{(r:m)}(y)\mathrm{d}y \quad \text{and} \quad \sigma^{*2}_{Y(r:m)} = \int (y - \mu^*_{Y(r:m)})^2 f^*_{(r:m)}(y)\mathrm{d}y,$$

where $f^*_{(r:m)}(y) = (\mathrm{d}/\mathrm{d}y)F^*_{(r:m)}(y)$. Similarly, the covariance between $Y^*_{r(r:m)j}$ and $Y^*_{r(s:m)j}$ is

$$\sigma^*_{Y(r,s:m)} = \int\int \left(y_r - \mu^*_{Y(r:m)}\right)\left(y_s - \mu^*_{Y(s:m)}\right) f^*_{(r,s:m)}(y_r, y_s)\mathrm{d}y_r\mathrm{d}y_s,$$

where $f^*_{(r,s:m)}(y_r, y_s) = (\mathrm{d}^2/\mathrm{d}y_s\mathrm{d}y_r)F^*_{(r,s:m)}(y_r, y_s)$. Using Eqs. (8.3) and (8.4), variances of the mean estimators given in Eqs. (8.1) and (8.2) can be easily computed.

Lemma 1. *Based on even and odd set sizes, the unbiased estimators of* $Var(\overline{Y}^{E}_{PPOJSS})$ *and* $Var(\overline{Y}^{O}_{PPOJSS})$ *are*

$$\hat{V}ar(\overline{Y}^{E}_{PPOJSS}) = \frac{1}{2nmt(t-1)}\sum_{j \neq j'}^{t}\left\{\sum_{r=1}^{m}(Y^*_{r(r:m)j} - Y^*_{r(r:m)j'})^2 \right.$$
$$\left. + 2\sum_{r=1}^{m/2}(Y^*_{r(r:m)j} - Y^*_{r(r:m)j'})(Y^*_{r(m-r+1:m)j} - Y^*_{r(m-r+1:m)j'})\right\}$$

and

$$\hat{V}ar(\overline{Y}^{O}_{PPOJSS}) = \frac{1}{2nmt(t-1)} \sum_{j\neq j'}^{t} \left\{ \sum_{r=1}^{m} \left(Y^*_{r(r:m)j} - Y^*_{r(r:m)j'} \right)^2 \right.$$

$$\left. + 2\sum_{r=1}^{(m-1)/2} (Y^*_{r(r:m)j} - Y^*_{r(r:m)j'})(Y^*_{r(m-r+1:m)j} - Y^*_{r(m-r+1:m)j'}) \right\},$$

respectively.

The proof is trivial.

8.3.2 L PARTIALLY ORDERED JUDGMENT SUBSET SAMPLING

As pointed out by Al-Naseer (2007), the LRSS scheme helps in selecting a more representative sample from a symmetric population (except uniform) than that using RSS, i.e., the mean estimator based on LRSS turns out to be more efficient than that based on RSS. On similar lines, in order to increase the efficiency of the POJSS-based mean estimator when sampling from a symmetric population, we propose an LPOJSS scheme for efficiently estimating the population mean.

The LPOJSS scheme works as follows: select the LPOJSS coefficient, say $k = [\alpha m]$. Identify wm^2 units from the underlying population and partition them into m sets, each comprising m units. The units within each set are further partitioned into m subsets, each of size w units. These subsets are then partially ranked ordered. Select one unit from the vth and one unit from the $(m-v+1)$th smallest ranked subsets of the first and last k sets, respectively, where $v \in 1,\ldots,[\mathrm{m}/2]$. Moreover, select one unit from the rth smallest ranked subset of the rth set, for $r = k+1,\ldots, m-k$. This completes one cycle of an L partially ordered judgment subset sample of size m. The whole procedure could be repeated t times to get t cycles with a total sample of size n units. Note that, given w and m, with different choices of k and v, several POJSS schemes could be constructed.

The sample mean and its variance under LPOJSS are, respectively, given by

$$\overline{Y}_{\text{LPOJSS}} = \frac{1}{n}\sum_{j=1}^{t}\left(\sum_{r=1}^{k} Y^*_{r(v:m)j} + \sum_{r=k+1}^{m-k} Y^*_{r(r:m)j} + \sum_{r=m-k+1}^{m} Y^*_{r(m-v+1:m)j}\right),$$

$$\text{Var}(\overline{Y}_{\text{LPOJSS}}) = \frac{1}{nm}\left(k(\sigma^{*2}_{Y(v:m)} + \sigma^{*2}_{Y(m-v+1:m)}) + \sum_{r=k+1}^{m-k}\sigma^{*2}_{Y(r:m)}\right).$$

Proposition 2. *For a symmetric population,* $(1 \le r \le m)$

i. $\mu^*_{Y(r:m)} + \mu^*_{Y(m-r+1:m)} = 2\mu_Y,$

ii. $\sigma^{*2}_{Y(r:m)} = \sigma^{*2}_{Y(m-r+1:m)}.$

Proof. To prove (i), using Eq. (8.3), we have

$$\begin{aligned}
\mu^*_{Y(r:m)} + \mu^*_{Y(m-r+1:m)} &= \frac{1}{w}\sum_{i=(r-1)w+1}^{rw} \mu_{Y(i:wm)} + \frac{1}{w}\sum_{i=(m-r)w+1}^{(m-r+1)w} \mu_{Y(i:wm)} \\
&= \frac{1}{w}\left(\mu_{Y((r-1)w+1:wm)} + \cdots + \mu_{Y(rw:wm)}\right) \\
&\quad + \frac{1}{w}\left(\mu_{Y((m-r)w+1:wm)} + \cdots + \mu_{Y((m-r+1)w:wm)}\right) \\
&= \frac{1}{w}\left\{\left(\mu_{Y((r-1)w+1:wm)} + \mu_{Y((m-r+1)w:wm)}\right) + \cdots + \left(\mu_{Y(rw:wm)} + \mu_{Y((m-r)w+1:wm)}\right)\right\}.
\end{aligned}$$

For a symmetric distribution, it is well known that $\mu_{Y(i:wm)} + \mu_{Y(wm-i+1:m)} = 2\mu_Y$, for $i = 1, \ldots, wm$. Using this result, we get

$$\mu^*_{Y(r:m)} + \mu^*_{Y(m-r+1:m)} = \frac{1}{w}(2\mu_Y + \cdots + 2\mu_Y) = 2\mu_Y.$$

To prove (ii), using Eq. (8.3), we have

$$\sigma^{*2}_{Y(r:m)} = \frac{1}{w}\sum_{i=(r-1)w+1}^{rw}\left(\mu^2_{Y(i:wm)} + \sigma^2_{Y(i:wm)}\right) - \left(\frac{1}{w}\sum_{i=(r-1)w+1}^{rw}\mu_{Y(r:wm)}\right)^2, \tag{8.5}$$

$$\sigma^{*2}_{Y(m-r+1:m)} = \frac{1}{w}\sum_{i=(m-r)w+1}^{(m-r+1)w}\left(\mu^2_{Y(i:wm)} + \sigma^2_{Y(i:wm)}\right) - \left(\frac{1}{w}\sum_{i=(m-r)w+1}^{(m-r+1)w}\mu_{Y(r:wm)}\right)^2. \tag{8.6}$$

Equate Eqs. (8.5) and (8.6), use symmetry relation of mean, to get

$$\sum_{i=(r-1)w+1}^{rw}\sigma^2_{Y(i:wm)} = \sum_{i=(m-r)w+1}^{(m-r+1)w}\sigma^2_{Y(i:wm)},$$

which always holds true for a symmetric distribution since $\sigma^2_{Y(i:wm)} = \sigma^2_{Y(mw-i+1:wm)}$, for $i = 1, \ldots, wm$.

Lemma 2. *For a symmetric population,*

i. $E(\overline{Y}_{LPOJSS}) = \mu_Y$.

ii. $Var(\overline{Y}_{LPOJSS}) \leq Var(\overline{Y}_{POJSS})$ *when* $\sum_{r=1}^{k}\sigma^{*2}_{Y(r:m)} \geq k\sigma^{*2}_{Y(v:m)}$.

iii. *An unbiased estimator of* $Var(\overline{Y}_{LPOJSS})$ *is*

$$\begin{aligned}
\hat{V}ar(\overline{Y}_{LPOJSS}) = \frac{1}{2nmt(t-1)}\sum_{j\neq j'}^{t}\Bigg[&k\left\{\left(Y^*_{r(v:m)j} - Y^*_{r(v:m)j'}\right)^2 + \left(Y^*_{r(m-v+1:m)j} - Y^*_{r(m-v+1:m)j'}\right)^2\right\} \\
&+ 2\sum_{r=1}^{m/2}\left(Y^*_{r(r:m)j} - Y^*_{r(r:m)j'}\right)\left(Y^*_{r(m-r+1:m)j} - Y^*_{r(m-r+1:m)j'}\right)\Bigg].
\end{aligned}$$

Proof. To prove (i), consider the expectation:

$$E(\overline{Y}_{LPOJSS}) = \frac{1}{m}\left(k(\mu^*_{Y(v:m)} + \mu^*_{Y(m-v+1:m)}) + \sum_{r=1}^{m}\mu^*_{Y(r:m)} - \sum_{r=1}^{k}\mu^*_{Y(r:m)} - \sum_{r=m-k+1}^{m}\mu^*_{Y(r:m)}\right)$$
$$= \frac{1}{m}\left(2k\mu_Y + (m-2k)\mu_Y\right) = \mu_Y,$$

using (i) symmetry relation of proposition 2 and (ii) identity of lemma 2.

To prove (ii), we have

$$\mathrm{Var}(\overline{Y}_{\mathrm{LPOJSS}}) = \frac{1}{nm}\left\{\sum_{r=1}^{m}\sigma^{*2}_{(r:m)} + k(\sigma^{*2}_{Y(v:m)} + \sigma^{*2}_{Y(m-v+1:m)}) \quad - \sum_{r=1}^{k}\sigma^{*2}_{Y(r:m)} - \sum_{r=m-k+1}^{m}\sigma^{*2}_{Y(r:m)}\right\}$$
$$= \mathrm{Var}(\overline{Y}_{\mathrm{POJSS}}) - \frac{2}{nm}\left(\sum_{r=1}^{k}\sigma^{*2}_{Y(r:m)} - k\sigma^{*2}_{Y(v:m)}\right),$$

using (ii) symmetry relation of proposition 2. We conjecture that the condition $\sum_{r=1}^{k}\sigma^{*2}_{Y(r:m)} \geq k\sigma^{*2}_{Y(v:m)}$ holds true for nonuniform (unimodal) distributions when $v = k+1$ and for the uniform distribution when $v = 1$ [cf., Ozturk and Wolfe, 2000]. The proof of (iii) is trivial.

In the case of an asymmetric population, $\overline{Y}_{\mathrm{LPOJSS}}$ is a biased estimator of μ_Y. The mean-squared error (MSE) of $\overline{Y}_{\mathrm{LPOJSS}}$ is

$$\mathrm{MSE}(\overline{Y}_{\mathrm{LPOJSS}}) = \mathrm{Var}(\overline{Y}_{\mathrm{LPOJSS}}) + \{E(\overline{Y}_{\mathrm{LPOJSS}}) - \mu_Y\}^2.$$

In Section 8.4, it is observed that LPOJSS leads to biased and imprecise estimates of the population mean when sampling from an asymmetric population. For a symmetric population, however, the mean estimates with LPOJSS are not only unbiased but more precise too.

8.3.3 RANKED PARTIALLY ORDERED JUDGMENT SUBSET SAMPLING

As figured out by Al-Saleh and Al-Kadiri (2000), the DRSS scheme helps in selecting a more representative sample than that using RSS, i.e., the mean estimator based on DRSS is always more efficient than that based on RSS. On similar lines, in order to increase the efficiency of the POJSS-based mean estimator, we propose an RPOJSS scheme for efficiently estimating the population mean.

The RPOJSS scheme works as follows: identify wm^3 units from the underlying population and partition them into m sets, each comprising wm^2 units. The POJSS scheme is then applied on each set to get m partially ordered judgment subset samples, each of size m units. Now apply the RSS scheme to get a ranked partially ordered judgment subset sample of size m. This completes one cycle of a ranked partially ordered judgment subset sample of size m. The whole procedure could be repeated t times to get a total sample of size n units. Note that DRSS is a special case of RPOJSS when $w = 1$.

Symbolically, under perfect ranking, let $(Y^{*(r)}_{1(1:m)j}, \ldots, Y^{*(r)}_{m(m:m)j})$ denote a partially ordered judgment subset sample of size m obtained from the rth set, $r = 1, \ldots, m$. Let $Y^{*(r)(r:m)}_{r(r:m)j} = r$th min of $(Y^{*(r)}_{1(1:m)j}, \ldots, Y^{*(r)}_{m(m:m)j})$, $r = 1, \ldots, m$, $j = 1, \ldots, t$, which represent a ranked partially ordered judgment subset sample of size n. Clearly, having fixed j, $Y^{*(r)(r:m)}_{r(r:m)j}$, $r = 1,\ldots,m$, are INID random variables. However, having fixed r, $Y^{*(r)(r:m)}_{r(r:m)j}$, $j = 1,\ldots,t$, are IID random variables. For brevity of discussion, let $Y^{*(r)(r:m)}_{r(r:m)j} \equiv Y^{*(r:m)}_{(r:m)}$, $j = 1,\ldots,t$.

The sample mean and its variance under RPOJSS are, respectively, given by

$$\overline{Y}_{\text{RPOJSS}} = \frac{1}{n}\sum_{j=1}^{t}\sum_{r=1}^{m} Y^{*(r)(r:m)}_{r(r:m)j} \text{ and } \operatorname{Var}(\overline{Y}_{\text{RPOJSS}}) = \frac{1}{nm}\sum_{r=1}^{m}\sigma^{*2(r:m)}_{Y(r:m)},$$

where $\operatorname{Var}(Y^{*(r)(r:m)}_{r(r:m)j}) = \operatorname{Var}(Y^{*(r:m)}_{(r:m)}) = \sigma^{*2(r:m)}_{Y(r:m)}$.

Let $F^{*(r:m)}_{(r:m)}(y)$, $f^{*(r:m)}_{(r:m)}(y)$, and $\mu^{*(r:m)}_{Y(r:m)}$ be the CDF, PDF, and mean of $Y^{*(r:m)}_{(r:m)}$, respectively. The identities in the following lemma are an analogue to those of Al-Saleh and Al-Kadiri (2000) for DRSS.

Lemma 3. *For any population,*

i. $f(y) = (1/m)\sum_{r=1}^{m} f^{*}_{(r:m)}(y) = (1/m)\sum_{r=1}^{m} f^{*(r:m)}_{(r:m)}(y)$.

ii. $\mu_Y = (1/m)\sum_{r=1}^{m} \mu^{*}_{Y(r:m)} = (1/m)\sum_{r=1}^{m} \mu^{*(r:m)}_{Y(r:m)}$.

iii. $\sigma^2_Y = (1/m)\sum_{r=1}^{m} \sigma^{*2(r:m)}_{Y(r:m)} + (1/m)\sum_{r=1}^{m} (\mu^{*(r:m)}_{Y(r:m)} - \mu_Y)^2$.

Proof. To prove (i), let us consider

$$Q = \sum_{r=1}^{m} Q_r \quad \text{and} \quad Q_r = \begin{cases} 1 & \text{if } Y^{*}_{(r:m)} \le y \\ 0 & \text{otherwise} \end{cases}.$$

Then

$$E(Q) = \sum_{r=1}^{m} F^{*}_{(r:m)}(y) = \sum_{r=1}^{m}\frac{1}{w}\sum_{i=(r-1)w+1}^{rw} F_{(i:wm)}(y) = mF(y),$$

by Takahasi and Wakimoto (1968).

Similarly, we can write

$$\begin{aligned}\sum_{r=1}^{m} F^{*(r:m)}_{(r:m)}(y) &= \sum_{r=1}^{m} P\left(Y^{*(r:m)}_{(r:m)} \le y\right) \\ &= \sum_{r=1}^{m} P\left(\text{at least } r \text{ of}(Y^{*}_{(1:m)},\ldots,Y^{*}_{(m:m)}) \le y\right) \\ &= \sum_{r=1}^{m} P(Q \ge r) = E(Q) = mF(y)\end{aligned}$$

and hence (i) follows. On similar lines, (ii) and (iii) could be proved.

As aforementioned, $Y^{*}_{(r:m)}$, $r = 1,\ldots,m$, are INID random variable and we consider order statistics from this sample to get $Y^{*(r:m)}_{(r:m)}$, $r = 1,\ldots,m$. In order to calculate the mean, variances, covariances of these random variables, the permanent function is used.

Let $\boldsymbol{A} = ((a_{i,j}))$ be a square matrix of order m. Then the permanent of $\boldsymbol{A}$ is

$$\text{Per}(\boldsymbol{A}) = \sum_P \sum_{j=1}^{m} a_{j,i_j},$$

where $\sum_P$ denotes the sum of over all $m!$ permutations $(i_1, \ldots, i_m)$ of $(1,\ldots, m)$. The definition of the permanent is very much similar to that of the determinant except that in the permanent we do not have the alternating sign whether the permutation is of even or odd order. For more details on this function, refer to Bapat and Beg (1989).

Following Vaughan and Venables (1972), the CDF of $Y^{*(r:m)}_{(r:m)}$ $(1 \le r \le m)$ is

$$F^{*(r:m)}_{(r:m)}(y) = \sum_{i=r}^{m} \frac{1}{i!(m-i)!}\text{Per}(\boldsymbol{A}_1), \qquad -\infty < y < \infty,$$

where

$$\boldsymbol{A}_1 = \begin{pmatrix} F^*_{(1:m)}(y) & F^*_{(2:m)}(y) & \cdots & F^*_{(m:m)}(y) \\ 1 - F^*_{(1:m)}(y) & 1 - F^*_{(2:m)}(y) & \cdots & 1 - F^*_{(m:m)}(y) \end{pmatrix} \begin{matrix} \}i \\ \}m-i \end{matrix},$$

where the first row is repeated i times and the second row is repeated $m - i$ times.

Similarly, the PDF of $Y^{*(r:m)}_{(r:m)}$ $(1 \le r \le m)$ is

$$f^{*(r:m)}_{(r:m)}(y) = \frac{1}{(r-1)!(m-r)!}\text{Per}(\boldsymbol{A}_2), \qquad -\infty < y < \infty, \tag{8.7}$$

where

$$\boldsymbol{A}_2 = \begin{pmatrix} F^*_{(1:m)}(y) & F^*_{(2:m)}(y) & \cdots & F^*_{(m:m)}(y) \\ f^*_{(1:m)}(y) & f^*_{(2:m)}(y) & \cdots & f^*_{(m:m)}(y) \\ 1 - F^*_{(1:m)}(y) & 1 - F^*_{(2:m)}(y) & \cdots & 1 - F^*_{(m:m)}(y) \end{pmatrix} \begin{matrix} \}r-1 \\ \}1 \\ \}m-r \end{matrix}. \tag{8.8}$$

Proceeding similarly, the joint CDF of $Y^{*(r:m)}_{(r:m)}$ and $Y^{*(s:m)}_{(s:m)}$ $(1 \le r < s \le m)$ is

$$f^{*(r,s:m)}_{(r,s:m)}(y_r, y_s) = \frac{1}{(r-1)!(s-r-1)!(m-s)!}\text{Per}(\boldsymbol{A}_3), \qquad -\infty < y_r < y_s < \infty,$$

where

$$\boldsymbol{A}_3 = \begin{pmatrix} F^*_{(1:m)}(y_r) & F^*_{(2:m)}(y_r) & \cdots & F^*_{(m:m)}(y_r) \\ f^*_{(1:m)}(y_r) & f^*_{(2:m)}(y_r) & \cdots & f^*_{(m:m)}(y_r) \\ F^*_{(1:m)}(y_s) - F^*_{(1:m)}(y_r) & F^*_{(2:m)}(y_s) - F^*_{(2:m)}(y_r) & \cdots & F^*_{(m:m)}(y_s) - F^*_{(m:m)}(y_r) \\ f^*_{(1:m)}(y_s) & f^*_{(2:m)}(y_s) & \cdots & f^*_{(m:m)}(y_s) \\ 1 - F^*_{(1:m)}(y_s) & 1 - F^*_{(2:m)}(y_s) & \cdots & 1 - F^*_{(m:m)}(y_s) \end{pmatrix} \begin{matrix} \}r-1 \\ \}1 \\ \}s-r-1 \\ \}1 \\ \}m-s \end{matrix}.$$

From Eqs. (8.7) and (8.8),

$$\mu^{*(r:m)}_{Y(r:m)} = \int y f^{*(r:m)}_{(r:m)}(y)\mathrm{d}y,$$

$$\sigma^{*2(r:m)}_{Y(r:m)} = \int \left(y - \mu^{*(r:m)}_{(r:m)}\right)^2 f^{*(r:m)}_{(r:m)}(y)\mathrm{d}y,$$

$$\sigma^{*(r,s:m)}_{Y(r,s:m)} = \int\int \left(y_s - \mu^{*(s:m)}_{(s:m)}\right)\left(y_r - \mu^{*(r:m)}_{(r:m)}\right) f^{*(r,s:m)}_{(r,s:m)}(y_r, y_s)\mathrm{d}y_r\mathrm{d}y_s.$$

Now it is shown that the mean estimator based on RPOJSS is not only unbiased but it is also more precise than the mean estimators based on SRS and POJSS schemes.

Lemma 4. *For any population,*

i. $E(\overline{Y}_{RPOJSS}) = \mu_Y$.
ii. $Var(\overline{Y}_{RPOJSS}) \leq Var(\overline{Y}_{SRS})$.
iii. $Var(\overline{Y}_{RPOJSS}) \leq Var(\overline{Y}_{POJSS})$.
iv. *An unbiased estimator of* $Var(\overline{Y}_{RPOJSS})$ *is*

$$\hat{V}ar(\overline{Y}_{RPOJSS}) = \frac{1}{2nmt(t-1)} \sum_{j \neq j'}^{t} \sum_{r=1}^{m} (Y^{*(r:m)}_{r(r:m)j} - Y^{*(r:m)}_{r(r:m)j'})^2.$$

Proof. Here, (i) and (ii) can be easily proved using (i) and (ii) of Lemma 2.

To prove (iii), we have

$$\mathrm{Var}(\overline{Y}_{\mathrm{POJSS}}) = \mathrm{Var}\left(\frac{1}{n}\sum_{j=1}^{t}\sum_{r=1}^{m} Y^{*}_{r(r:m)j}\right) = \frac{1}{nm}\mathrm{Var}\left(\sum_{r=1}^{m} Y^{*(r:m)}_{(r:m)}\right)$$

$$= \frac{1}{nm}\sum_{r=1}^{m}\mathrm{Var}\left(Y^{*(r:m)}_{(r:m)}\right) + \frac{1}{nm}\sum_{r \neq s}^{m}\mathrm{Cov}\left(Y^{*(r:m)}_{(r:m)}, Y^{*(s:m)}_{(s:m)}\right)$$

$$= \frac{1}{nm}\sum_{r=1}^{m}\sigma^{*2(r:m)}_{Y(r:m)} + \frac{1}{nm}\sum_{r \neq s}^{m}\sigma^{*(r,s:m)}_{Y(r,s:m)}$$

$$= \mathrm{Var}(\overline{Y}_{\mathrm{RPOJSS}}) + \frac{1}{nm}\sum_{r \neq s}^{m}\sigma^{*(r,s:m)}_{Y(r,s:m)},$$

where $\sigma^{*(r,s:m)}_{Y(r,s:m)} > 0$ is the covariance between $Y^{*(r:m)}_{(r:m)}$ and $Y^{*(s:m)}_{(s:m)}$. The proof of (iv) is trivial.

Based on the above formulas of the mean estimators under different sampling schemes, the relative efficiencies (REs) of mean estimators can be computed. The efficiency of a mean estimator, say $\overline{Y}_{\mathrm{D}}$, $(\mathrm{D} = \mathrm{PRSS}, \mathrm{RSS}, \mathrm{etc.})$, relative to $\overline{Y}_{\mathrm{SRS}}$ is given by

$$\mathrm{RE}(\overline{Y}_{\mathrm{D}}, \overline{Y}_{\mathrm{SRS}}) = \frac{\mathrm{Var}(\overline{Y}_{\mathrm{SRS}})}{\mathrm{MSE}(\overline{Y}_{\mathrm{D}})}.$$

For an unbiased estimator, the MSE is replaced by the variance.

Remark 1. *Note that in the case of imperfect rankings, the round brackets in the above estimators are replaced by square brackets to denote the judgment ranks of the order statistics. For example, replace* $(r{:}m)$ *in either subscript or superscript by* $[r{:}m]$ *when there are errors in the ranking procedure.*

8.4 EFFICIENCY COMPARISONS

In this section, we compare performances of the mean estimators under perfect and imperfect rankings in terms of their REs.

For a fair comparison of the mean estimators, both symmetric and asymmetric probability distributions are considered. These distributions include uniform $U(0, 1)$, normal $N(0, 1)$, exponential $G(1, 1)$, and gamma $G(5, 1)$ distributions. For brevity of discussion, consider different values of m and w with $t = 1$. Using the mathematical formulas presented in the previous section, REs of the mean estimators are computed for the considered distributions and are displayed in Table 8.1. For some sampling schemes, w, $k_,$ and v are the schemes' parameters. Thus, these schemes are abbreviated with their parameter choices in the tables. For example, given w, PPOJSS, POJSS, and RPOJSS are referred to as PPOJSS (w), POJSS (w), and RPOJSS (w), respectively. Similarly, given w, $v_,$ and k, LRSS and LPOJSS are referred to as LRSS (v, k) and LPOJSS $(v, k; w)$, respectively.

From Table 8.1, it is observed that all REs are greater than one—thus showing that the mean estimates with the proposed sampling schemes are more precise than those with SRS. Having fixed m, the REs increase as the value of w increases, and vice versa. The LPOJSS scheme provides most efficient mean estimates when $w = 3$, $v = 1$ for uniform distribution and $w = 3$, $v = k + 1$ for normal distribution. With $v = 1$ and $v = k + 1$, we usually do not recommend the use of LPOJSS when estimating the mean of an asymmetric distribution. But, given m, with $k = 1$ and $v = k + 1$, the mean estimates with this scheme outperform those with the existing schemes, when sampling from an asymmetric distribution. Interestingly, the mean estimates with PPOJSS are more precise than those with PRSS and RSS. When it is possible to ignore the ranking cost, RPOJSS provides better mean estimates than those with POJSS.

In usual practice, when using an RSS or POJSS scheme, the experimenter is forced to rank large set sizes for a greater efficiency of a mean estimator, the ranking errors are thus inevitable. However, when ranking the experimental units, we may not know when the judgment/ranking error occurs. Hence, we examine the effect of judgment error on the performances of the proposed mean estimators when sampling from symmetric and asymmetric populations.

For imperfect ranking, the simulation considered here is based on the method suggested by Dell and Clutter (1972). In the simulation, we consider $m = 3, 5$ and $w = 2, 3$ with different choices of v and k. For simplicity, the simulation method is explained for POJSS only, the same method implies for other RSS schemes. Given w, m, generate wm^2 values from the underlying distribution, say $Y_{r,i}$, $r = 1,\ldots, wm$, $i = 1,\ldots, m$. Also generate random errors, say $E_{r,i}$, of the same size from a normal distribution with the mean zero and the variance V, $E_{r,i} \sim N(0,V)$, where $Y_{r,i}$ is independent of $E_{r,i}$. Then, compute $X_{r,i} = Y_{r,i} + E_{r,i}$. Using the values of $X_{r,i}$, we select a partially ordered judgment subset sample of size m, denoted by $X^*_{r(r:m)}$, $r = 1,\ldots, m$. In fact, a pair $(Y^*_{r[r:m]}, X^*_{r(r:m)})$, $r = 1,\ldots, m$, is selected using the ranks of X, where the square bracket indicates that the rank of Y is induced by the rank of X. The above procedure is repeated t times. Hence, an imperfect partially ordered judgment subset sample of size n is obtained, denoted by $Y^*_{r[r:m]j}$ for $r = 1,\ldots, m$ and $j = 1,\ldots, t$. Under perfect ranking, i.e., $V = 0$, $Y^*_{r[r:m]j} = Y^*_{r(r:m)j}$, representing a perfectly partially ordered judgment

Table 8.1 REs of the Mean Estimators With Respect to the Mean Estimator Based on SRS

m	Scheme	$U(0,1)$	$N(0,1)$	$G(1,1)$	$G(5,1)$	m	Scheme	$U(0,1)$	$N(0,1)$	$G(1,1)$	$G(5,1)$
2	PRSS	1.000	1.000	1.000	1.000	4	LRSS(1,2)	2.083	2.774	2.441	2.678
	PPOJSS(2)	1.429	1.381	1.286	1.358		LPOJSS(1,2;2)	3.140	4.553	2.912	4.006
	PPOJSS(3)	1.750	1.627	1.439	1.581		LPOJSS(1,2;3)	4.051	6.105	3.080	4.946
	RSS	1.500	1.467	1.333	1.434		LRSS(2,1)	3.125	2.034	1.171	1.743
	POJSS(2)	1.923	1.785	1.516	1.717		LPOJSS(2,1;2)	5.000	2.579	1.211	2.053
	POJSS(3)	2.227	1.980	1.614	1.884		LPOJSS(2,1;3)	6.429	2.899	1.230	2.218
	DRSS	1.923	1.785	1.516	1.717		DRSS	4.281	3.526	2.523	3.239
	RPOJSS(2)	2.269	2.004	1.625	1.905		RPOJSS(2)	5.596	4.225	2.826	3.806
	RPOJSS(3)	2.513	2.141	1.688	2.020		RPOJSS(3)	6.591	4.679	3.005	4.163
3	PRSS	1.667	1.581	1.459	1.553	5	PRSS	2.333	2.219	1.947	2.152
	PPOJSS(2)	2.545	2.245	1.865	2.148		PPOJSS(2)	3.929	3.421	2.647	3.213
	PPOJSS(3)	3.235	2.684	2.088	2.525		PPOJSS(3)	5.313	4.290	3.062	3.940
	RSS	2.000	1.914	1.636	1.843		RSS	3.000	2.770	2.190	2.615
	POJSS(2)	2.882	2.547	1.983	2.395		POJSS(2)	4.840	4.027	2.840	3.684
	POJSS(3)	3.571	2.958	2.177	2.739		POJSS(3)	6.400	4.901	3.225	4.392
	LRSS(1,2)	1.667	2.229	2.250	2.225		LRSS(1,2)	2.561	3.262	2.620	3.074
	LPOJSS(1,2;2)	2.333	3.486	2.821	3.288		LPOJSS(1,2;2)	4.060	5.491	3.119	4.643
	LPOJSS(1,2;3)	2.895	4.560	3.058	4.071		LPOJSS(1,2;3)	5.375	7.463	3.299	5.738
	DRSS	3.026	2.633	2.024	2.467		LRSS(2,1)	3.621	2.407	1.322	2.029
	RPOJSS(2)	3.818	3.086	2.232	2.844		LPOJSS(2,1;2)	5.990	3.179	1.392	2.461
	RPOJSS(3)	4.403	3.376	2.353	3.078		LPOJSS(2,1;3)	7.907	3.648	1.424	2.692
4	PRSS	1.667	1.677	1.565	1.651		LRSS(2,3)	2.333	3.486	2.230	3.067
	PPOJSS(2)	2.727	2.568	2.138	2.458		LPOJSS(2,3;2)	3.667	6.020	2.274	4.337
	PPOJSS(3)	3.640	3.211	2.475	3.012		LPOJSS(2,3;3)	4.857	8.294	2.257	5.101
	RSS	2.500	2.347	1.920	2.235		DRSS	5.670	4.456	3.016	4.027
	POJSS(2)	3.857	3.293	2.422	3.048		RPOJSS(2)	7.574	5.411	3.412	4.787
	POJSS(3)	4.971	3.932	2.711	3.573		RPOJSS(3)	9.033	6.037	3.646	5.269

subset sample. In order to examine the effect of judgment error, we choose $V = 0.05, 0.50, 1, 3$. The size of the simulation is 100,000 replications. The REs of the mean estimators are computed when sampling from symmetric and asymmetric distributions and are reported in Tables 8.2 and 8.3, respectively.

From Tables 8.2 and 8.3, it is observed that the ranking error affects the REs of the mean estimators considered here. As expected, the RE of a mean estimator decreases as the value of V increases and vice versa. Unlike the REs under normal distribution, the REs with the uniform distribution quickly approach unity. The rest of the trends are the same as we were seen in Table 8.1.

8.5 AN EXAMPLE

A real dataset is considered here to investigate the performances of the mean estimators under the considered sampling schemes when sampling from a finite population.

The dataset comprises the heights of conifer trees (measured in feet), say Y (study variable), and the diameters of conifer trees (measured at breast height in centimeters), say X (auxiliary variable). Here, our interest lies in estimating the mean height of 399 trees. For more details and description on this dataset, we refer to Platt et al. (1988). The summary statistics of the data are given in Table 8.4, where ρ is the correlation between Y and X.

Using different values of m, w, v, and k, the REs of the mean estimators are computed under both perfect and imperfect rankings and are presented in Table 8.5. The simulation size is 100,000 replications. Under perfect ranking, the values of Y are ranked using its own ranks, while under imperfect ranking the values of Y are ranked using the ranks of X. From Table 8.5, we see that the REs in most cases are greater than one—thus showing the superiority of RSS and POJSS schemes over SRS. As might be anticipated, the REs under perfect ranking are greater than those with the imperfect ranking. Moreover, the REs are increasing with the set size m. The proposed schemes continue to perform better than the existing schemes in terms of giving more precise mean estimates under both perfect and imperfect rankings.

8.6 CONCLUSIONS

In this chapter, we have suggested three modified POJSS schemes for efficiently estimating the population mean, named PPOJSS, LPOJSS, and RPOJSS. The mean estimators with PPOJSS and RPOJSS are unbiased for any population, but the mean estimator with LPOJSS is unbiased only when the underlying population is symmetric. Moreover, it has been shown, both theoretically and numerically, that the mean estimators with RPOJSS are more precise than those with the SRS, RSS, and POJSS schemes. Besides, PPOJSS is an alternative to POJSS when the ranking cost is high or POJSS cannot be conducted with full confidence. Moreover, when sampling from a symmetric population, it has been observed that the mean estimator with LPOJSS surpasses the mean estimators with RSS and POJSS. Thus, when possible, we recommend using the proposed sampling schemes for precisely estimating the population mean.

Table 8.2 REs of the Mean Estimators With Respect to the Mean Estimator Based on SRS Under Imperfect Ranking for Symmetric Distributions

	Distribution	$U(0,1)$				$N(0,1)$			
m	Scheme	$V = 0.05$	$V = 0.50$	$V = 1.00$	$V = 3.00$	$V = 0.05$	$V = 0.50$	$V = 1.00$	$V = 3.00$
3	PRSS	1.347	1.063	1.022	0.997	1.551	1.320	1.230	1.095
	PPOJSS(2)	1.635	1.096	1.042	1.010	2.115	1.587	1.390	1.161
	PPOJSS(3)	1.803	1.104	1.060	1.011	2.484	1.703	1.457	1.190
	RSS	1.465	1.072	1.038	1.011	1.823	1.462	1.309	1.133
	POJSS(2)	1.716	1.105	1.047	1.003	2.359	1.673	1.441	1.178
	POJSS(3)	1.860	1.109	1.064	1.013	2.698	1.771	1.506	1.197
	LRSS(1,2)	1.233	1.037	1.019	1.011	2.113	1.594	1.390	1.147
	LPOJSS(1,2;2)	1.376	1.052	1.024	1.002	3.147	1.912	1.549	1.213
	LPOJSS(1,2;3)	1.442	1.056	1.018	1.013	3.918	2.081	1.654	1.247
	DRSS	1.733	1.100	1.052	1.011	2.437	1.712	1.450	1.189
	RPOJSS(2)	1.884	1.113	1.047	1.016	2.817	1.815	1.511	1.199
	RPOJSS(3)	1.989	1.118	1.068	1.021	3.040	1.870	1.549	1.219
5	PRSS	1.600	1.092	1.050	1.020	2.090	1.578	1.380	1.164
	PPOJSS(2)	1.919	1.112	1.051	1.029	3.099	1.892	1.563	1.222
	PPOJSS(3)	2.097	1.132	1.070	1.022	3.715	2.061	1.625	1.236
	RSS	1.739	1.104	1.050	1.019	2.537	1.739	1.484	1.197
	POJSS(2)	2.065	1.123	1.072	1.024	3.548	1.997	1.614	1.236
	POJSS(3)	2.188	1.132	1.081	1.022	4.109	2.140	1.669	1.252
	LRSS(1,2)	1.524	1.062	1.031	0.996	2.950	1.869	1.531	1.199
	LPOJSS(1,2;2)	1.728	1.076	1.041	1.003	4.568	2.211	1.703	1.266
	LPOJSS(1,2;3)	1.827	1.075	1.036	1.017	5.773	2.339	1.783	1.260
	LRSS(2,1)	2.062	1.146	1.066	1.013	2.251	1.644	1.423	1.176
	LPOJSS(2,1;2)	2.479	1.181	1.081	1.023	2.882	1.837	1.515	1.205
	LPOJSS(2,1;3)	2.678	1.182	1.102	1.036	3.214	1.942	1.572	1.221
	LRSS(2,3)	1.373	1.047	1.023	1.015	3.109	1.910	1.550	1.222
	LPOJSS(2,3;2)	1.544	1.056	1.023	1.021	4.822	2.257	1.713	1.264
	LPOJSS(2,3;3)	1.620	1.060	1.028	1.009	6.184	2.427	1.795	1.281
	DRSS	2.129	1.120	1.064	1.018	3.799	2.075	1.646	1.250
	RPOJSS(2)	2.263	1.127	1.064	1.024	4.460	2.186	1.707	1.261
	RPOJSS(3)	2.332	1.151	1.078	1.022	4.858	2.275	1.713	1.256

Table 8.3 REs of the Mean Estimators With Respect to the Mean Estimator Based on SRS Under Imperfect Ranking for Asymmetric Distributions

	Distribution	$G(1,1)$				$G(5,1)$			
m	Scheme	$V = 0.05$	$V = 0.50$	$V = 1.00$	$V = 3.00$	$V = 0.05$	$V = 0.50$	$V = 1.00$	$V = 3.00$
3	PRSS	1.454	1.255	1.194	1.100	1.551	1.482	1.419	1.284
	PPOJSS(2)	1.780	1.436	1.299	1.143	2.144	1.920	1.797	1.493
	PPOJSS(3)	1.948	1.499	1.321	1.162	2.514	2.209	2.016	1.595
	RSS	1.582	1.343	1.248	1.114	1.847	1.691	1.613	1.398
	POJSS(2)	1.887	1.466	1.324	1.152	2.389	2.102	1.923	1.574
	POJSS(3)	2.012	1.520	1.358	1.159	2.702	2.331	2.124	1.641
	LRSS(1,2)	2.109	1.803	1.657	1.410	2.196	1.978	1.837	1.563
	LPOJSS(1,2;2)	2.494	1.995	1.806	1.509	3.170	2.676	2.334	1.795
	LPOJSS(1,2;3)	2.583	1.995	1.870	1.569	3.952	3.089	2.616	1.921
	DRSS	1.911	1.486	1.340	1.154	2.431	2.165	1.987	1.579
	RPOJSS(2)	2.086	1.556	1.376	1.165	2.773	2.415	2.180	1.672
	RPOJSS(3)	2.167	1.561	1.365	1.171	3.042	2.579	2.277	1.712
5	PRSS	1.861	1.487	1.342	1.144	2.116	1.951	1.817	1.499
	PPOJSS(2)	2.464	1.711	1.455	1.202	3.145	2.654	2.349	1.754
	PPOJSS(3)	2.746	1.804	1.524	1.232	3.846	3.069	2.670	1.900
	RSS	2.046	1.566	1.372	1.166	2.577	2.253	2.054	1.626
	POJSS(2)	2.551	1.774	1.495	1.200	3.559	2.948	2.547	1.833
	POJSS(3)	2.895	1.844	1.559	1.234	4.279	3.351	2.790	1.956
	LRSS(1,2)	2.351	1.848	1.683	1.417	3.038	2.531	2.247	1.758
	LPOJSS(1,2;2)	2.644	1.961	1.771	1.483	4.453	3.376	2.791	1.977
	LPOJSS(1,2;3)	2.730	1.948	1.764	1.509	5.446	3.790	3.066	2.041
	LRSS(2,1)	1.254	1.048	0.999	0.936	2.009	1.793	1.668	1.379
	LPOJSS(2,1;2)	1.293	1.066	0.999	0.930	2.438	2.077	1.875	1.493
	LPOJSS(2,1;3)	1.332	1.082	0.978	0.921	2.597	2.265	1.990	1.540
	LRSS(2,3)	1.952	1.678	1.631	1.445	2.985	2.493	2.199	1.721
	LPOJSS(2,3;2)	1.878	1.638	1.633	1.489	4.125	3.133	2.608	1.859
	LPOJSS(2,3;3)	1.860	1.604	1.603	1.524	4.866	3.414	2.775	1.928
	DRSS	2.713	1.814	1.509	1.218	3.889	3.141	2.650	1.881
	RPOJSS(2)	3.020	1.903	1.552	1.230	4.596	3.535	2.924	1.983
	RPOJSS(3)	3.170	1.946	1.598	1.242	5.083	3.785	3.090	2.036

Table 8.4 Summary Statistics of 399 Trees Data

Variable	Mean	Variance	Skewness	Kurtosis	Median	ρ
Y	52.36	325.14	1.619	1.776	29	0.908
X	20.84	310.11	0.844	−0.423	14.5	

Table 8.5 REs of the mean estimators with respect to the mean estimator based on SRS under perfect and imperfect rankings

m	Scheme	Perfect	Imperfect	*m*	Scheme	Perfect	Imperfect
2	PRSS	1.000	1.000	4	LRSS(1,2)	1.964	1.752
	PPOJSS(2)	1.292	1.252		LPOJSS(1,2;2)	2.137	1.973
	PPOJSS(3)	1.431	1.390		LPOJSS(1,2;3)	2.142	2.035
	RSS	1.325	1.298		LRSS(2,1)	1.168	1.252
	POJSS(2)	1.487	1.449		LPOJSS(2,1;2)	1.131	1.224
	POJSS(3)	1.578	1.534		LPOJSS(2,1;3)	1.145	1.209
	DRSS	1.501	1.451		DRSS	2.639	2.263
	RPOJSS(2)	1.584	1.532		RPOJSS(2)	3.090	2.542
	RPOJSS(3)	1.637	1.601		RPOJSS(3)	3.289	2.736
3	PRSS	1.480	1.407	5	PRSS	1.995	1.797
	PPOJSS(2)	1.887	1.770		PPOJSS(2)	2.873	2.354
	PPOJSS(3)	2.140	1.963		PPOJSS(3)	3.462	2.703
	RSS	1.623	1.533		RSS	2.221	1.974
	POJSS(2)	2.000	1.824		POJSS(2)	3.027	2.466
	POJSS(3)	2.200	2.034		POJSS(3)	3.620	2.826
	LRSS(1,2)	1.892	1.658		LRSS(1,2)	2.065	1.839
	LPOJSS(1,2;2)	2.220	1.995		LPOJSS(1,2;2)	2.261	2.049
	LPOJSS(1,2;3)	2.280	2.119		LPOJSS(1,2;3)	2.290	2.133
	DRSS	2.044	1.874		LRSS(2,1)	1.372	1.439
	RPOJSS(2)	2.279	2.088		LPOJSS(2,1;2)	1.401	1.497
	RPOJSS(3)	2.397	2.216		LPOJSS(2,1;3)	1.416	1.535
4	PRSS	1.585	1.459		LRSS(2,3)	1.663	1.623
	PPOJSS(2)	2.247	1.966		LPOJSS(2,3;2)	1.510	1.576
	PPOJSS(3)	2.692	2.299		LPOJSS(2,3;3)	1.403	1.511
	RSS	1.929	1.747		DRSS	3.251	2.630
	POJSS(2)	2.513	2.200		RPOJSS(2)	3.864	2.969
	POJSS(3)	2.917	2.454		RPOJSS(3)	4.313	3.170

ACKNOWLEDGMENTS

The author is thankful to Dr. Zawar Hussain, Associate Professor, Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan, for his help and encouragement.

REFERENCES

Al-Naseer, A.D., 2007. L ranked set sampling: a generalized procedure for robust visual sampling. Commun. Stat.: Simul. Comput. 36 (1), 33−43.

Al-Omari, A.I., Raqab, M.Z., 2013. Estimation of the population mean and median using truncation-based ranked set samples. J. Stat. Comput. Simul. 83 (8), 1453−1471.

Al-Saleh, M.F., Al-Kadiri, M.A., 2000. Double-ranked set sampling. Stat. Probab. Lett. 48 (2), 205−212.

Bapat, R.B., Beg, M.I., 1989. Order statistics for nonidentically distributed variables and permanents. Sankhyā: Ind. J. Stat. Ser. A 51 (1), 79−93.

Chen, Z., 2007. Ranked set sampling: its essence and some new applications. Environ. Ecol. Stat. 14 (4), 355−363.

David, H.A., Nagaraja, H.N., 2003. Order Statistics, 3rd ed. John Wiley & Sons, Inc., Hoboken, New Jersey.

Dell, T.R., Clutter, J.L., 1972. Ranked set sampling theory with order statistics background. Biometrics 28 (2), 545−555.

Haq, A., 2017a. Two-stage cluster sampling with hybrid ranked set sampling in the secondary sampling frame. Commun. Stat.: Theory Methods 46 (17), 8450−8467.

Haq, A., 2017b. Estimation of the distribution function under hybrid ranked set sampling. J. Stat. Comput. Simul. 87 (2), 313−327.

Haq, A., Brown, J., Moltchanova, E., Al-Omari, A.I., 2013. Partial ranked set sampling design. Environmetrics 24 (3), 201−207.

Haq, A., Brown, J., Moltchanova, E., Al-Omari, A.I., 2014. Mixed ranked set sampling design. J. Appl. Stat. 41 (10), 2141−2156.

Haq, A., Brown, J., Moltchanova, E., Al-Omari, A.I., 2015. Varied L ranked set sampling scheme. J. Stat. Theory Pract. 9 (4), 741−767.

Haq, A., Brown, J., Moltchanova, E., Al-Omari, A.I., 2016a. Paired double-ranked set sampling. Commun. Stat.: Theory Methods 45 (10), 2873−2889.

Haq, A., Brown, J., Moltchanova, E., 2016b. Hybrid ranked set sampling scheme. J. Stat. Comput. Simul. 86 (1), 1−28.

Kaur, A., Patil, G.P., Taillie, C., 1997. Unequal allocation models for ranked set sampling with skew distributions. Biometrics 53 (1), 123−130.

McIntyre, G.A., 1952. A method for unbiased selective sampling, using ranked sets. Aust. J. Agric. Res. 3 (4), 385−390.

Muttlak, H.A., 1996. Pair rank set sampling. Biom. J. 38 (7), 879−885.

Muttlak, H.A., 1997. Median ranked set sampling. J. Appl. Stat. Sci. 6 (4), 245−255.

Muttlak, H.A., 2003. Investigating the use of quartile ranked set samples for estimating the population mean. Appl. Math. Comput. 146 (2−3), 437−443.

Ozturk, O., 2011. Sampling from partially rank-ordered sets. Environ. Ecol. Stat. 18 (4), 757−779.

Ozturk, O., Wolfe, D.A., 2000. Optimal allocation procedure in ranked set sampling for unimodal and multimodal distributions. Environ. Ecol. Stat. 7 (4), 343−356.

Patil, G.P., 1995. Editorial: ranked set sampling. Environ. Ecol. Stat. 2 (4), 271−285.

Patil, G.P., Sinha, A.K., Taillie, C., 1999. Ranked set sampling: a bibliography. Environ. Ecol. Stat. 6, 91−98.

Platt, W.J., Evans, G.W., Rathbun, S.L., 1988. The population dynamics of a long-lived conifer (pinus palustris). Am. Nat. 131 (4), 491–525.

Samawi, H.M., Ahmed, M.S., Abu-Dayyeh, W., 1996. Estimating the population mean using extreme ranked set sampling. Biom. J. 38 (5), 577–586.

Takahasi, K., Wakimoto, K., 1968. On unbiased estimates of the population mean based on the sample stratified by means of ordering. Ann. Inst. Stat. Math. 20 (1), 1–31.

Vaughan, R.J., Venables, W.N., 1972. Permanent expressions for order statistic densities. J. R. Stat. Soc. Ser. B 34 (2), 308–310.

Wolfe, D.A., 2012. Ranked set sampling: its relevance and impact on statistical inference. ISRN Probab. Stat. 2012, 32.

CHAPTER

9 RANKED SET SAMPLING WITH UNEQUAL SAMPLE SIZES

Dinesh S. Bhoj and Debashis Kushary

Department of Mathematical Sciences, Rutgers University, Camden, NJ, United States

9.1 INTRODUCTION

Ranked set sampling for estimating a population mean was first proposed by McIntyre (1952) as a cost-efficient alternative to simple random sampling (SRS) if the observations can be ranked according to the characteristic under investigation by means of visual inspection or other methods not requiring actual measurements. Mcintyre indicated that the use of RSS is more powerful and superior to the RSS procedure to estimate the population mean. Dave and Cutler (1972) and Takahashi and Wakimoto (1968) provided a mathematical foundation for RSS. Dave and Cutler (1972) showed that the estimator for population mean based on RSS is at least as efficient as the estimator based on SRS with the same number of measurements, even when there are ranking errors. RSS is a nonparametric procedure. However, it has been used recently in parametric settings (see Bhoj and Ahsanullah, 1996; Bhoj, 1997; Lam et al., 1994; Stokes, 1995). Most of the distributions considered by these investigators belong to the family of random variables with the cumulative distribution of the form $F((x-\mu)/\sigma)$, where μ and σ are the location and scale parameters, respectively. The various methods of estimation of parameters of the distribution with applications and an extensive list of references are given by Chen et al. (2004).

The selection of a ranked set sample of size m involves drawing m random samples with m units in each sample. Then units in each sample are ranked by using judgment or other methods not requiring actual measurements. The unit with the lowest rank is measured from the first sample, the unit with the second lowest rank is measured from the second sample, and this procedure is continued until the unit with the highest rank is measured from the last sample. The m^2 ordered observations in m samples produces a data set as follows:

$$\begin{array}{llll}
x_{[1]1} & x_{[1]2} & \cdots\cdot & x_{[1]m} \\
x_{[2]1} & x_{[2]2} & \cdots\cdot & x_{[2]m} \\
\ldots & \ldots\ldots & \cdots\cdot & \cdots\cdot \\
\ldots & \ldots\ldots & \cdots\cdot & \cdots\cdot \\
.x_{[m]1} & x_{[m]2} & \cdots\cdot & x_{[m]m}
\end{array}$$

We measure only m $(X_{[i]i}, i=1,2,\ldots,m)$ observations and they constitute RSS. It should be noted that m observations are independently but not identically distributed. In RSS, m is usually small and therefore in order to increase the sample size, the above procedure is repeated k times. For convenience, without loss of generality we usually assume that $k = 1$.

Ranked Set Sampling. DOI: https://doi.org/10.1016/B978-0-12-815044-3.00009-5

9.2 SOME RANKED SET SAMPLING PROCEDURES

There are various modifications of RSS to get a better estimator for the population mean, μ. One of the popular schemes is to use the median ranked set sampling (MRSS) (see Bhoj, 1997; Muttlack, 1997). MRSS performs very well when the distributions are unimodal and symmetric. In the MRSS procedure we rank m^2 observations, as in RSS. However, we measure only the observations with rank $(m+1)/2$ from each sample if m is odd. If $m = 2l$ is even, we use the l th order statistic from the first l samples and $(l+1)$ th order statistics from the last l samples. We compare the performance of the estimators based on ranked set sampling with unequal samples with those based on RSS and MRSS procedures.

9.2.1 RANKED SET SAMPLING WITH UNEQUAL SAMPLES

Bhoj (2001) proposed a ranked set sampling procedure with unequal sample sizes (RSSU). In RSSU we draw m samples where the size of ith sample is $m_i = 2i - 1$, $i = 1, 2,\ldots, m$. The steps in RSSU are the same as in RSS. In both sampling procedures we measure accurately m observations. However, in RSSU we rank only $(m^2 - 1)$ observations. When m is even, half the sample sizes are smaller than m and the other half are greater than m. In the case of odd m, one sample is of size m, $(m - 1)/2$ samples are greater than m and other $(m - 1)/2$ samples are smaller than m. Although the ranking error due to larger sample size is offset by the ranking error due to smaller sample size, it is important to keep m small in RSSU and the procedure is repeated to increase the sample size.

9.2.2 RANKED SET SAMPLING WITH UNEQUAL SAMPLES AND UNEQUAL REPLICATIONS

Bhoj and Kushary (2014) proposed ranked set sampling with unequal samples and unequal replications (RSSUR). In RSSUR, the ith sample of size $m_i = 2i - 1$ is repeated k_i times $i = 1, 2,\ldots, m$. In RSSUR $\sum_{i=1}^{m} k_i$ observations are measured and $\sum_{i=1}^{m} m_i k_i$ observations are ranked. It is noted that there is no ranking with $m_1 = 1$. In order to have fair comparisons of the estimators based on RSSUR we must have $\sum_{i=1}^{m} k_i = mk$ and $d = m^2 k - \sum_{i=2}^{m} k_i = 0$. However, it is not possible to achieve $d = 0$ with the appropriate integer values of k_i for $m = 2$. The authors chose $|d| \leq 1$ for $m = 2$, and $d = 0$, for $m = 3$, and $k = 2$, 3, and 4.

9.2.3 RANKED SET SAMPLING WITH UNEQUAL SAMPLES FOR SKEW DISTRIBUTIONS

Bhoj (2001) showed that the estimators for the population mean based on RSSU are superior to the estimators based on RSS and MRSS, when the distributions under consideration are symmetrical around μ or moderately skewed. However, the proposed estimators based on RSSU do not work well if the distributions are highly skewed. Therefore, Bhoj and Kushary (2016) proposed the ranked set sampling procedure with unequal samples for highly positive skew distributions (RSSUS). The authors proposed the estimators for μ which are weighted linear combinations of RSSU observations.

9.3 ESTIMATION OF THE POPULATION MEAN

McIntyre (1952) proposed the nonparametric estimator for a population mean, μ, based on RSS as

$$\hat{\mu}_{\text{RSS}} = \frac{1}{m}\sum_{i=1}^{m} x_{[i]i}$$

with variance $\text{Var}(\hat{\mu}_{\text{RSS}}) = \frac{1}{m^2}\sum_{i=1}^{m}\sigma^2_{[i]i}$ where $\sigma^2_{[i]i}$ is the variance of ith order statistic in a random sample of size m.

The estimator, $\hat{\mu}_{\text{MRSS}}$, for μ based on MRSS defined in Section 9.2 is

$$\hat{\mu}_{\text{MRSS}} = \begin{cases} \frac{1}{m}\left[\sum_{i=1}^{l} x_{[i]l} + \sum_{i=l+1}^{m} x_{[i](l+1)}\right], & \text{for even } m, \\ \frac{1}{m}\sum_{i=1}^{m} x_{[i]p}, \quad \text{where } p = (m+1)/2, & \text{for odd } m. \end{cases}$$

The variance of $\hat{\mu}_{\text{MRSS}}$ is

$$\text{Var}(\hat{\mu}_{\text{MRSS}}) = \begin{cases} \left[(\sigma^2_{[l]l} + \sigma^2_{[l+1]l+1})\right]/2m, & \text{for even } m, \\ (\sigma^2_{[p]p})/m, & \text{for odd } m. \end{cases}$$

The estimator $\hat{\mu}_{\text{MRSS}}$ is an unbiased estimator for μ when the distribution under consideration is symmetric around μ. When the distribution is skewed, $\hat{\mu}_{\text{MRSS}}$ is a biased estimator for μ. In this case, for comparison with other estimators, the mean square error (MSE) of $\hat{\mu}_{\text{RSS}}$, where $\text{MSE} = \text{Variance} + (\text{Bias})^2$ was used.

Bhoj (2001) proposed the following set of estimators for μ based on RSSU defined in Section 9.2:

$$\hat{\mu}_{r:\text{RSSU}} = \sum_{i=1}^{m} w_r X_{([i]i:m_i)}, \qquad r = 1, 2, \ldots, 6.$$

The variance of the estimator is

$$\text{Var}(\hat{\mu}_{r:\text{RSSU}}) = \sum_{i=1}^{m} w_r^2 \sigma^2_{([i]i:m_i)}$$

where $\sigma^2_{([i]i:m_i)}$ is the variance of ith order statistic in a random sample of size m_i.

He considered various values of the weights that are proportional to $(m_i + h)$ where $0 \le h \le 1$. The first four weights are derived under the assumption that w_r is proportional to $m_i, m_i + 1/4, m_i + 1/2, m_i + 3/4. w_5$ is the average of w_1 and w_3 while w_6 is obtained by taking the average of the weights that are proportional to m_i and $m_i + 1$. The main reason for the choice of this class of weights was that for some distributions the near optimal weights belonged to this class. For example, w_1, w_2, and w_3 are near optimal for Laplace, logistic, and normal distributions.

Now the estimators for μ based on RSSUR defined in Section 9.2 will be considered. In this case we have to repeat the sample $k > 1$ times to get the balanced ranked set samples. Bhoj and Kushary (2014) considered ranked set sampling with unequal samples and unequal replications where the ith sample is repeated k_i, $i = 1, 2, 3\ldots, m$ times, so that sample size $\sum_{i=1}^{m} k_i = mk$. The estimators for μ based on RSS, MRSS, and RSSU with k replications are given by

$$\hat{\mu}_{\text{RSS}} = \frac{1}{mk}\sum_{i=1}^{m}\sum_{j=1}^{k} x_{([i]i)j} \quad \text{and}$$

$$\hat{\mu}_{\text{MRSS}} = \begin{cases} \dfrac{1}{m^2k}\left(\sum_{i=1}^{l}\sum_{j=1}^{k} x_{([i]l)j} + \sum_{i=l+1}^{m} x_{([i]l+1)j}\right), & \text{for even } m, \\ \dfrac{1}{mk}\sum_{i=1}^{m}\sum_{j=1}^{k} x_{([i]p)j},\ p=(m+1)/2, & \text{for odd } m, \end{cases}$$

$$\hat{\mu}_{r:\text{RSSU}} = \frac{1}{k}\sum_{i=1}^{m}\sum_{j=1}^{k} w_r x_{([i]i)jm_i}, \qquad r = 1, 2, . . ., 6.$$

The six values of w_r are the same as in RSSU with $k = 1$. The variances of the above three estimators are given by

$$\text{Var}(\hat{\mu}_{\text{RSS}}) = \frac{1}{m^2r}\sum_{i=1}^{m}\sigma^2_{[i]i}$$

$$\text{Var}(\hat{\mu}_{\text{MRSS}}) = \begin{cases} \dfrac{1}{m^2r}\left(\sum_{i=1}^{l}\sigma^2_{[i]l} + \sum_{i=l+1}^{m}\sigma^2_{[i]l+1}\right), & \text{for even } m \\ \dfrac{1}{m^2r}\sum_{i=1}^{m}\sigma^2_{[i]p} \quad \text{where}, p=(m+1)/2, & \text{for odd } m, \text{ and} \end{cases}$$

$$\text{Var}(\hat{\mu}_{r:\text{RSSU}}) = \frac{1}{k}\sum_{i=1}^{m} w_r^2\,\sigma^2_{([i]i:m_i)}$$

Bhoj and Kushary (2014) proposed the following set of estimators for μ based on RSSUR discussed in Section 9.2.

$$\hat{\mu}_{r:\text{RSSUR}} = \sum_{i=1}^{m} w_{ru}\sum_{j=1}^{k_i}\frac{x_{[i]ijm_j}}{k_i}, \qquad r = 1, 2, ..., 6$$

The variance of $\hat{\mu}_{r:\text{RSSUR}}$ is

$$\text{Var}(\hat{\mu}_{r:\text{RSSUR}}) = \sum_{i=1}^{m} w_{ru}^2\frac{\sigma^2_{[i]i:m_i}}{k_i}$$

Bhoj and Kushary (2014) considered various weights that were proportional to $k_i(m_i + h)$ where $0 \le h \le 1.0$. The first four values of w_{ru} are obtained by using $h = 0$, 1/4, 1/2, and 3/4. w_{5u} is obtained by taking the average of w_{1u} and w_{3u}, and w_{6u} is derived by taking average of the weights that are proportional to k_im_i and $k_i(m_i + 1)$.

The estimators for μ based on the RSSU scheme do not work well if the distributions under consideration are highly skewed. Therefore, Bhoj and Kushary (2016) proposed the estimators for μ, which are weighted linear combinations of $x_{([i]i)m_i}$ for heavy right tail probability distributions. They proposed the following set of estimators for μ based on RSSUS.

$$\hat{\mu}_{r:\text{RSSUS}} = \sum_{i=1}^{m} w_r x_{[i]im_i} \quad r = 1, 2, ..., 6.$$

Bhoj and Kushary (2016) considered various weights that are based on the ratio w_i/w_1 and are given by

$$\frac{w_i}{w_1} = m_i + (m_i - m_{(i-1)} + d_i h_i)h,$$

where $d_i = \frac{i}{(i^2-1)}, h_2 = 1, h_3 = \frac{1}{h}, h_4 = h$ and $0 \le h \le 1$.

The values of w_1 for different values of m are determined so that the new set of estimators for μ based on RSSUS would perform better than the estimators for μ based on RSS and MRSS procedures for the chosen heavy right tail distributions. They used

$$
\begin{aligned}
w_1 &= \frac{m_1 + h_1}{D_i}, \quad i = 2, 3 \text{ and } 4 \text{ for } m = 2, 3 \text{ and } 4, \\
\text{where } D_i &= m^2 + (2m-3)h + (i-2), \quad \text{for } i = 2 \text{ and } 3 \\
D_4 &= m^2 + 1 + (2m-3)h + 0.4h^2, \quad \text{for } m = 4 \\
h_1 &= \frac{m(m-2) + |(i + c.h)/(i^2 - 6)|}{100}, \\
\text{where } c &= 0 \text{ for } i = 2, 3 \text{ and } c = 4 \text{ for } i = 4.
\end{aligned}
$$

In order to keep the number of weights within reasonable limits they used five values of w_r with $h = 0.75$, 0.8, 0.85, 0.90, and 0.95. The main reason for the choice of the values of h was, for some distributions, near optimal ratios of the weights belonged to some values of h. For example, $h = 0.75$ gives near optimum values of the ratios of weights for Weibull distribution for $n \leq 3$, and $h = 0.95$ gives near optimal values of the ratios of weights for Pareto (5) and lognormal distributions for $m = 2$.

9.4 COMPARISONS OF ESTIMATORS

In this section, the various estimators for μ based on RSS, MRSS, RSSUR, and RSSUS are compared. First the estimators based on RSS, MRSS, and RSSU are compared. For this purpose, the following nonparametric relative precisions (RPNs) are defined:

$$
\text{RPN}_\text{r} = \begin{cases} \text{Var}(\hat{\mu}_{\text{RSS}})/\text{Var}(\hat{\mu}_{r:\text{RSSU}}), & \text{if } \hat{\mu}_{\text{RSSU}} \text{ is an unbiased estimator} \\ \text{Var}(\hat{\mu}_{\text{RSS}})/\text{MSE}(\hat{\mu}_{r:\text{RSSU}}), & \text{if } \hat{\mu}_{\text{RSSU}} \text{ is a biased estimator} \end{cases}
$$

for $r = 1, 2, 3,\ldots,6$ and,

$$
\text{RPN}_7 = \begin{cases} \text{Var}(\hat{\mu}_{\text{RSS}})/\text{Var}(\hat{\mu}_{\text{MRSSU}}) & \text{if } \hat{\mu}_{\text{MRSS}} \text{ is an unbiased estimator} \\ \text{Var}(\hat{\mu}_{\text{RSS}})/\text{MSE}(\hat{\mu}_{\text{MRSSU}}) & \text{if } \hat{\mu}_{\text{MRSS}} \text{ is a biased estimator.} \end{cases}
$$

It is noted that RPN_r/RPN_7 can be used for comparison of the estimators based on RSSU and MRSS. Then $\hat{\mu}_{r:\text{RSSU}}$ is better than $\hat{\mu}_{\text{MRSS}}$ if $RPN_r > RPN_7$. Bhoj (2001) calculated the seven relative precisions for normal, logistic, Laplace, exponential, Weibull (2), Weibull (4), gamma (3), gamma (5), and extreme value distributions and $m = 2$, 3, and 4... These computations showed that $\hat{\mu}_{r:\text{RSSU}}$ $r = 1, 2, 3,\ldots.. 6$ are all superior to the estimator, $\hat{\mu}_{\text{RSS}}$, for all the distributions. However, all estimators based on RSSU are not better than $\hat{\mu}_{\text{MRSS}}$ for all distributions and sample sizes. $\hat{\mu}_{r:\text{RSSU}}$ for $r = 3$, 4, and 6 have substantial gain in relative precisions over $\hat{\mu}_{\text{RSS}}$ and $\hat{\mu}_{\text{MRSS}}$ for all nine distributions considered in the paper. Bhoj (2001) also considered the errors in ranking and showed the performance of $\hat{\mu}_{r:RSSU}$ is superior to the estimators $\hat{\mu}_{\text{RSS}}$ and $\hat{\mu}_{\text{MRSS}}$.

Now the estimators $\hat{\mu}_{r:\text{RSSUR}}$, $r = 1, 2, \ldots.., 6$ based on RSSUR are compared with the estimators based on RSS and MRSS. We define seven nonparametric relative precisions in this section which are similar to the relative precisions in the previous paragraph except $\hat{\mu}_{r:\text{RSSU}}$ is replaced by $\hat{\mu}_{r:\text{RSSUR}}$. Bhoj and Kushary (2014) calculated these relative precisions for normal, logistic, Laplace, exponential, Weibull (2), Weibull (4), gamma (3), gamma(5), and extreme value distribution, and $m = 2$ and 3 and $k = 2$, 3, 4, and the values of k_i. For given m and k there is one near optimal solution of k_i for all above eight distributions. $\hat{\mu}_{r:\text{RSSU}}$ is superior to $\hat{\mu}_{\text{RSS}}$. w_{1u}, w_{2u}, w_{3u} are nearly optimal for Laplace, logistic, and normal distributions, respectively. In general, w_{5u} works quite well for the three symmetrical distributions and w_{4u} performs better for all skewed distributions. The relative precisions based on $\hat{\mu}_{r:\text{RSSUR}}$, $r = 1, 2, ..., 6$ are all higher than RPU_7 for all the values of k and m and the eight distributions considered in the chapter. Hence $\hat{\mu}_{r:\text{RSSU}}$, $r = 1, 2, ..., 6$, is uniformly superior to $\hat{\mu}_{\text{MRSS}}$. Bhoj and Kushary (2014) tabulated the values of k_i for $m = 2$ and 3 and $k = 2$, 3, and 4. They computed relative precisions RPUR_r to compare the estimators for μ based on RSSU and RSSUR for $r = 1, 2, \ldots, 6$, $m = 2$ and3, $k = 2$, 3, and 4 for eight distributions. They tabulated the values of k_i and they showed that $\hat{\mu}_{r:\text{RSSUR}}$, $r = 1, 2, \ldots, 6$ is superior to $\hat{\mu}_{r:\text{RSSU}}$, for all eight distributions.

Now the estimators for μ based on RSS, MRSS, and RSSUS are compared. The estimator $\hat{\mu}_{r:\text{RSSUS}}$ is biased for highly skewed distributions. Therefore the following nonparametric relative precisions (RPNs) are defined:

$$\text{RPN}_r = \text{Var}(\hat{\mu}_{\text{RSS}})/\text{MSE}(\hat{\mu}_{r:\text{RSSUS}}), \quad \text{for } \text{r} = 1, 2, \ldots, 5$$

$$\text{RPN}_6 = \begin{cases} \text{Var}(\hat{\mu}_{\text{RSS}})/\text{Var}(\hat{\mu}_{\text{MRSS}}), & \text{if } \hat{\mu}_{\text{MRSS}} \text{ is an unbiased estimator,} \\ \text{Var}(\hat{\mu}_{\text{RSS}})/MSE(\hat{\mu}_{\text{MRSS}}), & \text{if } \hat{\mu}_{\text{MRSS}} \text{ is a biased estimator.} \end{cases}$$

Bhoj and Kushary (2016) computed RPN_r, $r = 1, 2, \ldots, 6$ for lognormal, Weibull (0.5), Pareto (2.5), and Pareto (5) distributions and $m = 2$, 3, and 4. They also tabulated variances and biases of the estimators based on RSSUS and MRSS. From the computations of relative precisions they noted that the estimators $\hat{\mu}_{r:\text{RSSUS}}$ are all superior to the estimators based on RSS and MRSS for the four distributions and three sample sizes. The gains in the relative precisions of the estimator of μ based on RSSUS over the estimators based on RSS are substantial. However, the gains in the precisions of $\hat{\mu}_{r:\text{RSSUS}}$ over the estimator based on MRSS are very good to marginal, depending on the value of m and the distribution. It was noted that the estimators based on RSSUS are adversely affected by the extreme values of means and variances of the extremely heavy tail distributions since RSSUS uses $m_1 = 1$. The estimator based on MRSS is not directly affected by the extreme values of the means and variances of the probability distributions.

9.5 MORE RANKED SET SAMPLING PROCEDURES WITH UNEQUAL SAMPLES

In this section, some more ranked set sampling procedures with unequal samples are discussed. Bhoj (2002) showed that the MRSS procedure does not perform well for even m, as compared to odd m. He computed the relative percentage increases (RPI) in RP where RPI is defined as{(RP for m − RP for $(m - 1)$) / RP for $(m - 1)\}$ × 100. These computations showed that the values of RPI are higher when we move from even to odd values of m, and they are lower when we switch from odd to even values of m. Therefore, he proposed a new median ranked set sampling (NMRSS) for even $m = 2l$. In this procedure, we draw first l samples of size $(m - 1)$ and the last l samples of size $(m + 1)$. Then the median is quantified from each sample to estimate the population mean. Bhoj (2002) showed that the relative precisions of the estimator based on the NMRSS procedure are better than the estimators based on MRSS when distributions are symmetric and unimodal for $m = 2$, 4, and 6. For moderate skew distributions MMRSS works reasonably well for $m = 2$ and 4. Most importantly the NMRSS procedure performs better than MRSS even for highly skew distributions when $m = 2$.

Biradar and Santosha (2014) proposed the maximum ranked set sampling procedure with unequal samples $m_i = i, \;\; i = 1, 2, 3, ., m$ to estimate the mean of exponential distribution. In this procedure, they quantified only the observations with maximum rank. Although they measured m observations, the number of observations ranked is only $m(m + 1)/2$. They derived maximum likelihood estimator and modified maximum likelihood estimator for the mean and showed that the relative precisions of these estimators are better than that based on SRS. By using simulation, they showed that the efficiency of the proposed estimator is better than the estimator based on RSS under ranking error.

The balanced RSS approach can not be used when the populations are not available at the time when the study was conducted. However, the entire population elements can be observed as batches of different sizes. For such situation, Samawi (2011) proposed varied set size ranked set sampling (VSRSS). In this procedure c sets of different sizes, say, $K_1^2, K_2^2, ., K_c^2$, are randomly selected. Next the RSS technique is applied to each set separately to obtain c ranked set samples of sizes, $K_1, K_2, ., K_c$, respectively. This produces a sample of size $\sum_{i=1}^{c} K_i$. Samawi (2011) showed that the estimator based on VSRSS is unbiased for the population mean and its variance is less than the variance of the estimator of μ based on SRS.

Sometimes the sets that arise naturally in the RSS applications are of unequal sizes. For example, commuters on different public buses or patients that have been waiting in doctors' waiting rooms that represent natural sets of different sizes. Germayel et al. (2010) proposed the estimator for the median of a symmetric population that combines medians of RSS samples of different sizes. The estimator is robust over a wide class of symmetric distributions, although it is not optimal for any specific symmetric distribution.

Some authors proposed an RSS procedure with random selection of the units for measurements. Li et al. (1999), Rahimov and Mutllak (2003a,b), and Amiri et al. (2015) proposed random ranked set sampling where the set size and/or the number of cycles are allowed to be random and unequal.

Zhang et al. (2014) considered a sign test under ranked set sampling with unequal set sizes (RSSU), and proposed weighted sign tests associated with judgment ranks. The optimal weight vector is shown to be distribution-free and RSSU proved to be more efficient than RSS.

9.6 APPLICATIONS TO REAL-WORLD DATA

Bhoj (2001) and Bhoj and Kushary (2014) applied their formulae derived under RSSU and RSSUR procedures to the longleaf pine data. The data consist of the coordinates and diameters (at breast height) of all longleaf pine trees at least 2 centimeters in diameter at breast height (dbs) in a 4-ha region on the Wade Tract in Thomas County, Georgia, in 1979. The data have 584 trees, with observations ranging from 2 to 75.9 cm dbh, indicating a large variability in the data. The Wade Tract contains all ages of trees up to 250 years. All observations on dbh are given in Cressie (1993). The data on 584 trees are considered as the population. The objective is to estimate the mean dbh value of longleaf pine trees in the 4-ha region by using RSS, MRSS, RSSU, and RSSUR.

Bhoj (2001) took random samples of size $m = 3$ from the given population. The cycle was repeated $k = 9$ times to estimate the variance within each rank. The computed values of the estimates based on mk observations are $\hat{\mu}_{\text{RSS}} = 21.80$, $\hat{\mu}_{\text{MRSS}} = 29.56$, and $\hat{\mu}_{r:\text{RSSU}} = 28.73$, 28.66, 28.59, 28.54, 28.66, and 28.61 for $r = 1, 2, 6$ and the estimated corresponding variances of the estimators are, respectively, 6.96, 6.11, 5.25, 5.10, 4.98, 5.11, and 5.01. It is observed that $\hat{\mu}_{3:\text{RSSU}}$, $\hat{\mu}_{4:\text{RSSU}}$, and $\hat{\mu}_{6:\text{RSSU}}$ are closer to the population mean $\mu = 26.84$ and the variances of these estimators are relatively small. We note that the variance of the sample mean $\overline{X}$ based on mk observations is $\sigma^2/mk = 334.238/27 = 12.38$. The estimated variances of the estimators based on the RSSU procedure are considerably smaller than variances of $\overline{X}$ based on the same number of quantified observations.

Bhoj and Kushary (2014) proposed ranked set sampling with unequal samples and unequal replications. They estimated the mean dbh value of longleaf pine trees by using various ranked set sampling procedures with equal replications and RSSU with unequal replications. Bhoj and Kushary (2014) took the random samples of size $m = 3$ from the given population. The cycle was repeated $k = 4$ times to estimate the variance within each rank. The estimators for the mean were also computed with unequal replications: $m_1 = 2$, $m_2 = 7$ and $m_3 = 3$. The computed values of the estimators are $\hat{\mu}_{\text{RSS}} = 25.39$, $\hat{\mu}_{r:\text{RSSU}}$, with $k = 4$, are 25.90, 26.00, 26.09, 26.17, 26.00, 26.07 and, $\hat{\mu}_{r:\text{RSSUR}} = 27.24$, 27.11, 27.00, 26.90, 27.12, 27.03 for $r = 1, 2, \ldots, 6$. The corresponding variances of the estimators are, respectively, $\text{Var}(\hat{\mu}_{\text{RSS}}) = 17.25$, $\text{Var}(\hat{\mu}_{r:\text{RSSU}}) = 14.5$, 14.63, 14.78, 14.94, 14.62, 14.74 and $\text{Var}(\hat{\mu}_{r:\text{RSSUR}}) = 12.63$, 12.73, 12.83, 12.91, 12.73, and 12.8. It can be easily seen that the estimators based on RSSUR are close to μ, with smaller variances as compared to the estimators based on RSSU with equal replications. Bhoj and Kushary (2016) also computed the MSE of the estimators and showed that the estimators based on RSSUS are better than those based on RSSU.

REFERENCES

Amiri, S., Modarres, R., Bhoj, D.S., 2015. Ranked set sampling with the random subsamples. J. Stat. Comput. Simul. 85 (5), 935–946.

Bhoj, D.S., 1997. Estimation of parameters using modied ranked set sampling. In: Ahsanullah, M. (Ed.), Appl. Stat. Sci., vol. II. Nova Science, New York, pp. 145–163.

Bhoj, D.S., 2001. Ranked set sampling with unequal samples. Biometrics 57, 957–962.

Bhoj, D.S., 2002. New Median ranked set sampling. Pak. J. Stat.. 18, 135–141.

Bhoj, D.S., Ahsanullah, M., 1996. Estimation of parameters of the generalized geometric distribution using ranked set sampling. Biometrics 52, 685–694.

Bhoj, D.S., Kushary, D., 2014. Ranked set sampling with unequal samples and unequal replications. Pak. J. Stat. 30, 361–372.

Bhoj, D.S., Kushary, D., 2016. Ranked set sampling with unequal samples for skew distributions. J. Stat. Comput. Simul. 86, 676–681.

Biradar, B.S., Santosha, C.D., 2014. Estimation of the mean of exponential distribution using maximum ranked set sampling with unequal samples. Open J. Stat. 4, 641–649.

Chen, Z., Bai, Z., Sinha, B.K., 2004. Ranked Set Sampling, Theory and Applications. Springer, p. 2004.

Cressie, N.A.C., 1993. Statistics for Spatial Data. Wiley, New York.

Dave, T.R., Cutler, J.L., 1972. Ranked set sampling theory with order statistics background. Biometrics 28, 545–555.

Germayel, N.M., Stasny, E.A., Wolf, D.A., 2010. Optimal ranked set sampling estimation based on medians from multiple set sizes. J. Nonparametr. Stat. 22 (4), 517–527.

Lam, K., Sinha, B.K., Zong, W., 1994. Estimation of parameters in the two-parameter exponential distribution using ranked set sample. Ann. Inst. Stat. Math. 46, 723–736.

Li, D., Sinha, B.K., Ferron, F., 1999. Random selection in ranked set sampling and its applications. J. Stat. Plan. Inference 76 (1–2), 185–201.

McIntyre, G.A., 1952. A method of unbiased selective sampling using ranked sets. J. Agric. Res. 3, 385–390.

Muttlack, H.A., 1997. Median ranked set sampling. J. Appl. Stat. Sci. 6, 245–255.

Rahimov, I., Mutllak, H.A., 2003a. Estimation of the population mean using random selection in ranked set sampling. Stat. Probab. Lett. 62, 203–209.

Rahimov, I., Mutllak, H.A., 2003b. Investigating the estimation of the population mean using random ranked set samples. J. Nonparametr. Stat. 15 (3), 311–324.

Samawi, H.M., 2011. Varied set size ranked set sampling with applications to mean and ratio estimation. Int. J. Model. Simul. 31 (1), 6–13.

Stokes, S.L., 1995. Parametric ranked set sampling. Ann. Inst. Stat. Math. 47, 465–482.

Takahashi, K., Wakimoto, W., 1968. On unbiased estimates of the population mean based on the sample stratified by means of ordering. Ann. Inst. Stat. Math. 20, 1–31.

Zhang, L., Dong, X., Xu, X., 2014. Sign tests using ranked set sampling with unequal set sizes. Stat. Probab. Lett. 85, 69–77.

CHAPTER

A NEW MORGENSTERN TYPE BIVARIATE EXPONENTIAL DISTRIBUTION WITH KNOWN COEFFICIENT OF VARIATION BY RANKED SET SAMPLING 10

Vishal Mehta

Department of Mathematics, Jaypee University of Information Technology, Waknaghat, Himachal Pradesh, India

10.1 INTRODUCTION

Cost-effective sampling methods are of major concern in some experiments, especially when the measurement of the characteristics is costly, painful, or time-consuming. The concept of ranked set sampling (RSS) was first introduced by McIntyre (1952) as a process of increasing the precision of sample mean as an unbiased estimator of population mean. The method of RSS provides an effective way to achieve observational economy or to achieve relatively more precision per unit of sampling. RSS as described by McIntyre (1952) is applicable whenever ranking of a set of sampling units can be done easily by judgment method. For a detailed discussion on theory and application of RSS, see Chen et al. (2004). In certain situations one may prefer exact measurements of some easily measurable variable X associated with the study variable Y to rank the units of samples rather than ranking them by a crude judgment method. Suppose the variable of interest Y, is difficult or much more expensive to measure, but an auxiliary variable X correlated with Y is readily measureable and can be ordered exactly. In this case as an alternative to McIntyre's (1952) method of ranked set sampling, Stokes (1977) used an auxiliary variable for the ranking of sampling units.

If $X_{(r)r}$ is the observation measured on the auxiliary variable X from the unit chosen from the rth set then we write $Y_{[r]r}$ to denote the corresponding measurement made on the study variable Y on this unit, then $Y_{[r]r}, r = 1, 2, \ldots, n$ from the ranked set sample. Clearly $Y_{[r]r}$ is the concomitant of the rth order statistic arising from the rth sample.

In many areas, especially in physical science, it is common to find the population standard deviation is proportional to the population mean, that is, the coefficient of variation (CV) is constant (e.g., Sen, 1978; Ebrahimi, 1984, 1985; Singh, 1986). In such cases it is possible to find a more efficient estimator of the mean assuming that the coefficient of variation (CV) is known than by using the sample mean.

Ranked Set Sampling. DOI: https://doi.org/10.1016/B978-0-12-815044-3.00010-1

Let X be a random variable having the two-parameter exponential distribution as

$$f_X(x) = \frac{1}{\sigma}\exp\left(-\frac{x-\theta}{\sigma}\right); x \geq \theta > 0, \sigma > 0. \tag{10.1}$$

Here θ is the location parameter (guarantee period) and σ is the scale parameter (measuring the mean life). Since $E(X) = \theta + \sigma$ and $Var(X) = \sigma^2$, therefore the $CV = \frac{\sigma}{\theta+\sigma}$. Using the fact that the CV is some known constant we get that $\sigma = a_1\theta$, where $a_1(>0)$ is known (see, Samanta, 1984, 1985; Joshi and Nabar, 1991) and therefore Eq. (10.1) reduces to

$$f_X(x) = \frac{1}{a_1\theta}\exp\left(-\frac{x-\theta}{a_1\theta}\right); x \geq \theta > 0, a_1 > 0, \tag{10.2}$$

which has mean $\theta(a_1+1)$ and variance $\theta^2 a_1^2$, therefore the $CV = \frac{a_1}{(a_1+1)}$ is the same for all $\theta(>0)$.

The cumulative density function (cdf) of Eq. (10.2) is given by

$$F_X(x) = 1 - \exp\left(-\frac{x-\theta}{a_1\theta}\right); x \geq \theta > 0, a_1 > 0. \tag{10.3}$$

Ali and Woo (2002) considered parametric estimation of a special case of the two-parameter exponential distribution in which both the threshold (location) and the scale parameters are equal. For $a_1 = 1$ the probability density function (pdf) $f_X(x)$ in Eq. (10.2) reduces to:

$$f_X(x) = \frac{1}{\theta}\exp\left(-\frac{x-\theta}{\theta}\right); x \geq \theta, \tag{10.4}$$

which is due to Ali and Woo (2002).

A general family of bivariate distributions is proposed by Morgenstern (1956) with specified marginal distributions $F_X(x)$ and $F_Y(y)$ as

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)[1 + \alpha(1-F_X(x))(1-F_Y(y))]; \ -1 \leq \alpha \leq 1, \tag{10.5}$$

where α is the association parameter between X and Y and $F_{X,Y}(x,y)$ is the joint distribution function (df) and $F_X(x)$ and $F_Y(y)$ are the marginal distribution function (df) of X and Y respectively (see Johnson and Kotz, 1972).

Also, the probability density function (pdf) of the Morgenstern family of distribution can be given as

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)[1 + \alpha(1-2F_X(x))(1-2F_Y(y))]; \ -1 \leq \alpha \leq 1. \tag{10.6}$$

The pdf of the concomitants of order statistics $Y_{[r]r}$ arising from MTBED is obtained as (see Scaria and Nair, 1999)

$$f_{Y_{[r]r}}(y) = f_Y(y)\left[1 + \alpha\left(\frac{n-2r+1}{n+1}\right)(1-2F_Y(y))\right]; \ -1 \leq \alpha \leq 1. \tag{10.7}$$

Now using Eqs. (10.2) and (10.3) in Eq. (10.6) we get a member of this family is Morgenstern type bivariate exponential distribution (MTBED) with the probability density function (pdf) as

$$f_{X,Y}(x,y) = \frac{\exp\left\{\left(-\frac{x-\theta_1}{a_1\theta_1}\right) + \left(-\frac{y-\theta_2}{a_2\theta_2}\right)\right\}\left[1 + \alpha\left(1 - 2\exp\left(-\frac{x-\theta_1}{a_1\theta_1}\right)\right)\left(1 - 2\exp\left(-\frac{y-\theta_2}{a_2\theta_2}\right)\right)\right]}{a_1a_2\theta_1\theta_2};$$

$$x \geq \theta_1, y \geq \theta_2, a_1, a_2 > 0, \ -1 \leq \alpha \leq 1. \tag{10.8}$$

Now the pdf of $Y_{[r]r}$ for $1 \le r \le n$ is given as (see Scaria and Nair, 1999)

$$f_{Y_{[r]r}}(y) = \frac{1}{a_2\theta_2}\exp\left(-\frac{y-\theta_2}{a_2\theta_2}\right)\left[1-\alpha\left(\frac{n-2r+1}{n+1}\right)\left(1-2\exp\left(-\frac{y-\theta_2}{a_2\theta_2}\right)\right)\right]; \tag{10.9}$$

$$y \ge \theta_2, a_2 > 0, \ -1 \le \alpha \le 1.$$

The mean and variance of $Y_{[r]r}$ for $1 \le r \le n$ are respectively given by

$$E\left(Y_{[r]r}\right) = \theta_2 \xi_r \quad \text{and} \quad Var\left(Y_{[r]r}\right) = \theta_2^2 \delta_r, \tag{10.10}$$

where

$$\xi_r = \left[1 + a_2\left(1 - \frac{\alpha}{2}\left(\frac{n-2r+1}{n+1}\right)\right)\right]$$

and

$$\delta_r = a_2^2\left[1 - \frac{\alpha}{2}\left(\frac{n-2r+1}{n+1}\right) - \frac{\alpha^2}{4}\left(\frac{n-2r+1}{n+1}\right)^2\right].$$

Stokes (1995) has considered the estimation of parameters of location-scale family of distributions using RSS. Lam et al. (1994, 1995) have obtained the BLUEs of location and scale parameters of exponential distribution and logistic distribution. Stokes (1980) has considered the method of estimation of correlation coefficient of bivariate normal distribution using RSS. Modarres and Zheng (2004) have considered the problem of estimation of the dependence parameter using RSS. A robust estimate of correlation coefficient for bivariate normal distribution has been developed by Zheng and Modarres (2006). Stokes (1977) has suggested the ranked set sample mean as an estimator for the mean of the study variate Y, when an auxiliary variable X is used for ranking the sample units, under the assumption that (X, Y) follows a bivariate normal distribution. Estimation of a parameter of Morgenstern type bivariate exponential distribution by using RSS was considered by Chacko and Thomas (2008). Barnett and Moore (1997) have improved the estimator of Stokes (1977) by deriving the best linear unbiased estimator (BLUE) of the mean of the study variate Y, based on ranked set sample obtained on the study variate Y. Lesitha et al. (2010) have considered application of RSS in estimating parameters of Morgenstern type bivariate logistic distribution. Tahmasebi and Jafari (2012) have considered upper RSS. For current references in this context the reader is referred to Sharma et al. (2016), Bouza (2001, 2002, 2005), Samawi and Muttlak (1996), Demir and Singh (2000); Singh and Mehta (2013, 2014a,b, 2015, 2016a,b,c, 2017), Mehta and Singh (2014, 2015), and Mehta (2017).

The remaining part of the chapter is organized as follows: Section 10.2.1 proposes an unbiased estimator $\hat{\theta}_2$ of the parameter θ_2 involved in Eq. (10.8) using ranked set sample mean along with its variance. In Section 10.2.2, we have derived BLUE θ_2^* of θ_2, when the association parameter α is known. We have also given the variance of BLUE θ_2^*. Section 10.2.3 deals with the problem of estimating the parameter θ_2 based on unbalanced multistage RSS. We have derived BLUE $\hat{\theta}_2^{n(r)}$ of θ_2 and obtained its variance. In Sections 10.2.4 and 10.2.5, we have discussed the problem of estimating the parameter θ_2 based on unbalanced single-stage and steady-state RSS, respectively, which are particular cases of the studies presented in Section 10.3.1. Section 10.3.2 compares the performance of the different estimators proposed in the chapter through a numerical illustration. In Section 10.4 we conclude the chapter with final remarks.

10.2 EXPERIMENTAL METHODS AND MATERIALS

10.2.1 RANKED SET SAMPLE MEAN AS AN ESTIMATOR OF θ_2

Let (X, Y) be a bivariate random variable which follows an MTBED with pdf defined by Eq. (10.8). Suppose RSS in the sense of Stokes (1977) has been carried out. Let $X_{(r)r}$ be the observation measured on the auxiliary variate X in the rth unit of the RSS and let $Y_{[r]r}$ be the measurement made on the Y variate of the same unit $r = 1, 2, \ldots, n$. Then clearly $Y_{[r]r}$ is distributed as the concomitant of rth order statistics of a random sample of n arising from Eq. (10.8). By using the expressions for mean and variances of concomitants of order statistics arising from MTBED obtained in Eq. (10.10), we propose an estimator $\hat{\theta}_2$ of θ_2 involved in Eq. (10.8) and proved that it is an unbiased estimator of θ_2.

Theorem 1.1: *Let $Y_{[r]r}, r = 1, 2, \ldots, n$ be the ranked set sample observations on a study variate Y obtained out of ranking made on an auxiliary variate X, when (X, Y) follows MTBED as defined in Eq. (10.8). Then the ranked set sample mean given by*

$$\hat{\theta}_2 = \frac{1}{n(a_2+1)} \sum_{r=1}^{n} Y_{[r]r}, \tag{10.11}$$

is an unbiased estimator of θ_2 and its variance is given by

$$Var\left(\hat{\theta}_2\right) = \frac{a_2^2\theta_2^2}{n(a_2+1)^2}\left[1 - \frac{\alpha^2}{4n}\sum_{r=1}^{n}\left(\frac{n-2r+1}{n+1}\right)^2\right]. \tag{10.12}$$

Proof: *Taking expectations of both sides of Eq. (10.11) we have*

$$E\left(\hat{\theta}_2\right) = \frac{1}{n(a_2+1)}\sum_{r=1}^{n} E\left(Y_{[r]r}\right) = \frac{\theta_2}{n(a_2+1)}\sum_{r=1}^{n}\left[1 + a_2\left(\frac{n-2r+1}{n+1}\right)\right]. \tag{10.13}$$

It is clear to note that

$$\sum_{r=1}^{n}(n-2r+1) = 0. \tag{10.14}$$

Using Eq. (10.14) in Eq. (10.13) we get

$$E\left(\hat{\theta}_2\right) = \theta_2.$$

Thus $\hat{\theta}_2$ is an unbiased estimator of θ_2.
The variance of $\hat{\theta}_2$ is given by

$$Var\left(\hat{\theta}_2\right) = \frac{1}{n^2(a_2+1)^2}\sum_{r=1}^{n} Var\left(Y_{[r]r}\right).$$

Now using Eq. (10.10) and Eq. (10.14) in the above sum we get,

$$\mathrm{Var}\left(\hat{\theta}_2\right) = \frac{a_2^2\theta_2^2}{n(a_2+1)^2}\left[1 - \frac{\alpha^2}{4n}\sum_{r=1}^{n}\left(\frac{n-2r+1}{n+1}\right)^2\right].$$

Thus the theorem is proved.♦

10.2.2 BEST LINEAR UNBIASED ESTIMATOR OF θ_2

In this section we provide a better estimator of θ_2 than that of $\hat{\theta}_2$ by deriving the BLUE θ_2^* of θ_2 provided the parameter α is known. Let $X_{(r)r}$ be the observation measured on the auxiliary variable X in the rth unit of ranked set samples and let $Y_{[r]r}$ be measurement made on the Y variable of the same unit, $r = 1, 2, \ldots, n$. Let $\mathbf{Y}_{[n]} = \left(Y_{[1]1}, Y_{[2]2}, \ldots, Y_{[n]n}\right)'$ and if the parameter α involved in ξ_r and δ_r is known, then proceeding as in David and Nagaraja (2003, p.185) the BLUE θ_2^* of θ_2 is obtained as

$$\theta_2^* = \left(\xi' G^{-1}\xi\right)^{-1}\xi' G^{-1}\mathbf{Y}_{[n]} \tag{10.15}$$

and

$$\mathrm{Var}\left(\theta_2^*\right) = \left(\xi' G^{-1}\xi\right)^{-1}\theta_2^2, \tag{10.16}$$

where $\xi = \left(\xi_1, \xi_2, \ldots, \xi_n\right)'$ and $G = diag(\delta_1, \delta_2, \ldots, \delta_n)$. On substituting the values of ξ and G in Eqs. (10.15) and (10.16) and simplifying we have

$$\theta_2^* = \frac{\sum_{r=1}^{n}\left(\xi_r/\delta_r\right)Y_{[r]r}}{\sum_{r=1}^{n}\left(\xi_r^2/\delta_r\right)} \tag{10.17}$$

and

$$\mathrm{Var}\left(\theta_2^*\right) = \frac{\theta_2^2}{\sum_{r=1}^{n}\left(\xi_r^2/\delta_r\right)}. \tag{10.18}$$

10.2.3 ESTIMATION OF θ_2 BASED ON UNBALANCED MULTISTAGE RANKED SET SAMPLING

Al-Saleh and Al-Kadiri (2000) have extended first the usual concept of RSS to double-stage ranked set sampling (DSRSS) with the objective of increasing the precision of certain estimators of the population when compared with those obtained based on usual RSS or using random sampling. Al-Saleh and Al-Omari (2002) have further extended DSRSS to multistage ranked set sampling (MSRSS) and shown that there is an increase in the precision of estimators obtained based on MSRSS when compared with those based on usual RSS and DSRSS. The MSRSS (in r stages) procedure is described below:

(1) Randomly select n^{r+1} sample units from the target population, where r is the number of stages of MSRSS.

(2) Allocate the n^{r+1} selected units randomly into n^{r-1} sets, each of size n^2.

(3) For each set in step (2), apply the procedure of RSS method to obtain a (judgment) ranked set, of size n; this step yields n^{r-1} (judgment) ranked sets, of size n each.

(4) Arrange n^{r-1} ranked sets of size n each, into n^{r-2} sets of n^2 units each and without doing any actual quantification, apply ranked set sampling method on each set to yield n^{r-2} second stage ranked sets of size n each.

(5) This process is continued, without any actual quantification, until we end up with the rth stage (judgment) ranked set of size n.

(6) Finally, the n identified elements in step (5) are now quantified for the variable of interest.

Instead of the judgment method of ranking at each stage if there exists an auxiliary variate on which one can make measurement very easily and exactly and if the auxiliary variate is highly correlated with the variable under study, then we can apply ranking based on these measurements to obtain the ranked set units at each stage of MSRSS. Then, on the finally selected units, one can make measurement on the study variable.

In this section we deal with the MSRSS by assuming that the random variable (X, Y) has an MTBED as defined in Eq. (10.8), where Y is the study variable and X is an auxiliary variable. In Section 10.2.2, we have considered a method for estimating θ_2 using the $Y_{[r]r}$ measured on the study variate Y on the unit having rth smallest value observed on the auxiliary variable X, of the rth sample $r = 1, 2, \ldots, n$, and hence the RSS considered there was balanced.

Abo-Eleneen and Nagaraja (2002) have shown that, in a bivariate sample of size n arising from MTBED, the concomitant of largest-order statistic possesses the maximum Fisher information on θ_2 whenever $\alpha > 0$ and the concomitant of smallest order statistic possesses the maximum Fisher information on θ_2 whenever $\alpha < 0$. Hence, in this section, first we considered $\alpha > 0$ and carry out an unbalanced MSRSS with the help of measurements made on an auxiliary variate to choose the ranked set and then estimate θ_2 involved in MTBED based on the measurements made on the study variable. At each stage and from each set we choose a unit of a sample with the largest value on the auxiliary variable as the units of ranked sets with an objective of exploiting the maximum Fisher information on the ultimately chosen ranked set sample.

Let $U_i^{(r)}, i = 1, 2, \ldots, n$ be the units chosen by the (r stage) MSRSS. Since the measurement of an auxiliary variable on each unit $U_i^{(r)}, i = 1, 2, \ldots, n$ has the largest value, we may write $Y_{[n]i}^{(r)}, i = 1, 2, \ldots, n$ to denote the value measured on the variable of primary interest on $U_i^{(r)}, i = 1, 2, \ldots, n$. Then it is easy to see that each $Y_{[n]i}^{(r)}$ is the concomitant of the largest-order statistic of n^r independently and identically distributed bivariate random variables with MTBED. Moreover $Y_{[n]i}^{(r)}, i = 1, 2, \ldots, n$ are also independently distributed with *pdf* given by (see Scaria and Nair, 1999)

$$f_{[n]i}^{(r)}(y) = \frac{1}{a_2\theta_2}\exp\left(-\frac{y-\theta_2}{a_2\theta_2}\right)\left[1+\alpha\left(\frac{n^r-1}{n^r+1}\right)\left(1-2\exp\left(-\frac{y-\theta_2}{a_2\theta_2}\right)\right)\right]; \tag{10.19}$$

$$y \geq \theta_2, a_2 > 0, \ -1 \leq \alpha \leq 1.$$

Thus the mean and variance of $Y_{[n]i}^{(r)}, i = 1, 2, \ldots, n$ are given below

$$E\left(Y_{[n]i}^{(r)}\right) = \theta_2\left[1 + a_2\left\{1 + \frac{\alpha}{2}\left(\frac{n^r-1}{n^r+1}\right)\right\}\right] = \theta_2\xi_{n^r}, \tag{10.20}$$

$$Var\left(Y_{[n]i}^{(r)}\right) = \theta_2^2 a_2^2\left[1 + \frac{\alpha}{2}\left(\frac{n^r-1}{n^r+1}\right) - \frac{\alpha^2}{4}\left(\frac{n^r-1}{n^r+1}\right)^2\right] = \theta_2^2\delta_{n^r}, \tag{10.21}$$

where

$$\xi_{n^r} = \left[1 + a_2\left\{1 + \frac{\alpha}{2}\left(\frac{n^r-1}{n^r+1}\right)\right\}\right] \tag{10.22}$$

and

$$\delta_{n^r} = a_2^2\left[1 + \frac{\alpha}{2}\left(\frac{n^r-1}{n^r+1}\right) - \frac{\alpha^2}{4}\left(\frac{n^r-1}{n^r+1}\right)^2\right]. \tag{10.23}$$

Let $\mathbf{Y}_{[n]}^{(r)} = \left(Y_{[n]1}^{(r)}, Y_{[n]2}^{(r)}, \ldots, Y_{[n]n}^{(r)}\right)'$, then by using Eqs. (10.20) and (10.21) we get the mean vector and dispersion matrix of $\mathbf{Y}_{[n]}^{(r)}$ as

$$E\left(\mathbf{Y}_{[n]}^{(r)}\right) = \theta_2\xi_{n^r}\mathbf{1} \tag{10.24}$$

and

$$D\left(\mathbf{Y}_{[n]}^{(r)}\right) = \theta_2^2\delta_{n^r}\mathbf{I}, \tag{10.25}$$

where $\mathbf{1}$ is the column vector of n ones and $\mathbf{I}$ is a unit matrix of order n.

If $\alpha > 0$ involved in ξ_{n^r} and δ_{n^r} is known then Eqs. (10.24) and (10.25) together define a generalized Gass–Markov setup and hence the BLUE of θ_2 is obtained as

$$\hat{\theta}_2^{n(r)} = \frac{1}{n\xi_{n^r}}\sum_{i=1}^{n} Y_{[n]i}^{(r)} \tag{10.26}$$

with variance given by

$$Var\left(\hat{\theta}_2^{n(r)}\right) = \frac{\theta_2^2\delta_{n^r}}{n\left(\xi_{n^r}\right)^2}. \tag{10.27}$$

10.2.4 ESTIMATION OF θ_2 BASED ON UNBALANCED SINGLE-STAGE RANKED SET SAMPLING

If we take $r = 1$ in the MSRSS method described above, then we get the usual single-stage unbalanced RSS. By putting $r = 1$ in Eqs. (10.26) and (10.27) we get the BLUE $\hat{\theta}_2^{n(1)}$ of θ_2 based on single-stage unbalanced ranked set sampling as

$$\hat{\theta}_2^{n(1)} = \frac{1}{n\xi_n}\sum_{i=1}^{n} Y_{[n]i} \tag{10.28}$$

with variance

$$\mathrm{Var}\left(\hat{\theta}_2^{n(1)}\right) = \frac{\theta_2^2\delta_n}{n\left(\xi_n\right)^2}, \tag{10.29}$$

where, we write $Y_{[n]i}$ instead of $Y_{[n]i}^{(1)}$ and it represents the measurement on the study variable of the unit selected in the RSS. Also ξ_n and δ_n are obtained by putting $r = 1$ in Eqs. (10.22) and (10.23), respectively i.e.,

$$\xi_n = \left[1 + a_2\left\{1 + \frac{\alpha}{2}\left(\frac{n-1}{n+1}\right)\right\}\right] \tag{10.30}$$

and

$$\delta_n = a_2^2\left[1 + \frac{\alpha}{2}\left(\frac{n-1}{n+1}\right) - \frac{\alpha^2}{4}\left(\frac{n-1}{n+1}\right)^2\right]. \tag{10.31}$$

10.2.5 ESTIMATION OF θ_2 BASED ON UNBALANCED STEADY-STATE RANKED SET SAMPLING

Al-Saleh (2004) has considered the steady-state RSS by letting r go to $+\infty$. For the steady-state RSS the problem considered in having the asymptotic distribution of $Y_{[n]i}^{(r)}$ is given by

$$f_{[n]i}^{(\infty)}(y) = \frac{1}{a_2\theta_2}\exp\left(-\frac{y-\theta_2}{a_2\theta_2}\right)\left[1 + \alpha\left(1 - 2\exp\left(-\frac{y-\theta_2}{a_2\theta_2}\right)\right)\right]; \tag{10.32}$$
$$y \ge \theta_2, a_2 > 0, \ -1 \le \alpha \le 1.$$

From the definition of unbalanced MSRSS it follows that $Y_{[n]i}^{(\infty)}, i = 1, 2, \ldots, n$ are independent and identically distributed random variables each with *pdf* as defined in Eq. (10.32). Then $Y_{[n]i}^{(\infty)}, i = 1, 2, \ldots, n$ may be regarded as an unbalanced steady-state ranked set sample of size n. The mean and variance of $Y_{[n]i}^{(\infty)}, i = 1, 2, \ldots, n$ are given below

$$E\left(Y_{[n]i}^{(r)}\right) = \theta_2\left[1 + a_2\left\{1 + \frac{\alpha}{2}\right\}\right], \tag{10.33}$$

$$\mathrm{Var}\left(Y_{[n]i}^{(r)}\right) = \theta_2^2 a_2^2\left[1 + \frac{\alpha}{2} - \frac{\alpha^2}{4}\right]. \tag{10.34}$$

Let $\mathbf{Y}_{[n]}^{(\infty)} = \left(Y_{[n]1}^{(\infty)}, Y_{[n]2}^{(\infty)}, \ldots, Y_{[n]n}^{(\infty)}\right)'$. Then the BLUE $\hat{\theta}_2^{n(\infty)}$ based on $\mathbf{Y}_{[n]}^{(\infty)}$ and the variance of $\hat{\theta}_2^{n(\infty)}$ is obtained by taking the limits as $r \to \infty$ in Eqs. (10.26) and (10.27), respectively, and are given by

$$\hat{\theta}_2^{n(\infty)} = \frac{1}{n\left[1 + a_2\left(1 + \frac{\alpha}{2}\right)\right]}\sum_{i=1}^{n} Y_{[n]i}^{(\infty)} \tag{10.35}$$

and

$$\mathrm{Var}\left(\hat{\theta}_2^{n(\infty)}\right) = \theta_2^2\frac{a_2^2\left(1 + \frac{\alpha}{2} - \frac{\alpha^2}{4}\right)}{n\left[1 + a_2\left(1 + \frac{\alpha}{2}\right)\right]}. \tag{10.36}$$

Remark 1: *As mentioned earlier for MTBED the concomitant of smallest-order statistic possesses the maximum Fisher information on θ_2 whenever $\alpha < 0$. Therefore when $\alpha < 0$ we consider an* *unbalanced MSRSS in which at each stage and from each set we choose a unit of a sample with the smallest value on the auxiliary variable as the units of ranked sets with an objective of exploiting the maximum Fisher information on the ultimately chosen ranked set sample.*

Let $Y_{[1]i}^{(r)}, i = 1, 2, \ldots, n$, be the value measured on the variable of primary interest on the units selected at the rth stage of the unbalanced MSRSS. Then it is easily to see that each $Y_{[1]i}^{(r)}, i = 1, 2, \ldots, n$ is the concomitant of the smallest-order statistic of n^r independently and identically distributed bivariate random variables with MTBED. Moreover $Y_{[1]i}^{(r)}, i = 1, 2, \ldots, n$ are also independently distributed with pdf given by

$$f_{[1]i}^{(r)}(y) = \frac{1}{a_2\theta_2}\exp\left(-\frac{y-\theta_2}{a_2\theta_2}\right)\left[1-\alpha\left(\frac{n^r-1}{n^r+1}\right)\left(1-2\exp\left(-\frac{y-\theta_2}{a_2\theta_2}\right)\right)\right]; \quad y \geq \theta_2, a_2 > 0, -1 \leq \alpha \leq 1. \tag{10.37}$$

Clearly from Eqs. (10.19) and (10.37) we have

$$f_{[1]i}^{(r)}(y;\alpha) = f_{[n]i}^{(r)}(y; -\alpha) \tag{10.38}$$

and hence $E\left(Y_{[n]i}^{(r)}\right)$ for $\alpha > 0$ and $E\left(Y_{[1]i}^{(r)}\right)$ for $\alpha < 0$ are identically equal. Similarly, $Var\left(Y_{[n]i}^{(r)}\right)$ for $\alpha > 0$ and $Var\left(Y_{[1]i}^{(r)}\right)$ for $\alpha < 0$ are identically equal. Consequently, if $\hat{\theta}_2^{1(1)}$ is the BLUE of θ_2, involved in MTBED for $\alpha < 0$, based on the unbalanced MSRSS observations $Y_{[1]i}^{(r)}, i = 1, 2, \ldots, n$ then the coefficients of $Y_{[1]i}^{(r)}, i = 1, 2, \ldots, n$ in the BLUE $\hat{\theta}_2^{1(1)}$ for $\alpha < 0$ is the same as the coefficients of $Y_{[n]i}^{(r)}, i = 1, 2, \ldots, n$ in the BLUE $\hat{\theta}_2^{n(r)}$ for $\alpha > 0$. Further we have $Var\left(\hat{\theta}_2^{1(1)}\right) = Var\left(\hat{\theta}_2^{n(r)}\right)$ and hence $Var\left(\hat{\theta}_2^{1(1)}\right) = Var\left(\hat{\theta}_2^{n(1)}\right)$ and $Var\left(\hat{\theta}_2^{1(\infty)}\right) = Var\left(\hat{\theta}_2^{n(\infty)}\right)$, where $\hat{\theta}_2^{1(1)}$ are the BLUE of θ_2 for $\alpha < 0$ based on the usual unbalanced single stage RSS observations $Y_{[n]i}, i = 1, 2, \ldots, n$ and $\hat{\theta}_2^{1(\infty)}$ are the BLUE of θ_2 for $\alpha < 0$ based on the unbalanced steady-state RSS observations $Y_{[n]i}^{(\infty)}, i = 1, 2, \ldots, n$.

Remark 2: *If we have a situation with α unknown, we introduce an estimator (moment type) for α as follows. For MTBED the correlation coefficient between the two variables is given by $\rho = \frac{\alpha}{4}$. If q is the sample correlation coefficient between $X_{(i)i}$ and $Y_{[i]i}, i = 1, 2, \ldots, n$ then the moment type estimator for α is obtained by equating with the population correlation coefficient ρ and is obtained as (see Chacko and Thomas, 2008):*

$$\hat{\alpha} = \begin{cases} -1 & \text{if } q < -1/4 \\ 4q & \text{if } -\frac{1}{4} \leq q \leq \frac{1}{4}. \\ 1 & \text{if } q > 1/4 \end{cases} \tag{10.39}$$

10.3 OBSERVATIONS, RESULTS, AND DISCUSSION

10.3.1 RELATIVE EFFICIENCY

We have obtained the relative efficiencies $e_1 = RE\left(\theta_2^*, \hat{\theta}_2\right) = \frac{\mathrm{Var}(\hat{\theta}_2)}{\mathrm{Var}(\theta_2^*)}, e_2 = RE\left(\hat{\theta}_2^{n(1)}, \hat{\theta}_2\right) = \frac{\mathrm{Var}(\hat{\theta}_2)}{\mathrm{Var}\left(\hat{\theta}_2^{n(1)}\right)}$ and $e_3 = RE\left(\hat{\theta}_2^{n(\infty)}, \hat{\theta}_2\right) = \frac{Var(\hat{\theta}_2)}{Var\left(\hat{\theta}_2^{n(\infty)}\right)}$ of $\theta_2^*, \hat{\theta}_2^{n(1)}$ and $\hat{\theta}_2^{n(\infty)}$ relative to $\hat{\theta}_2$ respectively, for $n = 2(2)20, \alpha = 0.25(0.25)1.00$ and $a_2 = 1(1)5$ and these are presented in Table 10.1.

Table 10.1 The Values of $e_i^{'}s, i = 1, 2, 3$

		$a_2 = 1$			$a_2 = 2$			$a_2 = 3$		
n	α	e_1	e_2	e_3	e_1	e_2	e_3	e_1	e_2	e_3
2	0.25	1.0008	1.0005	1.0160	1.0000	1.0138	1.0559	1.0004	1.0210	1.0766
	0.50	1.0016	1.0008	1.0581	1.0009	1.0280	1.1383	1.0004	1.0415	1.1793
	0.75	1.0041	1.0013	1.1241	1.0023	1.0416	1.2463	1.0014	1.0617	1.3093
	1.00	1.0075	1.0016	1.2150	1.0037	1.0537	1.3824	1.0022	1.0803	1.4703
4	0.25	1.0000	1.0034	1.0143	1.0009	1.0281	1.0549	1.0000	1.0401	1.0751
	0.50	1.0033	1.0118	1.0521	1.0018	1.0595	1.1316	1.0014	1.0842	1.1729
	0.75	1.0083	1.0235	1.1095	1.0047	1.0946	1.2304	1.0037	1.1307	1.2928
	1.00	1.0224	1.0388	1.1880	1.0125	1.1311	1.3517	1.0083	1.1782	1.4369
6	0.25	1.0000	1.0052	1.0135	1.0000	1.0343	1.0540	1.0000	1.0489	1.0743
	0.50	1.0024	1.0182	1.0487	1.0027	1.0762	1.1296	1.0022	1.1052	1.1704
	0.75	1.0126	1.0397	1.1049	1.0070	1.1225	1.2235	1.0044	1.1652	1.2852
	1.00	1.0316	1.0628	1.1760	1.0190	1.1730	1.3382	1.0138	1.2299	1.4230
8	0.25	1.0000	1.0060	1.0127	1.0000	1.0375	1.0530	1.0000	1.0536	1.0736
	0.50	1.0033	1.0225	1.0470	1.0037	1.0863	1.1285	1.0014	1.1172	1.1687
	0.75	1.0135	1.0481	1.1004	1.0075	1.1387	1.2190	1.0045	1.1853	1.2805
	1.00	1.0355	1.0771	1.1680	1.0236	1.1992	1.3312	1.0170	1.2615	1.4154
10	0.25	1.0000	1.0079	1.0135	1.0023	1.0417	1.0545	1.0000	1.0571	1.0736
	0.50	1.0082	1.0283	1.0487	1.0023	1.0920	1.1270	1.0018	1.1249	1.1674
	0.75	1.0127	1.0532	1.0967	1.0071	1.1495	1.2161	1.0056	1.2001	1.2791
	1.00	1.0402	1.0890	1.1650	1.0248	1.2151	1.3248	1.0195	1.2812	1.4090
12	0.25	1.0000	1.0063	1.0111	1.0000	1.0431	1.0540	1.0000	1.0603	1.0743
	0.50	1.0049	1.0312	1.0487	1.0028	1.0966	1.1265	1.0000	1.1291	1.1653
	0.75	1.0152	1.0593	1.0967	1.0114	1.1597	1.2167	1.0067	1.2093	1.2767
	1.00	1.0486	1.0987	1.1640	1.0269	1.2271	1.3210	1.0211	1.2974	1.4068
14	0.25	1.0000	1.0101	1.0143	1.0000	1.0436	1.0530	1.0000	1.0614	1.0736
	0.50	1.0057	1.0292	1.0445	1.0032	1.1005	1.1265	1.0025	1.1358	1.1674
	0.75	1.0179	1.0613	1.0940	1.0099	1.1663	1.2161	1.0052	1.2156	1.2743
	1.00	1.0506	1.1048	1.1620	1.0315	1.2394	1.3216	1.0219	1.3085	1.4041

Table 10.1 The Values of $e_i^{'}s, i = 1, 2, 3$ ***Continued***

		$a_2 = 1$			$a_2 = 2$			$a_2 = 3$		
n	α	e_1	e_2	e_3	e_1	e_2	e_3	e_1	e_2	e_3
16	0.25	1.0065	1.0122	1.0160	1.0036	1.0465	1.0549	1.0000	1.0628	1.0736
	0.50	1.0066	1.0300	1.0436	1.0037	1.1034	1.1265	1.0029	1.1374	1.1653
	0.75	1.0204	1.0676	1.0967	1.0076	1.1681	1.2122	1.0060	1.2227	1.2748
	1.00	1.0507	1.1091	1.1600	1.0280	1.2430	1.3158	1.0252	1.3176	1.4025
18	0.25	1.0000	1.0077	1.0111	1.0000	1.0465	1.0540	1.0000	1.0635	1.0732
	0.50	1.0074	1.0314	1.0436	1.0000	1.1027	1.1234	1.0033	1.1415	1.1666
	0.75	1.0153	1.0678	1.0940	1.0128	1.1753	1.2150	1.0067	1.2257	1.2725
	1.00	1.0574	1.1151	1.1610	1.0362	1.2533	1.3190	1.0248	1.3225	1.3988
20	0.25	1.0000	1.0064	1.0095	1.0000	1.0453	1.0521	1.0000	1.0648	1.0736
	0.50	1.0082	1.0376	1.0487	1.0046	1.1056	1.1244	1.0036	1.1425	1.1653
	0.75	1.0169	1.0728	1.0967	1.0095	1.1772	1.2133	1.0075	1.2295	1.2720
	1.00	1.0545	1.1183	1.1600	1.0302	1.2525	1.3120	1.0236	1.3289	1.3982

		$a_2 = 4$			$a_2 = 5$		
n	α	e_1	e_2	e_3	e_1	e_2	e_3
2	0.25	1.0000	1.0249	1.0887	1.0000	1.0277	1.0970
	0.50	1.0003	1.0498	1.2043	1.0003	1.0551	1.2210
	0.75	1.0006	1.0737	1.3477	1.0006	1.0818	1.3738
	1.00	1.0016	1.0964	1.5244	1.0012	1.1073	1.5611
4	0.25	1.0000	1.0474	1.0873	1.0000	1.0526	1.0957
	0.50	1.0006	1.0986	1.1975	1.0006	1.1082	1.2139
	0.75	1.0026	1.1523	1.3306	1.0018	1.1668	1.3561
	1.00	1.0066	1.2072	1.4896	1.0055	1.2265	1.5250
6	0.25	1.0000	1.0580	1.0870	1.0000	1.0635	1.0948
	0.50	1.0010	1.1223	1.1948	1.0009	1.1336	1.2110
	0.75	1.0039	1.1915	1.3233	1.0036	1.2094	1.3493
	1.00	1.0101	1.2642	1.4744	1.0093	1.2885	1.5107
8	0.25	1.0000	1.0639	1.0866	1.0012	1.0705	1.0951
	0.50	1.0013	1.1357	1.1929	1.0012	1.1486	1.2096
	0.75	1.0039	1.2150	1.3195	1.0024	1.2333	1.3440
	1.00	1.0136	1.2986	1.4661	1.0125	1.3251	1.5019
10	0.25	1.0000	1.0669	1.0856	1.0000	1.0733	1.0935
	0.50	1.0016	1.1446	1.1918	1.0015	1.1588	1.2093
	0.75	1.0033	1.2291	1.3156	1.0045	1.2507	1.3424
	1.00	1.0153	1.3212	1.4602	1.0141	1.3491	1.4959
12	0.25	1.0000	1.0700	1.0859	1.0000	1.0767	1.0938
	0.50	1.0019	1.1512	1.1914	1.0018	1.1660	1.2089
	0.75	1.0039	1.2405	1.3143	1.0054	1.2625	1.3408
	1.00	1.0185	1.3392	1.4582	1.0151	1.3671	1.4926

(*Continued*)

Table 10.1 The Values of $e_i^{\prime}s, i = 1, 2, 3$ *Continued*

		$a_2 = 4$			$a_2 = 5$		
n	α	e_1	e_2	e_3	e_1	e_2	e_3
14	0.25	1.0000	1.0718	1.0856	1.0000	1.0796	1.0945
	0.50	1.0022	1.1560	1.1910	1.0000	1.1698	1.2071
	0.75	1.0069	1.2503	1.3148	1.0042	1.2710	1.3392
	1.00	1.0168	1.3504	1.4543	1.0155	1.3793	1.4889
16	0.25	1.0000	1.0730	1.0853	1.0000	1.0807	1.0938
	0.50	1.0026	1.1604	1.1914	1.0000	1.1737	1.2068
	0.75	1.0052	1.2539	1.3109	1.0048	1.2771	1.3376
	1.00	1.0192	1.3620	1.4543	1.0177	1.3898	1.4871
18	0.25	1.0000	1.0750	1.0859	1.0000	1.0820	1.0938
	0.50	1.0029	1.1624	1.1903	1.0026	1.1782	1.2079
	0.75	1.0059	1.2617	1.3131	1.0054	1.2840	1.3384
	1.00	1.0186	1.3680	1.4509	1.0171	1.3982	1.4857
20	0.25	1.0000	1.0740	1.0839	1.0000	1.0844	1.0951
	0.50	1.0000	1.1647	1.1899	1.0029	1.1806	1.2075
	0.75	1.0033	1.2627	1.3092	1.0030	1.2852	1.3344
	1.00	1.0207	1.3750	1.4504	1.0190	1.4049	1.4843

It is observed from Table 10.1 that

- for fixed a_2, the values of $e_i{}'s, i = 1, 2, 3$ increase as n increase;
- for fixed n, the value of $e_i{}'s, i = 1, 2, 3$ increase as α increases;
- the values of $e_i{}'s, i = 1, 2, 3$ greater than "unity" for all values of (n, α, a_2), which follows that the estimators $\theta_2^*, \hat{\theta}_2^{n(1)}$ and $\hat{\theta}_2^{n(\infty)}$ are more efficient than unbiased estimator $\hat{\theta}_2$;
- when n is fixed, larger gain in efficiencies are observed for large values of α and all values of a_2;
- the values of $e_i{}'s, i = 2, 3$ increase as the value of a_2 increases. It follows that the larger gain in efficiency by using $\hat{\theta}_2^{n(1)}$ and $\hat{\theta}_2^{n(\infty)}$ over $\hat{\theta}_2$ can be obtained when the population is more heterogeneous. No trend is observed for e_1 as a_2 increases.

Therefore we conclude that the BLUE of steady-state RSS $\hat{\theta}_2^{n(\infty)}$ of θ_2 is a better estimator of $\hat{\theta}_2$, θ_2^* and $\hat{\theta}_2^{n(1)}$, respectively.

10.4 CONCLUSION

In this chapter, taking the motivation from Ebrahimi (1984, 1985), we have developed a new Morgenstern type bivariate exponential distribution (MTBED) with known coefficients of variation (CV) using the results due to Morgenstern (1956) and Scaria and Nair (1999). The mean and variance of newly developed MTBED with known CV have also been obtained. We have discussed the problem of estimating parameter θ_2 in MTBED in the presence of known CV. For estimating the parameter θ_2 of MTBED, we have derived an unbiased estimator $\hat{\theta}_2$ using ranked set sample mean and the BLUE θ_2^* based on RSS and their variances are given. We have further addressed the problem of estimating θ_2 using unbalanced RSS and its special cases known as unbalanced single-stage and steady-state RSS are also discussed. The reflective performance of the various proposed estimators of the parameter θ_2 are evaluated through numerical illustration and finally obtained that the BLUE of the steady-state RSS $\hat{\theta}_2^{n(\infty)}$ is more efficient among the estimators discussed in the chapter.

ACKNOWLEDGMENTS

The author is highly thankful to Prof. H.P. Singh, School of Studies in Statistics, Vikram University, Ujjain, Madhya Pradesh, India, for the guidance and constructive suggestions and Prof. C.N. Bouza, Department of Applied Mathematics, University of Havana, Havana, Cuba, for choosing my chapter in this book. Last, but not the least, I am also thankful to all teaching and nonteaching staff members of the Jaypee University of Information Technology (JUIT), Waknaghat, Solan, Himachal Pradesh, India, for providing all the necessary facilities.

REFERENCES

Abo-Eleneen, Z.A., Nagaraja, H.N., 2002. Fisher information in an order statistic and its concomitant. Ann. Inst. Stat. Math. 54, 667–680.

Al-Saleh, M.F., 2004. Steady-state ranked set sampling and parametric inference. J. Stat. Plan. Inference 123, 83–95.

Al-Saleh, M.F., Al-Kadiri, M., 2000. Double ranked set sampling. Stat. Probab. Lett. 48, 205–212.

Al-Saleh, M.F., Al-Omari, A., 2002. Multistage ranked set sampling. J. Stat. Plan. Inference 102, 273–286.

Ali, M.M., Woo, J., 2002. Estimation in an exponential distribution with common location and scale parameters. Calcutta Stat. Assoc. Bull. 53, 203–211.

Barnett, V., Moore, K., 1997. Best linear unbiased estimates in ranked set sampling with particular reference to imperfect ordering. J. Appl. Stat. 24, 697–710.

Bouza, C.N., 2001. Model assisted ranked survey sampling. Biom. J. 43, 249–259.

Bouza, C.N., 2002. Ranked set sampling the non-response stratum for estimating the difference of means. Biom. J. 44, 903–915.

Bouza, C.N., 2005. Sampling using ranked sets: concepts, results and perspectives. Rev. Investig. Oper. 26 (3), 275–293.

Chacko, M., Thomas, Y., 2008. Estimation of a parameter of Morgenstern type bivariate exponential by ranked set sampling. Ann. Inst. Stat. Math. 60, 273–300.

Chen, Z., Bai, Z., Sinha, B.K., 2004. Lecture Notes in Statistics, Ranked Set Sampling, Theory and Applications. Springer, New York.

David, H.A., Nagaraja, H.N., 2003. Order Statistics. Wiley, New York.

Demir, S., Singh, H., 2000. An application of the regression estimates to ranked set sampling. Hacit. Bull Nat. Sc. Eng. Ser. B 29, 93–101.

Ebrahimi, N., 1984. Maximum likelihood estimation, from double censored samples, of the mean of a normal distribution with known coefficient of variation. Commun. Stat. 651–661.

Ebrahimi, N., 1985. Estimation from censored samples the location of an exponential distribution with known coefficient of variation. Calcutta Stat. Assoc. Bull. 34, 169–177.

Johnson, N.L., Kotz, S., 1972. Distribution in Statistics, Continuous Multivariate distribution. John Wiley, New York.

Joshi, S.M., Nabar, S.P., 1991. Testing the scale parameter of the exponential distribution with known coefficient of variation. Commun. Stat.: Theory Methods 20, 747–756.

Lam, K., Sinha, B.K., Wu, Z., 1994. Estimation of a two-parameter exponential distribution using ranked set sample. Ann. Inst. Stat. Math. 46, 723–736.

Lam, K., Sinha, B.K., Wu, Z., 1995. Estimation of location and scale parameters of a logistic distribution using ranked set sample. In: Nagaraja, H.N., Sen, P.K., Morrison, D.F. (Eds.), Statistical Theory and Applications: Papers in Honor of Herbert A. David. Springer, New-York.

Lesitha, G., Thomas, P.Y., Chacko, M., 2010. Application of ranked set sampling in estimating parameters of Morgenstern Type bivariate logistic distribution. Calcutta Stat. Assoc. Bull. 62, 71–89.

McIntyre, G., 1952. A method for unbiased selective sampling using ranked set sampling. Aust. J. Agric. Res. 3, 385–390.

Mehta, V., 2017. Shrinkage Estimator of the Parameters of Normal Distribution Based On K- Record Values. Int. J. Sci. Res. Math. Stat. Sci. 4 (1), 1–5.

Mehta, V., Singh, H.P., 2014. Shrinkage estimators of parameters of Morgenstern type bivariate logistic distribution using ranked set sampling. J. Basic Appl. Eng. Res. 1 (13), 1–6.

Mehta, V., Singh, H.P., 2015. Minimum mean square error estimation of parameters in bivariate normal distribution using concomitants of record values. Statistics and Informatics in Agricultural Research. Indian Society of Agricultural Statistics, Excel India Publishers, Munirka, New Delhi, pp. 162–174.

Modarres, R., Zheng, G., 2004. Maximum likelihood estimation of dependence parameter using ranked set sampling. Stat. Probab. Lett. 68, 315–323.

Morgenstern, D., 1956. Einfache Beispiele Zweidimensionaler Verteilunge. Mitteilungs blatt fiir Mathematische Statistik 8, 234–235.

Samanta, M., 1984. Estimation of the location parameter of an exponential distribution with known coefficient of variation. Commun. Stat.: Theory Methods 13, 1357–1364.

Samanta, M., 1985. On estimation of the location parameter of an exponential distribution with known coefficient of variation. Calcutta Stat. Assoc. Bull. 34, 43–50.

Samawi, H.M., Muttlak, H.A., 1996. Estimation of a ratio using ranked set sampling. Biom. J. 36, 753–764.

Scaria, J., Nair, U., 1999. On concomitants of order statistics from Morgenstern family. Biom. J. 41, 483–489.

Sen, A.R., 1978. Estimation of the population mean when the coefficient of variation is known. Commun. Stat. 657–672.

Sharma, P., Bouza, C.N., Verma, H., Singh, R., Sautto, J.M., 2016. A generalized class of estimators for the finite population mean when the study variable is qualitative in nature. Rev. Investig. Oper. 37 (2), 163–172.

Singh, H.P., 1986. A note on the estimation of variance of sample mean using the knowledge of coefficient of variation in normal population. Commun. Stat.: Theory Methods 15 (12), 3737–3746.

Singh, H.P., Mehta, V., 2013. An improved estimation of parameters of Morgenstern type bivariate logistic distribution using ranked set sampling. Statistica 73 (4), 437–461.

Singh, H.P., Mehta, V., 2014a. Linear shrinkage estimator of scale parameter of Morgenstern type bivariate logistic distribution using ranked set sampling. Model Assist. Stat. Appl. 9, 295–307.

Singh, H.P., Mehta, V., 2014b. An alternative estimation of the scale parameter for morgenstern type bivariate log-logistic distribution using ranked set sampling. J. Reliab. Stat. Stud. 7 (1), 19–29.

Singh, H.P., Mehta, V., 2015. Estimation of scale parameter of a Morgenstern type bivariate uniform distribution using censored ranked set samples. Model Assist. Stat. Appl. 10, 139−153.

Singh, H.P., Mehta, V., 2016a. Improved estimation of scale parameter of Morgenstern type bivariate uniform distribution using ranked set sampling. Commun. Stat.: Theory Methods 45 (5), 1466−1476.

Singh, H.P., Mehta, V., 2016b. Some classes of shrinkage estimators in the Morgenstern type bivariate exponential distribution using ranked set sampling. Hacet. J. Math. Stat. 45 (2), 575−591.

Singh, H.P., Mehta, V., 2016c. A class of shrinkage estimators of scale parameter of uniform distribution based on K-record values. Natl. Acad. Sci. Lett. 39, 221−227.

Singh, H.P., Mehta, V., 2017. Improved estimation of the scale parameter for log-logistic distribution using balanced ranked set sampling. Stat. Transit. N. Ser. 18 (1), 53−74.

Stokes, S.L., 1977. Ranked set sampling with concomitant variables. Commun. Stat.: Theory Methods 6, 1207−1211.

Stokes, S.L., 1980. Inferences on the correlation coefficient in bivariate normal populations from ranked set samples. J. Am. Stat. Assoc. 75, 989−995.

Stokes, S.L., 1995. Parametric ranked set sampling. Ann. Inst. Stat. Math. 47, 465−482.

Tahmasebi, S., Jafari, A.A., 2012. Estimation of a scale parameter of Morgenstern type bivariate uniform distribution by ranked set sampling. J. Data Sci. 10, 129−141.

Zheng, G., Modarres, R., 2006. A robust estimate of correlation coefficient for bivariate normal distribution using ranked set sampling. J. Stat. Plan. Inference 136, 298−309.

CHAPTER

11 SHRINKAGE ESTIMATORS OF SCALE PARAMETER TOWARDS AN INTERVAL OF MORGENSTERN TYPE BIVARIATE UNIFORM DISTRIBUTION USING RANKED SET SAMPLING

Vishal Mehta

Department of Mathematics, Jaypee University of Information Technology, Waknaghat, Himachal Pradesh, India

11.1 INTRODUCTION

Ranked set sampling (RSS) is a method of sampling that can be advantageous when quantification of all sampling units is costly but a small set of units can be easily ranked, according to the character under investigation, without actual quantification. The technique was first introduced by McIntyre (1952) for estimating mean pasture and forage yields. The theory and applications of RSS are given by Chen et al. (2004). Suppose the variable of interest, Y, is difficult or much too expensive to measure, but an auxiliary variable X correlated with Y is readily measureable and can be ordered exactly. In this case, as an alternative to McIntyre's (1952) method of ranked set sampling, Stokes (1977) used an auxiliary variable for the ranking of sampling units. If $X_{(r)r}$ is the observation measured on the auxiliary variable X from the unit chosen from the rth set then we write $Y_{[r]r}$ to denote the corresponding measurement made on the study variable Y on this unit, then $Y_{[r]r}, r = 1, 2, \ldots, n$, from the ranked set sample. Clearly, $Y_{[r]r}$ is the concomitant of the rth order statistic arising from the rth sample. Stokes (1995) has obtained the estimation of parameters of the location-scale family of distribution by RSS. Lam et al. (1994) used RSS to estimate the two-parameter exponential distribution. Al-Saleh and Ananbeh (2005, 2007) estimated the means of the bivariate normal distribution using moving extremes RSS with a concomitant variable. Al-Saleh and Diab (2009) considered estimation of the parameters of Downton's bivariate exponential distribution using an RSS scheme. Barnett and Moore (1997) derived the best linear unbiased estimator (BLUE) for the mean of Y, based on a ranked set sample obtained using an auxiliary variable X for ranking the sample units.

In the estimation of an unknown parameter there often exists some prior knowledge about the parameter which one would like to utilize in order to get a better estimate. The Bayesian approach is a well-known example in which prior knowledge about the parameter is available in the form of

Ranked Set Sampling. DOI: https://doi.org/10.1016/B978-0-12-815044-3.00011-3

prior distribution. For current references in this context the reader is referred to Sharma et al. (2016), Bouza (2001, 2002, 2005), Samawi and Muttlak (1996), Demir and Singh (2000), Singh and Mehta (2013, 2014a,b, 2015, 2016a,b,c, 2017), Mehta and Singh (2015, 2014), and Mehta (2017).

The organization of this chapter is as follows. Section 11.2 introduces the general distribution theory, properties of Farlie–Gumbel–Morgenstern (FGM) distribution/Morgenstern distribution and a brief review of the estimators of the scale parameter θ_2 envisaged by Tahmasebi and Jafari (2012). In Section 11.3, some improved shrinkage toward interval estimators are described on the lines of Singh et al. (1973), Searls and Intarapanich (1960), Searls (1964), Jani (1991), and Kourouklis (1994), the expressions of bias and mean squared error (MSE) are obtained and compared with usual unbiased estimators. In Section 11.4, we have computed the relative efficiencies of different estimators numerically to evaluate their performance. Section 11.5 concludes the chapter with some final remarks.

11.2 REVIEW OF RSS IN FGM FAMILY OF DISTRIBUTION

A general family of bivariate distributions is proposed by Morgenstern (1956) with specified marginal distributions $F_X(x)$ and $F_Y(y)$ as

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)[1+\alpha(1-F_X(x))\,(1-F_Y(y))];\ -1\le\alpha\le 1, \tag{11.1}$$

where α is the association parameter between X and Y.

A member of this family is Morgenstern type bivariate uniform distribution (MTBUD) with the probability density function (pdf)

$$f_{X,Y}(x,y) = \frac{1}{\theta_1\theta_2}\left[1+\alpha\left(1-\frac{2x}{\theta_1}\right)\left(1-\frac{2y}{\theta_2}\right)\right];0<x<\theta_1, 0<y<\theta_2. \tag{11.2}$$

The pdf of $Y_{[r]r}$ for $1\le r\le n$ is given by [see Scaria and Nair (1999)]

$$g_{Y_{[r]r}}(y) = \int f_{Y|X}(y|x)\,f_r(x)\,dx = \frac{1}{\theta_2}\left[1+\alpha\left(\frac{n-2r+1}{n+1}\right)\left(1-\frac{2y}{\theta_2}\right)\right];\ 0<y<\theta_2,$$

where $f_r(x)$ is the density function of $X_{(r)r}$, i.e.,

$$f_r(x) = \frac{n!}{(r-1)!(n\text{-}r)!}\left[\frac{x^{r-1}(\theta_1-x)^{n-r}}{\theta_1^n}\right];\ 0<x<\theta_1,$$

and therefore, the mean and variance of $Y_{[r]r}$ for $1\le r\le n$ are, respectively, given by

$$E\left[Y_{[r]r}\right] = \theta_2\beta_r \quad \text{and} \quad \mathrm{Var}\left[Y_{[r]r}\right] = \theta_2^2\lambda_r, \tag{11.3}$$

where

$$\beta_r = \frac{1}{2}\left[1-\frac{\alpha}{3}\left(\frac{n-2r+1}{n+1}\right)\right] \quad \text{and} \quad \lambda_r = \frac{1}{12}\left[1-\frac{\alpha^2}{3}\left(\frac{n-2r+1}{n+1}\right)^2\right]$$

Let $Y_{[r]r}, r=1,2,\ldots,n$, be the RSS observations made on the units of the ranked set sampling regarding the study variable Y, which is correlated with the auxiliary variable X, when (X,Y) follows

MTBUD as defined in Eq. (11.2). Then an unbiased estimator for θ_2 based on RSS mean in Eq. (11.3) is given as [see Tahmasebi and Jafari (2012)]

$$t_1 = \hat{\theta}_{2,\mathrm{RSS}} = \frac{2}{n}\sum_{r=1}^{n} Y_{[r]r},$$

and its variance is

$$\mathrm{Var}(t_1) = \frac{\theta_2^2}{3n}\left[1 - \frac{\alpha^2}{3n}\sum_{r=1}^{n}\left(\frac{n-2r+1}{n+1}\right)^2\right] = \theta_2^2 V_1, \tag{11.4}$$

where

$$V_1 = \frac{1}{3n}\left[1 - \frac{\alpha^2}{3n}\sum_{r=1}^{n}\left(\frac{n-2r+1}{n+1}\right)^2\right].$$

When the parameter α is known, Tahmasebi and Jafari (2012) have suggested a BLUE θ_2^* of θ_2, which is more efficient than the estimator $\hat{\theta}_{2,\mathrm{RSS}}$ and is given as:

$$t_2 = \theta_2^* = \sum_{r=1}^{n}\left(\frac{\beta_r}{\lambda_r}\right)\left(\sum_{i=1}^{n}\left(\frac{\beta_i^2}{\lambda_i}\right)\right)^{-1} Y_{[r]r},$$

whose variance is

$$\mathrm{Var}(t_2) = \theta_2^2\left(\sum_{r=1}^{n}\left(\frac{\beta_r^2}{\lambda_r}\right)\right)^{-1} = \theta_2^2 V_2, \tag{11.5}$$

where

$$V_2 = \left(\sum_{r=1}^{n}\left(\frac{\beta_r^2}{\lambda_r}\right)\right)^{-1}.$$

Further, Tahmasebi and Jafari (2012) derived BLUE of θ_2 based on the upper ranked set sample (URSS) as

$$t_3 = \tilde{\theta}_2 = \frac{1}{n\beta_n}\sum_{r=1}^{n} Y_{[n]r},$$

and its variance is given by

$$\mathrm{Var}(t_3) = \theta_2^2\frac{\lambda_n}{n\beta_n^2} = \theta_2^2 V_3, \tag{11.6}$$

where

$$V_3 = \frac{\lambda_n}{n\beta_n^2}.$$

Using the extreme ranked set sampling (ERSS) method, Tahmasebi and Jafari (2012) also derived different estimators for θ_2 with concomitant variable for n. Below we have used the same notations ERSS$_1$, ERSS$_2$ and ERSS$_3$ as defined in Tahmasebi and Jafari (2012), pp. 134–135).

If n is even then the estimator of the θ_2 using ERSS$_1$ is defined as

$$t_4 = \hat{\theta}_{2,\mathrm{ERSS}_1} = \frac{2}{n}\sum_{r=1}^{n/2} Y_{[1]2r-1} + Y_{[n]2r},$$

and its variance is given by

$$\mathrm{Var}(t_4) = \frac{\theta_2^2}{3n}\left[1 - \frac{\alpha^2}{3}\left(\frac{n-1}{n+1}\right)^2\right] = \theta_2^2 V_4, \tag{11.7}$$

where

$$V_4 = \frac{1}{3n}\left[1 - \frac{\alpha^2}{3}\left(\frac{n-1}{n+1}\right)^2\right].$$

If n is odd then the estimators of θ_2 using ERSS$_2$ and ERSS$_3$ are obtained as

$$t_5 = \hat{\theta}_{2,\mathrm{ERSS}_2} = \frac{2\left(Y_{[1]1} + Y_{[n]2} + Y_{[1]3} + \ldots + Y_{[n](n-1)} + \dfrac{\left(Y_{[1]n} + Y_{[n]n}\right)}{2}\right)}{n},$$

and

$$t_6 = \hat{\theta}_{2,\mathrm{ERSS}_3} = \frac{2\left(Y_{[1]1} + Y_{[n]2} + Y_{[1]3} + \ldots + Y_{[n](n-1)} + Y_{[\frac{n+1}{2}]n}\right)}{n}.$$

The variances of the estimators t_5 and t_6 are, respectively, given by

$$\mathrm{Var}(t_5) = \frac{\theta_2^2}{3n}\left[1 - \frac{\alpha^2(n-1)^3}{3n(n+1)^2} - \frac{1}{2n} + \frac{\alpha^2(2-n)}{6n(n+2)}\right] = \theta_2^2 V_5, \tag{11.8}$$

$$\mathrm{Var}(t_6) = \frac{\theta_2^2}{3n}\left[1 - \frac{\alpha^2(n-1)^3}{3n(n+1)^2}\right] = \theta_2^2 V_6, \tag{11.9}$$

where

$$V_5 = \frac{1}{3n}\left[1 - \frac{\alpha^2(n-1)^3}{3n(n+1)^2} - \frac{1}{2n} + \frac{\alpha^2(2-n)}{6n(n+2)}\right],$$

and

$$V_6 = \frac{1}{3n}\left[1 - \frac{\alpha^2(n-1)^3}{3n(n+1)^2}\right].$$

Al-Saleh and Ananbeh (2007) proposed the concept of moving extreme ranked set sampling (MERSS) with a concomitant variable for the estimation of the means of the bivariate normal distribution. Now, suppose that the random vector (X, Y) has an MTBUD as defined in Eq. (11.2). An unbiased estimator of θ_2 based on MERSS is given by [see Tahmasebi and Jafari (2012)]

$$t_7 = \hat{\theta}_{2,\mathrm{MERSS}} = \frac{1}{n}\sum_{r=1}^{n}\left(Y_{[1]r} + Y_{[n]r}\right),$$

and its variance is

$$\mathrm{Var}(t_7) = \frac{\theta_2^2}{6n}\left[1 - \frac{\alpha^2}{3n}\left(\frac{n-1}{n+1}\right)^2\right] = \theta_2^2 V_7, \tag{11.10}$$

where

$$V_7 = \frac{1}{6n}\left[1 - \frac{\alpha^2}{3n}\left(\frac{n-1}{n+1}\right)^2\right].$$

11.3 THE SUGGESTED FAMILY OF ESTIMATORS FOR THE SCALE PARAMETER θ_2 BASED ON THE A PRIORI INTERVAL

The arithmetic mean (AM), the geometric mean (GM), and the harmonic mean (HM) are measures of location, which are used for suggesting different classes of shrinkage estimators for scale parameter θ_2. Let the prior information of θ_2 be available in the form of an interval whose end points are θ_{21} and θ_{22}, such that $\theta_{21} < \theta_{22}$. We define the following families of shrinkage estimators $\psi_{\theta_2}^{(i)}$ $(i = 1, 2, 3)$ of θ_2 as

$$\psi_{\theta_2}^{(i)} = \delta t_j + (1 - \delta)AGH(l,k) = \delta\left[t_j - AGH(l,k)\right] + AGH(l,k), \tag{11.11}$$

where $t_j, j = 1, 2, \ldots, 7$ is an unbiased estimator of the parameter θ_2,δ is a scalar such that $0 \le \delta \le 1$, and $AGH(l,k) = (\theta_{21}\theta_{22})^l\left(\frac{\theta_{21}+\theta_{22}}{2}\right)^k$ for $i = 1, 2, 3$ corresponding to (l,k) which should be taken as $(0, 1)$, $\left(\frac{1}{2}, 0\right)$ and $(1, -1)$ in $AGH(l,k)$. It is interesting to note that for different values of i we have formed the following classes of estimators:

i. For $i = 1$ and $(l,k) = (0, 1)$, we get the class of estimators as

$$\psi_{\theta_2}^{(1)} = \delta\left[t_j - AGH(0,1)\right] + AGH(0,1) = \delta\left[t_j - \left(\frac{\theta_{21} + \theta_{22}}{2}\right)\right] + \left(\frac{\theta_{21} + \theta_{22}}{2}\right), \tag{11.12}$$

ii. For $i = 2$ and $(l,k) = \left(\frac{1}{2}, 0\right)$, we obtain the class of estimators as

$$\psi_{\theta_2}^{(2)} = \delta\left[t_j - AGH\left(\frac{1}{2},0\right)\right] + AGH\left(\frac{1}{2},0\right) = \delta\left[t_j - \sqrt{\theta_{21}\theta_{22}}\right] + \sqrt{\theta_{21}\theta_{22}}, \tag{11.13}$$

iii. For $i = 3$ and $(l,k) = (1, -1)$, we get the class of estimators as

$$\psi_{\theta_2}^{(3)} = \delta\left[t_j - AGH(1,-1)\right] + AGH(1,-1) = \delta\left[t_j - \left(\frac{2\theta_{21}\theta_{22}}{\theta_{21} + \theta_{22}}\right)\right] + \left(\frac{2\theta_{21}\theta_{22}}{\theta_{21} + \theta_{22}}\right). \tag{11.14}$$

The bias and MSE of $\psi_{\theta_2}^{(i)}$ $(i = 1, 2, 3)$ are, respectively, given by

$$B\left[\psi_{\theta_2}^{(i)}\right] = \theta_2(1 - \delta)\left(\lambda_{(i)} - 1\right) \tag{11.15}$$

$$\mathrm{MSE}\left[\psi_{\theta_2}^{(i)}\right] = \theta_2^2\left[V_j\delta^2 + (1-\delta)^2\left(\lambda_{(i)} - 1\right)^2\right], \tag{11.16}$$

where $\lambda_{(i)} = \frac{AGH(l,k)}{\theta_2}$.

The minimum mean squared error (MMSE) estimators of the parameter θ_2 based on $t_j, j = 1, 2, \ldots, 7$ are given as

$$T_j^* = \frac{\theta_2}{(1 + V_j)}, j = 1, 2, \ldots, 7, \tag{11.17}$$

in the class of estimator $T_j = t_j A_j, j = 1, 2, \ldots, 7$, where $A_j^{'} s, j = 1, 2, \ldots, 7$ are suitably chosen constants such that the MSE of $T_j^{'} s, j = 1, 2, \ldots, 7$ are minimum.

The bias and MSE of $T_j^* s, j = 1, 2, \ldots, 7$ are, respectively, given by

$$B\left(T_J^*\right) = -\theta_2\left(\frac{V_j}{1 + V_j}\right), \tag{11.18}$$

$$\mathrm{MSE}\left(T_j^*\right) = \theta_2^2\left(\frac{V_j}{1 + V_j}\right). \tag{11.19}$$

Comparisons of the proposed shrinkage estimators $\psi_{\theta_2}^{(i)}$ $(i = 1, 2, 3)$ with that of corresponding usual unbiased estimators $t_j^{'} s, j = 1, 2, \ldots, 7$ are given in Theorem 1.1.

Theorem 1.1: *The proposed shrinkage estimators* $\psi_{\theta_2}^{(i)}$ $(i = 1, 2, 3)$ *are better than the corresponding usual unbiased estimators* $t_j^{'} s, j = 1, 2, \ldots, 7$ *if*

$$\frac{\left\{\left(\lambda_{(i)} - 1\right)^2 - V_j\right\}}{\left\{\left(\lambda_{(i)} - 1\right)^2 + V_j\right\}} < \delta < 1.$$

Proof: *From Eqs. (11.4)–(11.10) and (11.16), we have that*

$$Var\left(t_j\right) - MSE\left[\psi_{\theta_2}^{(i)}\right] > 0, \quad i = 1, 2, 3, \quad j = 1, 2, \ldots, 7 \; if$$

$$\theta_2^2 V_j - \theta_2^2 V_j \delta^2 - (1 - \delta)^2 \left(\lambda_{(i)} - 1\right)^2 \theta_2^2 > 0,$$

i.e., if $V_j\left(1 - \delta^2\right) > (1 - \delta)^2 \left(\lambda_{(i)} - 1\right)^2$, *i.e., if* $V_j(1 + \delta) > (1 - \delta)\left(\lambda_{(i)} - 1\right)^2$,

Now

$$(1 - \delta) > 0 \Rightarrow 1 > \delta \Rightarrow \delta < 1 \tag{11.20}$$

and $V_j + \delta\left\{V_j + \left(\lambda_{(i)} - 1\right)^2\right\} > \left(\lambda_{(i)} - 1\right)^2$, *or* $\delta\left\{V_j + \left(\lambda_{(i)} - 1\right)^2\right\} > \left\{\left(\lambda_{(i)} - 1\right)^2 - V_j\right\}$, *i.e., if*

$$\delta > \frac{\left\{\left(\lambda_{(i)} - 1\right)^2 - V_j\right\}}{\left\{\left(\lambda_{(i)} - 1\right)^2 + V_j\right\}}. \tag{11.21}$$

From Eqs. (11.20) and (11.21) we have

$$\frac{\left\{\left(\lambda_{(i)} - 1\right)^2 - V_j\right\}}{\left\{\left(\lambda_{(i)} - 1\right)^2 + V_j\right\}} < \delta < 1. \tag{11.22}$$

Hence the theorem.♦

Comparisons of the proposed shrinkage estimators $\psi_{\theta_2}^{(i)}$ $(i = 1, 2, 3)$ with that of corresponding MMSE estimators $T_j^* s, j = 1, 2, \ldots, 7$ are given in Theorem 1.2.

Theorem 1.2: *The proposed shrinkage estimators* $\psi_{\theta_2}^{(i)}$ $(i = 1, 2, 3)$ *are better than the corresponding MMSE estimators* $T_j^* s, j = 1, 2, \ldots, 7$ *if*

$$\left\{\frac{\left(\lambda_{(i)}-1\right)^2}{\left(\lambda_{(i)}-1\right)^2+V_j}-\frac{V_j\sqrt{\left\{1-\left(\lambda_{(i)}-1\right)^2\right\}}}{\sqrt{(1+V_j)\left\{\left(\lambda_{(i)}-1\right)^2+V_j\right\}}}\right\}<\delta<\left\{\frac{\left(\lambda_{(i)}-1\right)^2}{\left(\lambda_{(i)}-1\right)^2+V_j}+\frac{V_j\sqrt{\left\{1-\left(\lambda_{(i)}-1\right)^2\right\}}}{\sqrt{(1+V_j)\left\{\left(\lambda_{(i)}-1\right)^2+V_j\right\}}}\right\} \tag{11.23}$$

Proof: *From Eqs.* (11.16) *and* (11.19), *we have that*

$$MSE\left(T_j^*\right)-MSE\left[\psi_{\theta_2}^{(i)}\right]>0,\ \ i=1,2,3,\ \ j=1,2,\ldots,7\ \textit{if}$$

$$\theta_2^2\frac{V_j}{1+V_j}-\theta_2^2V_j\delta^2-(1-\delta)^2\left(\lambda_{(i)}-1\right)^2\theta_2^2>0,$$

i.e., if $-\frac{V_j}{1+V_j}+V_j\delta^2+\left(1+\delta^2-2\delta\right)\left(1-\delta^2\right)\left(\lambda_{(i)}-1\right)^2<0,$

i.e., if $\delta^2\left[-V_j+\left(\lambda_{(i)}-1\right)^2\right]-2\delta\left(\lambda_{(i)}-1\right)^2-\frac{V_j}{1+V_j}+\left(\lambda_{(i)}-1\right)^2>0,$

On solving the above quadratic equation with respect to δ *we have*

$$\left\{\frac{\left(\lambda_{(i)}-1\right)^2}{\left(\lambda_{(i)}-1\right)^2+V_j}-\frac{V_j\sqrt{\left\{1-\left(\lambda_{(i)}-1\right)^2\right\}}}{\sqrt{(1+V_j)\left\{\left(\lambda_{(i)}-1\right)^2+V_j\right\}}}\right\}<\delta<\left\{\frac{\left(\lambda_{(i)}-1\right)^2}{\left(\lambda_{(i)}-1\right)^2+V_j}+\frac{V_j\sqrt{\left\{1-\left(\lambda_{(i)}-1\right)^2\right\}}}{\sqrt{(1+V_j)\left\{\left(\lambda_{(i)}-1\right)^2+V_j\right\}}}\right\}.$$

Hence the theorem.♦

11.4 RELATIVE EFFICIENCY

We note here that among these seven estimators t_j, $j = 1, 2, \ldots, 7$ discussed above, the estimator t_2 is the best as we have observed numerically. Keeping this in view we have made an effort to compare the estimators $\psi_{\theta_2}^{(i)}$ $(i = 1, 2, 3)$ formulated based on the BLUE with that of the BLUE t_2 and its MMSE estimator T_2^* by using following the formula:

$$e_1^{(i)}=RE\left(\psi_{\theta_2}^{(i)},t_2\right)=\frac{V_2}{\left\{V_2\delta^2+(1-\delta)^2\left(\lambda_{(i)}-1\right)^2\right\}},\ \ i=1,2,3, \tag{11.24}$$

$$e_2^{(i)}=RE\left(\psi_{\theta_2}^{(i)},T_2^*\right)=\frac{V_2}{(1+V_2)\left\{V_2\delta^2+(1-\delta)^2\left(\lambda_{(i)}-1\right)^2\right\}},\ \ i=1,2,3. \tag{11.25}$$

The values of $e_1^{(i)}$ and $e_2^{(i)}$, $i = 1, 2, 3$ are shown in Table 11.1 for $n = 5(5)20$, $\alpha = 0.25(0.25)1.00$ and different values of $\psi_1 = \frac{\theta_{21}}{\theta_2} = 0.5(0.1)0.9$, $\psi_2 = \frac{\theta_{22}}{\theta_2} = 1.1(0.1)1.5$ and $\delta = 0.25(0.25)0.75$.

Table 11.1 The Values of $e_1^{(i)}$ and $e_2^{(i)'}s$, $i = 1, 2, 3$ for Different Values of $n, (\psi_1, \psi_2), \delta$ and Fixed $\alpha = 0.25$

$(\psi_1,\psi_2)\rightarrow$ $n\downarrow$	δ	(0.5,1.1)			(0.6,1.2)			(0.7,1.3)			(0.8,1.4)			(0.9,1.5)		
		$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$
5	0.25	2.4869	1.5882	1.1211	6.7842	3.8922	2.4869	16.0000	12.4488	7.6179	6.7842	10.9450	15.3735	2.4869	3.5095	5.1297
	0.50	2.4942	1.9918	1.6164	3.4754	2.9726	2.4942	4.0000	3.8771	3.5642	3.4754	3.8047	3.9820	2.4942	2.8665	3.2377
	0.75	1.6660	1.5987	1.5275	1.7485	1.7120	1.6660	1.7778	1.7715	1.7540	1.7485	1.7677	1.7769	1.6660	1.7030	1.7325
10	0.25	1.3465	0.8344	0.5801	4.3002	2.2128	1.3465	16.0000	10.1822	4.9940	4.3002	8.3114	14.7926	1.3465	1.9684	3.0509
	0.50	1.8106	1.3248	1.0117	3.0715	2.3637	1.8106	4.0000	3.7612	3.2132	3.0715	3.6272	3.9640	1.8106	2.2321	2.7181
	0.75	1.5672	1.4520	1.3385	1.7200	1.6508	1.5672	1.7778	1.7653	1.7307	1.7200	1.7577	1.7760	1.5672	1.6340	1.6893
15	0.25	0.9231	0.5658	0.3912	3.1475	1.5458	0.9231	16.0000	8.6137	3.7144	3.1475	6.6992	14.2539	0.9231	1.3677	2.1709
	0.50	1.4210	0.9924	0.7363	2.7516	1.9618	1.4210	4.0000	3.6520	2.9250	2.7516	3.4654	3.9463	1.4210	1.8276	2.3422
	0.75	1.4794	1.3299	1.1911	1.6925	1.5938	1.4794	1.7778	1.7592	1.7080	1.6925	1.7478	1.7751	1.4794	1.5704	1.6482
20	0.25	0.7023	0.4281	0.2952	2.4822	1.1878	0.7023	16.0000	7.4640	2.9569	2.4822	5.6110	13.7531	0.7023	1.0479	1.6850
	0.50	1.1695	0.7933	0.5787	2.4921	1.6767	1.1695	4.0000	3.5490	2.6843	2.4921	3.3175	3.9287	1.1695	1.5472	2.0577
	0.75	1.4010	1.2268	1.0730	1.6658	1.5406	1.4010	1.7778	1.7530	1.6860	1.6658	1.7380	1.7742	1.4010	1.5115	1.6090
$(\psi_1,\psi_2)\rightarrow$ $n\downarrow$	δ	(0.5,1.1)			(0.6,1.2)			(0.7,1.3)			(0.8,1.4)			(0.9,1.5)		
		$e_2^{(1)}$	$e_2^{(2)}$	$e_2^{(3)}$	$e_2^{(1)}$	$e_2^{(2)}$	$e_2^{(3)}$	$e_2^{(1)}$	$e_2^{(2)}$	$e_2^{(3)}$	$e_2^{(1)}$	$e_2^{(2)}$	$e_2^{(3)}$	$e_2^{(1)}$	$e_2^{(2)}$	$e_2^{(3)}$
5	0.25	2.3324	1.4895	1.0514	6.3627	3.6504	2.3324	15.0058	11.6753	7.1446	6.3627	10.2649	14.4182	2.3324	3.2915	4.8109
	0.50	2.3392	1.8680	1.5160	3.2595	2.7879	2.3392	3.7515	3.6362	3.3428	3.2595	3.5683	3.7345	2.3392	2.6884	3.0365
	0.75	1.5625	1.4993	1.4326	1.6398	1.6056	1.5625	1.6673	1.6615	1.6450	1.6398	1.6579	1.6665	1.5625	1.5971	1.6248
10	0.25	1.3033	0.8077	0.5615	4.1625	2.1420	1.3033	15.4877	9.8562	4.8341	4.1625	8.0453	14.3189	1.3033	1.9054	2.9532
	0.50	1.7526	1.2823	0.9794	2.9731	2.2880	1.7526	3.8719	3.6408	3.1103	2.9731	3.5110	3.8371	1.7526	2.1606	2.6311
	0.75	1.5170	1.4055	1.2957	1.6649	1.5979	1.5170	1.7209	1.7088	1.6753	1.6649	1.7014	1.7191	1.5170	1.5817	1.6352
15	0.25	0.9032	0.5536	0.3828	3.0796	1.5124	0.9032	15.6550	8.4279	3.6343	3.0796	6.5547	13.9465	0.9032	1.3382	2.1241
	0.50	1.3904	0.9710	0.7204	2.6922	1.9194	1.3904	3.9137	3.5733	2.8620	2.6922	3.3907	3.8612	1.3904	1.7881	2.2917
	0.75	1.4475	1.3012	1.1654	1.6560	1.5594	1.4475	1.7394	1.7212	1.6712	1.6560	1.7101	1.7368	1.4475	1.5365	1.6126
20	0.25	0.6909	0.4211	0.2904	2.4419	1.1684	0.6909	15.7399	7.3427	2.9088	2.4419	5.5197	13.5295	0.6909	1.0309	1.6576
	0.50	1.1505	0.7804	0.5693	2.4516	1.6494	1.1505	3.9350	3.4913	2.6407	2.4516	3.2636	3.8648	1.1505	1.5220	2.0242
	0.75	1.3782	1.2069	1.0555	1.6387	1.5155	1.3782	1.7489	1.7245	1.6586	1.6387	1.7098	1.7454	1.3782	1.4870	1.5829
(For Fixed $\alpha = 0.50$)																
$(\psi_1,\psi_2)\rightarrow$ $n\downarrow$	δ	(0.5, 1.1)			(0.6, 1.2)			(0.7, 1.3)			(0.8, 1.4)			(0.9, 1.5)		
		$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$
5	0.25	2.4473	1.5612	1.1014	6.7101	3.8365	2.4473	16.0000	12.3960	7.5421	6.7101	10.8790	15.3620	2.4473	3.4577	5.0636
	0.50	2.4763	1.9728	1.5981	3.4667	2.9580	2.4763	4.0000	3.8748	3.5568	3.4667	3.8012	3.9816	2.4763	2.8510	3.2259
	0.75	1.6640	1.5956	1.5234	1.7479	1.7108	1.6640	1.7778	1.7714	1.7535	1.7479	1.7675	1.7769	1.6640	1.7016	1.7316

Table 11.1 The Values of $e_1^{(i)}$ and $e_2^{(i)'}s$, $i = 1, 2, 3$ for Different Values of n, (ψ_1, ψ_2), δ and Fixed $\alpha = 0.25$ *Continued*

(For Fixed $\alpha = 0.50$)

$(\psi_1,\psi_2)\rightarrow$ $n\downarrow$	δ	(0.5, 1.1)			(0.6, 1.2)			(0.7, 1.3)			(0.8, 1.4)			(0.9, 1.5)		
		$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$
10	0.25	1.3178	0.8160	0.5671	4.2267	2.1684	1.3178	16.0000	10.0950	4.9137	4.2267	8.2176	14.7661	1.3178	1.9282	2.9933
	0.50	1.7873	1.3040	0.9941	3.0546	2.3409	1.7873	4.0000	3.7559	3.1982	3.0546	3.6192	3.9632	1.7873	2.2089	2.6976
	0.75	1.5628	1.4457	1.3307	1.7187	1.6480	1.5628	1.7778	1.7650	1.7296	1.7187	1.7572	1.7759	1.5628	1.6308	1.6873
15	0.25	0.9014	0.5523	0.3818	3.0844	1.5110	0.9014	16.0000	8.5136	3.6432	3.0844	6.6014	14.2144	0.9014	1.3366	2.1242
	0.50	1.3981	0.9737	0.7213	2.7299	1.9366	1.3981	4.0000	3.6440	2.9051	2.7299	3.4537	3.9449	1.3981	1.8026	2.3177
	0.75	1.4732	1.3214	1.1812	1.6904	1.5896	1.4732	1.7778	1.7587	1.7063	1.6904	1.7471	1.7750	1.4732	1.5657	1.6451
20	0.25	0.6850	0.4173	0.2877	2.4281	1.1594	0.6850	16.0000	7.3603	2.8946	2.4281	5.5163	13.7023	0.6850	1.0227	1.6461
	0.50	1.1480	0.7769	0.5659	2.4675	1.6514	1.1480	4.0000	3.5385	2.6612	2.4675	3.3026	3.9268	1.1480	1.5225	2.0316
	0.75	1.3932	1.2168	1.0619	1.6630	1.5352	1.3932	1.7778	1.7524	1.6837	1.6630	1.7370	1.7741	1.3932	1.5056	1.6050

$(\psi_1,\psi_2)\rightarrow$ $n\downarrow$	δ	(0.5, 1.1)			(0.6, 1.2)			(0.7, 1.3)			(0.8, 1.4)			(0.9, 1.5)		
		$e_2^{(1)}$	$e_2^{(2)}$	$e_2^{(3)}$	$e_2^{(1)}$	$e_2^{(2)}$	$e_2^{(3)}$	$e_2^{(1)}$	$e_2^{(2)}$	$e_2^{(3)}$	$e_2^{(1)}$	$e_2^{(2)}$	$e_2^{(3)}$	$e_2^{(1)}$	$e_2^{(2)}$	$e_2^{(3)}$
5	0.25	2.2979	1.4659	1.0342	6.3005	3.6023	2.2979	15.0234	11.6394	7.0817	6.3005	10.2150	14.4243	2.2979	3.2467	4.7546
	0.50	2.3251	1.8524	1.5006	3.2551	2.7774	2.3251	3.7558	3.6383	3.3397	3.2551	3.5692	3.7386	2.3251	2.6769	3.0290
	0.75	1.5624	1.4982	1.4304	1.6412	1.6064	1.5624	1.6693	1.6633	1.6465	1.6412	1.6596	1.6684	1.5624	1.5977	1.6259
10	0.25	1.2765	0.7905	0.5494	4.0944	2.1006	1.2765	15.4992	9.7791	4.7599	4.0944	7.9604	14.3039	1.2765	1.8679	2.8996
	0.50	1.7314	1.2632	0.9630	2.9590	2.2676	1.7314	3.8748	3.6383	3.0981	2.9590	3.5059	3.8392	1.7314	2.1397	2.6132
	0.75	1.5139	1.4004	1.2891	1.6649	1.5964	1.5139	1.7221	1.7098	1.6755	1.6649	1.7022	1.7204	1.5139	1.5798	1.6345
15	0.25	0.8825	0.5406	0.3737	3.0195	1.4792	0.8825	15.6633	8.3344	3.5665	3.0195	6.4625	13.9153	0.8825	1.3084	2.0795
	0.50	1.3687	0.9532	0.7061	2.6724	1.8959	1.3687	3.9158	3.5673	2.8440	2.6724	3.3810	3.8619	1.3687	1.7647	2.2690
	0.75	1.4422	1.2936	1.1563	1.6548	1.5561	1.4422	1.7404	1.7217	1.6704	1.6548	1.7103	1.7377	1.4422	1.5328	1.6105
20	0.25	0.6741	0.4107	0.2832	2.3896	1.1410	0.6741	15.7465	7.2437	2.8487	2.3896	5.4289	13.4851	0.6741	1.0065	1.6200
	0.50	1.1298	0.7646	0.5570	2.4284	1.6252	1.1298	3.9366	3.4824	2.6191	2.4284	3.2503	3.8646	1.1298	1.4984	1.9994
	0.75	1.3711	1.1976	1.0450	1.6367	1.5109	1.3711	1.7496	1.7246	1.6570	1.6367	1.7095	1.7460	1.3711	1.4817	1.5796

(For Fixed $\alpha = 0.75$)

$(\psi_1,\psi_2)\rightarrow$ $n\downarrow$	δ	(0.5, 1.1)			(0.6, 1.2)			(0.7, 1.3)			(0.8, 1.4)			(0.9, 1.5)		
		$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$
5	0.25	2.3798	1.5155	1.0682	6.5822	3.7413	2.3798	16.0000	12.3033	7.4110	6.5822	10.7637	15.3415	2.3798	3.3693	4.9504
	0.50	2.4451	1.9399	1.5667	3.4513	2.9324	2.4451	4.0000	3.8708	3.5437	3.4513	3.7949	3.9810	2.4451	2.8238	3.2051
	0.75	1.6605	1.5901	1.5161	1.7469	1.7087	1.6605	1.7778	1.7712	1.7527	1.7469	1.7672	1.7768	1.6605	1.6991	1.7301
10	0.25	1.2687	0.7846	0.5449	4.0995	2.0921	1.2687	16.0000	9.9403	4.7742	4.0995	8.0524	14.7182	1.2687	1.8592	2.8941
	0.50	1.7466	1.2680	0.9636	3.0245	2.3007	1.7466	4.0000	3.7462	3.1714	3.0245	3.6047	3.9617	1.7466	2.1679	2.6610
	0.75	1.5549	1.4344	1.3167	1.7163	1.6429	1.5549	1.7778	1.7645	1.7276	1.7163	1.7564	1.7759	1.5549	1.6252	1.6836

(*Continued*)

Table 11.1 The Values of $e_1^{(i)}$ and $e_2^{(i)'}s$, $i = 1, 2, 3$ for Different Values of n, (ψ_1, ψ_2), δ and Fixed $\alpha = 0.25$ *Continued*

(For Fixed $\alpha = 0.75$)

$(\psi_1,\psi_2) \rightarrow$ $n \downarrow$	δ	(0.5, 1.1)			(0.6, 1.2)			(0.7, 1.3)			(0.8, 1.4)			(0.9, 1.5)		
		$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$
15	0.25	0.8643	0.5290	0.3655	2.9750	1.4511	0.8643	16.0000	8.3359	3.5193	2.9750	6.4294	14.1425	0.8643	1.2830	2.0434
	0.50	1.3578	0.9413	0.6953	2.6910	1.8921	1.3578	4.0000	3.6292	2.8694	2.6910	3.4323	3.9425	1.3578	1.7586	2.2742
	0.75	1.4617	1.3062	1.1634	1.6866	1.5820	1.4617	1.7778	1.7578	1.7032	1.6866	1.7457	1.7749	1.4617	1.5572	1.6395
20	0.25	0.6553	0.3989	0.2749	2.3342	1.1106	0.6553	16.0000	7.1766	2.7864	2.3342	5.3502	13.6097	0.6553	0.9793	1.5790
	0.50	1.1105	0.7483	0.5438	2.4235	1.6066	1.1105	4.0000	3.5192	2.6197	2.4235	3.2755	3.9234	1.1105	1.4791	1.9853
	0.75	1.3791	1.1989	1.0420	1.6579	1.5253	1.3791	1.7778	1.7512	1.6795	1.6579	1.7351	1.7739	1.3791	1.4947	1.5976

$(\psi_1,\psi_2) \rightarrow$ $n \downarrow$	δ	(0.5, 1.1)			(0.6, 1.2)			(0.7, 1.3)			(0.8, 1.4)			(0.9, 1.5)		
		$e_2^{(1)}$	$e_2^{(2)}$	$e_2^{(3)}$	$e_2^{(1)}$	$e_2^{(2)}$	$e_2^{(3)}$	$e_2^{(1)}$	$e_2^{(2)}$	$e_2^{(3)}$	$e_2^{(1)}$	$e_2^{(2)}$	$e_2^{(3)}$	$e_2^{(1)}$	$e_2^{(2)}$	$e_2^{(3)}$
5	0.25	2.2390	1.4258	1.0050	6.1927	3.5199	2.2390	15.0531	11.5752	6.9724	6.1927	10.1267	14.4336	2.2390	3.1699	4.6574
	0.50	2.3004	1.8251	1.4739	3.2471	2.7589	2.3004	3.7633	3.6417	3.3340	3.2471	3.5703	3.7454	2.3004	2.6567	3.0154
	0.75	1.5622	1.4960	1.4264	1.6435	1.6075	1.5622	1.6726	1.6664	1.6490	1.6435	1.6626	1.6717	1.5622	1.5986	1.6277
10	0.25	1.2305	0.7610	0.5285	3.9762	2.0292	1.2305	15.5189	9.6414	4.6306	3.9762	7.8102	14.2756	1.2305	1.8033	2.8070
	0.50	1.6941	1.2299	0.9346	2.9335	2.2315	1.6941	3.8797	3.6336	3.0761	2.9335	3.4963	3.8425	1.6941	2.1027	2.5810
	0.75	1.5081	1.3913	1.2771	1.6647	1.5935	1.5081	1.7243	1.7114	1.6757	1.6647	1.7036	1.7225	1.5081	1.5763	1.6330
15	0.25	0.8469	0.5183	0.3581	2.9151	1.4219	0.8469	15.6777	8.1680	3.4485	2.9151	6.2999	13.8576	0.8469	1.2572	2.0023
	0.50	1.3305	0.9223	0.6813	2.6368	1.8540	1.3305	3.9194	3.5561	2.8116	2.6368	3.3632	3.8631	1.3305	1.7232	2.2284
	0.75	1.4323	1.2799	1.1400	1.6526	1.5501	1.4323	1.7420	1.7224	1.6689	1.6526	1.7105	1.7391	1.4323	1.5259	1.6065
20	0.25	0.6453	0.3929	0.2708	2.2989	1.0938	0.6453	15.7578	7.0680	2.7442	2.2989	5.2692	13.4036	0.6453	0.9644	1.5551
	0.50	1.0937	0.7370	0.5356	2.3868	1.5823	1.0937	3.9394	3.4660	2.5800	2.3868	3.2260	3.8640	1.0937	1.4567	1.9553
	0.75	1.3582	1.1808	1.0262	1.6328	1.5022	1.3582	1.7509	1.7247	1.6540	1.6328	1.7089	1.7471	1.3582	1.4721	1.5734

(For Fixed $\alpha = 1.00$)

$(\psi_1,\psi_2) \rightarrow$ $n \downarrow$	δ	(0.5, 1.1)			(0.6, 1.2)			(0.7, 1.3)			(0.8, 1.4)			(0.9, 1.5)		
		$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$
5	0.25	2.2820	1.4494	1.0202	6.3926	3.6023	2.2820	16.0000	12.1617	7.2159	6.3926	10.5892	15.3097	2.2820	3.2405	4.7840
	0.50	2.3982	1.8908	1.5201	3.4276	2.8935	2.3982	4.0000	3.8645	3.5234	3.4276	3.7851	3.9801	2.3982	2.7826	3.1733
	0.75	1.6550	1.5817	1.5050	1.7454	1.7053	1.6550	1.7778	1.7709	1.7515	1.7454	1.7666	1.7768	1.6550	1.6954	1.7278
10	0.25	1.1966	0.7387	0.5125	3.9092	1.9795	1.1966	16.0000	9.6998	4.5645	3.9092	7.7988	14.6414	1.1966	1.7576	2.7468
	0.50	1.6845	1.2137	0.9179	2.9769	2.2384	1.6845	4.0000	3.7308	3.1290	2.9769	3.5815	3.9592	1.6845	2.1048	2.6040
	0.75	1.5422	1.4165	1.2948	1.7124	1.6348	1.5422	1.7778	1.7636	1.7244	1.7124	1.7550	1.7757	1.5422	1.6161	1.6778

Table 11.1 The Values of $e_1^{(i)}$ and $e_2^{(i)'}s$, $i = 1, 2, 3$ for Different Values of n, (ψ_1, ψ_2), δ and Fixed $\alpha = 0.25$ *Continued*

(For Fixed $\alpha = 1.00$)

$(\psi_1,\psi_2)\rightarrow$		(0.5, 1.1)			(0.6, 1.2)			(0.7, 1.3)			(0.8, 1.4)			(0.9, 1.5)		
$n\downarrow$	δ	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$	$e_1^{(1)}$	$e_1^{(2)}$	$e_1^{(3)}$
15	0.25	0.8096	0.4948	0.3417	2.8115	1.3626	0.8096	16.0000	8.0600	3.3336	2.8115	6.1659	14.0261	0.8096	1.2039	1.9236
	0.50	1.2967	0.8925	0.6566	2.6295	1.8235	1.2967	4.0000	3.6054	2.8126	2.6295	3.3979	3.9384	1.2967	1.6909	2.2062
	0.75	1.4434	1.2819	1.1354	1.6805	1.5696	1.4434	1.7778	1.7564	1.6981	1.6805	1.7434	1.7747	1.4434	1.5436	1.6305
20	0.25	0.6115	0.3719	0.2562	2.1945	1.0385	0.6115	16.0000	6.8933	2.6248	2.1945	5.0973	13.4598	0.6115	0.9152	1.4796
	0.50	1.0538	0.7056	0.5109	2.3543	1.5381	1.0538	4.0000	3.4880	2.5540	2.3543	3.2319	3.9178	1.0538	1.4128	1.9135
	0.75	1.3564	1.1705	1.0108	1.6497	1.5093	1.3564	1.7778	1.7492	1.6726	1.6497	1.7320	1.7736	1.3564	1.4772	1.5857

$(\psi_1,\psi_2)\rightarrow$		(0.5, 1.1)			(0.6, 1.2)			(0.7, 1.3)			(0.8, 1.4)			(0.9, 1.5)		
$n\downarrow$	δ	$e_2^{(1)}$	$e_2^{(2)}$	$e_2^{(3)}$	$e_2^{(1)}$	$e_2^{(2)}$	$e_2^{(3)}$	$e_2^{(1)}$	$e_2^{(2)}$	$e_2^{(3)}$	$e_2^{(1)}$	$e_2^{(2)}$	$e_2^{(3)}$	$e_2^{(1)}$	$e_2^{(2)}$	$e_2^{(3)}$
5	0.25	2.1530	1.3675	0.9626	6.0315	3.3988	2.1530	15.0960	11.4746	6.8082	6.0315	9.9909	14.4447	2.1530	3.0574	4.5137
	0.50	2.2627	1.7840	1.4342	3.2340	2.7300	2.2627	3.7740	3.6461	3.3243	3.2340	3.5712	3.7552	2.2627	2.6254	2.9940
	0.75	1.5614	1.4924	1.4199	1.6468	1.6090	1.5614	1.6773	1.6708	1.6525	1.6468	1.6668	1.6764	1.5614	1.5996	1.6301
10	0.25	1.1627	0.7178	0.4980	3.7986	1.9235	1.1627	15.5476	9.4255	4.4355	3.7986	7.5783	14.2274	1.1627	1.7079	2.6691
	0.50	1.6368	1.1794	0.8920	2.8928	2.1751	1.6368	3.8869	3.6253	3.0405	2.8928	3.4802	3.8472	1.6368	2.0453	2.5304
	0.75	1.4986	1.3764	1.2581	1.6640	1.5886	1.4986	1.7275	1.7138	1.6757	1.6640	1.7054	1.7255	1.4986	1.5704	1.6304
15	0.25	0.7943	0.4855	0.3352	2.7586	1.3369	0.7943	15.6988	7.9083	3.2708	2.7586	6.0498	13.7621	0.7943	1.1812	1.8874
	0.50	1.2722	0.8757	0.6442	2.5800	1.7892	1.2722	3.9247	3.5375	2.7596	2.5800	3.3339	3.8643	1.2722	1.6591	2.1646
	0.75	1.4162	1.2577	1.1140	1.6488	1.5401	1.4162	1.7443	1.7234	1.6662	1.6488	1.7106	1.7413	1.4162	1.5145	1.5998
20	0.25	0.6029	0.3666	0.2525	2.1636	1.0239	0.6029	15.7743	6.7961	2.5878	2.1636	5.0254	13.2699	0.6029	0.9023	1.4587
	0.50	1.0389	0.6956	0.5037	2.3211	1.5164	1.0389	3.9436	3.4388	2.5180	2.3211	3.1863	3.8626	1.0389	1.3928	1.8865
	0.75	1.3373	1.1540	0.9965	1.6264	1.4880	1.3373	1.7527	1.7246	1.6490	1.6264	1.7076	1.7486	1.3373	1.4564	1.5633

11.5 CONCLUSION

It is observed from Table 11.1 that:

- when $(\psi_1, \psi_2) \in (0.7, 1.3)$ the proposed classes of estimators $\psi_{\theta_2}^{(i)}$ $(i = 1, 2, 3)$ is always better than the usual unbiased estimator t_2 and MMSE estimator T_2^*;
- the gain in efficiency by using $\psi_{\theta_2}^{(i)}$ $(i = 1, 2, 3)$ over MMSE estimator T_2^* is fewer than by using $\psi_{\theta_2}^{(i)}$ $(i = 1, 2, 3)$ over the BLUE t_2;
- for $(\psi_1, \psi_2) \in (0.7, 1.3)$, the developed class of estimators $\psi_{\theta_2}^{(1)}$ (based on AM) is the best (best in the sense of having smaller MSE) among $\psi_{\theta_2}^{(i)}(i = 1, 2, 3)$, while for $(\psi_1, \psi_2) \in (0.9, 1.5)$ the developed class of estimator $\psi_{\theta_2}^{(3)}$ (based on HM) is the best among $\psi_{\theta_2}^{(i)}$ $(i = 1, 2, 3)$.

In general the proposed estimator $\psi_{\theta_2}^{(1)}$ is recommended when $(\psi_1, \psi_2) \in (0.5, 1.3)$ and $\psi_{\theta_2}^{(3)}$ is recommended when $(\psi_1, \psi_2) \in (0.8, 1.5)$ and the sample size n is small. In practice, when the observations are expensive such small sizes may be all that are available, particularly in defense weapon testing problems.

ACKNOWLEDGMENTS

The author is highly thankful to Prof. H.P. Singh, School of Studies in Statistics, Vikram University, Ujjain, Madhya Pradesh, India, for the guidance and constructive suggestions and Prof. C.N. Bouza, Department of Applied Mathematics, University of Havana, Havana, Cuba, for choosing my chapter in this book. Last, but not the least, I am also thankful to all teaching and nonteaching staff members of the Jaypee University of Information Technology (JUIT), Waknaghat, Solan, Himachal Pradesh, India, for providing all the necessary facilities.

REFERENCES

Al-Saleh, M.F., Ananbeh, A., 2005. Estimating the correlation coefficient in a bivariate normal distribution using moving extreme ranked set sampling with a concomitant variable. J. Korean Stat. Soc. 34, 125–140.

Al-Saleh, M.F., Ananbeh, A., 2007. Estimation of the means of the bivariate normal distribution using moving extreme ranked set sampling with concomitant variable. Stat. Papers 48, 179–195.

Al-Saleh, M.F., Diab, Y.A., 2009. Estimation of the parameters of Downton's bivariate exponential distribution using ranked set sampling scheme. J. Stat. Plan. Inference 139, 277–286.

Barnett, V., Moore, K., 1997. Best linear unbiased estimates in ranked set sampling with particular reference to imperfect ordering. J. Appl. Stat. 24, 697–710.

Bouza, C.N., 2001. Model assisted ranked survey sampling. Biom. J. 43, 249–259.

Bouza, C.N., 2002. Ranked set sampling the non-response stratum for estimating the difference of means. Biom. J. 44, 903–915.

Bouza, C.N., 2005. Sampling using ranked sets: concepts, results and perspectives. Rev. Investig. Oper. 26 (3), 275–293.

Chen, Z., Bai, Z., Sinha, B.K., 2004. Lecture Notes in Statistics, Ranked Set Sampling, Theory and Applications. Springer, New York.

Demir, S., Singh, H., 2000. An application of the regression estimates to ranked set sampling. Hacit. Bull Nat. Sc. Eng. Ser. B 29, 93–101.

Jani, P.N., 1991. A class of shrinkage estimators for the scale parameter of the exponential distribution. IEEE Trans. Reliab. 40 (1), 68−70.

Kourouklis, S., 1994. Estimation in the two-parameter exponential distribution with prior information. IEEE Trans. Reliab. 43 (3), 446−450.

Lam, K., Sinha, B.K., Wu, Z., 1994. Estimation of a two-parameter exponential distribution using ranked set sample. Ann. Inst. Stat. Math. 46, 723−736.

McIntyre, G., 1952. A method for unbiased selective sampling using ranked set sampling. Aust. J. Agric. Res. 3, 385−390.

Mehta, V., 2017. Shrinkage estimator of the parameters of normal distribution based on K-record values. Int. J. Sci. Res. Math. Stat. Sci. 4 (1), 1−5.

Mehta, V., Singh, H.P., 2014. Shrinkage estimators of parameters of Morgenstern type bivariate logistic distribution using ranked set sampling. J. Basic Appl. Eng. Res. 1 (13), 1−6.

Mehta, V., Singh, H.P., 2015. Minimum mean square error estimation of parameters in bivariate normal distribution using concomitants of record values, edited book entitled "*Statistics and Informatics in Agricultural Research*". Indian Society of Agricultural Statistics, Excel India Publishers, New Delhi, India, pp. 162−174.

Morgenstern, D., 1956. Einfache Beispiele Zweidimensionaler Verteilunge. Mitt. Bl. Math. Stat. 8, 234−235.

Samawi, H.M., Muttlak, H.A., 1996. Estimation of a ratio using ranked set sampling. Biom. J. 36, 753−764.

Scaria, J., Nair, U., 1999. On concomitants of order statistics from Morgenstern family. Biom. J. 41, 483−489.

Searls, D.T., 1964. The utilization of a known coefficient of variation in the estimation procedure. J. Am. Stat. Assoc. 59, 1225−1226.

Searls, D.T., Intarapanich, P., 1960. A note on the estimator for the variance that utilizes the kurtosis. Am. Stat. 44, 295−296.

Sharma, P., Bouza, C.N., Verma, H., Singh, R., Sautto, J.M., 2016. A generalized class of estimators for the finite population mean when the study variable is qualitative in nature. Rev. Investig. Oper. 37 (2), 163−172.

Singh, H.P., Mehta, V., 2013. An improved estimation of parameters of Morgenstern type bivariate logistic distribution using ranked set sampling. Statistica 73 (4), 437−461.

Singh, H.P., Mehta, V., 2014a. Linear shrinkage estimator of scale parameter of Morgenstern type bivariate logistic distribution using ranked set sampling. Model Assist. Stat. Appl. 9, 295−307.

Singh, H.P., Mehta, V., 2014b. An alternative estimation of the scale parameter for Morgenstern type bivariate log-logistic distribution using ranked set sampling. J. Reliab. Stat. Stud. 7 (1), 19−29.

Singh, H.P., Mehta, V., 2015. Estimation of scale parameter of a Morgenstern type bivariate uniform distribution using censored ranked set samples. Model Assist. Stat. Appl. 10, 139−153.

Singh, H.P., Mehta, V., 2016a. Improved estimation of scale parameter of Morgenstern type bivariate uniform distribution using ranked set sampling. Commun. Stat: Theory Methods 45 (5), 1466−1476.

Singh, H.P., Mehta, V., 2016b. Some classes of shrinkage estimators in the Morgenstern type bivariate exponential distribution using ranked set sampling. Hacet. J. Math. Stat. 45 (2), 575−591.

Singh, H.P., Mehta, V., 2016c. A class of shrinkage estimators of scale parameter of uniform distribution based on K-record values. Natl. Acad. Sci. Lett. 39, 221−227.

Singh, H.P., Mehta, V., 2017. Improved estimation of the scale parameter for log-logistic distribution using balanced ranked set sampling. Stat. Trans: New Ser. 18 (1), 53−74.

Singh, J., Pandey, B.N., Hirano, K., 1973. On the utilization of known coefficient of kurtosis in the estimation procedure of variance. Ann. Inst. Stat. Math. 25, 51−55.

Stokes, S.L., 1977. Ranked set sampling with concomitant variables. Commun. Stat.: Theory Methods 6, 1207−1211.

Stokes, S.L., 1995. Parametric ranked set sampling. Ann. Inst. Stat. Math. 47, 465−482.

Tahmasebi, S., Jafari, A.A., 2012. Estimation of a scale parameter of Morgenstern type bivariate uniform distribution by ranked set sampling. J. Data Sci. 10, 129−141.

CHAPTER

STATISTICAL INFERENCE USING STRATIFIED RANKED SET SAMPLES FROM FINITE POPULATIONS 12

Omer Ozturk[1] **and Konul Bayramoglu Kavlak**[2]

[1]*Department of Statistics, The Ohio State University, Columbus, OH, United States* [2]*Department of Actuarial Sciences, Hacettepe University, Ankara, Turkey*

12.1 INTRODUCTION

In many survey sampling studies, the population is often divided into exclusively disjointed subpopulations using supplementary or auxiliary information on population units. If these subpopulations have different mean and variance values, one can select a stratified sample to construct more precise estimators for population quantities. Stratified sampling is well understood and studied in survey sampling literature. For settings, where auxiliary information is available for all population units, in addition to stratum structure, one can induce a second layer of structure within each stratum sample by grouping the observations based on their relative positions in small sets. This second layer structure can be induced by selecting independent ranked set samples across strata populations. Stratified ranked set sample (SRSS) controls the variation in the sample in a two-stage process. The first stage divides the population into disjointed subpopulations and selects ranked set samples (RSSs) from each stratum. It partitions the total variation in the sample as between- and within-stratum variation. The construction of the RSS sample from each stratum in the second-stage further partitions the within-stratum variation into between- and within-ranking group variations. Due to this two-layer stratification, SRSS controls the total variation better than a stratified SRS and ranked set sample alone. Hence, stratified RSS yields better informative samples than its competitor samples.

In a finite population, the construction of a ranked set sample of size n requires a set size H and cycle size d. Once we determine the set and cycle sizes, we select nH units from the population without replacement and partition them at random into n sets, each having H units. We rank the units in each set with respect to the characteristic of interest. In these sets, we measure the units with rank 1 in the first d sets, the units with rank 2 in the next d sets and so on. This yields samples of H different sets of judgment order statistics, each of which has d independent and identically distributed judgment order statistics. These measured observations are called a ranked set sample from a finite population.

Ranked Set Sampling. DOI: https://doi.org/10.1016/B978-0-12-815044-3.00012-5

In this chapter, we consider the case when the entire population is divided into L *layers* or strata. The l-th stratum population is denoted with

$$P^{N_l} = \{y_{1l}, \ldots, y_{N_l l}\}; \quad l = 1, \ldots, L,$$

where y_{il} are unknown nonrandom fixed quantities and N_l is the population size for stratum l. We assume that N_l; $l = 1, \ldots, L$ are all known and $N = N_1 + \ldots + N_L$, where N is the total number of units in the entire population.

Construction of SRSS requires L independent RSS samples, one from each stratum. Let H_l and d_l be the set and cycle sizes for the lth stratum. Stratified RSS then consists of L independent RSS samples RSS$_l$, where RSS$_l$ is selected from the lth stratum, $l = 1, \ldots, L$. Since we are in a finite population setting, the distributional properties of the RSS$_l$ sample depend on whether the sample is constructed with or without replacement and whether we use a design-based or model-based inference. A detailed description of the construction of RSS$_l$ is provided in Section 12.2.

Ranked set sampling was first suggested by McIntyre (1952) to increase the efficiency of the estimator of the population mean. The theoretical foundation of stratification based on ordering of sample units is considered in Takahasi and Wakimoto (1968). They showed that if the stratification is done based on a balanced ranked set sample in which each judgment class has an equal number of measured observations, the estimators are unbiased. They also showed that the ranked set sample mean is more efficient than a simple random sample mean of comparable size and provided an upper and lower bound for its relative efficiency. This upper bound is achieved for uniform distribution. Patil, Sinha, and Taillie (1995) used a without replacement RSS sample to estimate the mean of a finite population of size N. Deshpande, Frey, and Ozturk (2006) introduced three different without replacement sampling policies for RSS designs from finite populations. The first design constructs a sample with replacement, the second design constructs a sample by replacing only the measured observations, and the third design constructs a sample by replacing none of the units in each ranked set regardless of the measurement status. They provide a computational algorithm to construct confidence intervals for the population quantiles based on these three designs.

Over the last two decades, research effort in RSS sampling in finite populations has concentrated in two areas. Many researchers computed inclusion probabilities of sample units and constructed Horwitz−Thompson type estimators (Al-Saleh and Samawi, 2007; Frey, 2011; Gokpinar and Ozdemir, 2010; Ozturk and Jafari Jozani, 2013; Jafari Jozani and Johnson, 2011). In the other direction, researchers applied RSS methods in well-established survey sampling techniques in finite populations. Sroka (2008) and Samawi (1996) used ranked set sampling design to stratify populations. Both researchers used with replacement sampling design to construct ranked set samples from each stratum. Wang et al. (2016) used ranked set sampling in cluster randomized designs to estimate the treatment effect in two-sample problems. They fit a mixed effect model to RSS data assuming the cluster effect is random. Ozturk (2017) developed RSS sampling designs for finite clustered populations. Nematollahi, Salehi, and Aliakbari Saba (2008) used an RSS sampling design only in the second stage of a two-stage sampling. Sud and Mishra (2006) used RSS sampling in a clustered population under the assumption that all cluster populations have the same size. Samawi and Siam (2003) and Mandowara and Mehta (2014) applied with replacement RSS samples to ratio estimators. Most of these published papers assume that the population has infinite size or that the RSS sample is constructed with replacement. Recently, Ozturk (2014, 2016a,b) developed statistical inference based on without replacement RSS sampling designs in finite populations. He showed that without replacement RSS sampling designs provide additional benefits to improve the efficiency due to negative correlations among measured units.

To our knowledge, all published work in RSS in finite population settings, with the exception of Ozturk and Bayramoglu Kavlak (2017), develops inference using design-based randomized ranked set samples. In this chapter, we focus on finite stratified population setting and develop inference using both design- and model-based approaches. In Section 12.2, we clearly explain the construction of RSS samples from each stratum population. The estimators for the population mean and total are given using a stratified RSS sample. Section 12.3 investigates the distributional properties of the SRSS mean estimator under design- and model-based sampling methods. Section 12.4 provides unbiased estimators for the variance and mean square prediction error (MSPE) of the sample mean estimator. These unbiased estimators are used to construct confidence and prediction intervals for the population mean under design- and model-based inference, respectively. Section 12.5 provides empirical evidence to evaluate the properties of the estimators, confidence, and prediction intervals. Section 12.6 provides an example. Finally, Section 12.7 provides concluding remarks.

12.2 STRATIFIED RANKED SET SAMPLE

To construct a stratified RSS sample, for each l, we first determine the set size H_l and cycle size d_l, and select a set of size H_l experimental units at random without replacement,$Y_{1l}, \ldots, Y_{H_l l}$, from P^{N_l}. Units in this set are ranked based on the variable of interest Y in an increasing magnitude without actual measurement, $\{Y_{[1]l}, Y^*_{[2]l}, \ldots, Y^*_{[H]l}\}$. The ranking process can be performed either using visual inspection or some auxiliary variables. Hence, it is subjected to ranking error. In the ranked set, we identify and measure the unit that corresponds to the smallest Y, $Y_{[1]l}$. The remaining unmeasured units are marked as $Y^*_{[2]l}, \ldots, Y^*_{[H]l}$. After $Y_{[1]l}$ is measured, none of the H_l units in the set $\{Y_{[1]l}, Y^*_{[2]l}, \cdots, Y^*_{[H_l]l}\}$ is returned to the population P^{N_l}. Hence, the new population $P^{N_l - H_l}$ contains $N_l - H_l$ units prior to selection of the next set. We now select another set of size H_l from the population $P^{N_l - H_l}$, rank the units, and measure the second smallest unit $Y_{[2]l}$ in $\{Y^*_{[1]l}, Y_{[2]l}, Y^*_{[3]l}, \ldots, Y^*_{[H_l]l}\}$. We continue the process in this way until we select a set from population $P^{N_l - H_l(H_l - 1)}$ and measure $Y_{[H_l]l}$ in $\{Y^*_{[1]l}, Y^*_{[2]l}, \ldots, Y^*_{[H_l-1]l}, Y_{[H_l]l}\}$. The measured observations $Y_{[h]l}$; $h = 1, \ldots, h_l$, are called a cycle in the RSS sample from stratum l. To increase the sample size to $n_l = d_l H_l$, we repeat this process d_l times and obtain an RSS sample $Y_{[h]il}$; $i = 1, \ldots, d_l$; $h = 1, \ldots, H_l$ from stratum $l = 1, \ldots, L$. For notational convenience, a capital letter ($Y_{[h]il}$) is used to denote the random variables and a lowercase letter (y_{il}) is used to denote the value of the ith unit in the population P^{N_l}.

The estimator of the population mean based on SRSS data can be constructed as follows:

$$\overline{Y}_{\text{SRSS}} = \sum_{l=1}^{L} \frac{N_l}{N} \left(\frac{1}{d_l H_l} \sum_{i=1}^{d_l} \sum_{h=1}^{H_l} Y_{[h]il} \right) = \sum_{l=1}^{L} \frac{N_l}{N} \overline{Y}_{\text{RSS}_l}$$

where $\overline{Y}_{\text{RSS}_l}$ is the mean of the ranked set sample from stratum l. It is immediately observed that the estimator ($\overline{Y}_{\text{SRSS}}$) is the weighted average of the RSS sample strata means. The estimator for the population total can easily be established

$$T_{\text{SRSS}} = N \overline{Y}_{\text{SRSS}}.$$

The distributional properties of $\overline{Y}_{\text{SRSS}}$ and T_{SRSS} depend on whether the inference is developed based on a randomization theory or super population model. In the next section we investigate these two models in detail.

12.3 STATISTICAL INFERENCE

Statistical inference in finite population setting can be developed either using a randomization theory or model-based approach. The randomization principal treats the value of each unit in the population as a fixed quantity. The variation in the sample is accounted for by the probability that the unit is included in the sample which is controlled by how the sample units are selected. The bias and variance of the estimator, and the coverage probability of a confidence interval are computed over the probability distribution of sampling design that governs the selection of all possible samples. There is no need to make an assumption on the distribution of random variables. Hence, the randomization theory provides a nonparametric inference for the finite population. The randomization principal is also referred to as a design-based approach since the survey statistician designs his/her own selection probabilities.

In a design-based approach, the population is divided into L mutually exclusive subpopulations $P^{N_l} = \{y_{1l}, \ldots, y_{N_l l}\}, \;\; l = 1,\ldots,L$. In this population y_{il} is a nonrandom fixed value. The mean and variance of subpopulations are defined by

$$\bar{y}_l = \frac{1}{N_l}\sum_{i=1}^{N_l} y_{il}, \quad S_l^2 = \frac{1}{N_l}\sum_{i=1}^{N_l}(y_{il}-\bar{y}_l)^2, \;\; l = 1,\ldots,L.$$

The overall population contains all units in all subpopulations

$$\mathscr{P} = \{y_{11}, \ldots, y_{N_1 1}, \ldots, y_{1L}, \ldots, y_{N_L L}.$$

The total and mean of population $\mathscr{P}$ is given by:

$$t_N = \sum_{l=1}^{L} t_l, \quad t_l = \sum_{i=1}^{N_l} y_{il}, \quad \bar{y}_N = t_N/N.$$

Under a design-based approach, SRSS observations $Y_{[h]il}$ are independent only if they are from different strata. Since sets are constructed without replacements, any two observations $Y_{[h]il}, Y_{[h]i'l}$ from the same stratum are correlated. We first look at the marginal and joint probability distributions of $Y_{[h]il}$ and $(Y_{[h]il}, Y_{[h]jl})$ in RSS$_l$ obtained from subpopulation P^{N_l}. The proof of the following lemma is given in Ozturk (2016b).

Lemma 1: *Let* $Y_{[h]il}$; $h = 1,\ldots,H_l$; $\;i = 1,\ldots,d_l$, be a ranked set sample from population $\mathscr{P}^{N_l}$.

1. The marginal and joint probability mass functions of $Y_{[h]1l}$ and $(Y_{[h]1l}, Y_{[h']2l})$, respectively, given by

$$\beta(k;h|l) = P(Y_{(h)il} = y_{kl}) = \frac{\binom{k-1}{h-1}\binom{N_l-k}{H_l-h}}{\binom{N_l}{H_l}}, y_{kl} \in \mathscr{P}^{N_l}$$

and

$$\beta(k,k';h,h'|l) = P(Y_{(h)il} = y_{kl}, Y_{(h')jl} = y_{k'l}), k < k', (y_{kl}, y_{k'l}) \in P^{N_l}$$

$$= \sum_{\lambda=0}^{k'-k-1} \frac{\binom{k-1}{h-1}\binom{k'-k-1}{\lambda}\binom{N_l-k'}{H_l-\lambda-h}\binom{k'-1-h-\lambda}{h^{'}-1}\binom{N_l-k'-H_l+\lambda+h}{H_l\ -\ h'}}{\binom{N_l}{H_l}\binom{N_l-H_l}{H_l}}$$

2. The mean and variance of $Y_{(h)1l}$ and covariance of $(Y_{(h)1l}, Y_{(h')2l})$ are given by

$$y_{(h)l} = E(Y_{(h)1l}) = \sum_{k=1}^{N_l} y_{kl}\beta(k;h|l)$$

$$S^2_{(h)l} = \mathrm{Var}(Y_{(h)1l}) = \sum_{k=1}^{N} y^2_{kl}\beta(k;h|l) - y^2_{(h)l}$$

$$S^2_{(h,h')l} = \mathrm{Cov}(Y_{(h)1l}, Y_{(h')2l}) = \sum_{k=1}^{N_l}\sum_{k'=1}^{N_l} y_{kl}y_{k'l}\beta(k,k';h,h'|l) - y_{(h)l}y_{(h')l}.$$

We note that in Lemma 1 we assume that the ranking process in each set is perfect. Hence, we replace the square brackets with round ones to indicate that Lemma 1 holds only under perfect ranking. Under imperfect ranking, we replace the round parentheses in $\bar{y}_{(h)l}$, $S^2_{(h)l}$ and $S^2_{(h,h')l}$ with square brackets and write $\bar{y}_{[h]l}$, $S^2_{[h]l}$, and $S^2_{[h,h']l}$. Under imperfect ranking, there is no closed form expression for the mean, variance, and covariance of judgment order statistics. For notational convenience, SRSS mean under design-based inference will be denoted as $\overline{Y}_D = \overline{Y}_{\mathrm{SRSS}}$.

Theorem 1: *Let* $Y_{[h]il}$; $l = 1, 2, \ldots, L$; $h = 1, \ldots, H_l$; $i = 1, \ldots, d_l$ *be a stratified ranked set sample. The estimator* $\overline{Y}_D$ *is unbiased for* $\bar{y}_N$ *and its variance is given by* $\sigma^2_D = Var(\overline{Y}_D)$

$$\sigma^2_D = \sum_{l=1}^{L}\left(\frac{N_l}{N}\right)^2\left[\left(\frac{N_l-1-n_l}{(N_l-1)n_l}\right)S^2_l - \frac{1}{n_lH_l}\sum_{h=1}^{H_l}(y_{[h]l} - y_l)^2 - \frac{1}{n_lH_l}\sum_{h=1}^{H_l}S^2_{[h,h]l}\right],$$

where the subscript "D" is used to highlight that the variance is computed under a design-based approach.

Using Theorem 1 one can easily establish that the estimator $T_D = T_{\mathrm{RSSS}}$ is unbiased for t_N and its variance equals $\mathrm{Var}(T_{\mathrm{RSSS}}) = N^2\sigma^2_D$.

Model-based inference treats the value y_{il} on a finite population unit in P^{N_l} as a realization from a larger population, a *super population*. In this case the finite population unit i has a random variable Y_{il} that has some probability distribution with mean μ_l and variance σ^2_l. The actual values $y_{1l}, \ldots, y_{N_l l}$ of the finite population P^{N_l} are one realization of the random variables $Y_{il}; i = 1, \ldots, N_l$ from a distribution with mean μ_l and variance σ^2_l. The joint distribution of Y_{il}; $i = 1, \ldots, N_l$ provides the link between the units in the sample and units not in the sample. This link does not exist in a design-based approach. In model-based inference, we observe the sample from the finite population, and use these data and the model to predict the unobserved values in the population. Thus, a model-based approach can be put in the framework of a prediction model. A model structure of stratified sample can be framed as a one-way ANOVA model with fixed effects

$$Y_{il} = \mu_l + \epsilon_{il}, \quad E(\epsilon_{il}) = 0, \quad \mathrm{Var}(\epsilon_{il}) = \sigma_l^2; \quad i = 1,\ldots,N_l; \quad l = 1,\ldots,L. \tag{12.1}$$

In a model-based approach, one can easily establish the following equalities

$$\mu_{[h]l} = E_M(Y_{[h]il}), \quad \sigma^2_{[h]l} = \mathrm{Var}_M(Y_{[h]il}), \quad \sigma^2_{[h,h']l} = \mathrm{Cov}_M(Y_{[h]il}, Y_{[h']2l}).$$

We again use subscript "M" to denote that the mean, variance, and covariance are computed under super population model in Eq. (12.1). Let

$$\overline{Y}_N = \frac{1}{N}\sum_{l=1}^{L} N_l \overline{Y}_l, \quad \overline{Y}_l = \frac{1}{N_l}\sum_{i=1}^{N_l} Y_{il}; \;\; l = 1,\cdots,L.$$

For notational simplicity, under the super population model, we denote the SRSS mean $\overline{Y}_{\mathrm{SRSS}}$ with $\overline{Y}_M = \overline{Y}_{\mathrm{SRSS}}$. We can show that the estimator $\overline{Y}_M$ is model unbiased

$$E_M(\overline{Y}_M - \overline{Y}_N) = \sum_{l=1}^{L} \frac{N_l}{N} E_M(\overline{Y}_{\mathrm{RSS},l} - \overline{Y}_l) = 0.$$

The last equality in the above equation follows from the fact that $\overline{Y}_{\mathrm{RSS},l}$ is an unbiased estimator for μ_l. The mean square prediction error (MSPE) under model (1) is given by

$$\sigma_M^2 = \mathrm{MSPE}(\overline{Y}_M) = E_M(\overline{Y}_M - \overline{Y}_N)^2.$$

Theorem 2: *Let $Y_{[h]il}$, $l = 1, 2,\ldots,L$, $h = 1,\ldots,H_l$, $i = 1,\ldots,d_l$ be a stratified ranked set sample from a finite population. Under the super population model in* Eq. (12.1)*, the MSPE of the estimator $\overline{Y}_M = \overline{Y}_{SRSS}$ is given by*

$$\sigma_M^2 = \sum_{l=1}^{L} \left(\frac{N_l}{N}\right)^2 \left[\left(\frac{N_l - n_l}{N_l n_l}\right)\sigma_l^2 - \frac{1}{n_l H_l}\sum_{h=1}^{H_l}(\mu_{[h]l} - \mu_l)^2\right].$$

It is immediately observed that

$$\sigma_M^2 = \sigma^2_{\mathrm{SSRS},M} - \sum_{l=1}^{L}\left(\frac{N_l}{N}\right)^2 \frac{1}{n_l H_l}\sum_{h=1}^{H_l}(\mu_{[h]l} - \mu_l)^2,$$

where

$$\sigma^2_{\mathrm{SSRS},M} = \sum_{l=1}^{L}\left(\frac{N_l}{N}\right)^2\left[\left(\frac{N_l - n_l}{N_l n_l}\right)\sigma_l^2\right]$$

is the MSPE of the estimator of the population mean using stratified simple random sample (SSRS) under super population model in Eq. (12.1). Thus, it can be concluded that the MSPE of the estimator $\overline{Y}_M$ is never greater than the MSPE of the SSRS estimator.

12.4 ESTIMATORS OF VARIANCE AND MSPE

In this section, we construct unbiased estimators for σ_M^2 and σ_D^2. We first rewrite the estimators in slightly different forms

$$
\begin{aligned}
\sigma_M^2 &= \sum_{l=1}^{L}\left(\frac{N_l}{N}\right)^2\left[\left(\frac{N_l-n_l}{N_l n_l}\right)\sigma_l^2-\frac{1}{n_l H_l}\sum_{h=1}^{H_l}(\mu_{[h]l}-\mu_l)^2\right]\\
&= \sum_{l=1}^{L}\left(\frac{-N_l}{N^2}\sigma_l^2+\frac{N_l^2}{N^2 H_l n_l}\sum_{h=1}^{H_l}\sigma_{[h]l}^2\right)
\end{aligned}
$$

and

$$
\begin{aligned}
\sigma_D^2 &= \sum_{l=1}^{L}\left(\tfrac{N_l}{N}\right)^2\left[\left(\frac{N_l-1-n_l}{(N_l-1)n_l}\right)S_l^2-\frac{1}{n_l H_l}\sum_{h=1}^{H_l}(y_{[h]l}-y_l)^2-\frac{1}{n_l H_l}\sum_{h=1}^{H_l}S_{[h,h]l}^2\right]\\
&= \sum_{l=1}^{L}\left(\tfrac{N_l}{N}\right)^2\left[\frac{1}{n_l H_l}\sum_{h=1}^{H_l}\left(S_{[h]l}^2-S_{[h,h]l}^2\right)-\frac{S_l^2}{(N_l-1)}\right].
\end{aligned}
$$

Let

$$
\begin{aligned}
T_{1l}^{*} &= \frac{1}{2d_l^2 H_l^2}\sum_{h=1}^{H_l}\sum_{h\neq h'}^{H_l}\sum_{i=1}^{d_l}\sum_{j=1}^{d_l}(Y_{[h]il}-Y_{[h']jl})^2\\
T_{2l}^{*} &= \frac{1}{2d_l(d_l-1)H_l^2}\sum_{h=1}^{H_l}\sum_{i=1}^{d_l}\sum_{j\neq i}^{d_l}(Y_{[h]il}-Y_{[h]jl})^2.
\end{aligned}
$$

Using these definitions, one can easily establish the following result.

Theorem 3: *Let* $Y_{[h]il};\ i=1,\ldots,d_l;\ \ h=1,\ldots,H;\ \ l=1,\ldots,L$, *be an SRSS of set size* H_l *from a finite population. The unbiased estimator of* σ_M^2 *and* σ_D^2 *is given by*

$$
\hat{\sigma}_M^2=\hat{\sigma}_D^2=\sum_{l=1}^{L}\left[\frac{N_l^2}{N^2 d_l}T_{2l}^{*}-\left(T_{1l}^{*}+T_{2l}^{*}\right)\frac{N_l}{N^2}\right]. \tag{12.2}
$$

Theorem 3 indicates that the estimators $\hat{\sigma}_M^2$ and $\hat{\sigma}_D^2$ are unbiased for any sample and set sizes, regardless of the quality of ranking information, as long as $d_l>1$ for $l=1,\ldots,L$. These unbiased estimators allow us to construct approximate $(1-\alpha)100\%$ confidence and prediction intervals under randomization design and super population model in Eq. (12.1), respectively. Using normal approximation, a $(1-\alpha)100\%$ confidence interval for the population mean $\overline{y}_N$ is given by

$$
\overline{Y}_D \pm t_{n-L,\alpha/2}\hat{\sigma}_D^2, \tag{12.3}
$$

where $t_{df,a}$ is the a-th upper quantile of t-distribution with df degrees of freedom. The degrees of freedom $df=n-L$ are suggested to account for the heterogeneity among L stratum populations. In a similar fashion, an approximate prediction interval for $\overline{Y}_N$ is given by

$$
\overline{Y}_M \pm t_{n-L,\alpha/2}\hat{\sigma}_M^2. \tag{12.4}
$$

12.5 EMPIRICAL RESULTS

In this section we investigate the finite sample properties of the SRSS mean to estimate the population mean $\overline{y}_N$ from a stratified population with three strata. Strata populations are generated from discrete normal populations for population sizes, $N_1 = 200$, $N_2 = 300$, and $N_3 = 400$. The discrete normal population is generated from $y_{li} = F^{-1}(i/(N_l + 1), \mu_l, \sigma_l)$; $i = 1, \ldots, N_l$, where $F(u, \mu_l, \sigma_l)$ is the cumulative distribution function of a normal distribution with mean μ_l and standard deviation σ_l. Location and scale parameters for the three strata populations are selected to be $\mu_1 = 5$, $\mu_2 = 10$, $\mu_3 = 13$ and $\sigma_1 = 2$, $\sigma_2 = 5$, $\sigma_3 = 7$, respectively. In the simulation study two different set and cycle sizes are considered. The set sizes H_l are taken to be $2, 3, 5$ and $4, 5, 7$. The cycle sizes are selected as $d_l = 5, 8$; $l = 1, 2, 3$. Units are ranked based on an auxiliary variable X. The quality of the ranking information is controlled by the correlation coefficient between Y and X, $\rho = \text{corr}(Y, X)$. The correlation coefficient $\rho = 1$ yields perfect ranking, the correlation coefficients $\rho = 0.90, 0.75, 0.50$ yield imperfect ranking. For each combination of simulation parameters, we generated 50,000 SRSS and SSRS. Relative efficiencies of SRSS mean estimators for both design- and model-based approaches are compared with stratified simple random sample mean estimator of the population mean. We use the following expressions to obtain the relative efficiencies

$$RE_M = \frac{\sigma^2_{\text{SSRS},M}}{\sigma^2_M}, \quad RE_D = \frac{\sigma^2_{\text{SSRS},D}}{\sigma^2_D},$$

where $\sigma^2_{\text{SSRS},M}$ and $\sigma^2_{\text{SSRS},D}$ are the variance of SSRS mean

$$\sigma^2_{\text{SSRS},M} = \sigma^2_{\text{SSRS},D} = \sum_{l=1}^{L} \left(\frac{N_l}{N}\right)^2 \left[\left(\frac{N_l - n_l}{N_l n_l}\right)\sigma_l^2\right].$$

The simulation study also investigated the properties of the estimators $\hat{\sigma}^2_M$, $\hat{\sigma}^2_D$ and coverage probabilities of confidence and prediction intervals in Eqs. (12.3) and (12.4).

Table 12.1 presents empirical results for selected simulation parameters. In Table 12.1, the headings Var($\overline{Y}_M$) and Var($\overline{Y}_D$) give the variances of 50,000 simulated $\overline{Y}_D$ and $\overline{Y}_M$, respectively. Under perfect ranking, unbiased variance estimates $\hat{\sigma}^2_M$, $\hat{\sigma}^2_D$ and simulated variance estimates Var($\overline{Y}_M$) and Var($\overline{Y}_D$) are all close to the theoretical variances σ^2_D and σ^2_M. Under imperfect ranking, there are no available analytic expressions to compute σ^2_D and σ^2_M, hence these entries are left blank in Table 12.1. Under imperfect ranking, simulated and unbiased variance estimates are very close to each other within the simulation variation.

The efficiencies of the estimators $\overline{Y}_M$ and $\overline{Y}_D$ with respect to the same estimators based on stratified simple random samples are all greater than one, and increase with ρ and cycle size d as expected. The coverage probabilities of the confidence ($C(\overline{Y}_D)$) and prediction ($C(\overline{Y}_M)$) intervals of population mean are very close to the nominal coverage probability 0.95.

12.6 EXAMPLE

In this section we apply the proposed stratified ranked set sampling design to apple production data in Turkey. The data set was collected by the Turkish Statistical Institute. Apples in Turkey are produced in seven different geographical regions: Marmara, Aegean, Mediterranean, Central Anatolia, Black Sea, Eastern Anatolia, and Southeastern Anatolia regions. These regions have different climate patterns and apple production varies from region to region. The data set contains two variables, apple production (Y) (in tons, 1 ton = 1000 kg) and the number of apple trees (X) in each

Table 12.1 Variance Estimates, Relative Efficiencies of the Estimators ($\overline{Y}_M$, $\overline{Y}_M$), Coverage Probabilities ($C(\overline{Y}_D)$, $C(\overline{Y}_M)$) of 95% Confidence and Prediction Intervals of Population Mean. Data sets are Generated From Discrete Normal Population With Strata Population Means $\mu_1 = 5,\ \mu_2 = 10,\ \mu_3 = 15$, Strata Population Standard Deviations $\sigma_1 = 2,\ \sigma_2 = 5,\ \sigma_3 = 7$, and Strata Population Sizes $N_1 = 200,\ N_2 = 300,\ N_3 = 400$

					Simulated		Unbiased		Coverage		Efficiency	
d	ρ	σ^2_M	σ^2_D	σ^2_{SSRS}	$V(\overline{Y}_M)$	$V(\overline{Y}_D)$	$\hat{\sigma}^2_M$	$\hat{\sigma}^2_D$	$C(\overline{Y}_M)$	$C(\overline{Y}_D)$	RE_M	RE_D
$H_1 = 2,\ H_2 = 3,\ H_3 = 5$												
5	0.50	–	–	0.558	0.473	0.473	0.472	0.472	0.947	0.947	1.180	1.180
	0.75	–	–	0.558	0.376	0.371	0.374	0.374	0.947	0.948	1.481	1.504
	0.90	–	–	0.558	0.288	0.291	0.287	0.287	0.945	0.945	1.937	1.917
	1.00	0.216	0.207	0.558	0.206	0.209	0.208	0.208	0.946	0.945	2.701	2.668
	0.50	–	–	0.336	0.282	0.282	0.282	0.282	0.949	0.945	1.191	1.189
8	0.75	–	–	0.336	0.221	0.224	0.221	0.221	0.948	0.947	1.516	0.500
	0.90	–	–	0.336	0.167	0.168	0.166	0.167	0.947	0.946	2.004	1.994
	1.00	0.122	0.117	0.336	0.117	0.116	0.117	0.117	0.947	0.947	2.869	2.881
$\boldsymbol{H_1 = 4,\ H_2 = 5,\ H_3 = 7}$												
5	0.50	–	–	0.363	0.294	0.296	0.296	0.296	0.949	0.948	1.235	1.227
	0.75	–	–	0.363	0.218	0.216	0.217	0.218	0.947	0.948	1.663	1.680
	0.90	–	–	0.363	0.148	0.147	0.147	0.147	0.946	0.947	2.446	2.473
	1.00	0.087	0.083	0.363	0.083	0.083	0.083	0.083	0.943	0.943	4.375	4.379
	0.50	–	–	0.214	0.172	0.172	0.173	0.172	0.950	0.950	1.246	1.253
8	0.75	–	–	0.214	0.124	0.121	0.123	0.123	0.947	0.951	1.725	1.762
	0.90	–	–	0.214	0.080	0.079	0.079	0.079	0.947	0.946	2.676	2.696
	1.00	0.041	0.039	0.214	0.039	0.039	0.039	0.039	0.943	0.941	5.525	5.447

Table 12.2 Population Characteristics of Apple Production (in tons, 1 ton = 1000 kg) Data

Strata (l)	μ_l	σ_l	N_l	ρ_l
Marmara ($l=1$)	1536.8	6425	106	0.816
Aegean ($l=2$)	2212.6	11551.5	106	0.856
Mediterranean ($l=3$)	9384.31	29907.5	94	0.901
Black Sea ($l=4$)	967	2389.7	204	0.713
Central Anatolia ($l=5$)	5588	28643.4	171	0.986
Eastern Anatolia ($l=6$)	625.4	1167	104	0.886
Southeastern Anatolia ($l=7$)	71.4	110.9	69	0.917

township in each region. The X-values in all townships are available in the data frame prior to sampling. Hence they can be used for ranking the townships for their apple production in sets. In this population, we treat these seven regions as a stratified population. Table 12.2 gives the parameters of strata populations. As we observe from Table 12.2, strata populations have different means and variances. There is a strong positive correlation, ρ_l, between the X and Y variables. The sets of small townships can be ranked fairly accurately using the number of apple trees in each locality. The entire population has $N = 854$ townships and its mean is 2930.126 tons. Readers are referred to Kadilar and Cingi (2003) for further details about this population.

To illustrate the use of the proposed sampling design, we generated SRSS and SSRS from the apple production data. For the SRSS, we use $H_l = 3$ and $d_l = 4$, $l = 1,\ldots,7$. With these choices strata sample sizes become $n_l = 12$; $l = 1,\ldots,7$. For the SSRS, we constructed simple random samples of size $n_l = 12$ from each stratum population, so that both SRSS and SSRS have the same number of observations. The samples are presented in Table 12.3.

For the data set in Table 12.3, the estimated population means based on SRSS and SSRS are 3106.454 and 4747.012 tons, respectively. Estimates of the standard errors of these estimators are $\hat{\sigma}_D = \hat{\sigma}_M = 763.80$ tons and $\hat{\sigma}_{\text{SSRS}} = 2674.714$ tons. For these particular samples SRSS estimators have a smaller standard error as expected.

12.7 CONCLUDING REMARKS

In this chapter, we have constructed a stratified ranked set sample from a finite stratified population. Samples are constructed without replacement. Hence, measured observations are correlated. In a finite population setting, the statistical inference can be drawn either using design-based sampling techniques or a super population model. We constructed unbiased estimators for the population mean and total and their estimates using both approaches.

We show that the SRSS estimators are unbiased and they have higher efficiencies than the corresponding SSRS estimators. Confidence and prediction intervals for the population mean are reasonably close to nominal coverage probabilities. The proposed sampling scheme and estimators are applied to apple production data in a stratified populations.

Table 12.3 Stratified Ranked Set Sample From Apple Production Data

$l = 1$			$l = 2$			$l = 3$			$l = 4$			$l = 5$			$l = 9$			$l = 7$		
$Y_{[h]i1}$	h	SRS	$Y_{[h]i2}$	h	SRS	$Y_{[h]i3}$	h	SRS	$Y_{[h]i4}$	h	SRS	$Y_{[h]i5}$	h	SRS	$Y_{[h]i6}$	h	SRS	$Y_{[h]i7}$	h	SRS
100	1	540	90	1	201	14	1	110	83	1	115	85	1	21	22	1	497	14	1	50
488	2	495	390	2	8	308	2	22149	2900	2	132	816	2	21	293	2	35	53	2	190
1311	3	76	320	3	637	32340	3	183680	23389	3	668	37244	3	16	999	3	973	210	3	14
290	1	297	45	1	36	800	1	40	70	1	1723	81	1	342	11	1	730	5	1	53
193	2	111	90	2	6488	47183	2	18	1020	2	600	222	2	541	730	2	27	40	2	53
4476	3	152	2250	3	12	21200	3	6960	5070	3	470	1540	3	132	172	3	75	190	3	5
66	1	1800	2	1	67	121	1	1040	668	1	308	378	1	135621	16	1	230	44	1	6
20	2	14	828	2	65	550	2	1666	307	2	234	2625	2	660	63	2	6	115	2	77
297	3	193	15400	3	3671	563	3	2412	1358	3	230	230	3	54	840	3	265	198	3	1
90	1	24	94	1	248	14	1	133	132	1	980	144	1	1092	0	1	84	8	1	90
33	2	107	48	2	88	370	2	4959	130	2	1611	85	2	25	525	2	465	5	2	25
6500	3	275	1330	3	1587	24010	3	237	1411	3	63	4500	3	1668	667	3	50	53	3	5

ACKNOWLEDGMENT

The second author's research is supported by The Scientific and Technological Research Council of Turkey (TUBITAK).

REFERENCES

Al-Saleh, M.F., Samawi, H., 2007. A note on inclusion probability in ranked set sampling for finite population. Test 16, 198−209.

Deshpande, J.V., Frey, J., Ozturk, O., 2006. Nonparametric ranked set-sampling confidence intervals for a finite population. Environ. Ecol. Stat. 13, 25−40.

Frey, J., 2011. A note on ranked-set sampling using a covariate. J. Stat. Plan. Inference 141, 809−816.

Gokpinar, F., Ozdemir, Y.A., 2010. Generalization of inclusion probabilities in ranked set sampling. Hacet. J. Math. Stat. 39, 89−95.

Jafari Jozani, M., Johnson, B.C., 2011. Design based estimation for ranked set sampling in finite population. Environ. Ecol. Stat. 18, 663−685.

Kadilar, C., Cingi, H., 2003. Ratio estimators in stratified random sampling. Biom. J. 45, 218−225.

Mandowara, V.L., Mehta, N.R., 2014. Modified ratio estimators using stratified ranked set sampling. Hacet. J. Stat. 43, 461−471.

McIntyre, G., 1952. *A method for unbiased selective sampling using ranked set sampling*. Australian. J. Agric. Res. 3, 385−390.

Nematollahi, N., Salehi, M.M., Aliakbari Saba, R., 2008. Two-stage cluster sampling with ranked set sampling in the secondary sampling frame. Commun. Stat: Theory Methods 37, 2402−2415.

Ozturk, O., 2014. Estimation of population mean and total in finite population setting using multiple auxiliary variables. J. Agric. Biol. Environ. Stat. 19, 161−184.

Ozturk, O., 2016a. Estimation of a finite population mean and total using population ranks of sample units. J. Agric. Biol. Environ. Stat. 21, 181−202.

Ozturk, O., 2016b. Statistical inference based on judgment post-stratified samples in finite population. Surv. Methodol. 42, 239−262.

Ozturk, O., 2017. Two-stage cluster samples with ranked set sampling designs. Ann. Inst. Stat. Mathem. 69, 1−29.

Ozturk, O., Bayramoglu Kavlak, K. (2017). Model based inference using ranked set samples. Survey Methodology (in print).

Ozturk, O., Jafari Jozani, M., 2013. Inclusion probabilities in partially rank ordered set sampling. Comput. Stat. Data Anal. 69, 122−132.

Patil, G.P., Sinha, A.K., Taillie, C., 1995. Finite population corrections for ranked set sampling. Ann. Inst. Stat. Math. 47, 621−636.

Samawi, H.M., 1996. Stratified ranked set sample. Pak. J. Stat. 12, 9−16.

Samawi, H.M., Siam, M.I., 2003. Ratio estimation using stratified ranked set sample. Metron: Int. J. Stat. LXI, 75−90.

Sroka, C.J., 2008. *Extending Ranked Set Sampling to Survey Methodology*. PhD Thesis. Department of Statistics, Ohio State University, OH, United States.

Sud, V., Mishra, D.C., 2006. Estimation of finite population mean using ranked set two stage sampling designs. J. Indian Soc. Agric. Stat. 60, 108−117.

Takahasi, K., Wakimoto, K., 1968. On unbiased estimates of the population mean based on the sample stratified by means of ordering. Ann. Inst. Stat. Math. 20, 1−31.

Wang, X., Lim, J., Stokes, L., 2016. Using ranked set sampling with cluster randomized designs for improved inference on treatment effects. J. Am. Stat. Assoc. 516, 1576−1590.

APPENDIX

Proof of Theorem 1: The variance of $\overline{Y}_D$ can be written as

$$\begin{aligned}\mathrm{Var}_D(\overline{Y}_{\mathrm{SRSS}}) &= E_D\left(\sum_{l=1}^{L}\tfrac{N_l}{N}\overline{Y}_{\mathrm{RSS},l}-\sum_{l=1}^{L}\tfrac{N_l}{N}\overline{y}_l\right)^2\\ &=\sum_{l=1}^{L}\frac{N_l^2}{N^2}\sum_{l=1}^{L}E_D(\overline{Y}_{\mathrm{RSS},l}-\overline{y}_l)^2.\end{aligned}$$

The last equality follows from the fact that samples from different strata are independent. Adopting our notation in Theorem 4.1 or Eq. 4.5 in Patil et al. (1995), we write

$$\mathrm{Var}_D(Y_{\mathrm{SRSS}}) = \sum_{l=1}^{L}\frac{N_l^2}{N^2}\sum_{l=1}^{L}\left[\left(\frac{N_l-1-n_l}{(N_l-1)n_l}\right)S_l^2-\frac{1}{n_lH_l}\sum_{h=1}^{H_l}(y_{[h]l}-y_l)^2-\frac{1}{n_lH_l}\sum_{h=1}^{H_l}S_{[h,h]l}^2\right]$$

and complete the proof.

Proof of Theorem 2: We write the mean square prediction error (MSPE) as

$$\begin{aligned}\mathrm{MSPE}(\overline{Y}_{\mathrm{SRSS}}) &= E_M\left\{\overline{Y}_{\mathrm{SRSS}}-\overline{Y}_N\right\}^2\\ &= E_M\left\{\sum_{l=1}^{L}\tfrac{N_l}{N}\left(\tfrac{1}{d_lH_l}\sum_{i=1}^{d_l}\sum_{h=1}^{H_l}Y_{[h]il}\right)-\tfrac{1}{N}\sum_{l=1}^{L}\sum_{i=1}^{N_l}Y_{il}\right\}^2\\ &=\sum_{l=1}^{L}\frac{N_l^2}{N^2}E_M\left\{\overline{Y}_{\mathrm{RSS},l}-\overline{Y}_l\right\}^2.\end{aligned}$$

We use Theorem 1 in Ozturk and Bayramoglu Kavlak (2017) to complete the proof

$$\mathrm{MSPE}(\overline{Y}_{\mathrm{SRSS}}) = \sum_{l=1}^{L}\left(\frac{N_l}{N}\right)^2\left[\left(\frac{N_l-n_l}{N_ln_l}\right)\sigma_l^2-\frac{1}{n_lH_l}\sum_{h=1}^{H_l}(\mu_{[h]l}-\mu_l)^2\right].$$

Proof of Theorem 3: We first look at the expected values of T_{1l}^* and T_{2l}^* under super population and design-based models

$$E\left(T_{1l}^*\right) = \begin{cases}\dfrac{1}{H_l}\sum_{h=1}^{H_l}\left(\mu_{[h]l}-\mu_l\right)^2+\dfrac{H_l-1}{H_l^2}\sum_{h=1}^{H_l}\sigma_{[h]l}^2 & \text{for model based}\\ \dfrac{1}{H_l}\sum_{h=1}^{H_l}\left(y_{[h]l}-y_l\right)^2+\dfrac{H_l-1}{H_l^2}\sum_{h=1}^{H_l}S_{[h]l}^2-\dfrac{1}{H_l^2}\sum_{h=1}^{H_l}\sum_{h'\neq h}^{H_l}S_{[h,h']l} & \text{for design based}\end{cases}$$

and

$$E\left(T_{2l}^{*}\right)=\left(\begin{array}{ll}\frac{1}{H_l^2}\sum_{h=1}^{H_l}\sigma_{[h]l}^2 & \text{for model based}\\ \frac{1}{H_l^2}\sum_{h=1}^{H_l}S_{[h]l}^2-\frac{1}{H_l^2}\sum_{h=1}^{H_l}S_{[h,h]l}^2 & \text{for design based.}\end{array}\right.$$

It is now easy to establish that $E\left(T_{1l}^{*}+T_{2l}^{*}\right)=\sigma_l^2$ for model-based approach and $E\left(T_{1l}^{*}+T_{2l}^{*}\right)=\frac{N_l S_l^2}{(N_l-1)}$ for the design-based approach. The proof is then completed by inserting T_{1l}^{*} and T_{2l}^{*} in Eq. (12.2) and computing the expected values.

CHAPTER

SIMULTANEOUS ESTIMATION OF MEANS OF TWO SENSITIVE VARIABLES USING RANKED SET SAMPLING 13

Kumar Manikanta Pampana, Stephen A. Sedory and Sarjinder Singh

Department of Mathematics, Texas A&M University-Kingsville, Kingsville, TX, United States

13.1 INTRODUCTION

The problem of estimating the population mean of a sensitive variable, such as income, underreported tax, and number of induced abortions, etc., is well known in the field of randomized response sampling. Horvitz et al. (1967) and Greenberg et al. (1971) extended the Warner (1965) model to the case where the responses to the sensitive question are quantitative rather than a simple "yes" or "no," as when estimating the proportion of a sensitive attribute. One could refer to Fox (2016) that Fox and Tracy (1986) used the unrelated question model to estimate the correlation between two quantitative sensitive attributes. For estimating the mean of a sensitive variable, say Y, an additive model was introduced by Himmelfarb and Edgell (1980). In the additive model each interviewee scrambles a response Y by adding it to a random scrambling variable S and only then reports the scrambled value $Z = Y + S$ to the interviewer. The authors showed that the mean of the true values can be estimated from a sample of scrambled values by making use of knowledge of the distribution of the scrambling variable S.

Another variation of scrambled responses, with the name "multiplicative model," was introduced by Eichhorn and Hayre (1983) to estimate the population mean of a sensitive quantitative variable. In the multiplicative model each interviewee scrambles a response Y by multiplying it by a random scrambling variable S and only then reports the scrambled result $Z = Y \ S$ to the interviewer. The mean of the true variable given by $E(Y)$ can be estimated from a sample of scrambled values Z, again by making the use of knowledge of the distribution of the scrambling variable S.

Ahmed et al. (2018) pointed out that in both the additive and multiplicative models there are concerns about the choice of scrambling variables used while collecting data from the respondents. Both the additive and multiplicative models assume that the distribution of S is known, so a good guess about the maximum and minimum values of the scrambling variable S would also be known. Thus an interviewee may be suspicious that his/her true value of the sensitive variable can be discovered.

McIntyre (1952) was the first to introduce the idea of ranked set sampling (RSS) and claimed that it is more efficient than simple random and with replacement sampling (SRSWR).

Ranked Set Sampling. DOI: https://doi.org/10.1016/B978-0-12-815044-3.00013-7

To our knowledge, Bouza (2009) was the first to introduce the ingenious idea of using RSS while estimating the population mean of a sensitive quantitative variable. The units are ranked based on a judgment ranking but, for the purpose of analysis, the judgment ranking is assumed to be accurate.

Recently, Ahmed et al. (2018) considered a different approach which can be used to estimate the means of two sensitive variables simultaneously by making use of scrambled responses. They also claimed that a respondent would likely be more cooperative in responding because the proposed method makes use of one scrambled response and another fake response that is free from the true sensitive variables. In the next section, we discuss the Ahmed et al. (2018) procedure in brief.

13.2 AHMED, SEDORY, AND SINGH MODEL

Ahmed et al. (2018) introduced a new ingenious model, which we refer to as the Ahmed et al. (2018) model, where they consider the simultaneous estimation of means of two sensitive variables in a population Ω consisting of finite number of N persons. In their model, they consider selecting a sample s of n persons from the population Ω by using simple random and with replacement sampling (SRSWR). In the population of interest, the ith values of the variables of interest are labeled as Y_{1i} and Y_{2i} for the two quantitative sensitive variables. Assume population means of the first and the second variables Y_{1i} and Y_{2i} are μ_{y_1} and μ_{y_2} which are to be estimated. In the Ahmed et al. (2018) model, each respondent selected in the simple random and with replacement sample (SRSWR) is asked to generate two values of scrambling variables S_1 and S_2 one from each of two known distributions. Further, they assume that the scrambling variables S_1 and S_2 are independent, which helps to maintain the privacy of respondents and $E(S_1) = \theta_1$, $V(S_1) = \gamma_{20}$, $E(S_2) = \theta_2$ and $V(S_2) = \gamma_{02}$ are known.

In the Ahmed et al. (2018) randomized response model, each respondent selected in the sample is asked to report the scrambled response:

$$Z_{1i} = S_1 Y_{1i} + S_2 Y_{2i} \tag{13.1}$$

The authors claim that mixing two sensitive variables with two scrambling variables will certainly makes it difficult for an interviewer to guess the individual values of two sensitive variables. Further, they assume that there is no restriction on the scrambling variables to take any negative values, which will certainly increase respondents' cooperation while doing a face-to-face survey. Since the main theme of a randomized response survey is to protect a respondent during a face-to-face survey, the use of simple random sampling is highly recommended. Note that any other, more complex design making use of a highly correlated auxiliary variable at the selection stage may threaten the privacy of a respondent.

In the Ahmed et al. (2018) model, each respondent is also requested to rotate a spinner which consists of two outcomes, similar to the Warner (1965) spinner. If the pointer lands in a shaded area then the respondent is asked to report the value of the scrambling variable S_1, and if the pointer lands in the nonshaded area then the respondent is asked to report the value of the scrambling variable S_2. Let P be the proportion of shaded area and $(1-P)$ be the proportion of nonshaded area of the spinner. Thus the second response from the ith respondent is given by:

$$Z_i = \begin{cases} S_1 \text{ with probability } P \\ S_2 \text{ with probability } (1-P) \end{cases} \tag{13.2}$$

where

$$P \neq \frac{\theta_1\gamma_{02}}{\theta_1\gamma_{02} + \theta_2\gamma_{20}}.$$

Taking the expected value on both sides of Eq. (13.1) we have

$$E(Z_{1i}) = E[S_1Y_{1i} + S_2Y_{2i}] = \theta_1\mu_{y_1} + \theta_2\mu_{y_2} \tag{13.3}$$

From Eqs. (13.1) and (13.2), we generate the response Z_{2i} as follows:

$$Z_{2i} = Z_iZ_{1i} = \begin{cases} S_1^2Y_{1i} + S_1S_2Y_{2i} \text{ with probability } P \\ \\ S_1S_2Y_{1i} + S_2^2Y_{2i} \text{ with probabilty } (1-P) \end{cases} \tag{13.4}$$

Taking the expected value on both sides of Eq. (13.4) we have

$$E(Z_{2i}) = P\left[\mu_{y_1}\left(\theta_1^2 + \gamma_{20}\right) + \theta_1\theta_2\mu_{y_2}\right] + (1-P)\left[\theta_1\theta_2\mu_{y_1} + \left(\theta_2^2 + \gamma_{02}\right)\mu_{y_2}\right] \tag{13.5}$$

From Eqs. (13.3) and (13.5), by the method of moments, we have:

$$\theta_1\hat{\mu}_{y_1} + \theta_2\hat{\mu}_{y_2} = \frac{1}{n}\sum_{i=1}^{n} Z_{1i} \tag{13.6}$$

and

$$\left[P\left(\theta_1^2 + \gamma_{20}\right) + (1-P)\theta_1\theta_2\right]\hat{\mu}_{y_1} + \left[P\theta_1\theta_2 + (1-P)\left(\theta_2^2 + \gamma_{02}\right)\right]\hat{\mu}_{y_2} = \frac{1}{n}\sum_{i=1}^{n} Z_{2i} \tag{13.7}$$

Based on the Ahmed et al. (2018) model, unbiased estimators of μ_{y_1} and μ_{y_2} are, respectively, given by

$$\hat{\mu}_{y_1} = \frac{\{P\theta_1\theta_2 + (1-P)(\gamma_{02} + \theta_2^2)\}\overline{Z}_1 - \theta_2\overline{Z}_2}{(1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20}} \tag{13.8}$$

and

$$\hat{\mu_{y_2}} = \frac{\theta_1\overline{Z}_2 - \left\{P(\gamma_{20} + \theta_1^2) + (1-P)\theta_1\theta_2\right\}\overline{Z}_1}{(1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20}} \tag{13.9}$$

where

$$\overline{Z}_1 = \frac{1}{n}\sum_{i=1}^{n} Z_{1i} \text{ and } \overline{Z}_2 = \frac{1}{n}\sum_{i=1}^{n} Z_{2i}.$$

The variance of the estimator $\hat{\mu}_{y_1}$ is given by

$$V(\hat{\mu}_{y_1}) = \frac{\{P\theta_1\theta_2 + (1-P)(\gamma_{02} + \theta_2^2)\}^2\sigma_{Z_1}^2 + \theta_2^2\sigma_{Z_2}^2 - 2\theta_2\{P\theta_1\theta_2 + (1-P)(\gamma_{02} + \theta_2^2)\}\sigma_{Z_1Z_2}}{n\{(1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20}\}^2} \tag{13.10}$$

where

$$\sigma_{Z_1}^2 = \gamma_{20}(\sigma_{y_1}^2 + \mu_{y_1}^2) + \gamma_{02}(\sigma_{y_2}^2 + \mu_{y_2}^2) + \theta_1^2\sigma_{y_1}^2 + \theta_2^2\sigma_{y_2}^2 + 2\theta_1\theta_2\sigma_{y_1y_2} \tag{13.11}$$

$$\begin{aligned}\sigma_{Z_2}^2 = {} & (\sigma_{y_1}^2 + \mu_{y_1}^2)\left[P(\gamma_{40} + 4\gamma_{30}\theta_1 + 6\gamma_{20}\theta_1^2 + \theta_1^4) + (1-P)(\gamma_{20} + \theta_1^2)(\gamma_{02} + \theta_2^2)\right] \\ & + (\sigma_{y_2}^2 + \mu_{y_2}^2)\left[(1-P)(\gamma_{04} + 4\gamma_{03}\theta_2 + 6\gamma_{02}\theta_2^2 + \theta_2^4) + P(\gamma_{20} + \theta_1^2)(\gamma_{02} + \theta_2^2)\right] \\ & + 2(\sigma_{y_1y_2} + \mu_{y_1}\mu_{y_2})\left[P\theta_2(\gamma_{30} + 3\theta_1\gamma_{20} + \theta_1^3) + (1-P)\theta_1(\gamma_{03} + 3\theta_2\gamma_{02} + \theta_2^3)\right] \\ & - \left[\mu_{y_1}\{P(\gamma_{20}+\theta_1^2)+(1-P)\theta_1\theta_2\} + \mu_{y_2}\{P\theta_1\theta_2+(1-P)(\gamma_{02}+\theta_2^2)\}\right]^2\end{aligned} \tag{13.12}$$

and

$$\begin{aligned}\sigma_{Z_1Z_2} = {} & (\sigma_{y_1}^2 + \mu_{y_1}^2)\{P(\gamma_{30} + 3\theta_1\gamma_{20} + \theta_1^3) + (1-P)\theta_2(\gamma_{20} + \theta_1^2)\} \\ & + (\sigma_{y_2}^2 + \mu_{y_2}^2)\{P\theta_1(\gamma_{02} + \theta_2^2) + (1-P)(\gamma_{03} + 3\theta_2\gamma_{02} + \theta_2^3)\} \\ & + 2(\sigma_{y_1y_2} + \mu_{y_1}\mu_{y_2})\{P\theta_2(\gamma_{20} + \theta_1^2) + (1-P)\theta_1(\gamma_{02} + \theta_2^2)\} \\ & - (\theta_1\mu_{y_1} + \theta_2\mu_{y_2})[\mu_{y_1}\{P(\gamma_{20} + \theta_1^2) + (1-P)\theta_1\theta_2\} \\ & + \mu_{y_2}\{P\theta_1\theta_2 + (1-P)(\gamma_{02} + \theta_2^2)\}]\end{aligned} \tag{13.13}$$

The variance of the estimator $\hat{\mu}_{y_2}$ is given by

$$V(\hat{\mu}_{y_2}) = \frac{\theta_1^2\sigma_{Z_2}^2 + \{P(\gamma_{20}+\theta_1^2)+(1-P)\theta_1\theta_2\}^2\sigma_{Z_1}^2 - 2\theta_1\{P(\gamma_{20} + \theta_1^2) + (1-P)\theta_1\theta_2\}\sigma_{Z_1Z_2}}{n((1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20})^2} \tag{13.14}$$

where

$$\gamma_{ab} = E[S_1 - \theta_1]^a[S_2 - \theta_2]^b \tag{13.15}$$

In the next section, we consider an interesting extension of the Ahmed et al. (2018) model to a situation where the sample is taken by RSS. It becomes more interesting to consider which sensitive variable, the first or the second variable, should be considered for judgment ranking.

13.3 PROPOSED RANKED SET SAMPLING RANDOMIZED RESPONSE MODEL

Assume a population of interest Ω has two sensitive quantitative variables, Y_{1i} and Y_{2i}, $i = 1, 2, .., N$. Note that the precise values of both variables Y_{1i} and Y_{2i}, are unobservable for the ith unit in the population, Ω. Now the judgment ranking could be made either on the basis of the first sensitive variable Y_{1i} or on the basis of the second sensitive variable Y_{2i}. It may not be practical to make judgment ranking by considering both variables at the same time. For simplicity, let us consider judgment ranking based on only the first unobserved sensitive variable Y_{1i} while the second variable Y_{2i} is considered as a ranked auxiliary variable. One could refer to the recent works of Santiago et al. (2016) and Singh et al. (2014) where they considered the problem of estimation of mean of a study variable in the presence of an auxiliary variable. We may imagine arranging the

Table 13.1 Ranked Set Sampling Procedure

***t*th Cycle**				
$\{Y_{[11]t}, Y_{(21)t}\}$	$\{Y_{[12]t}, Y_{(22)t}\}$			$\{Y_{[1m]t}, Y_{(2m)t}\}$
$\{Y_{[12]t}, Y_{(22)t}\}$	$\{Y_{[12]t}, Y_{(22)t}\}$			$\{Y_{[1m]t}, Y_{(2m)t}\}$
$\vdots$	$\vdots$			$\vdots$
$\{Y_{[1m]t}, Y_{(2m)t}\}$	$\{Y_{[1m]t}, Y_{(2m)t}\}$			$\{Y_{[1m]t}, Y_{(2m)t}\}$

ranked values of the first sensitive variable Y_{1i} and the second sensitive variable Y_{2i} in the tth cycle, $t = 1, 2, \ldots, r$, of the proposed RSS as shown in Table 13.1.

Note that the diagonal entries are selected in the sample from the tth and each subsample consists of m units such that $n = mr$. Further note that there could be a little confusion while reading the suffixes, so read the above judgment ranking carefully.

In the proposed procedure, the observed ranked response $X_{[1i]}$ can be written as:

$$X_{[1i]} = S_1 Y_{[1i]} + S_2 Y_{(2i)} \tag{13.16}$$

where the square parentheses indicate that the first variable is arranged based on judgment ranking and the open parentheses indicate that the second variable is treated as an auxiliary variable and has not been ranked. It may be worth pointing out that we have considered the simplest case of a multiplicative model in Eq. (13.16) due to Eichhorn and Hayre (1983). In case of some special types of sensitive variables one should either follow their remark in Section 6 on page 315 or use another more general model due to Ahmed et al. (2018).

Again, following Ahmed et al. (2018), each respondent selected in the ranked set sample is also requested to experience a randomization device, say a deck of cards, and having two possible outcomes S_1 and S_2 with probabilities P and $(1 - P)$, respectively. We denote the second observed response in RSS as:

$$X_i = \begin{bmatrix} S_1 & \text{with probability} & P \\ S_2 & \text{with probability} & (1-P) \end{bmatrix} \tag{13.17}$$

Note that the observed second response cannot be ranked, because it is free from the true values of the sensitive variables. From Eqs. (13.16) and (13.17), we generate the second observed response from the ith person in the ranked set sample as

$$X_{[2i]} = \begin{bmatrix} S_1 X_{[1i]} & \text{with probability} & P \\ S_2 X_{[1i]} & \text{with probability} & (1-P) \end{bmatrix} \tag{13.18}$$

Taking the expected value on both sides of Eq. (13.16), we have

$$E\left[X_{[1i]}\right] = E\left[S_1 Y_{[1i]} + S_2 Y_{(2i)}\right] = \theta_1 \mu_{Y_1} + \theta_2 \mu_{Y_2} \tag{13.19}$$

Taking the expected value on both sides of Eq. (13.18), we have

$$\begin{aligned} E\left[X_{[2i]}\right] &= P\left[\mu_{Y_1} E(S_1^2) + E(S_1)E(S_2)\mu_{Y_2}\right] + (1-P)\left[\mu_{Y_1} E(S_1)E(S_2) + E(S_2^2)\mu_{Y_2}\right] \\ &= P\left[\mu_{Y_1}(\gamma_{20} + \theta_1^2) + \theta_1\theta_2\mu_{Y_2}\right] + (1-P)\left[\mu_{Y_1}\theta_1\theta_2 + (\gamma_{02} + \theta_2^2)\mu_{Y_2}\right] \end{aligned} \tag{13.20}$$

On solving Eqs. (13.19) and (13.20) for μ_{Y_1} and μ_{Y_2}, and by the method of moments, we have the following theorems:

Theorem 3.1: *Unbiased estimators of μ_{Y_1} and μ_{Y_2} using ranked set sampling are, respectively, given by:*

$$\hat{\mu}_{Y_{[1]}} = \frac{\{P\theta_1\theta_2 + (1-P)(\gamma_{02} + \theta_2^2))\}\overline{X}_{[1]} - \theta_2\overline{X}_{[2]}}{(1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20}} \tag{13.21}$$

and

$$\hat{\mu}_{Y_{[2]}} = \frac{\theta_1\overline{X}_{[2]} - \{P(\gamma_{20} + \theta_1^2) + (1-P)\theta_1\theta_2\}\overline{X}_{[1]}}{(1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20}} \tag{13.22}$$

where
$\overline{X}_{[1]} = \frac{1}{n}\sum_{i=1}^{n} X_{[1i]}$ *and* $\overline{X}_{[2]} = \frac{1}{n}\sum_{i=1}^{n} X_{[2i]}$ *are the means of the observed responses using ranked set sampling.*

Proof: Taking the expected value on both sides of $\hat{\mu}_{Y_{[1]}}$ we have

$$\begin{aligned}
E\left[\hat{\mu}_{Y_{[1]}}\right] &= E\left[\frac{\{P\theta_1\theta_2 + (1-P)(\gamma_{02} + \theta_2^2))\}\overline{X}_{[1]} - \theta_2\overline{X}_{[2]}}{(1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20}}\right] \\
&= \frac{\{P\theta_1\theta_2 + (1-P)(\gamma_{02} + \theta_2^2))\}E\left[\overline{X}_{[1]}\right] - \theta_2 E\left[\overline{X}_{[2]}\right]}{(1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20}} \\
&= \mu_{Y_1}
\end{aligned}$$

In the same way, taking the expected value on both sides of $\hat{\mu}_{Y_{[2]}}$ we have

$$\begin{aligned}
E\left[\hat{\mu}_{Y_{[2]}}\right] &= E\left[\frac{\theta_1\overline{X}_{[2]} - \{P(\gamma_{20} + \theta_1^2) + (1-P)\theta_1\theta_2\}\overline{X}_{[1]}}{(1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20}}\right] \\
&= \frac{\theta_1 E\left[\overline{X}_{[2]}\right] - \{P(\gamma_{20} + \theta_1^2) + (1-P)\theta_1\theta_2\}E\left[\overline{X}_{[1]}\right]}{(1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20}} \\
&= \mu_{Y_2}
\end{aligned}$$

which proves the theorem.

Theorem 3.2: *The variances of the unbiased estimators of $\hat{\mu}_{Y_{[1]}}$ and $\hat{\mu}_{Y_{[2]}}$ using ranked set sampling are, respectively, given by*

$$V\left[\hat{\mu}_{Y_{[1]}}\right] = \frac{\{P\theta_1\theta_2 + (1-P)(\gamma_{02} + \theta_2^2)\}^2\sigma^2_{\overline{X}_{[1]}} + \theta_2^2\sigma^2_{\overline{X}_{[2]}} - 2\theta_2\{P\theta_1\theta_2 + (1-P)(\gamma_{02} + \theta_2^2)\}\sigma_{\overline{X}_{[1]}\overline{X}_{[2]}}}{[(1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20}]^2} \tag{13.23}$$

and

$$V\left[\hat{\mu}_{Y_{[2]}}\right] = \frac{\theta_1^2\sigma^2_{\overline{X}_{[2]}} + \{P(\gamma_{20} + \theta_1^2) + (1-P)\theta_1\theta_2\}^2\sigma^2_{\overline{X}_{[1]}} - 2\theta_1\{P(\gamma_{20} + \theta_1^2) + (1-P)\theta_1\theta_2\}\sigma_{\overline{X}_{[1]}\overline{X}_{[2]}}}{\left[(1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20}\right]^2} \tag{13.24}$$

where

$$\begin{aligned}\sigma^2_{\overline{X}_{[1]}} = \frac{1}{n}\Bigg[&\sigma^2_{Z_1} - \frac{(\gamma_{20}+\theta_1^2)}{m}\sum_{j=1}^{m}(\mu_{Y_1[j]}-\mu_{Y_1})^2 - \frac{(\gamma_{02}+\theta_2^2)}{m}\sum_{j=1}^{m}(\mu_{Y_2(j)}-\mu_{Y_2})^2\\ &-2\frac{\theta_1\theta_2}{m}\sum_{j=1}^{m}(\mu_{Y_1[j]}-\mu_{Y_1})(\mu_{Y_2(j)}-\mu_{Y_2})\Bigg]\end{aligned} \tag{13.25}$$

$$\begin{aligned}\sigma^2_{\overline{X}_{[2]}} = \frac{1}{n}\Bigg[&\sigma^2_{Z_2} - \frac{1}{m}\left\{P(\gamma_{40}+4\gamma_{30}\theta_1+6\gamma_{20}\theta_1^2+\theta_1^4)+(1-P)(\gamma_{20}+\theta_1^2)(\gamma_{02}+\theta_2^2)\right\}\sum_{j=1}^{m}(\mu_{Y_1[j]}-\mu_{Y_1})^2\\ &-\frac{1}{m}\left\{P(\gamma_{20}+\theta_1^2)(\gamma_{02}+\theta_2^2)+(1-P)(\gamma_{04}+4\gamma_{03}\theta_2+6\gamma_{02}\theta_2^2+\theta_2^4\right\}\sum_{j=1}^{m}(\mu_{Y_2(j)}-\mu_{Y_2})^2\\ &-\frac{2}{m}\left\{P(\gamma_{30}+3\theta_1\gamma_{20}+\theta_1^3)\theta_2+(1-P)(\gamma_{03}+3\theta_2\gamma_{02}+\theta_2^3)\theta_1\right\}\sum_{j=1}^{m}(\mu_{Y_1[j]}-\mu_{Y_1})(\mu_{Y_2(j)}-\mu_{Y_2})\Bigg]\end{aligned} \tag{13.26}$$

and

$$\begin{aligned}\sigma_{\overline{X}_{[1]}\overline{X}_{[2]}} = \frac{1}{n}\Bigg[&\sigma_{Z_1Z_2} - \frac{1}{m}\left\{P(\gamma_{30}+3\theta_1\gamma_{20}+\theta_1^3)+(1-P)(\gamma_{20}+\theta_1^2)\theta_2\right\}\sum_{j=1}^{m}(\mu_{Y_1[j]}-\mu_{Y_1})^2\\ &-\frac{1}{m}\left\{P(\gamma_{02}+\theta_2^2)\theta_1+(1-P)(\gamma_{03}+3\theta_2\gamma_{02}+\theta_2^3)\right\}\sum_{j=1}^{m}(\mu_{Y_2(j)}-\mu_{Y_2})^2\\ &-\frac{2}{m}\left\{P(\gamma_{20}+\theta_1^2)\theta_2+(1-P)(\gamma_{02}+\theta_2^2)\theta_1\right\}\sum_{j=1}^{m}(\mu_{Y_1[j]}-\mu_{Y_1})(\mu_{Y_2(j)}-\mu_{Y_2})\Bigg]\end{aligned} \tag{13.27}$$

Proof: Note that the responses are independent, thus the variance of $\overline{X}_{[1]}$ is given by

$$\sigma^2_{\overline{X}_{[1]}} = V\left[\overline{X}_{[1]}\right] = V\left[\frac{1}{n}\sum_{i=1}^{n}X_{[1i]}\right] = \frac{1}{n^2}\sum_{i=1}^{n}V\left[X_{[1i]}\right] \tag{13.28}$$

Now the variance of $X_{[1i]}$ is given by

$$\begin{aligned}V\left[X_{[1i]}\right] &= E\left[X^2_{[1i]}\right] - \left[E\left(X_{[1i]}\right)\right]^2\\ &= (\gamma_{20}+\theta_1^2)E\left[Y^2_{[1i]}\right] + (\gamma_{02}+\theta_2^2)E\left[Y^2_{(2i)}\right] + 2\theta_1\theta_2\left[E\left(Y_{[1i]}Y_{(2i)}\right)\right] - \left[\theta_1\mu_{Y_1}+\theta_2\mu_{Y_2}\right]^2\\ &= (\gamma_{20}+\theta_1^2)\left[\sigma^2_{Y_1} - \frac{1}{m}\sum_{j=1}^{m}(\mu_{Y_1[j]}-\mu_{Y_1})^2 + \mu^2_{Y_1}\right]\\ &\quad + (\gamma_{02}+\theta_2^2)\left[\sigma^2_{Y_2} - \frac{1}{m}\sum_{j=1}^{m}(\mu_{Y_2(j)}-\mu_{Y_2})^2 + \mu^2_{Y_2}\right]\\ &\quad + 2\theta_1\theta_2\left[\sigma_{Y_1Y_2} - \frac{1}{m}\sum_{j=1}^{m}(\mu_{Y_1[j]}-\mu_{Y_1})(\mu_{Y_2[j]}-\mu_{Y_2}) + \mu_{Y_1}\mu_{Y_2}\right] - \left[\theta_1\mu_{Y_1}+\theta_2\mu_{Y_2}\right]^2\end{aligned} \tag{13.29}$$

Note that the responses are independent, thus the variance of $\overline{X}_{[2]}$ is given by

$$\sigma^2_{\overline{X}_{[2]}} = V\left[\overline{X}_{[2]}\right] = V\left[\frac{1}{n}\sum_{i=1}^{n} X_{[2i]}\right] = \frac{1}{n^2}\sum_{i=1}^{n} V\left[X_{[2i]}\right] \tag{13.30}$$

Now the variance of $X_{[2i]}$ is given by

$$\begin{aligned}
V\left[X_{[2i]}\right] &= E\left[X^2_{[2i]}\right] - \left[E\left(X_{[2i]}\right)\right]^2 \\
&= PE\left[S_1^2 Y_{[1i]} + S_1 S_2 Y_{(2i)}\right]^2 + (1-P)E\left[S_1 S_2 Y_{[1i]} + S_2^2 Y_{(2i)}\right]^2 \\
&\quad - \left[PE\{S_1^2 Y_{[1i]} + S_1 S_2 Y_{(2i)}\} + (1-P)E\{S_1 S_2 Y_{[1i]} + S_2^2 Y_{(2i)}\right]^2 \\
&= PE\left[S_1^4 Y^2_{[1i]} + S_1^2 S_2^2 Y^2_{(2i)} + 2S_1^3 S_2 Y_{[1i]} Y_{[2i]}\right] \\
&\quad + (1-P)E\left[S_1^2 S_2^2 Y^2_{[1i]} + S_2^4 Y^2_{(2i)} + 2S_1 S_2^2 Y_{[1i]} Y_{[2i]}\right] \\
&\quad - \left[PE\{S_1^2 Y_{[1i]} + S_1 S_2 Y_{(2i)}\} + (1-P)E\{S_1 S_2 Y_{[1i]} + S_2^2 Y_{(2i)}\right]^2 \\
&= P\left[\left(\gamma_{40} + 4\gamma_{30}\theta_1 + 6\gamma_{20}\theta_1^2 + \theta_1^4\right)\left\{\sigma^2_{Y_1} - \frac{1}{m}\sum_{j=1}^{m}\left(\mu_{Y_1[j]} - \mu_{Y_1}\right)^2 + \mu^2_{Y_1}\right\}\right. \\
&\quad + \left(\gamma_{20} + \theta_1^2\right)\left(\gamma_{02} + \theta_2^2\right)\left\{\sigma^2_{Y_2} - \frac{1}{m}\sum_{j=1}^{m}\left(\mu_{Y_2(j)} - \mu_{Y_2}\right)^2 + \mu^2_{Y_2}\right\} \\
&\quad \left. + 2\left(\gamma_{30} + 3\theta_1\gamma_{20} + \theta_1^2\right)\theta_2\left\{\sigma_{Y_1Y_2} - \frac{1}{m}\sum_{j=1}^{m}\left(\mu_{Y_1[j]} - \mu_{Y_1}\right)\left(\mu_{Y_2(j)} - \mu_{Y_2}\right) + \mu_{Y_1}\mu_{Y_2}\right\}\right] \\
&\quad + (1-P)\left[\left(\gamma_{20} + \theta_1^2\right)\left(\gamma_{02} + \theta_2^2\right)\left\{\sigma^2_{Y_1} - \frac{1}{m}\sum_{j=1}^{m}\left(\mu_{Y_1[j]} - \mu_{Y_1}\right)^2 + \mu^2_{Y_1}\right\}\right. \\
&\quad + \left(\gamma_{04} + 4\gamma_{03}\theta_2 + 6\gamma_{02}\theta_2^2 + \theta_2^4\right)\left\{\sigma^2_{Y_2} - \frac{1}{m}\sum_{j=1}^{m}\left(\mu_{Y_2(j)} - \mu_{Y_2}\right)^2 + \mu^2_{Y_2}\right\} \\
&\quad \left. + 2\left(\gamma_{03} + 3\theta_2\gamma_{02} + \theta_2^3\right)\theta_1\left\{\sigma_{Y_1Y_2} - \frac{1}{m}\sum_{j=1}^{m}(\mu_{Y_1[j]} - \mu_{Y_1})(\mu_{Y_2(i)} - \mu_{Y_2}) + \mu_{Y_1}\mu_{Y_2}\right\}\right] \\
&\quad - \left[P\left\{\left(\gamma_{20} + \theta_1^2\right)\mu_{Y_1} + \theta_1\theta_2\mu_{Y_2}\right\} + (1-P)\left\{\theta_1\theta_2\mu_{Y_1} + (\gamma_{02} + \theta_2^2)\mu_{Y_2}\right\}\right]^2
\end{aligned} \tag{13.31}$$

Note that the responses are independent, thus the covariance between $\overline{X}_{[1]}$ and $\overline{X}_{[2]}$ is given by

$$\sigma_{\overline{X}_{[1]}\overline{X}_{[2]}} = \mathrm{Cov}\left[\overline{X}_{[1]}, \overline{X}_{[2]}\right] = \mathrm{Cov}\left[\frac{1}{n}\sum_{i=1}^{n} X_{[1i]}, \frac{1}{n}\sum_{i=1}^{n} X_{[2i]}\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Cov}\left[X_{[1i]}, X_{[2i]}\right] \tag{13.32}$$

Now the covariance between $X_{[1i]}$ and $X_{[2i]}$ is given by

$$\begin{aligned}
\mathrm{Cov}\left[\mathrm{X}_{[1\mathrm{i}]}, X_{[2\mathrm{i}]}\right] &= E\left[X_{[1i]}X_{[2i]}\right] - \left[E\left(X_{[1i]}\right)\right]\left[E\left(X_{[2i]}\right)\right] \\
&= PE\left[S_1^3 Y_{[1i]}^2 + 2S_1^2 S_2 Y_{[1i]} Y_{[2i]} + S_1 S_2^2 Y_{(2i)}^2\right] \\
&\quad + (1-P)E\left[S_1^2 S_2 Y_{[1i]}^2 + 2S_1 S_2^2 Y_{[1i]} Y_{[2i]} + S_2^3 Y_{(2i)}^2\right] \\
&\quad - E(S_1 Y_{[1i]} + S_2 Y_{(2i)}) E\{P(Y_{[1i]} S_1^2 + S_1 S_2 Y_{(2i)}) + (1-P)(S_1 S_2 Y_{[1i]} + S_2^2 Y_{(2i)}\} \\
&= P\left[(\gamma_{30} + 3\theta_1\gamma_{20} + \theta_1^3)\left\{\sigma_{Y_1}^2 - \frac{1}{m}\sum_{j=1}^{m}(\mu_{Y_1[j]} - \mu_{Y_1})^2 + \mu_{Y_1}^2\right\}\right. \\
&\quad + 2\left(\gamma_{20} + \theta_1^2\right)\theta_2\left\{\sigma_{Y_1Y_2} - \frac{1}{m}\sum_{j=1}^{m}(\mu_{Y_1[j]} - \mu_{Y_1})(\mu_{Y_2(j)} - \mu_{Y_2}) + \mu_{Y_1}\mu_{Y_2}\right\} \\
&\quad \left. + \theta_1(\gamma_{02} + \theta_2^2)\left\{\sigma_{Y_2}^2 - \frac{1}{m}\sum_{j=1}^{m}(\mu_{Y_2(j)} - \mu_{Y_2})^2 + \mu_{Y_2}^2\right\}\right] \\
&\quad + (1-P)\left[(\gamma_{20} + \theta_1^2)\theta_2\left\{\sigma_{Y_1}^2 - \frac{1}{m}\sum_{j=1}^{m}(\mu_{Y_1[j]} - \mu_{Y_1})^2 + \mu_{Y_1}^2\right\}\right. \\
&\quad + 2\left(\gamma_{02} + \theta_2^2\right)\theta_1\left\{\sigma_{Y_1Y_2} - \frac{1}{m}\sum_{j=1}^{m}(\mu_{Y_1[j]} - \mu_{Y_1})(\mu_{Y_2(j)} - \mu_{Y_2}) + \mu_{Y_1}\mu_{Y_2}\right\} \\
&\quad \left. + (\gamma_{03} + 3\theta_2\gamma_{02} + \theta_2^3)\left\{\sigma_{Y_2}^2 - \frac{1}{m}\sum_{j=1}^{m}(\mu_{Y_2(j)} - \mu_{Y_2})^2 + \mu_{Y_2}^2\right\}\right] \\
&\quad - \left(\theta_1\mu_{Y_1} + \theta_2\mu_{Y_2}\right)\left[P\left\{\mu_{Y_1}\left(\gamma_{20} + \theta_1^2\right) + \theta_1\theta_2\mu_{Y_2}\right\} + (1-P)\left\{\mu_{Y_1}\theta_1\theta_2 + (\gamma_{02} + \theta_2^2)\mu_{Y_2}\right\}\right]
\end{aligned} \tag{13.33}$$

Now the variance of the estimator $\hat{\mu}_{Y[1]}$ is given by

$$\begin{aligned}
V\left[\hat{\mu}_{Y_{[1]}}\right] &= V\left[\frac{\{P\theta_1\theta_2 + (1-P)(\gamma_{02} + \theta_2^2)\}\overline{X}_{[1]} - \theta_2\overline{X}_{[2]}}{(1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20}}\right] \\
&= \frac{\{P\theta_1\theta_2 + (1-P)(\gamma_{02} + \theta_2^2)\}^2 V\{\overline{X}_{[1]}\} + \theta_2^2 V\{\overline{X}_{[2]}\} - 2\theta_2\{P\theta_1\theta_2 + (1-P)(\gamma_{02} + \theta_2^2)\}\mathrm{Cov}(\overline{X}_{[1]}, \overline{X}_{[2]})}{\{(1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20}\}^2}
\end{aligned} \tag{13.34}$$

and the variance of the estimator $\hat{\mu}_{Y[2]}$ is given by

$$V\left[\hat{\mu}_{Y_{[2]}}\right] = V\left[\frac{\theta_1\overline{X}_{[2]} - \{P(\gamma_{20}+\theta_1^2)+(1-P)\theta_1\theta_2\}\overline{X}_{[1]}}{(1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20}}\right]$$

$$= \frac{\theta_1^2 V\{\overline{X}_{[2]}\} + \{P(\gamma_{20}+\theta_1^2)+(1-P)\theta_1\theta_2\}^2 V\{\overline{X}_{[1]}\} - 2\theta_1\{P(\gamma_{20}+\theta_1^2)+(1-P)\theta_1\theta_2\}\mathrm{Cov}(\overline{X}_{[1]},\overline{X}_{[2]})}{\{(1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20}\}^2} \tag{13.35}$$

On using Eqs. (13.25) to (13.33) in Eqs. (13.34) and (13.35), we have the theorem.

In the next section, we consider the comparison of the RSS based estimators with respect to the one with simple random sampling.

13.4 EFFICIENCY OF RANKED SET SAMPLING

It is a well-known fact that the use of RSS leads to more efficient estimators than the use of simple random sampling and with replacement scheme. Also it would be worthwhile investigating the usefulness of RSS when estimating means of the two sensitive variables at the same time. For illustration purposes we considered the population listed in the Appendix of Singh (2003) where we considered the first sensitive variable Y_{1i} as the amount ($000) of nonreal estate farm loans in different states during 1997, and the second sensitive variable is Y_{2i} as the amount ($000) of real estate farm loans in different states during 1997. As mentioned in Singh et al. (2008), a data set could be considered as sensitive in one situation and nonsensitive in another situation. Thus we consider these variables as sensitive variables for the purpose of testing the newly proposed methodology. A graphical representation of such variables associated with each other is shown in Fig. 13.1.

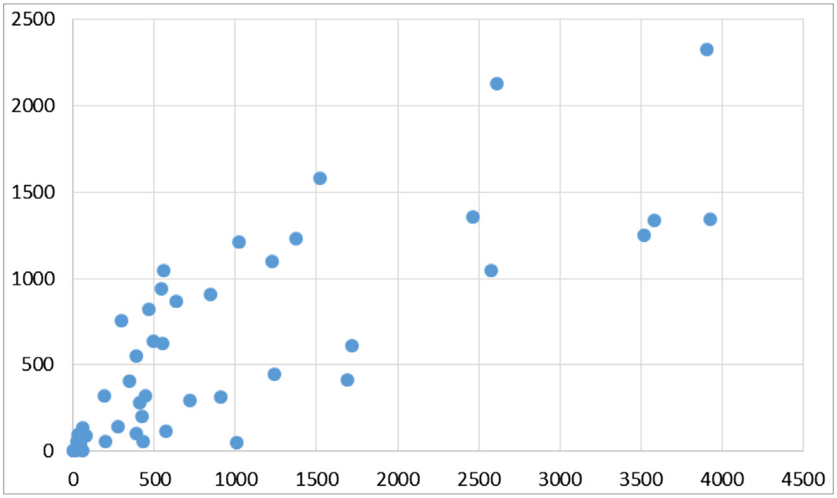


FIGURE 13.1

Scatterplot of the two variables considered as sensitive variables.

A brief description of the parameters of both variables is given below: $N = 50$, $\mu_{Y_1} = 878.16$, $\mu_{Y_2} = 555.43$, $\sigma_{Y_1} = 1084.67$,$\sigma_{Y_2} = 584.82$, $S_{kY_1} = 1.66$, $S_{kY_2} = 1.14$, $K_{urtY_1} = 1.92$, and $K_{urtY_2} = 0.85$.

We wrote the SAS code to investigate the percent relative efficiency values (see Appendix A). The percent relative efficiency of the RSS over the simple random sampling is defined as:

$$RE(1) = \frac{V(\hat{\mu}_{Y_1})}{V(\hat{\mu}_{[1]})} \times 100\% \tag{13.36}$$

and

$$RE(2) = \frac{V(\hat{\mu}_{Y_2})}{V\left(\hat{\mu}_{[2]}\right)} \times 100\% \tag{13.37}$$

Following Singh et al. (2014), we also defined realized ratios of the judgment-based ranked values to that of true mean values for the first and second variables as:

$$RD_1[i] = \frac{\mu_{Y_1[i]}}{\mu_{Y_1}} \text{ and } RD_2[i] = \frac{\mu_{Y_{2(i)}}}{\mu_{Y_2}}$$

for $i = 1, 2, 3, .., m$ in each cycle. In this simulation study we considered several values of

$$RD_1[i] = A_1 + 0.08e_i \tag{13.38}$$

and

$$RD_2[i] = A_2 + 0.08e_i \tag{13.39}$$

where $e_i \sim N(0,\ 1)$. Then different values of

$$A_1 = \{0.75,\ 1.0,\ 1.25\} \text{ and}$$

$$A_2 = \{0.25,\ 0.50,\ 0.75,\ 1.00,\ 1.25,\ 1.50,\ 1.75,\ 2.0,\ 2.25,\ 2.50\}$$

are investigated through a simulation study. The choice of A_1 is made that the judgment ranking could be 75% of the original true value, could be perfect ranking, or could be 25% higher judgment ranking. More that this variation in judgment ranking is not considered, because then judgment ranking will introduce a lot of measurement errors in the first sensitive variable Y_{1i}. The value of A_2 is given a wider range from 0.25 to 2.50 with a step of 0.25 because it is not in the hands of the investigator to control the value of the second sensitive variable Y_{2i}. Recall that judgment ranking was made only for the first sensitive variable. We used two scrambling variables S_1 and S_2. The scrambling variable S_1 consists of 5000 random numbers generated from the chi-squared distribution with five degrees of freedom. The second scrambled variable S_2 is generated from the gamma distribution with shape parameter $\alpha = 0.5$ and scale parameter $\beta = 1.5$. Then, all the required first-, second-, third-, and fourth-ordered moments for both the scrambling variables were calculated from those 5000 random numbers. The percent relative efficiency values $RE(1)$ and $RE(2)$ for different choices of A_1 and A_2 are given in Table 13.2.

It has been observed that the choice of the scrambling variables leads to a sampling scheme such that RSS is more efficient than simple random sampling with replacement when estimating the two sensitive variables simultaneously, as noted in Ahmed et al. (2018). For $A_1 = 0.75$, while the value of A_2 varies between 0.25 and 2.50, both with standard deviations of 0.08, the value of $RE(1)$ varies from 101.00% to 102.30% and the value of $RE(2)$ varies from 100.62% to 102.50%.

Table 13.2 Percent Relative Efficiency Values

Obs	A_1	A_2	RE_1	RE_2
1	0.75	0.25	102.300	102.503
2	0.75	0.50	101.301	101.413
3	0.75	0.75	101.000	101.092
4	0.75	1.00	101.943	102.133
5	0.75	1.25	101.032	101.137
6	0.75	1.50	100.647	100.719
7	0.75	1.75	101.536	101.709
8	0.75	2.00	101.134	101.273
9	0.75	2.25	100.540	100.617
10	0.75	2.50	101.066	101.211
11	1.00	0.25	100.108	100.119
12	1.00	0.50	100.055	100.060
13	1.00	0.75	100.072	100.079
14	1.00	1.00	100.040	100.044
15	1.00	1.25	100.112	100.122
16	1.00	1.50	100.123	100.136
17	1.00	1.75	100.130	100.142
18	1.00	2.00	100.177	100.196
19	1.00	2.25	100.249	100.279
20	1.00	2.50	100.216	100.241
21	1.25	0.25	101.056	101.178
22	1.25	0.50	101.048	101.160
23	1.25	0.75	100.882	100.973
24	1.25	1.00	101.056	101.158
25	1.25	1.25	101.402	101.532
26	1.25	1.50	101.314	101.431
27	1.25	1.75	101.036	101.123
28	1.25	2.00	102.562	102.780
29	1.25	2.25	101.705	101.838
30	1.25	2.50	102.828	103.055

If one has very perfect judgment ranking $A_1 = 1.0$ with a standard deviation of 0.08, and the value of A_2 changes from 0.25 to 2.50 with a step of 0.25, the value of $RE(1)$ changes from 100.04% to 100.25%, and the value of $RE(2)$ changes from 100.04% and 100.28%. In the same way, for the value of $A_1 = 1.25$ as the value of A_2 changes from 0.25 to 2.50, the value of $RE(1)$ changes from 100.88% to 102.83%, and that of $RE(2)$ changes from 100.97% to 103.05%.

It seems that there is potentially a much wider scope of application of this study to other randomized response models, such as that due to Arcos et al. (2015), by making use of RSS along the lines of Bouza (2009), where he investigates the Chaudhuri and Stenger (1992) randomized response model. Also, further note that the other situations when the ranking can be made based on the second sensitive variable, and/or both variables, could also be of worth in future studies.

Nevertheless pending investigations in future studies by following Bouza (2016) for other complex designs are duly acknowledgeable and it seems that there is potentially a much wider scope of application of this study, but that is beyond the scope of this chapter.

ACKNOWLEDGMENTS

The authors are thankful to Prof. Dr. Carlos N. Bouza-Herrera, and two reviewers for their valuable comments/suggestions on the original version of the chapter. Regular help from the Editorial Project Manager Susan Ikeda is also duly acknowledged. This chapter is a part of the first author's project work for the completion of an MS in Statistical Analytics, Computing, and Modelling.

REFERENCES

Ahmed, S., Sedory, S.A., Singh, S., 2018. Simultaneous estimation of means of two sensitive quantitative variables. Commun. Stat.: Theory Methods 47 (2), 324–343.

Arcos, A., Rueda, M., Singh, S., 2015. A generalized approach to randomized response for quantitative variables. Qual. Quant.: Int. J. Methodol. 49 (3), 1239–1256.

Bouza, C.N., 2009. Ranked set sampling and randomized response procedure for estimating the mean of a sensitive quantitative character. Metrika 70, 267–277.

Bouza, C.N., 2016. Behavior of some scrambled randomized response models under simple random sampling, ranked set sampling and Rao-Hartley-Cochran designs. In: Chaudhuri, A., Christofides, T.C., Rao, C.R. (Eds.), Handbook of Statistics 34, Data Gathering, Analysis and Protection of Privacy Through Randomized Response Techniques: Qualitative and Quantitative Human Traits. Elsevier, B.V., pp. 209–220.

Chaudhuri, A., Stenger, H., 1992. Sampling Survey. Marcel Dekker, New York.

Eichhorn, B.H., Hayre, L.S., 1983. Scrambled randomized response methods for obtaining sensitive quantitative data. J. Statist. Planning Infer. 7, 307–316.

Fox, J.A., 2016. Randomized Response and Related Methods. SAGE, Los Angeles978-1-4833-8103-9, ISBN.

Fox, J.A., Tracy, P.E., 1986. Randomized Response: A Method for Sensitive Surveys. SAGE Publications, California.

Greenberg, B.G., Kuebler, R.R., Abernathy, J.R., Horvitz, D.G., 1971. Application of the randomized response technique in obtaining quantitative data. J. Amer. Stat. Assoc. 66, 243–250.

Himmelfarb, S., Edgell, S.E., 1980. Additive constant model: a randomized response technique for eliminating evasiveness to quantitative response questions. Psychological Bulletin 87, 525–530.

Horvitz, D.G., Shah, B.V., Simmons, W.R., 1967. The unrelated question randomized response model. Proc. Social Stat. Sect., Am. Stat. Assoc. 65–72.

McIntyre, G.A., 1952. A method of unbiased selective sampling using ranked sets. Aust. J. Agric. Res. 3, 385–390.

Santiago, A., Bouza, C.N., Sautto, M., Al-Omari, A.I., 2016. Randomized response procedure for the estimation of the population ratio using ranked set sampling. J. Math. Stat. 12 (2), 107–114.

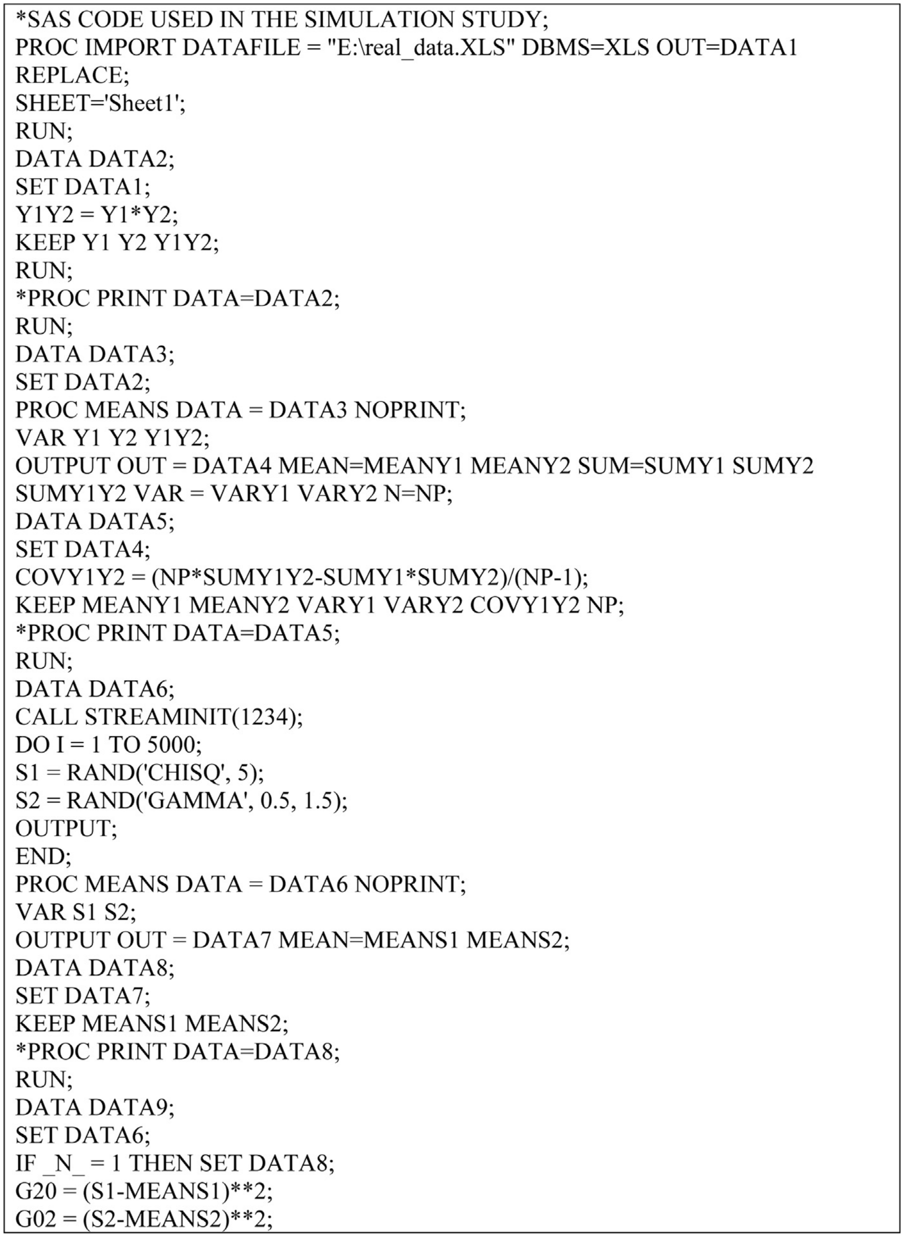
Singh, H.P., Tailor, R., Singh, S., 2014. General procedure for estimating the population mean using ranked set sampling. J. Stat. Comput. Simul. 84 (5), 931–945.

Singh, S., 2003. Advanced Sampling Theory with Applications: How Michael "Selected" Amy. Kluwer Academic Publishers, The Netherlands.

Singh, S., Kim, J.-M., Grewal, I.S., 2008. Imputing and jackknifing scrambled responses. Metron LXVI (2), 183–204.

Warner, S.L., 1965. Randomized response: a survey technique for eliminating evasive answer bias. J. Am. Stat. Assoc. 60, 63–69.

APPENDIX A



```
*SAS CODE USED IN THE SIMULATION STUDY;
PROC IMPORT DATAFILE = "E:\real_data.XLS" DBMS=XLS OUT=DATA1
REPLACE;
SHEET='Sheet1';
RUN;
DATA DATA2;
SET DATA1;
Y1Y2 = Y1*Y2;
KEEP Y1 Y2 Y1Y2;
RUN;
*PROC PRINT DATA=DATA2;
RUN;
DATA DATA3;
SET DATA2;
PROC MEANS DATA = DATA3 NOPRINT;
VAR Y1 Y2 Y1Y2;
OUTPUT OUT = DATA4 MEAN=MEANY1 MEANY2 SUM=SUMY1 SUMY2
SUMY1Y2 VAR = VARY1 VARY2 N=NP;
DATA DATA5;
SET DATA4;
COVY1Y2 = (NP*SUMY1Y2-SUMY1*SUMY2)/(NP-1);
KEEP MEANY1 MEANY2 VARY1 VARY2 COVY1Y2 NP;
*PROC PRINT DATA=DATA5;
RUN;
DATA DATA6;
CALL STREAMINIT(1234);
DO I = 1 TO 5000;
S1 = RAND('CHISQ', 5);
S2 = RAND('GAMMA', 0.5, 1.5);
OUTPUT;
END;
PROC MEANS DATA = DATA6 NOPRINT;
VAR S1 S2;
OUTPUT OUT = DATA7 MEAN=MEANS1 MEANS2;
DATA DATA8;
SET DATA7;
KEEP MEANS1 MEANS2;
*PROC PRINT DATA=DATA8;
RUN;
DATA DATA9;
SET DATA6;
IF _N_ = 1 THEN SET DATA8;
G20 = (S1-MEANS1)**2;
G02 = (S2-MEANS2)**2;
```

```
G30 = (S1-MEANS1)**3;
G03 = (S2-MEANS2)**3;
G40 = (S1-MEANS1)**4;
G04 = (S2-MEANS2)**4;
PROC MEANS DATA = DATA9 NOPRINT;
VAR S1 S2 G20 G02 G30 G03 G40 G04;
OUTPUT OUT = DATA11 SUM=SUMS1 SUMS2 SUMG20 SUMG02 SUMG30
SUMG03 SUMG40 SUMG04 N=NITR;
DATA DATA12;
SET DATA11;
TH1 = SUMS1/NITR;
TH2 = SUMS2/NITR;
GAM20 = SUMG20/(NITR-1);
GAM02 = SUMG02/(NITR-1);
GAM03 = SUMG03/(NITR-1);
GAM30 = SUMG30/(NITR-1);
GAM04 = SUMG04/(NITR-1);
GAM40 = SUMG40/(NITR-1);
KEEP TH1 TH2 GAM20 GAM02 GAM30 GAM03 GAM04 GAM40;
*PROC PRINT DATA=DATA12;
RUN;
DATA DATA13;
SET DATA5;
IF _N_ = 1 THEN SET DATA12;
PROC PRINT DATA=DATA13;
RUN;
%MACRO KUMAR(III, A1, A2);
DATA DATA14;
DO I = 1 TO 5;
RD1 = &A1 + 0.08*RAND('NORMAL');
RD2 = &A2 + 0.08*RAND('NORMAL');
*RD1 = 1.25 + 0.08*RAND('NORMAL');
*RD2 = 1.00 + 0.08*RAND('NORMAL');
OUTPUT;
END;
*PROC PRINT DATA=DATA14;
RUN;
DATA DATA15;
SET DATA14;
RD1_SQ = (RD1-1)**2;
RD2_SQ = (RD2-1)**2;
RD1RD2 = (RD1-1)*(RD2-1);
PROC MEANS DATA = DATA15 NOPRINT;
VAR RD1_SQ RD2_SQ RD1RD2;
OUTPUT OUT = DATA16 MEAN = MRD1_SQ MRD2_SQ MRD1RD2;
```

```
*PROC PRINT DATA=DATA16;
RUN;
DATA DATA17;
SET DATA13;
IF _N_ = 1 THEN SET DATA16;
P = 0.7;
VARZ1 = GAM20*(VARY1+MEANY1**2)+ GAM02*(VARY2+MEANY2**2)+
TH1**2*VARY1+TH2**2*VARY2+2*TH1*TH2*COVY1Y2;
VARX1 = VARZ1 - (GAM20+TH1**2)*MRD1_SQ*MEANY1**2-
(GAM02+TH2**2)* MRD2_SQ*MEANY2**2-2*TH1*TH2* MRD1RD2*
MEANY1*MEANY2;
VARZ2 = (VARY1+MEANY1**2)* (P*(GAM40+4*GAM30*TH1+
6*GAM20*TH1**2+TH1**4)+(1-P)*(GAM20+TH1**2)*(GAM02+TH2**2))
+ (VARY2+MEANY2**2)*((1-P)*(GAM04+4*GAM03*TH2 +6*GAM02*
TH2**2+TH2**4)+P*(GAM20+TH1**2)*(GAM02+TH2**2))
+ 2*(COVY1Y2+MEANY1*MEANY2)*(P*TH2*(GAM30+3*TH1*GAM20
+TH1**3)+(1-P)*TH1*(GAM03+3*TH2*GAM02+TH2**3))
-(MEANY1*(P*(GAM20+TH1**2)+(1-P)*TH1*TH2)+ MEANY2*(P*TH1*TH2+(1-
P)*(GAM02+TH2**2)))**2;
VARX2 = VARZ2 - (P*(GAM40+4*GAM30*TH1+6*GAM20*TH1**2+TH1**4)+(1-
P)*(GAM20+TH1**2)*(GAM02+TH2**2))*MRD1_SQ*MEANY1**2-(P*
(GAM20+TH1**2)*(GAM02+TH2**2) + (1-P)* (GAM04+ 4*GAM03*TH2+
6*GAM02*TH2**2+TH2**4))* MRD2_SQ*MEANY2**2                    -
2*(P*(GAM30+3*TH1*GAM20+TH1**3)*TH2+(1-P)* (GAM03+3*TH2*GAM02
+TH2**3)*TH1) *MRD1RD2*MEANY1*MEANY2;
COVZ1Z2 = (VARY1+MEANY1**2)*(P*(GAM30+3*TH1*GAM20+TH1**3)+(1-
P)*TH2*(GAM20+TH1**2)) +(VARY2+MEANY2**2)*
(P*TH1*(GAM02+TH2**2)+(1-P)*(GAM03+3*TH2*GAM02+TH2**3))
+2*(COVY1Y2+MEANY1*MEANY2)*(P*TH2*(GAM20+TH1**2)+(1-
P)*TH1*(GAM02+TH2**2)) -(TH1*MEANY1 +TH2*MEANY2)* ( MEANY1*
(P*(GAM20+TH1**2)+(1-P)*TH1*TH2) + MEANY2*  (P*TH1*TH2+(1-
P)*(GAM02+TH2**2)));
COVX1X2 = COVZ1Z2- (P*(GAM30+3*TH1*GAM20+TH1**3)+(1-P)*(GAM20+
+TH1**2)*TH2)*MRD1_SQ*MEANY1**2 -(P*(GAM02+ TH2**2)*TH1 +(1-
P)*(GAM03+3* TH2*GAM02+TH2**3))*MRD2_SQ*MEANY2**2
 -2*(P*(GAM20+TH1**2)*TH2+(1-P)* (GAM02+TH2**2) *TH1)* MRD1RD2*
MEANY1*MEANY2;
VMUY1_SRS =( ( P*TH1*TH2+(1-P)*(GAM02+TH2**2) )**2*VARZ1 +
TH2**2*VARZ2-2*TH2*(P*TH1*TH2+(1-P)*(GAM02+TH2**2))*COVZ1Z2)
/((1-P)*TH1*GAM02-P*TH2*GAM20)**2;
VMUY2_SRS = ((TH1**2*VARZ2+(P*(GAM20+TH1**2)+(1-P)*TH1*TH2)**2
*VARZ1-2*TH1*(P*(GAM20+TH1**2)+(1-P)*TH1*TH2)*COVZ1Z2))
 /((1-P)*TH1*GAM02-P*TH2*GAM20)**2;
VMUY1_RSS =( ( P*TH1*TH2+(1-P)*(GAM02+TH2**2) )**2*VARX1 +
TH2**2*VARX2-2*TH2*(P*TH1*TH2+(1-P)*(GAM02+TH2**2))*COVX1X2)
```

```
/((1-P)*TH1*GAM02-P*TH2*GAM20)**2;
VMUY2_RSS = ((TH1**2*VARX2+(P*(GAM20+TH1**2)+(1-P)*TH1*TH2)**2*
VARX1-2*TH1*(P*(GAM20+TH1**2)+(1-P)*TH1*TH2)*COVX1X2))
/((1-P)*TH1*GAM02-P*TH2*GAM20)**2;
KEEP P VMUY1_SRS VMUY2_SRS VMUY1_RSS VMUY2_RSS;
DATA DATA18&III;
SET DATA17;
RE1 = VMUY1_SRS*100/VMUY1_RSS;
RE2 = VMUY2_SRS*100/VMUY2_RSS;
A1RD1=&A1;
A2RD2=&A2;
KEEP A1RD1 A2RD2 RE1 RE2;
PROC PRINT DATA=DATA18&III;
RUN;
%MEND KUMAR;
%KUMAR(1,  0.75,0.25)
%KUMAR(2,  0.75,0.50)
%KUMAR(3,  0.75,0.75)
%KUMAR(4,  0.75,1.00)
%KUMAR(5,  0.75,1.25)
%KUMAR(6,  0.75, 1.50)
%KUMAR(7,  0.75,1.75)
%KUMAR(8,  0.75,2.00)
%KUMAR(9,  0.75,2.25)
%KUMAR(10, 0.75,2.50)
%KUMAR(11, 1.00,0.25)
%KUMAR(12, 1.00,0.50)
%KUMAR(13, 1.00,0.75)
%KUMAR(14, 1.00,1.00)
%KUMAR(15, 1.00,1.25)
%KUMAR(16, 1.00, 1.50)
%KUMAR(17, 1.00,1.75)
%KUMAR(18, 1.00,2.00)
%KUMAR(19, 1.00,2.25)
%KUMAR(20, 1.00,2.50)
%KUMAR(21, 1.25,0.25)
%KUMAR(22, 1.25,0.50)
%KUMAR(23, 1.25,0.75)
%KUMAR(24, 1.25,1.00)
%KUMAR(25, 1.25,1.25)
%KUMAR(26, 1.25, 1.50)
%KUMAR(27, 1.25,1.75)
%KUMAR(28, 1.25,2.00)
%KUMAR(29, 1.25,2.25)
%KUMAR(30, 1.25,2.50)
```

```
RUN;
DATA DATA19;
SET DATA181 DATA182 DATA183 DATA184 DATA185 DATA186 DATA187
DATA188 DATA189 DATA1810 DATA1811  DATA1812 DATA1813  DATA1814
DATA1815 DATA1816 DATA1817 DATA1818 DATA1819 DATA1820 DATA1821
DATA1822 DATA1823 DATA1824 DATA1825 DATA1826 DATA1827 DATA1828
DATA1829 DATA1830
;
PROC PRINT DATA = DATA19;
VAR A1RD1 A2RD2 RE1 RE2;
RUN;
DATA DATA20;
SET DATA2;
PROC MEANS DATA=DATA20;
VAR Y1 Y2;
OUTPUT OUT=DATA21 MEAN = MEANY1 MEANY2 STD=SDY1 SDY2
SKEW=SKY1 SKY2 KURT=KURTY1 KURTY2;
RUN;
PROC PRINT DATA=DATA21;
RUN;
```

CHAPTER

FORCED QUANTITATIVE RANDOMIZED RESPONSE MODEL USING RANKED SET SAMPLING 14

Vaishnavi Bollaboina, Stephen A. Sedory and Sarjinder Singh

Department of Mathematics Texas A&M University-Kingsville, Kingsville, TX, United States

14.1 INTRODUCTION

Warner (1965) was the first to estimate the proportion of the prevalence of a sensitive attribute with the use of a randomization device. Warner considered only the situation when the attribute of interest has only two possible outcomes, one with stigma and another without. His pioneer method was capable of estimating the proportion of persons in a population who bear a stigmatizing characteristic without disclosing the privacy of the respondents while being interviewed. However, the problem of estimating the population mean of a sensitive quantitative variable, such as income, number of induced abortions, and amount of illegal use of drug is also well-known. Horvitz et al. (1967) and Greenberg et al. (1971) were the first to extend the Warner (1965) pioneer model for qualitative variables to the situation of quantitative variables. Himmelfarb and Edgell (1980) introduced the idea of an additive scrambled randomized response model, which they used to estimate the population mean of a sensitive variable by making use of the known distribution of a scrambling variable. Eichhorn and Hayre (1983) came up with the idea of a multiplicative randomized response model which could also be used to estimate the population mean of a sensitive variable. Later, Chaudhuri and Stenger (1992) proposed an ingenious idea of combining both the additive and multiplicative model together to estimate the population mean of a sensitive variable. Let us describe their method, which is also adopted by Bouza (2009), while considering the use of ranked set sampling. For the *i*th person selected in the sample, a set of two randomization devices are given, say two boxes: Box − I and Box-II. Box-I contains T cards labeled with numbers $\{A_1, A_2, ..., A_T\}$ and Box-II contains S cards labeled with numbers $\{B_1, B_2, ..., B_S\}$. The mean and variances of the numbers written on the cards in Box-I and Box-II are assumed to be known, and are computed as:

$$\mu_A = \frac{1}{T}\sum_{i=1}^{T} A_i, \quad \sigma_A^2 = \frac{1}{T}\sum_{i=1}^{T}(A_i - \mu_A)^2$$

$$\mu_B = \frac{1}{S}\sum_{i=1}^{S} B_i, \quad \sigma_B^2 = \frac{1}{S}\sum_{i=1}^{S}(B_i - \mu_B)^2$$

Ranked Set Sampling. DOI: https://doi.org/10.1016/B978-0-12-815044-3.00014-9

Assume Y_i is the value of the study variable for the ith unit in the population consisting of N units, say persons. Then the ultimate goal is to estimate the population mean of the sensitive quantitative variable Y_i given by

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i \tag{14.1}$$

Chaudhuri and Stenger (1992) considered the selection of the n persons by using simple random and with replacement sampling (SRSWR). The ith selected person in the sample is requested to draw a card, say A_i from Box-I and another card, say B_i, from Box-II, and report the scrambled response as:

$$Z_i = A_i Y_i + B_i \tag{14.2}$$

Chaudhuri and Stenger (1992) proposed an unbiased estimator of the population mean $\overline{Y}$, based on n observed responses, as:

$$\overline{y}_{CS} = \frac{\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu_B}{\mu_A} \tag{14.3}$$

where $\mu_A \neq 0$, with variance

$$V(\overline{y}_{CS}) = \frac{\sigma_y^2}{n} + \frac{(\sigma_y^2 + \overline{Y}^2)}{n} C_A^2 + \frac{\mu_B^2 C_B^2}{n\mu_A^2} \tag{14.4}$$

where $C_A = \frac{\sigma_A}{\mu_A}$ and $C_B = \frac{\sigma_B}{\mu_B}$ are the known values of the coefficient of variations of the numbers in Box-I and Box-II, respectively.

McIntyre (1952) felt that it could be possible to rank a sample of a few trees taken from an orchard by eye inspection or judgment ranking. The information used to rank trees before taking them in a sample could be useful in the estimation process which became popularly known as ranked set sampling (RSS). Likewise, Bouza (2009) felt that respondents selected in a simple random sample can be ranked based on the value of the sensitive variable. Bouza (2009) introduced an ingenious idea assuming the sensitive variable can be ranked based on some kind of judgment before collecting information from the respondents. Bouza (2009) considered the use of ranked set sampling (RSS) which involves first selecting m independent SRSWR samples each of size m. Then from the ith respondent selected in the ranked set sample, a scrambled response is collected, which we denote by

$$Z_{(i)} = A_i Y_{(i)} + B_i \tag{14.5}$$

Without use of generality, the process is repeated r times so that the total effective RSS sample size is given by $n = mr$. Bouza (2009) considered the following unbiased estimator of the population mean

$$\overline{y}_{\text{Bouza}} = \frac{\frac{1}{n}\sum_{i=1}^{n} Z_{(i)} - \mu_B}{\mu_A} \tag{14.6}$$

Let E_d and V_d denote the design expectation and design variance, respectively. Also let E_R and V_R denote the randomization expectation and randomization variance, respectively. Then the variance of the estimator $\bar{y}_{\text{Bouza}}$ is given by

$$\begin{aligned} V(\bar{y}_{\text{Bouza}}) &= E_d V_R \left[\frac{\frac{1}{n}\sum_{i=1}^{n} Z_{(i)} - \mu_B}{\mu_A} \right] + V_d E_R \left[\frac{\frac{1}{n}\sum_{i=1}^{n} Z_{(i)} - \mu_B}{\mu_A} \right] \\ &= E_d \left[\frac{\frac{1}{n^2}\sum_{i=1}^{n} \left\{ Y_{(i)}^2 \sigma_A^2 + \sigma_B^2 \right\}}{\mu_A^2} \right] + V_d \left[\frac{1}{n}\sum_{i=1}^{n} Y_{(i)} \right] \\ &= \left[\frac{\frac{1}{n^2}\sum_{i=1}^{n} \left\{ E_d\left\{ Y_{(i)}^2 \right\} \sigma_A^2 + \sigma_B^2 \right\}}{\mu_A^2} \right] + V_d\left[\bar{y}_{\text{RSS}}\right] \end{aligned} \tag{14.7}$$

Now, we have

$$\begin{aligned} V_d\left(\bar{y}_{\text{RSS}}\right) &= \frac{1}{mr}\sigma_y^2 - \frac{1}{m^2 r}\sum_{i=1}^{m}(\mu_{Y(i)} - \bar{Y})^2 \\ &= \frac{1}{n}\left[\sigma_Y^2 - \frac{1}{m}\sum_{i=1}^{m}(\mu_{Y(i)} - \bar{Y})^2 \right] \end{aligned} \tag{14.8}$$

and we also have

$$\begin{aligned} E_d\left[Y_{(i)}^2\right] &= V_d\left(Y_{(i)}\right) + \left[E_d\left(Y_{(i)}\right)\right]^2 \\ &= \sigma_Y^2 - \frac{1}{m}\sum_{i=1}^{m}\left[\mu_{Y(i)} - \bar{Y}\right]^2 + \bar{Y}^2 \end{aligned} \tag{14.9}$$

On substituting Eqs. (14.8) and (14.9) into Eq. (14.7), we have

$$\begin{aligned} V(\bar{y}_{\text{Bouza}}) &= \frac{\frac{1}{n^2}\sum_{i=1}^{n}\left\{ E_d\left\{Y_{(i)}^2\right\}\sigma_A^2 + \sigma_B^2 \right\}}{\mu_A^2} + V_d\left[\bar{y}_{\text{RSS}}\right] \\ &= \frac{\sigma_A^2}{n^2\mu_A^2}\sum_{i=1}^{n}\left[\left(\sigma_Y^2 - \frac{1}{m}\sum_{i=1}^{m}(\mu_{Y(i)} - \bar{Y})^2 \right) + \bar{Y}^2 \right] + \frac{\sigma_B^2}{n\mu_A^2} + \frac{1}{n}\left[\sigma_Y^2 - \frac{1}{m}\sum_{i=1}^{m}(\mu_{Y(i)} - \bar{Y})^2 \right] \\ &= \frac{\sigma_A^2}{n\mu_A^2}\left[\sigma_Y^2 - \frac{1}{m}\sum_{i=1}^{m}(\mu_{Y(i)} - \bar{Y})^2 + \bar{Y}^2 \right] + \frac{\sigma_B^2}{n\mu_A^2} + \frac{1}{n}\left\{ \sigma_Y^2 - \frac{1}{m}\sum_{i=1}^{m}(\mu_{Y(i)} - \bar{Y})^2 \right\} \\ &= \frac{1}{n}\left[\sigma_Y^2 - \frac{1}{m}\sum_{i=1}^{m}(\mu_{Y(i)} - \bar{Y})^2 \right]\left(1 + \frac{\sigma_A^2}{\mu_A^2}\right) + \frac{\sigma_A^2}{n\mu_A^2}\bar{Y}^2 + \frac{\sigma_B^2}{n\mu_B^2} \\ &= \frac{1}{n}\left[\sigma_Y^2 - \frac{1}{m}\sum_{i=1}^{m}(\mu_{Y(i)} - \bar{Y})^2 \right]\left(1 + C_A^2\right) + \frac{C_A^2\bar{Y}^2}{n} + \frac{\mu_B^2}{n\mu_A^2}C_B^2 \end{aligned} \tag{14.10}$$

In the next section, we derive an estimator of the population mean with a forced quantitative randomized response (FQRR) model and then find its variance expression. The reason for considering the use of RSS is based on the pioneering idea of Bouza (2009) that the use of RSS sampling is practical and more efficient than SRS. In the present study, we investigate the same idea of RSS in the case of the FQRR model. One can also refer to Al-Omari and Bouza (2014) for a detailed review of ranked set sampling to learn about its simplicity and practicability.

14.2 PROPOSED FORCED QUANTITATIVE RANDOMIZED RESPONSE MODEL

In this section, we combine the thinking of Bouza (2009) and Gjestvang and Singh (2007) as follows. Each respondent selected in the ranked set sample (RSS) is requested to experience a randomization device, say a spinner, with three possible outcomes. Let P_1 be the shaded area of the spinner with "salmon" color, P_2 be the shaded area of the spinner with "thistle" color, and P_3 be the shaded area of the spinner with "firebrick" color. The spinner is rotated by the interviewee unobserved by the interviewer. If the pointer lands in the "salmon" area, the respondent is requested to report the scrambled response by using the two boxes of Chaudhuri and Stenger (1992), if the pointer lands in the "thistle" color, the respondent is requested to report the true response, and if the pointer lands in the "firebrick" color then the respondent is requested to report a fixed response which is already printed on the spinner. The names of the colors are chosen such that it is easy to remember the scrambled response for "salmon," true response for "thistle," and forced response for "firebrick" outcome of the spinner. A graphical representation of such a spinner is given in Fig. 14.1.

It may be worth pointing out here that the forced randomized response model due to Liu and Chow (1976a,b) is applicable only for estimating the population proportion of a sensitive characteristic. In the model considered here, if Y_i is a qualitative variable taking a value of 1 or a value of 0

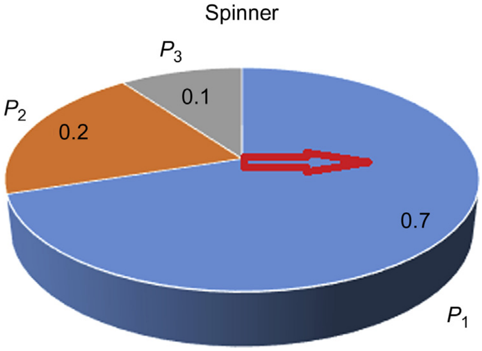


FIGURE 14.1

Spinner for the FQRR model.

for a sensitive and nonsensitive attribute in the population, set $A_i = 0$ and $B_i = 0$ as forced "no" answer, and set $F = 1$ as forced "yes" answer, then the Stem and Steinhorst (1984) model becomes a special case of the proposed model for RSS sampling, which are obviously improvements over the use of SRS sampling as explained in Fox and Tracy (1986). No doubt if $P_1 = 1$ and $P_2 = P_3 = 0$ then the proposed model reduces to the Bouza (2009) model for RSS sampling. If $B_i = 0$, $P_3 = 0$, $P_1 = P$, and $P_2 = (1 - P)$ then the proposed model leads to the Bar-Lev, Bobovitch, and Boukai (2004) model for the RSS scheme. Further, note that in the Bar-Lev, Bobovitch, and Boukai (2004) model, an amendment of change of origin is sometimes needed by adopting the remark (page 315, Section 6, in Eichhorn and Hayre, 1983) while handling special types of sensitive variables, and the present model will be free from such amendments.

Thus from the ranked set sample (RRS), the observed response from the ith respondent is given by

$$Z^*_{(i)} = \begin{cases} A_i Y_{(i)} + B_i & \text{with probability} \quad P_1 \\ Y_{(i)} & \text{with probability} \quad P_2 \\ F & \text{with probability} \quad P_3 \end{cases} \tag{14.11}$$

Note that if the ith person reports a fixed response then that value cannot be based on any ranking.

The expected value of the observed response $Z^*_{(i)}$ in the RSS is given by

$$\begin{aligned} E\left[Z^*_{(i)}\right] &= P_1 E\left[A_i Y_{(i)} + B_i\right] + P_2 E(Y_{(i)}) + P_3 E(F) \\ &= P_1\left[\mu_A \overline{Y} + \mu_B\right] + P_2 \overline{Y} + P_3 F \\ &= \left[P_1 \mu_A + P_2\right]\overline{Y} + P_1 \mu_B + P_3 F \end{aligned} \tag{14.12}$$

Now we propose a new estimator of the population mean $\overline{Y}$ using the proposed FQRR model as:

$$\overline{y}_{\mathrm{RSS}(F)} = \frac{\frac{1}{n}\sum_{i=1}^{n} Z^*_{(i)} - P_1 \mu_B - P_3 F}{P_1 \mu_A + P_2} \tag{14.13}$$

where $\mu_A \neq -P_2/P_1$. Now we have the following theorems:

Theorem 2.1: *The estimator $\overline{y}_{\mathrm{RSS}(F)}$ is an unbiased estimator of the population mean $\overline{Y}$.*

Proof: *Taking the expected value on both sides of $\overline{y}_{\mathrm{RSS}(F)}$, we have*

$$E\left[\overline{y}_{\mathrm{RSS}(F)}\right] = E\left[\frac{\frac{1}{n}\sum_{i=1}^{n} Z^*_{(i)} - P_1 \mu_B - P_3 F}{P_1 \mu_A + P_2}\right] = \frac{\frac{1}{n}\sum_{i=1}^{n} E\{Z^*_{(i)}\} - P_1 \mu_B - P_3 F}{P_1 \mu_A + P_2} = \overline{Y}$$

which proves the theorem.

Theorem 2.2: *The minimum variance of the estimator* $\overline{y}_{\mathrm{RSS}(F)}$ *is given by*

$$
\begin{aligned}
\mathrm{Min}V(\overline{y}_{\mathrm{RSS}(F)}) \quad &= \frac{1}{n(P_1\mu_A+P_2)^2}\Bigg[(\sigma_y^2+\overline{Y}^2)\{P_1(\sigma_A^2+\mu_A^2)+P_2\}+P_1\{\sigma_B^2+\mu_B^2+2\mu_A\mu_B\overline{Y}\}\\
&\quad -\{P_1(\mu_A\overline{Y}+\mu_B)+P_2\overline{Y}\}^2-\frac{\{P_1(\sigma_A^2+\mu_A^2)+P_2\}}{m}\sum_{i=1}^{m}(\mu_{Y(i)}-\overline{Y})^2\\
&\quad -\frac{P_3\{P_2\overline{Y}+P_1(\mu_A\overline{Y}+\mu_B)\}^2}{(1-P_3)}\Bigg]
\end{aligned}
\tag{14.14}
$$

Proof: *The variance of the estimator* $\overline{y}_{\mathrm{RSS}(F)}$ *is given by*

$$
V\left[\overline{y}_{\mathrm{RSS}(F)}\right]=V\left[\frac{\frac{1}{n}\sum_{i=1}^{n}Z_{(i)}^{*}-P_1\mu_B-P_3F}{P_1\mu_A+P_2}\right]=\frac{\frac{1}{n^2}\sum_{i=1}^{n}V(Z_{(i)}^{*})}{(P_1\mu_A+P_2)^2}=\frac{\sigma_{Z*}^2}{n(P_1\mu_A+P_2)^2}
\tag{14.15}
$$

Now the variance σ_{Z*}^2 *is given by*

$$
\begin{aligned}
\sigma_{Z*}^2 \quad &= E\left[Z_{(i)}^{*2}\right]-\left[E\left(Z_{(i)}^{*}\right)\right]^2\\
&= P_1E\left[(A_iY_{(i)}+B_i)^2\right]+P_2E\left(Y_{(i)}^2\right)+P_3E(F^2)-\left[P_1E(A_iY_{(i)}+B_i)+P_2E(Y_{(i)})+P_3E(F)\right]^2\\
&= P_1E\left[A_i^2Y_{(i)}^2+B_i^2+2A_iB_iY_{(i)}\right]+P_2E\left(Y_{(i)}^2\right)+P_3F^2-\left[P_1(\mu_A\overline{Y}+\mu_B)+P_2\overline{Y}+P_3F\right]^2\\
&= P_1\left[(\sigma_A^2+\mu_A^2)\left\{\sigma_y^2-\frac{1}{m}\sum_{i=1}^{m}(\mu_{Y(i)}-\overline{Y})^2+\overline{Y}^2\right\}+(\sigma_B^2+\mu_B^2)+2\mu_A\mu_B\overline{Y}\right]\\
&\quad +P_2\left[\sigma_y^2-\frac{1}{m}\sum_{i=1}^{m}(\mu_{Y(i)}-\overline{Y})^2+\overline{Y}^2\right]+P_3F^2\\
&\quad -\left[P_1^2(\mu_A\overline{Y}+\mu_B)^2+P_2^2\overline{Y}^2+P_3^2F^2+2P_1P_2\overline{Y}(\mu_A\overline{Y}+\mu_B)+2P_2P_3F\overline{Y}+2P_1P_3F(\mu_A\overline{Y}+\mu_B)\right]\\
&= (\sigma_y^2+\overline{Y}^2)\{P_1(\sigma_A^2+\mu_A^2)+P_2\}+P_1(\sigma_B^2+\mu_B^2+2\mu_A\mu_B\overline{Y})\\
&\quad -\{P_1(\mu_A\overline{Y}+\mu_B)+P_2\overline{Y}\}^2+P_3(1-P_3)F^2-2F\{P_2P_3\overline{Y}+P_1P_3(\mu_A\overline{Y}+\mu_B)\}\\
&\quad -\frac{\{P_1(\sigma_A^2+\mu_A^2)+P_2\}}{m}\sum_{i=1}^{m}(\mu_{Y(i)}-\overline{Y})^2
\end{aligned}
\tag{14.16}
$$

On substituting *Eq. (14.16)* *into* *Eq. (14.15)*, *the variance of the estimator* $\overline{y}_{\mathrm{RSS}(F)}$ *is given by*

$$
\begin{aligned}
V(\overline{y}_{\mathrm{RSS}(F)}) \quad &= \frac{1}{n(P_1\mu_A+P_2)^2}\Bigg[(\sigma_y^2+\overline{Y}^2)\{P_1(\sigma_A^2+\mu_A^2)+P_2\}+P_1(\sigma_B^2+\mu_B^2+2\mu_A\mu_B\overline{Y})\\
&\quad -\{P_1(\mu_A\overline{Y}+\mu_B)+P_2\overline{Y}\}^2+P_3(1-P_3)F^2-2F\{P_2P_3\overline{Y}+P_1P_3(\mu_A\overline{Y}+\mu_B)\}\\
&\quad -\frac{\{P_1(\sigma_A^2+\mu_A^2)+P_2\}}{m}\sum_{i=1}^{m}(\mu_{Y(i)}-\overline{Y})^2\Bigg]
\end{aligned}
\tag{14.17}
$$

On differentiating $V(\bar{y}_{\mathrm{RSS}(F)})$ *in Eq. (14.17) with respect to* F *and equating to zero, we have the optimum value of* F given by

$$F = \frac{P_2\bar{Y} + P_1(\mu_A\bar{Y} + \mu_B)}{(1 - P_3)} \tag{14.18}$$

On substituting the optimal value of F *from Eq. (14.18) into Eq. (14.17), we have the theorem.*

14.3 PRACTICABLE ASPECT OF THE PROPOSED FQRR MODEL

Note that the optimum value of F in Eq. (14.18) depends on the unknown parameter, the population mean $\bar{Y}$ which we wish to estimate. Following Singh and Gorey (2017), it is advisable to use an estimate of F, so one can use a spinner with "firebrick" color for the forced response. Note that we are using ranked set sampling, thus either of the following two possible estimators of F could be used in the spinner while collecting information from the interviewees.

$$\hat{F} = \frac{P_2Y_{(i)} + P_1(\mu_A Y_{(i)} + \mu_B)}{(1 - P_3)} \tag{14.19}$$

or

$$\hat{\hat{F}} = \frac{P_2Y_{(i)} + P_1(A_i Y_{(i)} + B_i)}{(1 - P_3)} \tag{14.20}$$

The resulting estimators, after replacing the estimator of F, would be investigated in future studies.

14.4 RELATIVE EFFICIENCY

It is a well-known fact that the use of RSS leads to more efficient estimators than the use of the SRSWR scheme. Also, it would be worth investigating the usefulness of ranked set sampling while using the forced quantitative randomized response model for estimating the mean of a sensitive variable. For illustration purposes we considered the population listed in the appendix of Singh (2003) where we considered the first sensitive variable Y_i as the amount ($000) of nonreal-estate farm loans in various states during 1997. We consider this variable as a sensitive variable for the purpose of testing the new proposed methodology. As reported in Singh et al. (2008), one dataset could be regarded as sensitive in one situation and nonsensitive in another. A boxplot showing the nature of the dataset nonreal-estate farm loan is shown in Fig. 14.2.

From Fig. 14.2, one can see that the distribution of the dataset is skewed to the right, which is typical of the distribution of income as well. Thus it is not unreasonable to consider such a dataset as representing a sensitive variable in real practice. The higher values would be more sensitive than the other data values. We define the percent relative efficiencies of the proposed FQRR model over the Chaudhuri and Stenger (1992) and Bouza (2009) estimators, respectively, as:

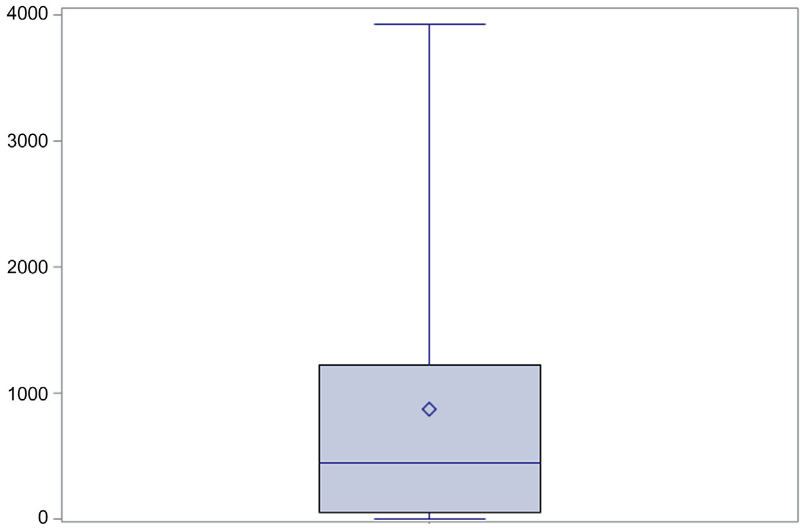


FIGURE 14.2

Box plot showing the distribution of the nonreal-estate farm loan.

$$RE(1) = \frac{V(\bar{y}_{CS})}{\text{Min.}V(\bar{y}_{\text{RSS}(F)})} \times 100\% \tag{14.21}$$

and

$$RE(2) = \frac{V(\bar{y}_{Bouza})}{Min.V(\bar{y}_{RSS(F)})} \times 100\% \tag{14.22}$$

In the investigation, we considered using two boxes, each consisting of 5000 cards. The values of A_i and B_i were generated from a gamma distribution with parameters as shown in the SAS codes (see Appendix A). After executing the SAS code the percent relative efficiency values are given in Table 14.1.

From Table 14.1, we conclude that the Bouza (2009) estimator is more efficient than the Chaudhuri and Stenger (1992) estimator, and the proposed FQRR estimator is more efficient than the Bouza (2009) estimator. Followed Singh et al. (2014), we compute

$$RDY_{(i)} = \left[\frac{\overline{Y}_{(i)}}{\overline{Y}} - 1\right]^2 \tag{14.23}$$

For $P_1 = 0.65$, $P_2 = 0.325$ and the value of C_1 in the range from 0.25 and 1.50, the value of $RE(1)$ varies from 100.21% to 127.82%; and the value of $RE(2)$ varies from 185.79% to 188.95%, with a maximum of 188.95% for $C_1 = 0.25$. In the same way the rest of the results in Table 14.1 can be interpreted. Thus, interestingly, the proposed FQRR model is found to be more beneficial if the values of $\mu_{Y(i)}$ are closer to $\overline{Y}$ than the Bouza (2009) estimator. The efficiency of the proposed estimator provides a statistical reason that if someone has a good guess about the fixed response, it would be beneficial in producing efficient estimates in the case of RSS in relation to its competitors.

Table 14.1 Percent *RE* Values for the Different Choices of Parameters

Obs	P_1	P_2	C_1	$RE(1)$	$RE(2)$
1	0.650	0.325	0.25	127.82	188.95
2	0.650	0.325	0.50	110.83	186.82
3	0.650	0.325	0.75	102.46	185.79
4	0.650	0.325	1.00	100.21	185.52
5	0.650	0.325	1.25	104.25	186.01
6	0.650	0.325	1.50	111.61	186.92
7	0.600	0.300	0.25	134.18	175.16
8	0.600	0.300	0.50	112.97	172.69
9	0.600	0.300	0.75	103.26	171.59
10	0.500	0.250	1.00	100.20	142.71
11	0.500	0.250	1.25	103.63	143.03
12	0.500	0.250	1.50	113.32	143.95
13	0.500	0.250	0.25	136.62	146.21
14	0.500	0.250	0.50	110.89	143.72
15	0.500	0.250	0.75	103.10	142.98
16	0.500	0.250	1.00	100.77	142.76
17	0.500	0.250	1.25	103.14	142.98
18	0.500	0.250	1.50	109.35	143.57

14.5 CONCLUSIONS

In this chapter, we investigate the use of RSS for the case of the forced quantitative randomized response (FQRR) model introduced by Gjestvang and Singh (2007) for estimating the mean of a sensitive quantitative variable. We note that the findings match with the observations of Bouza (2009), in that the use of RSS for a sensitive variable performs better than the use of SRS while also using the proposed FQRR model.

ACKNOWLEDGMENTS

The authors are thankful to Prof. Dr. Carlos N. Bouza-Herrera and two reviewers for their valuable comments/suggestions on the original version of this chapter. Regular help from the Editorial Project Manager Susan Ikeda is also duly acknowledged. This chapter is a part of the first author's project work for the completion of an MS in Statistical Analytics, Computing and Modelling.

REFERENCES

Al-Omari, A.I., Bouza, C.N., 2014. Review of ranked set sampling: modifications and applications. Rev. Investig. Oper. 35 (3), 215–240.

Bar-Lev, S.K., Bobovitch, E., Boukai, B., 2004. A note on randomized response models for quantitative data. Metrika 255–260.

Bouza, C.N., 2009. Ranked set sampling and randomized response procedure for estimating the mean of a sensitive quantitative character. Metrika 70, 267–277.

Chaudhuri, A., Stenger, H., 1992. Sampling Survey.. Marcel Dekker, New York.

Eichhorn, B.H., Hayre, L.S., 1983. Scrambled randomized response methods for obtaining sensitive quantitative data. J. Stat. Plan Inference 7, 307–316.

Fox, J.A., Tracy, P.E., 1986. Randomized Response: A Method for Sensitive Surveys. SAGE Publications, CA.

Gjestvang, C., Singh, S., 2007. Forced quantitative randomized response model: a new device. Metrika 66 (2), 243–257.

Greenberg, B.G., Kuebler, R.R., Abernathy, J.R., Horvitz, D.G., 1971. Application of the randomized response technique in obtaining quantitative data. J. Am. Stat. Assoc. 66, 243–250.

Himmelfarb, S., Edgell, S.E., 1980. Additive constant model: a randomized response technique for eliminating evasiveness to quantitative response questions. Psychol. Bull. 87, 525–530.

Horvitz, D.G., Shah, B.V., Simmons, W.R., 1967. The unrelated question randomized response model. Proc. Soc. Stat. Sect., Am. Stat. Assoc. 65–72.

Liu, P.T., Chow, L.P., 1976a. A new discrete quantitative randomized response model. J. Am. Stat. Assoc. 71, 72–73.

Liu, P.T., Chow, L.P., 1976b. The efficiency of the multiple trial randomized response technique. Biometrics 32, 607–618.

McIntyre, G.A., 1952. A method of unbiased selective sampling using ranked sets. Aust. J. Agric. Res. 3, 385–390.

Singh, H.P., Gorey, S.M., 2017. An alternative to Odumade and Singh's generalized forced quantitative randomized response mode: a unified approach. Model Assist. Stat. Appl. 12 (2), 163–177.

Singh, H.P., Tailor, R., Singh, S., 2014. General procedure for estimating the population mean using ranked set sampling. J. Stat. Comput. Simul. 84 (5), 931–945.

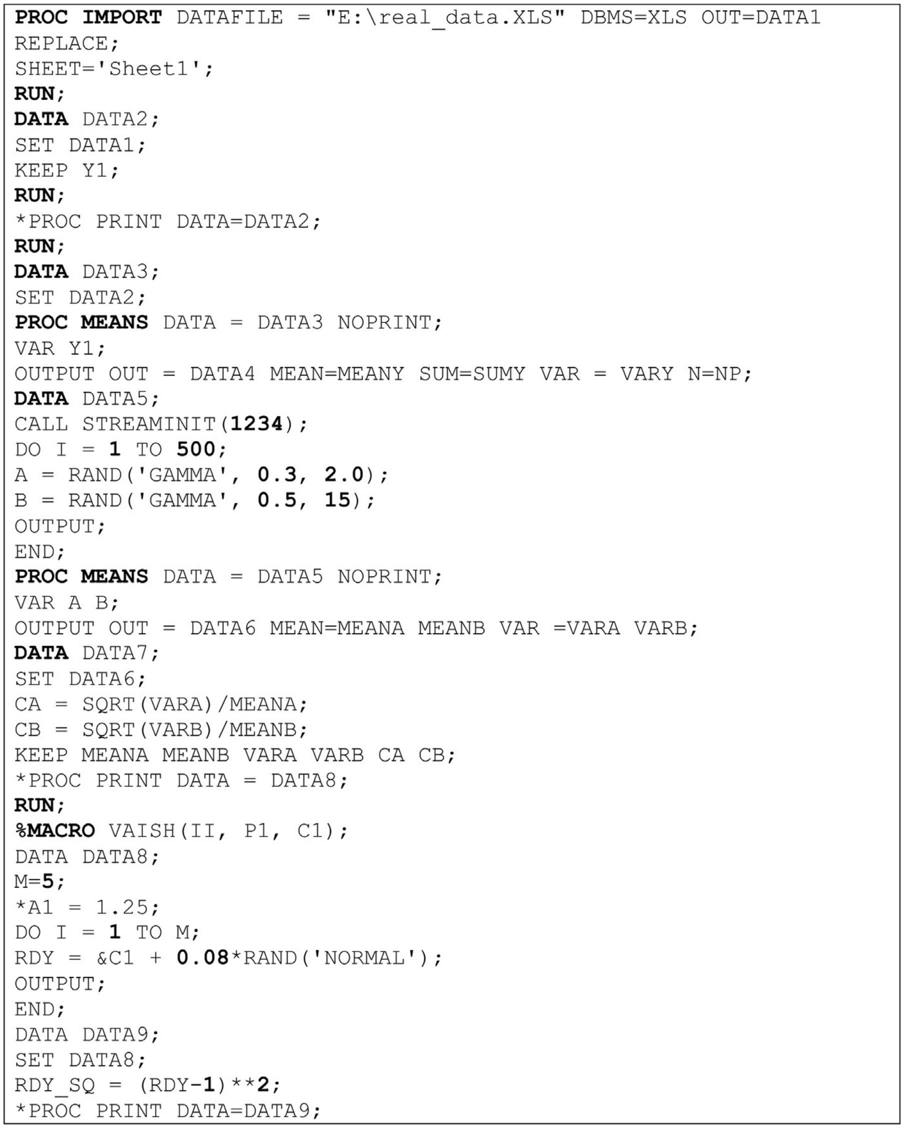
Singh, S., 2003. Advanced Sampling Theory with Applications: How Michael "Selected" Amy. Kluwer Academic Publishers, The Netherlands.

Singh, S., Kim, J.-M., Grewal, I.S., 2008. Imputing and jackknifing scrambled responses. Metron LXVI (2), 183–204.

Stem, D.E., Steinhorst, R.K., 1984. Telephone interview and mail questionnaire applications of the randomized response model. J. Am. Stat. Assoc. 79, 555–564.

Warner, S.L., 1965. Randomized response: a survey technique for eliminating evasive answer bias. J. Am. Stat. Assoc. 60, 63–69.

APPENDIX A



```
PROC IMPORT DATAFILE = "E:\real_data.XLS" DBMS=XLS OUT=DATA1
REPLACE;
SHEET='Sheet1';
RUN;
DATA DATA2;
SET DATA1;
KEEP Y1;
RUN;
*PROC PRINT DATA=DATA2;
RUN;
DATA DATA3;
SET DATA2;
PROC MEANS DATA = DATA3 NOPRINT;
VAR Y1;
OUTPUT OUT = DATA4 MEAN=MEANY SUM=SUMY VAR = VARY N=NP;
DATA DATA5;
CALL STREAMINIT(1234);
DO I = 1 TO 500;
A = RAND('GAMMA', 0.3, 2.0);
B = RAND('GAMMA', 0.5, 15);
OUTPUT;
END;
PROC MEANS DATA = DATA5 NOPRINT;
VAR A B;
OUTPUT OUT = DATA6 MEAN=MEANA MEANB VAR =VARA VARB;
DATA DATA7;
SET DATA6;
CA = SQRT(VARA)/MEANA;
CB = SQRT(VARB)/MEANB;
KEEP MEANA MEANB VARA VARB CA CB;
*PROC PRINT DATA = DATA8;
RUN;
%MACRO VAISH(II, P1, C1);
DATA DATA8;
M=5;
*A1 = 1.25;
DO I = 1 TO M;
RDY = &C1 + 0.08*RAND('NORMAL');
OUTPUT;
END;
DATA DATA9;
SET DATA8;
RDY_SQ = (RDY-1)**2;
*PROC PRINT DATA=DATA9;
```

```
RUN;
PROC MEANS DATA = DATA9 NOPRINT;
VAR RDY_SQ;
OUTPUT OUT = DATA10 MEAN = MRDY_SQ;
DATA DATA11;
SET DATA10;
*PROC PRINT DATA=DATA11;
RUN;
DATA DATA12;
SET DATA4;
IF _N_ = 1 THEN SET DATA7;
DATA DATA13;
SET DATA12;
IF _N_=1 THEN SET DATA11;
*PROC PRINT DATA=DATA13;
DATA DATA14;
SET DATA13;
P1 = &P1;
P2 = P1/2;
P3 = 1-P1-P2;
VARCS = VARY+(VARY+MEANY**2)*CA**2+MEANB**2*CB**2/MEANA**2;
VARBOUZA = (VARY-MEANY**2*MRDY_SQ)*(1+CA**2)
+(MEANY**2)*CA**2+MEANB**2*CB**2/MEANA**2;
RE1 = VARCS*100/VARBOUZA;
VARNEW = (VARY+MEANY**2)*(P1*(VARA+MEANA**2)+P2)
+P1*(VARB+MEANB**2+2*MEANA*MEANB*MEANY)
-(P1*(MEANA*MEANY+MEANB)+P2*MEANY)**2
-(P1*(VARA+MEANA**2)+P2)*MRDY_SQ*MEANY**2
-P3*(P2*MEANY+P1*(MEANA*MEANY+MEANB))**2/(1-P3);
VARNEW1 = VARNEW/(P1*MEANA+P2)**2;
RE2 = VARBOUZA*100/VARNEW1;
C1 = &C1;
KEEP P1 P2 C1 RE1 RE2;
DATA DATA15&II;
SET DATA14;
*PROC PRINT DATA=DATA14;
RUN;
%MEND VAISH;
%VAISH(1, 0.65, 0.25);
%VAISH(2, 0.65, 0.50);
%VAISH(3, 0.65, 0.75);
%VAISH(4, 0.65, 1.00);
%VAISH(5, 0.65, 1.25);
%VAISH(6, 0.65, 1.50);
%VAISH(7, 0.6, 0.25);
%VAISH(8, 0.6, 0.50);
```

```
%VAISH(9, 0.6, 0.75);
%VAISH(10, 0.5, 1.00);
%VAISH(11, 0.5, 1.25);
%VAISH(12, 0.5, 1.50);
%VAISH(13, 0.5, 0.25);
%VAISH(14, 0.5, 0.50);
%VAISH(15, 0.5, 0.75);
%VAISH(16, 0.5, 1.00);
%VAISH(17, 0.5, 1.25);
%VAISH(18, 0.5, 1.50);
DATA DATA16;
SET DATA151 DATA152 DATA153 DATA154 DATA155 DATA156 DATA157
DATA158 DATA159 DATA1510 DATA1511 DATA1512 DATA1513 DATA1514
DATA1515 DATA1516 DATA1517 DATA1518;
RUN;
PROC PRINT DATA=DATA16;
VAR P1 P2 C1 RE1 RE2;
RUN;
DATA DATA17;
SET DATA2;
BY=1;
PROC SORT DATA=DATA17;
BY Y1;
PROC BOXPLOT DATA=DATA17;
PLOT Y1*BY;
RUN;
```

CHAPTER

CONSTRUCTION OF STRATA BOUNDARIES FOR RANKED SET SAMPLING 15

Ruiqiang Zong, Stephen A. Sedory and Sarjinder Singh

Department of Mathematics, Texas A&M University-Kingsville, Kingsville, TX, United States

15.1 INTRODUCTION

The problem of estimating the population mean of a study variable, y, has been widely addressed in the field of survey sampling. There are numerous sampling schemes, such as simple random sampling, probability proportional to size, two-phase sampling, two-stage sampling, systematic sampling, and stratified random sampling. The use of stratified random sampling has gained popularity due to its simplicity and it almost ensures a gain in efficiency of the estimators if used properly. The main gains in stratified random sampling result from the construction of homogeneous strata. If strata are homogeneous, then stratified random sampling has been found to be efficient in so far as the precision of an estimator is considered. In stratified random sampling, there are two issues: allocation of sample size to each stratum and the construction of strata boundaries which could as much as possible contribute to forming homogeneous groups. Such strata boundaries, which lead to homogeneous strata, are also called optimum strata boundaries (OSB). The construction of OSB for simple random and with replacement sampling for Neyman allocation is a famous example. In addition, researchers have also approached it as a mathematical programming problem (MPP), which minimizes the variance of the estimator of the population mean while the total sample size and the cost of the survey remain the same.

In this project, we consider the construction of OSB while using ranked set sampling within each stratum.

In the next section, we provide commonly used notation in a study of stratified random sampling.

15.2 STRATIFIED RANDOM SAMPLING

Consider a population Ω of N units that is divided into l homogeneous groups, called strata, each of size N_h, $h = 1, 2, 3, \cdots l$. From the hth stratum of size N_h, assume a simple random and with replacement (SRSWR) sample of n_h units is selected, such that $\sum_{l=1}^{l} n_h = n$, the total fixed sample size. Let y_{hi}: $i = 1, 2, 3, \cdots n_h, h = 1, 2, 3 \cdots l$, be the values of the selected ith unit from the hth stratum.

Ranked Set Sampling. DOI: https://doi.org/10.1016/B978-0-12-815044-3.00015-0

Let

$$\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi} \tag{15.1}$$

be the unbiased estimator of the hth stratum population mean

$$\bar{Y}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} y_{hi} \tag{15.2}$$

Then an unbiased estimator of the population mean

$$\bar{Y} = \frac{1}{N}\sum_{h=1}^{l}\sum_{i=1}^{N_h} y_{hi} \tag{15.3}$$

is given by

$$\bar{y}_{st} = \sum_{h=1}^{l} W_h \bar{y}_h \tag{15.4}$$

when $W_h = \frac{N_h}{N}$ is the population proportion of data value in the hth stratum. Obviously, the variance of $\bar{y}_{st}$ is given by

$$V\left(\bar{y}_{st}\right) = \sum_{h=1}^{l} W_h^2 \frac{\sigma_{hy}^2}{n_h} \tag{15.5}$$

when

$$\sigma_{hy}^2 = \frac{1}{N_h}\sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2 \tag{15.6}$$

is the population variance of the hth stratum.

Under Neyman (1934) allocation, the optimum sample size is given by

$$n_h = n \frac{W_h \sigma_{hy}}{\sum_{h=1}^{l} W_h \sigma_{hy}} \tag{15.7}$$

and the minimum variance is given by

$$V\left(\bar{y}_{st}\right) = \frac{1}{n}\Big(\sum_{h=1}^{l} W_h \sigma_{hy}\Big)^2 \tag{15.8}$$

The set of point of stratification $y_1, y_2, \ldots y_{l-1}$ which minimize the $V(\bar{y}_{st})_N$ should give the best stratification for the Neyman allocation. Following Sukhatme and Sukhatme (1970), the optimum points of stratification with Neyman allocation are obtained by solving a set of $(l-1)$ simultaneous equations:

$$\frac{\sigma_{hy}^2 + \left(y_h - \bar{y}_h\right)^2}{\sigma_{hy}} = \frac{\sigma_{h+1}^2 + \left(y_h - \bar{y}_{h+1}\right)^2}{\sigma_{h+1}}, \quad h = 1, 2, \ldots l-1 \tag{15.9}$$

Under the assumption, $\sigma_{hy} = constant$ for all $h = 1, 2, \ldots l - 1$, the optimum points of stratification are obtained by solving $l - 1$ simultaneous equations

$$y_h = \frac{\overline{y}_h + \overline{y}_{h+1}}{2}, \quad h = 1, 2, \cdots l - 1 \tag{15.10}$$

Eq. (15.9) is not easy to solve for y_h if the number of strata become large, even when Eq. (15.10) is free from strata variances.

Dalenius and Hodges (1957) came up with the idea of using the cumulated value of $\sqrt{f(y)}$. Let $H = \int_a^b \sqrt{f(y)}dy$ be the total cumulated value of $\sqrt{f(y)}$; then the first approximation of the optimum point of stratification is given by

$$y_h = \frac{hH}{l}, l = 1, 2, \ldots l - 1 \tag{15.11}$$

Dalenius (1950) was the first to introduce the process of constructing OSB while using the same study variable for both estimation and stratification. We are also considering using the same variable for estimation and for construction of strata boundaries. There are many studies by different researchers, such as Mahalanobis (1952), Sethi (1963), Serfling (1968), Singh and Sukhatme (1969), and Singh (1971) among others, which deal with constructing OSB under different situation.

Sharma (2017) and Khan et al. (2009) considered a different approach that yields both the OSB and optimum sample size by making use of MPP. Following this direction, let y_0 and y_1 be the minimum and maximum values of the study variable y. The problem of determining the strata boundaries is to divide the range

$$d = y_l - y_0 \tag{15.12}$$

at intermediate points $y_1 \le y_2 \le \cdots \le y_{l-1}$ and find the optimum sample sizes n_h, such that the variance:

$$V(\overline{y}_{st}) = \sum_{h=1}^{l} W_h^2 \frac{\sigma_{hy}^2}{n_h} \tag{15.13}$$

is minimum.

Thus the problem of determining OSB and the optimum allocation can be formulated as:

Minimize

$$V(\overline{y}_{st}) = \sum_{h=1}^{l} W_h^2 \frac{\sigma_{hy}^2}{n_h} \tag{15.14}$$

Subject to the constraints:

$$\sum_{h=1}^{l} d_h = d \tag{15.15}$$

$$\sum_{h=1}^{l} n_h = n \tag{15.16}$$

and

$$d_h = (y_h - y_{h-1}) \geq 0 \tag{15.17}$$

$n_h \geq 1$ *is an integer.*

In the next section, we consider the use of ranked set sampling.

15.3 STRATIFIED RANKED SET SAMPLING

McIntyre (1952) introduced the idea of ranked set sampling for estimating the population mean yield of a crop in a field. He provides a clear and insightful introduction to ranked set sampling. Douglas (2012) has contributed an introductory review article on the use of ranked set sampling and its modifications since 1952.

The goal is to select n_h data values by selecting R_h groups of m_h data values. Each group of m_h data comes from m_h sets of m_h units each. For each set, units are ranked according to estimated y-values. Note that it would be pure judgmental ranking, which could involve the use of auxiliary information experience, etc., but the actual data values remain unknown. Here we assume that the unit that is judged to have the smallest y-value in the ranked set in fact has the smallest value, and the value of the study variable is actually measured only for this unit. This observation from the hth stratum is denoted by $y_{h[1]}$. In other words, we select an SRS sample of m_h units with judged data values y_{h1}, $y_{h2}, \ldots,$ y_{hm_h} and then ranked as

$$y_{h[1]} \leq y_{h[2]} \leq \cdots y_{h[m_h]}$$

This is the way the first observation $y_{h[1]}$ is selected from the hth stratum and the rest of the $(m_h - 1)$ units are discarded.

Now, we select another independent SRS of m_h units from the hth stratum, and rank them based on judgment as

$$y_{h[1]} \leq y_{h[2]} \leq \cdots y_{h[m_h]}$$

and this time $y_{h[2]}$ is retained in the RSS and the other $(m_h - 1)$ units are discarded. Repeat the process until m_h data values are included in the first group of data value in the RSS from the hth stratum. It is called the first cycle in the hth stratum of the SRSS and generated this over the first group of m_h data values.

Then we use such R_h cycles in the hth stratum such that $n_h = m_h R_h$. The ultimate SRSS sample of n_h units can be visualized as in Table 15.1.

Under SRSS, we propose an unbiased estimator of the population mean $\overline{Y}$ as

$$\overline{y}_{\text{SRSS}} = \sum_{h=1}^{l} W_h \sum_{j=1}^{Rh} \sum_{i=1}^{m_h} \frac{y_{h[i]j}}{m_h R_h} \tag{15.18}$$

or equivalently,

$$\overline{y}_{\text{SRSS}} = \sum_{h=1}^{l} W_h \frac{1}{n_h} \sum_{j=1}^{R_h} \sum_{i=1}^{m_h} y_{h[i]j} \tag{15.19}$$

The variance of the estimate $\overline{y}_{\text{SRSS}}$ is given by

Table 15.1 Stratified RSS Scheme

Cycle	In hth Stratum, $h = 1, 2 \cdots l$
1	$y_{h[1]1} y_{h[2]1} \cdots\cdots y_{h[m_h]1}$
2	$y_{h[1]2} y_{h[2]2} \cdots\cdots y_{h[m_h]2}$
3	$y_{h[1]3} y_{h[2]3} \cdots\cdots y_{h[m_h]3}$
$\cdots\cdots$	$\cdots\cdots\cdots$
R_h	$y_{h[1]R_h} y_{h[2]R_h} \cdots\cdots y_{h[m_h]R_h}$

$$\begin{aligned} V(\overline{y}_{\text{SRSS}}) &= \sum\nolimits_{h=1}^{l} W_h^2 \left\{ \frac{\sigma_{hy}^2}{R_h m_h} - \frac{1}{m_h R_h^2} \sum\nolimits_{i=1}^{R_h} \left(\overline{Y}_{h[i]} - \overline{Y} \right)^2 \right\} \\ &= \sum\nolimits_{h=1}^{l} W_h^2 \left\{ \frac{\sigma_{hy}^2}{n_h} - \frac{1}{n_h R_h} \sum\nolimits_{i=1}^{R_h} \left(\overline{Y}_{h[i]} - \overline{Y} \right)^2 \right\} \\ &= \sum\nolimits_{h=1}^{l} \frac{W_h^2}{n_h} \left\{ \sigma_{hy}^2 - \frac{1}{R_h} \sum\nolimits_{i=1}^{R_h} \left(\overline{Y}_{h[i]} - \overline{Y} \right)^2 \right\} \end{aligned} \tag{15.20}$$

for fixed R_h, and where $\overline{Y}_{h[i]}$ is the expected value of the ith units selected in the sample. The Neyman allocation in SRSS is

$$n_h = n \frac{W_h \sqrt{\sigma_{hy}^2 - \frac{1}{R_h} \sum_{i=1}^{R_h} \left(\overline{Y}_{h[i]} - \overline{Y}_h \right)^2}}{\sum_{h=1}^{l} W_h \sqrt{\sigma_{hy}^2 - \frac{1}{R_h} \sum_{i=1}^{R_h} \left(\overline{Y}_{h[i]} - \overline{Y}_h \right)^2}} \tag{15.21}$$

The minimum variance of $\overline{y}_{SRSS}$ with Neyman allocation is given by

$$\text{Min. } V(\overline{y}_{\text{SRSS}})_N = \frac{1}{n} \left[\sum\nolimits_{h=1}^{l} W_h \sqrt{\sigma_{hy}^2 - \frac{1}{R_h} \sum\nolimits_{i=1}^{R_h} \left(\overline{Y}_{h[i]} - \overline{Y}_h \right)^2} \right]^2 \tag{15.22}$$

On differentiating $V(\overline{y}_{\text{SRSS}})_N$ with respect to y_h and setting it equal to zero, we get

$$\begin{aligned} & W_h \frac{\partial \sqrt{\sigma_{hy}^2 - \frac{1}{R_h} \sum_{i=1}^{R_h} \left(\overline{Y}_{h[i]} - \overline{Y}_h \right)^2}}{\partial y_h} + \sqrt{\sigma_{hy}^2 - \frac{1}{R_h} \sum\nolimits_{i=1}^{R_h} \left(\overline{Y}_{h[i]} - \overline{Y}_h \right)^2} \frac{\partial W_h}{\partial y_h} \\ & + W_{h+1} \frac{\partial \sqrt{\sigma_{(h+1)y}^2 - \frac{1}{R_{(h+1)}} \sum_{i=1}^{R_{h+1}} \left(\overline{Y}_{(h+1)[i]} - \overline{Y}_{(h+1)} \right)^2}}{\partial y_h} \\ & + \sqrt{\sigma_{(h+1)y}^2 - \frac{1}{R_{(h+1)}} \sum\nolimits_{i=1}^{R_{h+1}} \left(\overline{Y}_{(h+1)[i]} - \overline{Y}_{(h+1)} \right)^2} \frac{\partial W_{(h+1)}}{\partial y_h} = 0 \end{aligned} \tag{15.23}$$

Now it is very difficult to find the following derivative:

$$\frac{\partial \sqrt{\sigma_{(h+1)y}^2 - \frac{1}{R_{(h+1)}} \sum_{i=1}^{R_{h+1}} \left(\overline{Y}_{(h+1)[i]} - \overline{Y}_{(h+1)} \right)^2}}{\partial y_h} \tag{15.24}$$

From the above we see that the problem of constructing strata boundaries in the case of ranked set sampling by following Dalenius (1950) becomes complicated. The other methods due to

Dalenius and Hodges (1957), called the cumulative square root of frequencies approach, would provide the same strata boundaries for SRSS, and hence is not useful. Thus we consider the MPP approach to find the OSB and sampling n_h as follows:

$$\text{Min. } V\left(\overline{y}_{\text{SRSS}}\right) = \sum\nolimits_{h=1}^{L} \frac{W_h^2}{n_h}\left\{\sigma_{hy}^2 - \frac{1}{R_h}\sum\nolimits_{i=1}^{R_h}\left(\overline{Y}_{h[i]} - \overline{Y}_h\right)^2\right\} \tag{15.25}$$

where $n_h = m_h R_h$

Subject to:

$$\sum_{h=1}^{l} d_h = d \tag{15.26}$$

$$\sum_{h=1}^{l} n_h = n \tag{15.27}$$

$$d_h = (y_h - y_{h-1}) \geq 0,$$

and $n_h \geq$ *is an integer.*

Also, we can use mathematical program:

$$\text{Min. } V\left(\overline{y}_{\text{SRSS}}\right) = \sum\nolimits_{h=1}^{L} \frac{W_h^2}{n_h}\left\{\sigma_{hy}^2 - \frac{1}{R_h}\sum\nolimits_{i=1}^{R_h}\left(\overline{Y}_{h[i]} - \overline{Y}_h\right)^2\right\} \tag{15.28}$$

Subject to:

$$\sum_{h=1}^{l} d_h = d \tag{15.29}$$

and $d_h = (y_h - y_{h-1}) \geq 0$

If we define $RDY_{[i]}$ as the ratio $\frac{\overline{Y}_{h[i]}}{\overline{Y}_h}$ as in Singh et al. (2014), then we can reformulate the problem as

$$\text{Min. } V\left(\overline{y}_{\text{SRSS}}\right) = \sum_{h=1}^{l} \frac{W_h^2}{n_h}\left\{\sigma_{hy}^2 - \frac{\overline{Y}_h^2}{R_h}\sum_{i=1}^{R_h}(RDY_{[i]} - 1)^2\right\} \tag{15.30}$$

Subject to:

$$\sum_{h=1}^{l} d_h = d \tag{15.31}$$

and $d_h = (y_h - y_{h-1}) \geq 0$

For comparing our proposed method to stratified random sampling we consider the percent *RE* of $\overline{y}_{\text{SRSS}}$ with respect to $\overline{y}_{st}$ is given by

$$R.E. = \frac{\text{min.}V(\overline{y}_{st})}{\text{min.}V(\overline{y}_{\text{SRSS}})} \times 100\% \tag{15.32}$$

In the next section, we show the performance of the proposed method through numerical illustrations.

15.4 NUMERICAL ILLUTRATIONS

In order to find the rank set sampling boundaries and variance, different sets of populations that follow uniform, triangular and exponential distributions are considered.

Population 1: Uniform distribution

The uniform distribution is a continuous distribution that has equal probability of observations over a given range. Two parameters, maximum and minimum values, define the distribution. Assume the maximum value is b, minimum value is a, then the density function of the uniform distribution is

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases} \tag{15.33}$$

We generated a population of size $N = 1000$ units from the uniform distribution with $a = 0.0$ and $b = 50$ using a random number generator in LINGO/PYTHON. We note that the minimum value of the study variable is $y_0 = 0.001$ and the largest value is $y_l = 49.883$. The range of the study variable in the population is given by $d = y_l - y_0 = 49.882$.

We determined the OSB and appropriate sample sizes to minimize the variance by using the programming language LINGO/PYTHON.

As the study variable y has uniform distribution with density function $f(y)$ in Eq. (15.33), we obtain W_h (stratum weight), $\overline{Y}_h$ (stratum mean), and σ_{hy}^2 (stratum variance), respectively. After integration of and organizing the results, we obtain the followings equations

$$W_h = \frac{d_h}{b-a} \tag{15.34}$$

where

$$d_h = y_h - y_{h-1} \tag{15.35}$$

$$\overline{Y}_h = \frac{y_h + y_{h-1}}{2} \tag{15.36}$$

$$\sigma_{hy}^2 = \frac{d_h^2}{12} \tag{15.37}$$

In the case of ranked set sampling, let $F = \frac{\overline{Y}_h^2}{R_h} \sum_{i=1}^{R_h} (RDY_{[i]} - 1)^2$, where $RDY_{[i]}$ is as defined in Singh et al. (2014).

We consider the use of ranked set sampling in each stratum and note the reduction in variance as given in Table 15.2.

After executing the LINGO/PYTHON code, the stratum sample size does not change much, but the variance of the estimator of the population mean is reduced. The value of F changes the variance of the population mean. From Table 15.2 it is noted that the value of percent relative efficiency is between 100% and 100.99% as the value of F changes from 0.0 to 0.508. As soon as the value of F becomes 0.509 then there is a drastic decrease in the value of percent relative efficiency to 62.03%. In the case of the uniform distribution the optimum sample sizes remain nearly the same, i.e., 25, in all four strata as the value of F ranges between 0.0 and 0.508. It is interesting to

Table 15.2 Uniform Distribution

F Values	$V(\bar{y}_{\text{SRSS}})$	Stratum Sample Size	*RE*
0	0.1166304	$n_1 = 25.00009$	100%
		$n_2 = 25.00005$	
		$n_3 = 24.99987$	
		$n_4 = 25.00000$	
0.05	0.1165179	$n_1 = 24.99999$	100.1%
		$n_2 = 25.00001$	
		$n_3 = 24.99999$	
		$n_4 = 25.00000$	
0.1	0.1164054	$n_1 = 24.99998$	100.19%
		$n_2 = 25.00002$	
		$n_3 = 24.99997$	
		$n_4 = 25.00003$	
0.2	0.1161804	$n_1 = 25.00000$	100.39%
		$n_2 = 25.00000$	
		$n_3 = 25.00000$	
		$n_4 = 25.00000$	
0.3	0.1159554	$n_1 = 24.99999$	100.58%
		$n_2 = 25.00001$	
		$n_3 = 24.99999$	
		$n_4 = 25.00000$	
0.4	0.1157304	$n_1 = 25.00000$	100.78%
		$n_2 = 25.00001$	
		$n_3 = 24.99999$	
		$n_4 = 24.99999$	
0.5	0.1155054	$n_1 = 25.00009$	100.97%
		$n_2 = 24.99996$	
		$n_3 = 24.99997$	
		$n_4 = 24.99998$	
0.505	0.1154941	$n_1 = 25.00003$	100.98%
		$n_2 = 24.99999$	
		$n_3 = 24.99998$	
		$n_4 = 25.00000$	
0.506	0.1154919	$n_1 = 25.00000$	100.99%
		$n_2 = 25.00000$	
		$n_3 = 25.00000$	
		$n_4 = 25.00000$	
0.507	0.1154896	$n_1 = 25.00000$	100.99%
		$n_2 = 25.00000$	
		$n_3 = 25.00000$	
		$n_4 = 25.00000$	

Table 15.2 Uniform Distribution *Continued*

F Values	$V(\bar{y}_{\text{SRSS}})$	Stratum Sample Size	*RE*
0.508	0.1154874	$n_1 = 25.00000$ $n_2 = 24.99998$ $n_3 = 25.00001$ $n_4 = 25.00000$	100.99%
0.509	0.1880247	$n_1 = 1.00000$ $n_2 = 33.00000$ $n_3 = 33.00000$ $n_4 = 33.00000$	62.03%

note that if F becomes 0.509 then the optimum sample size for the first stratum become 1, while for the other three strata it becomes 33.

Population 2: Right triangular distribution

The right triangular distribution is defined by two variables, which are its maximum and minimum values, say b and a, respectively. The general formula for the probability density function of right triangular distribution is given by

$$f(y) = \begin{cases} \dfrac{2(b-y)}{(b-a)^2}, & a \le y \le b \\ 0, & \text{otherwise} \end{cases} \tag{15.38}$$

We obtain W_h (stratum weight), $\overline{Y}_h$ (stratum mean), and σ^2_{hy} (stratum variance), respectively. After integration and organizing the results, we obtain the following equations

$$W_h = \frac{d_h(2a_h - d_h)}{(b-a)^2} \tag{15.39}$$

where $a_h = b - y_{h-1}$

$$\overline{Y}_h = \frac{3b(d_h + 2y_{h-1}) - 2(d_h^2 + 2d_h y_{h-1} + 3y_{h-1}^2)}{3(2a_h - d_h)} \tag{15.40}$$

$$\sigma^2_{yh} = \frac{d_h^2(d_h^2 - 6a_h d_h + 6a_h^2)}{18(2a_h - d_h)^2} \tag{15.41}$$

where $y_h = d_h + y_{h-1}$

We generated a population of size $N = 1000$ units from the right triangular distribution, and predetermined the sample size $n = 100$ from the population. We chose the minimum value of the study variable to be $y_0 = 0$ and the largest value to be $y_l = 5$. The range of the study variable in the population is then $d = y_1 - y_0 = 5$. Table 15.3 shows the results.

Again it is interesting to note that in the case of triangular distribution the value of percent relative efficiency lies between 100% and 100.009% as the value of F changes from 0.0 to 0.0000012.

Table 15.3 Results for right triangular distribution

F-values	$V(\bar{y}_{\text{SRSS}})$	Stratum Boundaries	Stratum Sample Size	Stratum Weight	*RE*
0	0.0002893406	$y_0 = 0.001$			100%
		$y_1 = 0.2892344$	$n_1 = 24.00$	$w_1 = 0.2700801$	
		$y_2 = 0.6208286$	$n_2 = 24.00$	$w_2 = 0.2582297$	
		$y_3 = 1.030881$	$n_3 = 24.00$	$w_3 = 0.2416786$	
		$y_4 = 1.980000$	$n_4 = 28.00$	$w_4 = 0.2300117$	
0.000001	0.0002893383	$y_0 = 0.001$			100.008%
		$y_1 = 0.2892346$	$n_1 = 24.00$	$w_1 = 0.2700802$	
		$y_2 = 0.6208288$	$n_2 = 24.00$	$w_2 = 0.2582297$	
		$y_3 = 1.030881$	$n_3 = 24.00$	$w_3 = 0.2416785$	
		$y_4 = 1.980000$	$n_4 = 28.00$	$w_4 = 0.2300116$	
0.00000101	0.0002893383	$y_0 = 0.001$			100.008%
		$y_1 = 0.2892345$	$n_1 = 24.00$	$w_1 = 0.2700801$	
		$y_2 = 0.6208286$	$n_2 = 24.00$	$w_2 = 0.2582297$	
		$y_3 = 1.030881$	$n_3 = 24.00$	$w_3 = 0.2416786$	
		$y_4 = 1.980000$	$n_4 = 28.00$	$w_4 = 0.2300116$	
0.00000104	0.0002893383	$y_0 = 0.001$			100.008%
		$y_1 = 0.2892346$	$n_1 = 24.00$	$w_1 = 0.2700802$	
		$y_2 = 0.6208288$	$n_2 = 24.00$	$w_2 = 0.2582297$	
		$y_3 = 1.030881$	$n_3 = 24.00$	$w_3 = 0.2416785$	
		$y_4 = 1.980000$	$n_4 = 28.00$	$w_4 = 0.2300116$	
0.00000106	0.0002893382	$y_0 = 0.001$			100.008%
		$y_1 = 0.2892346$	$n_1 = 24.00$	$w_1 = 0.2700802$	
		$y_2 = 0.6208288$	$n_2 = 24.00$	$w_2 = 0.2582297$	
		$y_3 = 1.030881$	$n_3 = 24.00$	$w_3 = 0.2416785$	
		$y_4 = 1.980000$	$n_4 = 28.00$	$w_4 = 0.2300116$	
0.00000108	0.0002893382	$y_0 = 0.001$			100.009%
		$y_1 = 0.2892345$	$n_1 = 24.00$	$w_1 = 0.2700802$	
		$y_2 = 0.6208287$	$n_2 = 24.00$	$w_2 = 0.2582297$	
		$y_3 = 1.030881$	$n_3 = 24.00$	$w_3 = 0.2416786$	
		$y_4 = 1.980000$	$n_4 = 28.00$	$w_4 = 0.2300116$	
0.00000109	0.0002893381	$y_0 = 0.001$			100.009%
		$y_1 = 0.2892345$	$n_1 = 24.00$	$w_1 = 0.2700802$	
		$y_2 = 0.6208287$	$n_2 = 24.00$	$w_2 = 0.2582297$	
		$y_3 = 1.030881$	$n_3 = 24.00$	$w_3 = 0.2416785$	
		$y_4 = 1.980000$	$n_4 = 28.00$	$w_4 = 0.2300116$	
0.0000011	0.0002893381	$y_0 = 0.001$			100.009%
		$y_1 = 0.2892345$	$n_1 = 24.00$	$w_1 = 0.2700802$	
		$y_2 = 0.6208287$	$n_2 = 24.00$	$w_2 = 0.2582297$	
		$y_3 = 1.030881$	$n_3 = 24.00$	$w_3 = 0.2416786$	
		$y_4 = 1.980000$	$n_4 = 28.00$	$w_4 = 0.2300116$	

Table 15.3 Results for right triangular distribution *Continued*

F-values	$V(\bar{y}_{\text{SRSS}})$	Stratum Boundaries	Stratum Sample Size	Stratum Weight	*RE*
0.0000012	0.0002893379	$y_0 = 0.001$			100.009%
		$y_1 = 0.2892346$	$n_1 = 24.00$	$w_1 = 0.2700802$	
		$y_2 = 0.6208288$	$n_2 = 24.00$	$w_2 = 0.2582297$	
		$y_3 = 1.030881$	$n_3 = 24.00$	$w_3 = 0.2416785$	
		$y_4 = 1.980000$	$n_4 = 28.00$	$w_4 = 0.2300116$	
0.0000013	0.0005075044	$y_0 = 0.001$			57.0124%
		$y_1 = 0.3882379$	$n_1 = 31.00$	$w_1 = 0.3530589$	
		$y_2 = 0.3894686$	$n_2 = 1.00$	$w_2 = 0.001$	
		$y_3 = 0.8718532$	$n_3 = 31.00$	$w_3 = 0.3323938$	
		$y_4 = 1.980000$	$n_4 = 37.00$	$w_4 = 0.3135473$	

In the first three strata the optimum sample size is 24 and in the fourth strata the sample size is 28. As soon as the value of F changes to 0.0000013 then there is a drastic decrease in the value of percent relative efficiency to 57.0124% and the optimum sample size in the first and third strata is 31, in the second stratum it is 1, in the fourth stratum it is 37.

Population 3: Exponential distribution

The exponential distribution is a continuous distribution with density given by

$$f(x;y) = \begin{cases} \lambda e^{-\lambda y}, & y \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{15.42}$$

For convenience, we set $\lambda = 1$ and $y \in [0,\ 5]$, therefore the density function for population 3 is:

$$f(y) = \begin{cases} \dfrac{1}{1-e^{-5}} e^{-y}, & y \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{15.43}$$

Because $\frac{1}{1-e^{-5}} \approx 1$, for convenience, in generating samples, we simplify Eq. (15.43) as:

$$f(y) = \begin{cases} e^{-y}, & y \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{15.44}$$

We generated a population of size $N = 1000$ units from the exponential distribution, and predetermined the sample size $n = 100$ from the population. We note that the minimum value of the study variable is $y_0 = 0.00673795$ and the largest value is $y_l = 5$. The range of the study variable in the population is given by $d = y_l - y_0 = 4.9932605$.

As the study variable y has exponential distribution with density function $f(y)$ in Eq. (15.42), we obtain W_h (stratum weight), $\overline{Y}_h$ (stratum mean), and σ^2_{hy} (stratum variance), respectively. After integration and organization, we obtain the following equations

$$W_h = \frac{e^{y_h} - e^{y_{h-1}}}{e^{y_h} e^{y_{h-1}}} \tag{15.45}$$

$$\overline{Y}_h = \frac{e^{y_h} y_{h-1} - e^{y_{h-1}} y_h}{e^{y_h} - e^{y_{h-1}}} + 1 \tag{15.46}$$

$$\sigma^2_{yh} = \frac{e^{y_h}\left(y^2_{h-1} + 2y_{h-1} + 2\right) - e^{y_{h-1}}\left(y^2_h + 2y_h + 2\right)}{e^{y_h} - e^{y_{h-1}}} - \overline{Y}^2_h \tag{15.47}$$

where

$$y_0 = 0, y_4 = 5 \ \text{ and } \ d_h = y_h - y_{h-1}$$

The minimum sample size has been set to 5, meaning $n_h \geq 5$ (not like the illustration for population 1 and 2), for the purpose of sample distribution in each stratum. Table 15.4 shows the results for exponential distribution.

The results in Table 15.4 are more promising in the case of exponential distribution as long as the problem of constructing strata boundaries for ranked set sampling is concerned. The value of percent relative efficiency value changes from 100% to 1483.08% as the value of F increases from 0.00 to 0.009 with steps of 0.001. For each value of F, there is quite a variation in the sample allocations among the four strata. For example, if F is 0.001 then the optimum allocation to the first stratum is 28 units, the second stratum is 34 units, the third stratum is 5 units, and the fourth stratum is 33 units. On the other hand, if F is 0.003 then the optimum allocation to the first stratum is

Table 15.4 Results for exponential distribution

***F*-values**	**$V(\bar{y}_{\text{SRSS}})$ $(\times 10^{-4})$**	**Stratum Boundaries**	**Stratum Sample Size**	**Stratum Weight**	***RE***
0	0.1324658	$y_0 = 0.00$			100%
		$y_1 = 0.2490131$	$n_1 = 32.00$	$w_1 = 0.2204302$	
		$y_2 = 0.5311369$	$n_2 = 31.00$	$w_2 = 0.1916336$	
		$y_3 = 2.280201$	$n_3 = 5.00$	$w_3 = 0.4856725$	
		$y_4 = 5.00$	$n_4 = 32.00$	$w_4 = 0.031$	
0.001	0.1186701	$y_0 = 0.00$			111.63%
		$y_1 = 0.2415763$	$n_1 = 28.00$	$w_1 = 0.2146112$	
		$y_2 = 0.5283637$	$n_2 = 34.00$	$w_2 = 0.1958199$	
		$y_3 = 2.302915$	$n_3 = 5.00$	$w_3 = 0.486019$	
		$y_4 = 5.00$	$n_4 = 33.00$	$w_4 = 0.033$	
0.002	0.1032451	$y_0 = 0.00$			128.3%
		$y_1 = 0.229411$	$n_1 = 22.00$	$w_1 = 0.2049983$	
		$y_2 = 0.5208214$	$n_2 = 37.00$	$w_2 = 0.2009693$	
		$y_3 = 2.368153$	$n_3 = 5.00$	$w_3 = 0.5003789$	
		$y_4 = 5.00$	$n_4 = 36.00$	$w_4 = 0.036$	
0.003	0.1070703	$y_0 = 0.00$			123.72%
		$y_1 = 0.3243645$	$n_1 = 51.00$	$w_1 = 0.270134$	
		$y_2 = 0.5142723$	$n_2 = 5.00$	$w_2 = 0.1250155$	
		$y_3 = 2.429445$	$n_3 = 5.00$	$w_3 = 0.5098499$	
		$y_4 = 5.00$	$n_4 = 39.00$	$w_4 = 0.039$	

Table 15.4 Results for exponential distribution *Continued*

F-values	$V(\bar{y}_{\text{SRSS}})$ $(\times 10^{-4})$	Stratum Boundaries	Stratum Sample Size	Stratum Weight	*RE*
0.004	0.07373335	$y_0 = 0.00$			179.66%
		$y_1 = 0.2929248$	$n_1 = 50.00$	$w_1 = 0.2539218$	
		$y_2 = 0.5122774$	$n_2 = 5.00$	$w_2 = 0.1469486$	
		$y_3 = 2.449082$	$n_3 = 5.00$	$w_3 = 0.5127569$	
		$y_4 = 5.00$	$n_4 = 40.00$	$w_4 = 0.04$	
0.005	0.02888716	$y_0 = 0.00$			458.56%
		$y_1 = 0.2453175$	$n_1 = 5.00$	$w_1 = 0.2175439$	
		$y_2 = 0.5103681$	$n_2 = 49.00$	$w_2 = 0.1821815$	
		$y_3 = 2.46834$	$n_3 = 5.00$	$w_3 = 0.5155493$	
		$y_4 = 5.0$	$n_4 = 41.00$	$w_4 = 0.041$	
0.006	0.03490493	$y_0 = 0.00$			379.5%
		$y_1 = 0.2362862$	$n_1 = 46.00$	$w_1 = 0.2104453$	
		$y_2 = 0.5050990$	$n_2 = 5.00$	$w_2 = 0.1861089$	
		$y_3 = 2.524018$	$n_3 = 5.00$	$w_3 = 0.5233089$	
		$y_4 = 5.0$	$n_4 = 44.00$	$w_4 = 0.044$	
0.007	0.02363396	$y_0 = 0.00$			560.4%
		$y_1 = 0.2130424$	$n_1 = 45.00$	$w_1 = 0.1918782$	
		$y_2 = 0.5034811$	$n_2 = 5.00$	$w_2 = 0.2036989$	
		$y_3 = 2.541919$	$n_3 = 5.00$	$w_3 = 0.5257079$	
		$y_4 = 5.0$	$n_4 = 45.00$	$w_4 = 0.045$	
0.008	0.007393389	$y_0 = 0.00$			1791.68%
		$y_1 = 0.3105857$	$n_1 = 5.00$	$w_1 = 0.2669825$	
		$y_2 = 0.5034812$	$n_2 = 45.00$	$w_2 = 0.1285946$	
		$y_3 = 2.541919$	$n3 = 5.00$	$w_3 = 0.5257078$	
		$y_4 = 5.0$	$n_4 = 45.00$	$w_4 = 0.045$	
0.009	0.008931817	$y_0 = 0.00$			1483.08%
		$y_1 = 0.2276341$	$n_1 = 17.00$	$w_1 = 0.2035844$	
		$y_2 = 0.4752243$	$n_2 = 5.00$	$w_2 = 0.17467$	
		$y_3 = 2.946215$	$n_3 = 5.00$	$w_3 = 0.5292074$	
		$y_4 = 5.0$	$n_4 = 73.00$	$w_4 = 0.073$	
0.04	0.0001542432	$y_0 = 0.00$			85,881.13%
		$y_1 = 0.2682696$	$n_1 = 32.00$	$w_1 = 0.2352985$	
		$y_2 = 0.746305$	$n_2 = 16.00$	$w_2 = 0.2905863$	
		$y_3 = 1.383136$	$n_3 = 46.00$	$w_3 = 0.2233243$	
		$y_4 = 5.0$	$n_4 = 6.00$	$w_4 = 0.006$	

51 units, to the second and third strata it is 5 units each, and the fourth stratum allocation is 39 units. It is very interesting to note that as the value of F becomes 0.04 then there is a substantial jump in the value of percent relative efficiency to 85881.13% with an optimum allocation of 32

units to the first stratum, 16 units to the second stratum, 46 units to the third stratum, and 6 units to the fourth stratum.

15.5 CONCLUSIONS

We conclude that, when creating optimum stratum boundaries for the uniform distribution, there is no difference between the use of simple random with replacement sampling and ranked set sampling. The optimum allocation for both sampling schemes also remains the same. However, there is a slight gain in the relative efficiency due to the use of ranked set sampling. Similar findings are observed in the case of right triangular distribution. In contrast, when considering the exponential distribution, we note that there is a change in strata boundaries, and optimum allocations, and that there is a substantial gain in relative efficiency while making use of ranked set sampling.

In future studies, we suggest the possible extension for the construction of strata boundaries and optimal allocation by following Mahajan et al. (2007), on similar lines for randomized response sampling due to Bouza (2009) for ranked set sampling. We also suggest that other RSS schemes cited in the review by Al-Omari and Bouza (2014) can also be investigated.

ACKNOWLEDGMENTS

The authors are thankful to Prof. Dr. Carlos N. Bouza-Herrera and two reviewers for their valuable comments/suggestions on the original version of this chapter. This chapter is a part of the first author's project work for the completion of MS in Statistical Analytics, Computing and Modelling. The LINGO/PYTHON code can be requested from the authors.

REFERENCES

Al-Omari, A.I., Bouza, C.N., 2014. Review of ranked set sampling: modifications and applications. Rev. Investig. Oper. 35 (3), 215–240.

Bouza, C.N., 2009. Ranked set sampling and randomized response procedure for estimating the mean of a sensitive quantitative character. Metrika 70, 267–277.

Dalenius, T., 1950. The problem of optimum stratification. Skand. Akd. 33, 203–213.

Dalenius, T., Hodges, J.L., 1957. The choice of stratification points. Skand. Aktuarietid Skrift 40, 198–203.

Douglas, A.W., 2012. Ranked set sampling: its relevance and impact on statistical inference. International Scholarly Research Network, ISRN Probability and Statistics 2012, . Available from: https://doi.org/10.5402/2012/568385Article ID 568385, 32 pages.

Khan, M.G.M., Ahmad, N., Khan, S., 2009. Determining the optimum stratum boundaries using mathematical programming. J. Math. Model. Algorithms 8 (4), 409–423.

Mahajan, P.K., Sharma, P., Gupta, R.K., 2007. Optimum stratification for allocation proportional to strata totals for scrambled response. Model Assist. Stat. Appl. 2 (2), 81–88.

Mahalanobis, P.C., 1952. Some aspects of the design of sample surveys. Sankhya 12, 1–7.

McIntyre, G.A., 1952. A method for unbiased selective sampling, using ranked set sampling and stratified simple random sampling. J. Appl. Stat. 23, 231–255.

Neyman, J., 1934. On the two different aspects of the representative's methods: the method stratified sampling and the method of purposive selection. J. Roy. Stat. Soc. 97, 558–606.

Serfling, R.J., 1968. Approximately optimal stratification. J. Am. Stat. Assoc. 63 (324), 1298–1309.

Sethi, V.K., 1963. A note on optimum stratification of population for estimating the population mean. Aust. J. Stat. 5, 20–33.

Sharma, S., 2017. Use of Mathematical Programming in Sampling. Unpublished MS Thesis, Submitted to the University of the South Pacific, Suva, Fiji Islands.

Singh, H.P., Tailor, R., Singh, S., 2014. General procedure for estimating the population mean using ranked set sampling. J. Stat. Comput. Simul. 84 (5), 931–945.

Singh, R., 1971. An alternate method of stratification on the auxiliary variable. Sankhya C 37, 100–108.

Singh, R., Sukhatme, B.V., 1969. Optimum stratification for equal allocation. Ann. Inst. Math. 27, 273–280.

Sukhatme, P.V., Sukhatme, B.V., 1970. Sampling Theory of Surveys with Applications.. Iowa State University Press, Ames.

CHAPTER

CALIBRATED ESTIMATOR OF POPULATION MEAN USING TWO-STAGE RANKED SET SAMPLING

16

Veronica I. Salinas, Stephen A. Sedory and Sarjinder Singh

Department of Mathematics, Texas A&M University-Kingsville, Kingsville, TX, United States

16.1 INTRODUCTION

The use of auxiliary information in estimating population mean or total is well known in the field of survey sampling. Various survey sampling schemes such as stratified sampling, cluster sampling, and multistage sampling are frequently used, among them two-stage sampling has the benefit of saving time, cost, and effort. As mentioned in Salinas et al. (2018), the two-stage sampling method is an improvement over cluster sampling when it is not possible or easy to enumerate all the units from the selected clusters. A solution to this difficulty is to select clusters, called first-stage units (FSUs), from the given population of interest and select subsamples from the selected clusters called second-stage units (SSUs). Assuming heterogeneous groups, this technique of sampling helps to increase the precision of the resultant estimates. It is easy to collect information from a few units within the selected FSUs, saving the cost of survey. Assume the population of interest $\Omega = \{1, 2, \ldots, N\}$ consists of N nonoverlapping clusters, called FSUs. The whole population is divided as $\Omega = \{\Omega_1, \Omega_2, \ldots, \Omega_N\}$, where Ω_i denotes the ith cluster of size M_i, for $i = 1, 2, \ldots, N$ such that $\Omega = \cup_{i=1}^{N} \Omega_i$ and $M = \sum\limits_{i=1}^{N} M_i$. Särndal et al. (1992) consider three situations with the auxiliary information in two-stage sampling. For the first situation, the auxiliary variable is available for all the FSUs, the second situation has the auxiliary variable for all the units in the population. Lastly, the third situation has the auxiliary variable available for all elements in the selected FSUs. For clarity, assume the simplest and most practical design where the FSUs are selected by simple random and without replacement (SRSWOR) and the SSUs are selected by simple random and with replacement (SRSWR) sampling schemes. Also assume that the population means of the auxiliary variable for the selected FSUs are known or available. The auxiliary information at the individual level may or may not be known. For simplicity of results, focus is put on the use of a single auxiliary variable. The application of two-stage sampling can involve various situations to the interest of the investigator. For example, in the agricultural sectors selecting villages as FSUs, and farmers at the SSUs; in education, selecting departments as FSUs, and faculty as SSUs. In politics, selecting blocks as FSUs and dwellings as

Ranked Set Sampling. DOI: https://doi.org/10.1016/B978-0-12-815044-3.00016-2

SSUs. In the health sector, FSUs could be hospitals and SSUs could be doctors. At a city level study, FSUs could be households and SSUs could be family members.

In the next section, we introduce notations and some basic results related to two-stage sampling.

16.2 NOTATIONS AND BASIC RESULTS

As stated earlier, consider a population Ω with N FSUs where the ith FSU Ω_i contains M_i SSUs, for $i = 1, 2, \ldots N$.

Let y_{ij} and x_{ij} denote the value of the study variable y and the auxiliary variable x respectively, for the jth SSU of the ith FSU, for $j = 1, 2, \ldots M_i$.

Let

$M = \sum_{i=1}^{N} M_i$ be the total number of SSUs in the population,

$\overline{M} = \frac{1}{N}\sum_{i=1}^{N} M_i$ be the average number of SSUs per FSU in the population, and

$\mu_i = \frac{M_i}{\overline{\overline{M}}} = \frac{NM_i}{M}$ be the expected number of units in the ith FSU.

Let

$Y_i = \sum_{j=1}^{M_i} y_{ij}$ be the population total of the study variable in the ith FSU,

$\overline{Y}_i = \frac{1}{M_i}\sum_{j=1}^{M_i} y_{ij} = \frac{Y_i}{M_i}$ be the population mean of the study variable in ith FSU,

$Y = \sum_{i=1}^{N} Y_i$ be population total of the study variable, and

$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} \mu_i \overline{Y}_i = \frac{1}{N}\sum_{i=1}^{N} \frac{M_i}{\overline{\overline{M}}} \cdot \frac{1}{M_i}\sum_{j=1}^{M_i} y_{ij} = \frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij} = \frac{Y}{M}$ be the population mean per SSU of the study variable, which is the focus of estimation.

Let

$X_i = \sum_{j=1}^{M_i} x_{ij}$ be the population total of the auxiliary variable in the ith FSU,

$\overline{X}_i = \frac{1}{M_i}\sum_{j=1}^{M_i} x_{ij} = \frac{X_i}{M_i}$ be the population mean of the auxiliary variable in ith FSU, which is assumed to be known,

$X = \sum_{i=1}^{N} X_i$ be the population total of the auxiliary variable, and

$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} \mu_i \overline{X}_i = \frac{1}{N}\sum_{i=1}^{N} \frac{M_i}{\overline{\overline{M}}} \cdot \frac{1}{M_i}\sum_{j=1}^{M_i} x_{ij} = \frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{M_i} x_{ij} = \frac{X}{M}$ be the population mean per SSU of the auxiliary variable, which is assumed to be known.

Let

$\sigma_{iy}^2 = \frac{1}{M_i}\sum_{j=1}^{M_i} \left(y_{ij} - \overline{Y}_i\right)^2$, be the population variance for the ith FSU, and

$S_{by}^2 = \frac{1}{N-1}\sum_{i=1}^{N} \left(\mu_i \overline{Y}_i - \overline{Y}\right)^2$ be the population variance of between the weighted FSUs population means.

Suppose an SRSWOR of n FSUs is selected from N FSUs. A sample of m_i SSUs from the ith selected FSU of size M_i is selected by SRSWR sampling.

Let

$\bar{y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}$, be the sample mean of the study variable in ith FSU, and

$\bar{x}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} x_{ij}$, be the sample mean of the auxiliary variable in ith FSU.

Then we have the following lemmas:

Lemma 16.1: An unbiased estimator of the population mean $\overline{Y}$ is given by

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} \mu_i \bar{y}_i, \tag{16.1}$$

Proof: Let E_2 denote the expected value over all possible second-stage samples, each of size m_i taken using SRSWR sampling from a given FSU of size M_i.

Let E_1 denote the expected value over all possible first-stage samples each of size n taken using SRSWOR sampling from a given population of N FSUs.

Taking the expected value of the sample mean $\bar{y}$, we have

$$E(\bar{y}) = E_1 E_2\left[\frac{1}{n}\sum_{i=1}^{n}\mu_i\bar{y}_i\right] = E_1\left[\frac{1}{n}\sum_{i=1}^{n}\mu_i E_2(\bar{y}_i)\right] = E_1\left[\frac{1}{n}\sum_{i=1}^{n}\mu_i\overline{Y}_i\right] = \frac{1}{N}\sum_{i=1}^{N}\mu_i\overline{Y}_i = \overline{Y}$$

which proves the lemma. Now we have the following corollary:

Lemma 16.2: An unbiased estimator of the population mean $\overline{X}$ is given by:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} \mu_i \bar{x}_i \tag{16.2}$$

Proof: Following Lemma 16.1, it is obvious that

$$E(\bar{x}) = E_1 E_2\left[\frac{1}{n}\sum_{i=1}^{n}\mu_i\bar{x}_i\right] = E_1\left[\frac{1}{n}\sum_{i=1}^{n}\mu_i E_2(\bar{x}_i)\right] = E_1\left[\frac{1}{n}\sum_{i=1}^{n}\mu_i\overline{X}_i\right] = \frac{1}{N}\sum_{i=1}^{N}\mu_i\overline{X}_i = \overline{X}$$

which proves the lemma.

Lemma 16.3: The variance of the sample mean estimator $\bar{y}$ is given by

$$V(\bar{y}) = \frac{1}{nN}\sum_{i=1}^{N}\mu_i^2\left(\frac{\sigma_{iy}^2}{m_i}\right) + \left(\frac{1-f}{n}\right)S_{by}^2 \tag{16.3}$$

were $f = \frac{n}{N}$ is the finite population correction factor while n FSUs are selected from the N FSUs by SRSWOR, where $\sigma_{iy}^2 = \frac{1}{M_i}\sum_{j=1}^{M_i}\left(y_{ij}-\overline{Y}_i\right)^2$, and $S_{by}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\mu_i\overline{Y}_i-\overline{Y}\right)^2$ have their usual meanings.

Proof: Let V_2 denote the variance over all possible second-stage samples each of size m_i taken using SRSWR sampling from a given FSU of size M_i. Let V_1 denote the variance value over all possible first-stage samples each of size n taken using SRSWOR sampling from a given population of N FSUs.

By the definition of variance, the variance of the sample mean $\bar{y}$ is given by:

$$\begin{aligned}
V(\bar{y}) &= E_1 V_2(\bar{y}) + V_1 E_2(\bar{y}) \\
&= E_1 V_2\left(\frac{1}{n}\sum_{i=1}^{n}\mu_i \bar{y}_i\right) + V_1 E_2\left(\frac{1}{n}\sum_{i=1}^{n}\mu_i \bar{y}_i\right) \\
&= E_1\left[\frac{1}{n^2}\sum_{i=1}^{n}\mu_i^2 V_2(\bar{y}_i)\right] + V_1\left[\frac{1}{n}\sum_{i=1}^{n}\mu_i E_2(\bar{y}_i)\right] \\
&= E_1\left[\frac{1}{n^2}\sum_{i=1}^{n}\mu_i^2\left(\frac{\sigma_{iy}^2}{m_i}\right)\right] + V_1\left[\frac{1}{n}\sum_{i=1}^{n}\mu_i \overline{Y}_i\right] \\
&= \frac{1}{nN}\sum_{i=1}^{N}\mu_i^2\left(\frac{\sigma_{iy}^2}{m_i}\right) + \left(\frac{1-f}{n}\right)\left(\frac{1}{N-1}\right)\sum_{i=1}^{N}(\mu_i\overline{Y}_i - \overline{Y})^2 \\
&= \frac{1}{nN}\sum_{i=1}^{N}\mu_i^2\left(\frac{\sigma_{iy}^2}{m_i}\right) + \left(\frac{1-f}{n}\right)S_{by}^2
\end{aligned}$$

which proves the lemma.

Lemma 16.4: The variance of the sample mean estimator $\bar{x}$ is given by

$$V(\bar{x}) = \frac{1}{nN}\sum_{i=1}^{N}\mu_i^2\left(\frac{\sigma_{ix}^2}{m_i}\right) + \left(\frac{1-f}{n}\right)S_{bx}^2 \tag{16.4}$$

where $\sigma_{ix}^2 = \frac{1}{M_i}\sum_{j=1}^{M_i}\left(x_{ij} - \overline{X}_i\right)^2$ and $S_{bx}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\mu_i\overline{X}_i - \overline{X}\right)^2$ have their usual meanings.

Proof: It follows from the previous lemma.

Lemma 16.5: The covariance between the sample mean estimators $\bar{y}$ and $\bar{x}$ is given by

$$\mathrm{Cov}(\bar{y}, \bar{x}) = \frac{1}{nN}\sum_{i=1}^{N}\mu_i^2\left(\frac{\sigma_{ixy}}{m_i}\right) + \left(\frac{1-f}{n}\right)S_{bxy} \tag{16.5}$$

where $\sigma_{ixy} = \frac{1}{M_i}\sum_{j=1}^{M_i}\left(y_{ij} - \overline{Y}_i\right)\left(x_{ij} - \overline{X}_i\right)$ and $S_{bxy} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\mu_i\overline{Y}_i - \overline{Y}\right)\left(\mu_i\overline{X}_i - \overline{X}\right)$ have their usual meanings.

Proof: Let C_2 denote the covariance over all possible second-stage samples each of size m_i taken using SRSWOR sampling from a given FSU of size M_i. Let C_1 denote the covariance value over all possible first-stage samples each of size n taken using SRSWOR sampling from a given population of N FSUs. By the definition of covariance, the covariance between the sample means $\bar{y}$ and $\bar{x}$ is given by:

$$
\begin{aligned}
\operatorname{Cov}(\bar{y},\bar{x}) &= E_1[C_2(\bar{y},\bar{x})] + C_1[E_2(\bar{y}), E_2(\bar{x})] \\
&= E_1\left[C_2\left(\frac{1}{n}\sum_{i=1}^{n}\mu_i\bar{y}_i, \frac{1}{n}\sum_{i=1}^{n}\mu_i\bar{x}_i\right)\right] + C_1\left[E_2\left(\frac{1}{n}\sum_{i=1}^{n}\mu_i\bar{y}_i\right), E_2\left(\frac{1}{n}\sum_{i=1}^{n}\mu_i\bar{x}_i\right)\right] \\
&= E_1\left[\frac{1}{n^2}\sum_{i=1}^{n}\mu_i^2 C_2(\bar{y}_i,\bar{x}_i)\right] + C_1\left[\frac{1}{n}\sum_{i=1}^{n}\mu_i\bar{Y}_i, \frac{1}{n}\sum_{i=1}^{n}\mu_i\bar{X}_i\right] \\
&= \frac{1}{nN}\sum_{i=1}^{N}\mu_i^2\left(\frac{1}{m_i}\right)\sigma_{ixy} + \left(\frac{1-f}{n}\right)\left(\frac{1}{N-1}\right)\sum_{i=1}^{N}(\mu_i\bar{Y}_i - \bar{Y})(\mu_i\bar{X}_i - \bar{X}) \\
&= \frac{1}{nN}\sum_{i=1}^{N}\mu_i^2\left(\frac{1}{m_i}\right)\sigma_{ixy} + \left(\frac{1-f}{n}\right)S_{bxy}
\end{aligned}
$$

which proves the lemma.

Sukhatme, Sukhatme, Sukhatme, and Asok (1984) suggested a regression-type estimator of the population mean in two-stage sampling as

$$\bar{y}_{lr} = \frac{1}{n}\sum_{i=1}^{n}\mu_i\bar{y}_i + \hat{\beta}(\bar{X} - \bar{x}) = \bar{y} + \hat{\beta}(\bar{X} - \bar{x}) \tag{16.6}$$

The variance of the regression-type estimator $\bar{y}_{lr}$ can be approximated as:

$$
\begin{aligned}
V(\bar{y}_{lr}) &= V\left[\bar{y} + \hat{\beta}(\bar{X} - \bar{x})\right] \\
&\approx V(\bar{y}) + \beta^2 V(\bar{x}) - 2\beta\ \operatorname{Cov}(\bar{y},\ \bar{x}) \\
&= \frac{1}{nN}\sum_{i=1}^{N}\mu_i^2\left(\frac{\sigma_{iy}^2}{m_i}\right) + \left(\frac{1-f}{n}\right)S_{by}^2 + \beta^2\left[\frac{1}{nN}\sum_{i=1}^{N}\mu_i^2\left(\frac{\sigma_{ix}^2}{m_i}\right) + \left(\frac{1-f}{n}\right)S_{bx}^2\right] \\
&\quad - 2\beta\left[\frac{1}{nN}\sum_{i=1}^{N}\mu_i^2\left(\frac{\sigma_{ixy}}{m_i}\right) + \left(\frac{1-f}{n}\right)S_{bxy}\right] \\
&= \left(\frac{1-f}{n}\right)\left[S_{by}^2 + \beta^2 S_{bx}^2 - 2\beta S_{bxy}\right] + \left(\frac{1}{nN}\right)\sum_{i=1}^{N}\mu_i^2\left(\frac{1}{m_i}\right)\left[\sigma_{iy}^2 + \beta^2\sigma_{ix}^2 - 2\beta\sigma_{ixy}\right]
\end{aligned} \tag{16.7}
$$

Sukhatme et al. (1984) substituted the value of the true regression coefficient as

$$\beta = \frac{S_{bxy}}{S_{bx}^2}. \tag{16.8}$$

The value of minimum variance for the above optimum choice of the regression coefficient β is given by

$$\text{Min. } V(\bar{y}_{lr}) = \left(\frac{1-f}{n}\right)S_{by}^2\left[1 - \rho_{bxy}^2\right] + \left(\frac{1}{nN}\right)\sum_{i=1}^{N}\mu_i^2\left(\frac{1}{m_i}\right)\left[\sigma_{iy}^2 + \beta^2\sigma_{ix}^2 - 2\beta\sigma_{ixy}\right] \tag{16.9}$$

where

$$\rho_{bxy} = \frac{S_{bxy}}{S_{bx}S_{by}} \tag{16.10}$$

denotes the population correlation coefficient between the population means of the FSUs.

In the next section, we consider the problem of estimation of the population mean using ranked set sampling (RSS) at the second stage of sampling.

16.3 TWO-STAGE RANKED SET SAMPLING

Suppose an SRSWOR sample of n FSUs is selected from N FSUs. A sample of m_i SSUs from the ith selected FSU of size M_i is selected by a RSS. From the ith FSU of M_i units, we select an SRSWR of h_i units (y_{i1}, x_{i1}), (y_{i2}, x_{i2}), $\ldots$,(y_{ih_i}, x_{ih_i}). Then rank the units based on the study variable by a judgment ranking as $(y_{i[1]}, x_{i(1)})$, $(y_{i[2]}, x_{i(2)})$, .. $\ldots$, $(y_{i[u_i]}, x_{i(h_i)})$. Retain only the first ranked ordered pair $(y_{i[1]}, x_{i(1)})$. Again, select an SRSWR sample of h_i units as (y_{i1}, x_{i1}), (y_{i2}, x_{i2}), $\ldots$,(y_{ih_i}, x_{ih_i}), then rank the study variable based on judgment ranking as $(y_{i[1]}, x_{i(1)})$, $(y_{i[2]}, x_{i(2)})$,$\ldots$,$(y_{i[h_i]}, x_{i(h_i)})$. Then retain the second ranked ordered pair $(y_{i[2]}, x_{i(2)})$. Repeat the process r_i times within the ith selected FSU of M_i units such that $m_i = h_i r_i$. Details about improvements and applications can be found and observed from Al-Omari and Bouza (2014).

Let

$\bar{y}_{[i]} = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{i[j]}$, be the RSS mean of the study variable in ith FSU, and

$\bar{x}_{(i)} = \frac{1}{m_i}\sum_{j=1}^{m_i} x_{i(j)}$, be the RSS mean of the auxiliary variable in ith FSU.

Then we have the following lemmas:

Lemma 16.6: An unbiased estimator of the population mean $\overline{Y}$ is given by

$$\bar{y}_{\text{RSS}} = \frac{1}{n}\sum_{i=1}^{n} \mu_i \bar{y}_{[i]}, \tag{16.11}$$

Proof: Let E_2 denote the expected value over all possible second-stage samples each of size m_i taken using RSS sampling from a given FSU of size M_i.

Let E_1 denote the expected value over all possible first-stage samples each of size n taken using SRSWOR sampling from a given population of N FSUs.

Taking the expected value of the sample mean $\bar{y}_{\text{RSS}}$, we have

$$E\left(\bar{y}_{\text{RSS}}\right) = E_1 E_2\left[\frac{1}{n}\sum_{i=1}^{n} \mu_i \bar{y}_{[i]}\right] = E_1\left[\frac{1}{n}\sum_{i=1}^{n} \mu_i E_2\left(\bar{y}_{[i]}\right)\right] = \frac{1}{N}\sum_{i=1}^{N} \mu_i \overline{Y}_i = \overline{Y}$$

which proves the lemma.

Now we have the following corollary

Lemma 16.7: An unbiased estimator of the population mean $\overline{X}$ is given by:

$$\bar{x}_{\text{RSS}} = \frac{1}{n}\sum_{i=1}^{n}\mu_i\bar{x}_{(i)} \tag{16.12}$$

Proof: It is obvious that

$$E(\bar{x}_{\text{RSS}}) = E_1E_2\left[\frac{1}{n}\sum_{i=1}^{n}\mu_i\bar{x}_{(i)}\right] = E_1\left[\frac{1}{n}\sum_{i=1}^{n}\mu_iE_2\left(\bar{x}_{(i)}\right)\right] = E_1\left[\frac{1}{n}\sum_{i=1}^{n}\mu_i\overline{X}_i\right] = \frac{1}{N}\sum_{i=1}^{N}\mu_i\overline{X}_i = \overline{X}$$

which proves the lemma.

Lemma 16.8: The variance of the sample mean estimator $\bar{y}_{\text{RSS}}$ is given by

$$V\left(\bar{y}_{\text{RSS}}\right) = \frac{1}{nN}\sum_{i=1}^{N}\mu_i^2\left(\frac{\sigma_{iy}^2}{m_i} - \frac{1}{m_ir_i}\sum_{t=1}^{r_i}\left(\overline{Y}_{i[t]}-\overline{Y}_i\right)^2\right) + \left(\frac{1-f}{n}\right)S_{by}^2 \tag{16.13}$$

where $f = \frac{n}{N}$ is the finite population correction factor while selecting n FSUs from the N FSUs by SRSWOR, where $\sigma_{iy}^2 = \frac{1}{M_i}\sum\limits_{j=1}^{M_i}\left(y_{ij}-\overline{Y}_i\right)^2$, and $S_{by}^2 = \frac{1}{N-1}\sum\limits_{i=1}^{N}\left(\mu_i\overline{Y}_i-\overline{Y}\right)^2$ have their usual meanings.

Proof: Let V_2 denote the variance over all possible second-stage samples each of size m_i taken using RSS from a given FSU of size M_i. Let V_1 denote the variance value over all possible first-stage samples each of size n taken using SRSWOR sampling from a given population of N FSUs.

By the definition of variance, the variance of the sample mean $\bar{y}_{\text{RSS}}$ is given by

$$\begin{aligned}
V\left(\bar{y}_{\text{RSS}}\right) &= E_1V_2\left(\bar{y}_{\text{RSS}}\right) + V_1E_2\left(\bar{y}_{\text{RSS}}\right)\\
&= E_1V_2\left(\frac{1}{n}\sum_{i=1}^{n}\mu_i\bar{y}_{[i]}\right) + V_1E_2\left(\frac{1}{n}\sum_{i=1}^{n}\mu_i\bar{y}_{[i]}\right)\\
&= E_1\left[\frac{1}{n^2}\sum_{i=1}^{n}\mu_i^2V_2\left(\bar{y}_{[i]}\right)\right] + V_1\left[\frac{1}{n}\sum_{i=1}^{n}\mu_iE_2\left(\bar{y}_{[i]}\right)\right]\\
&= E_1\left[\frac{1}{n^2}\sum_{i=1}^{n}\mu_i^2\left(\frac{\sigma_{iy}^2}{m_i} - \frac{1}{m_ir_i}\sum_{t=1}^{r_i}(\overline{Y}_{i[t]}-\overline{Y}_i)^2\right)\right] + V_1\left[\frac{1}{n}\sum_{i=1}^{n}\mu_i\overline{Y}_i\right]\\
&= \frac{1}{nN}\sum_{i=1}^{N}\mu_i^2\left(\frac{\sigma_{iy}^2}{m_i} - \frac{1}{m_ir_i}\sum_{t=1}^{r_i}(\overline{Y}_{i[t]}-\overline{Y}_i)^2\right) + \left(\frac{1-f}{n}\right)\left(\frac{1}{N-1}\right)\sum_{i=1}^{N}\left(\mu_i\overline{Y}_i-\overline{Y}\right)^2\\
&= \frac{1}{nN}\sum_{i=1}^{N}\mu_i^2\left(\frac{\sigma_{iy}^2}{m_i} - \frac{1}{m_ir_i}\sum_{t=1}^{r_i}(\overline{Y}_{i[t]}-\overline{Y}_i)^2\right) + \left(\frac{1-f}{n}\right)S_{by}^2
\end{aligned}$$

which proves the lemma.

Lemma 16.9: The variance of the ranked set sample mean estimator $\overline{x}_{\text{RSS}}$ is given by

$$V(\overline{x}_{\text{RSS}}) = \frac{1}{nN}\sum_{i=1}^{N}\mu_i^2\left(\frac{\sigma_{ix}^2}{m_i} - \frac{1}{m_i r_i}\sum_{t=1}^{r_i}(\overline{X}_{i(t)} - \overline{X}_i)^2\right) + \left(\frac{1-f}{n}\right)S_{bx}^2 \tag{16.14}$$

where $\sigma_{ix}^2 = \frac{1}{M_i}\sum_{j=1}^{M_i}\left(x_{ij} - \overline{X}_i\right)^2$ and $S_{bx}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\mu_i\overline{X}_i - \overline{X}\right)^2$ have their usual meanings.

Proof: It follows from the previous lemma.

Lemma 16.10: The covariance between the sample mean estimators $\overline{y}_{\text{RSS}}$ and $\overline{x}_{\text{RSS}}$ is given by

$$\text{Cov}\left(\overline{y}_{\text{RSS}}, \overline{x}_{\text{RSS}}\right) = \frac{1}{nN}\sum_{i=1}^{N}\mu_i^2\left(\frac{\sigma_{ixy}}{m_i} - \frac{1}{m_i r_i}\sum_{t=1}^{r_i}(\overline{X}_{i(t)} - \overline{X})(\overline{Y}_{i[t]} - \overline{Y}_i)\right) + \left(\frac{1-f}{n}\right)S_{bxy} \tag{16.15}$$

where $\sigma_{ixy} = \frac{1}{M_i}\sum_{j=1}^{M_i}\left(y_{ij} - \overline{Y}_i\right)\left(x_{ij} - \overline{X}_i\right)$ and $S_{bxy} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\mu_i\overline{Y}_i - \overline{Y}\right)\left(\mu_i\overline{X}_i - \overline{X}\right)$ have their usual meanings.

Proof: Let C_2 denote the covariance over all possible second-stage samples each of size m_i taken using RSS from a given FSU of size M_i. Let C_1 denote the covariance value over all possible first-stage samples each of size n taken using SRSWOR sampling from a given population of N FSUs. By the definition of covariance, the covariance between the sample means $\overline{y}_{\text{RSS}}$ and $\overline{x}_{\text{RSS}}$ is given by

$$\begin{aligned}
\text{Cov}\left(\overline{y}_{\text{RSS}}, \overline{x}_{\text{RSS}}\right) &= E_1\left[C_2\left(\overline{y}_{\text{RSS}}, \overline{x}_{\text{RSS}}\right)\right] + C_1\left[E_2\left(\overline{y}_{\text{RSS}}\right), E_2(\overline{x}_{\text{RSS}})\right] \\
&= E_1\left[C_2\left(\frac{1}{n}\sum_{i=1}^{n}\mu_i\overline{y}_{[i]}, \frac{1}{n}\sum_{i=1}^{n}\mu_i\overline{x}_{(i)}\right)\right] + C_1\left[E_2\left(\frac{1}{n}\sum_{i=1}^{n}\mu_i\overline{y}_{[i]}\right), E_2\left(\frac{1}{n}\sum_{i=1}^{n}\mu_i\overline{x}_{(i)}\right)\right] \\
&= E_1\left[\frac{1}{n^2}\sum_{i=1}^{n}\mu_i^2 C_2\left(\overline{y}_{[i]}, \overline{x}_{(i)}\right)\right] + C_1\left[\frac{1}{n}\sum_{i=1}^{n}\mu_i\overline{Y}_i, \frac{1}{n}\sum_{i=1}^{n}\mu_i\overline{X}_i\right] \\
&= \frac{1}{nN}\sum_{i=1}^{N}\mu_i^2\left(\frac{\sigma_{ixy}}{m_i} - \frac{1}{m_i r_i}\sum_{t=1}^{r_i}(\overline{Y}_{i[t]} - \overline{Y}_i)(\overline{X}_{i(t)} - \overline{X}_i)\right) + \left(\frac{1-f}{n}\right)\left(\frac{1}{N-1}\right)\sum_{i=1}^{N}\left(\mu_i\overline{Y}_i - \overline{Y}\right)\left(\mu_i\overline{X}_i - \overline{X}\right) \\
&= \frac{1}{nN}\sum_{i=1}^{N}\mu_i^2\left(\frac{\sigma_{ixy}}{m_i} - \frac{1}{m_i r_i}\sum_{t=1}^{r_i}(\overline{Y}_{i[t]} - \overline{Y}_i)\left(\overline{X}_{i(t)} - \overline{X}_i\right)\right) + \left(\frac{1-f}{n}\right)S_{bxy}
\end{aligned}$$

which proves the lemma.

In the next section, we define a new calibrated estimator in two-stage RSS.

16.4 CALIBRATED ESTIMATOR IN TWO-STAGE RANKED SET SAMPLING

We consider a new calibrated estimator of the population mean $\overline{Y}$ in two-stage RSS as

$$\overline{y}_{\text{RSS}(c)} = \sum_{i=1}^{n} w_i\overline{y}_{[i]}, \tag{16.16}$$

where w_i are calibrated weights, and are obtained by minimizing chi-squared distance function defined as

$$D = \frac{1}{2}\sum_{i=1}^{n} \frac{\left(w_i - u_i/n\right)^2}{(q_i u_i/n)}, \tag{16.17}$$

where q_i is another set of known weights, subject to the calibration constraints given by

$$\sum_{i=1}^{n} w_i = \frac{1}{n}\sum_{i=1}^{n} \mu_i \tag{16.18}$$

and

$$\sum_{i=1}^{n} w_i \bar{x}_{(i)} = \overline{X}. \tag{16.19}$$

The Lagrange function is given by

$$L = \frac{1}{2}\sum_{i=1}^{n} \frac{\left(w_i - u_i/n\right)^2}{(q_i u_i/n)} - \lambda_1\left[\sum_{i=1}^{n} w_i - \frac{1}{n}\sum_{i=1}^{n} u_i\right] - \lambda_2\left[\sum_{i=1}^{n} w_i \bar{x}_{(i)} - \overline{X}\right] \tag{16.20}$$

On setting

$$\frac{\partial L}{\partial w_i} = 0$$

We get

$$w_i = \frac{u_i}{n} + \lambda_1 \frac{q_i u_i}{n} + \lambda_2 \frac{q_i u_i}{n} \bar{x}_{(i)} \tag{16.21}$$

On substituting Eq. (16.21) into Eq. (16.18) and into Eq. (16.19), we get

$$\lambda_1 \sum_{i=1}^{n} q_i u_i + \lambda_2 \sum_{i=1}^{n} q_i u_i \bar{x}_{(i)} = 0 \tag{16.22}$$

and

$$\frac{\lambda_1}{n}\sum_{i=1}^{n} q_i u_i \bar{x}_{(i)} + \frac{\lambda_2}{n}\sum_{i=1}^{n} q_i u_i \{\bar{x}_{(i)}\}^2 = \overline{X} - \frac{1}{n}\sum_{i=1}^{n} u_i \bar{x}_{(i)} \tag{16.23}$$

On solving Eqs. (16.22) and (16.23) for λ_1 and λ_2, we get

$$\lambda_1 = \frac{-\left(\sum_{i=1}^{n} q_i u_i \bar{x}_{(i)}\right)\left[\overline{X} - \frac{1}{n}\sum_{i=1}^{n} u_i \bar{x}_{(i)}\right]}{\frac{1}{n}\left(\sum_{i=1}^{n} q_i u_i\right)\left(\sum_{i=1}^{n} q_i u_i \bar{x}_{(i)}^2\right) - \frac{1}{n}\left(\sum_{i=1}^{n} q_i u_i \bar{x}_{(i)}\right)^2} \tag{16.24}$$

and

$$\lambda_2 = \frac{\left(\sum_{i=1}^{n} q_i u_i\right)\left[\overline{X} - \frac{1}{n}\sum_{i=1}^{n} u_i \bar{x}_{(i)}\right]}{\frac{1}{n}\left(\sum_{i=1}^{n} q_i u_i\right)\left(\sum_{i=1}^{n} q_i u_i \bar{x}_{(i)}^2\right) - \frac{1}{n}\left(\sum_{i=1}^{n} q_i u_i \bar{x}_{(i)}\right)^2} \tag{16.25}$$

Substituting the values of λ_1 and λ_2 into Eq. (16.21), the calibrated weights are given by

$$w_i = \frac{u_i}{n} + \frac{q_i u_i \overline{x}_{(i)}\left(\sum_{i=1}^{n} q_i u_i\right) - q_i u_i \left(\sum_{i=1}^{n} q_i u_i \overline{x}_{(i)}\right)}{\left(\sum_{i=1}^{n} q_i u_i\right)\left(\sum_{i=1}^{n} q_i u_i \overline{x}_{(i)}^2\right) - \left(\sum_{i=1}^{n} q_i u_i \overline{x}_{(i)}\right)^2}\left[\overline{X} - \frac{1}{n}\sum_{i=1}^{n} u_i \overline{x}_{(i)}\right] \tag{16.26}$$

Substituting Eq. (16.26) into Eq. (16.16), the calibrated estimator of the population mean in two-stage RSS is given by

$$\overline{y}_{\mathrm{RSS}(c)} = \frac{1}{n}\sum_{i=1}^{n} u_i \overline{y}_{[i]} + \hat{\beta}_{\mathrm{cal}}\left[\overline{X} - \frac{1}{n}\sum_{i=1}^{n} u_i \overline{x}_{(i)}\right], \tag{16.27}$$

where

$$\hat{\beta}_{\mathrm{cal}} = \frac{\left(\sum_{i=1}^{n} q_i u_i\right)\sum_{i=1}^{n} q_i u_i \overline{x}_{(i)} \overline{y}_{[i]} - \left(\sum_{i=1}^{n} q_i u_i \overline{y}_{[i]}\right)\left(\sum_{i=1}^{n} q_i u_i \overline{x}_{(i)}\right)}{\left(\sum_{i=1}^{n} q_i u_i\right)\left(\sum_{i=1}^{n} q_i u_i \overline{x}_{(i)}^2\right) - \left(\sum_{i=1}^{n} q_i u_i \overline{x}_{(i)}\right)^2} \tag{16.28}$$

It may be worth mentioning if all the N FSUs are included in the survey, then the new proposed calibrated estimator in Eq. (16.27) would behave like the stratified sampling estimator recently studied by Koyuncu (2017).

Since $\hat{\beta}_{\mathrm{cal}}$ is a consistent estimator of the regression coefficient β, the estimator (16.27) can be approximated as

$$\overline{y}_{\mathrm{RSS}(c)} \approx \overline{y}_{\mathrm{RSS}} + \beta\left[\overline{X} - \overline{x}_{\mathrm{RSS}}\right] + \text{Higher order terms} \tag{16.29}$$

The variance of the calibrated estimator is approximated as

$$\begin{aligned}
V(\overline{y}_{\mathrm{RSS}(c)}) &\approx V(\overline{y}_{\mathrm{RSS}}) + \beta^2 V(\overline{x}_{\mathrm{RSS}}) - 2\beta\ Cov\left(\overline{y}_{\mathrm{RSS}}, \overline{x}_{\mathrm{RSS}}\right)\\
&= \frac{1}{nN}\sum_{i=1}^{N}\mu_i^2\left(\frac{\sigma_{iy}^2}{m_i} - \frac{1}{m_i r_i}\sum_{t=1}^{r_i}\left(Y_{i[t]} - \overline{Y}_i\right)^2\right) + \left(\frac{1-f}{n}\right)S_{by}^2\\
&+ \beta^2\left[\frac{1}{nN}\sum_{i=1}^{N}\mu_i^2\left(\frac{\sigma_{ix}^2}{m_i} - \frac{1}{m_i r_i}\sum_{t=1}^{r_i}\left(\overline{X}_{i(t)} - \overline{X}_i\right)^2\right) + \left(\frac{1-f}{n}\right)S_{bx}^2\right]\\
&- 2\beta\left[\frac{1}{nN}\sum_{i=1}^{N}\mu_i^2\left(\frac{\sigma_{ixy}}{m_i} - \frac{1}{m_i r_i}\sum_{t=1}^{r_i}\left(Y_{i[t]} - \overline{Y}_i\right)\left(\overline{X}_{i(t)} - \overline{X}_i\right)\right) + \left(\frac{1-f}{n}\right)S_{bxy}\right]\\
&= \left(\frac{1-f}{n}\right)\left[S_{by}^2 + \beta^2 S_{bx}^2 - 2\beta S_{bxy}\right] + \left(\frac{1}{nN}\right)\sum_{i=1}^{N}\mu_i^2\left(\frac{1}{m_i}\right)\left[\sigma_{iy}^2 + \beta^2\sigma_{ix}^2 - 2\beta\sigma_{ixy}\right]\\
&- \frac{1}{nN}\sum_{i=1}^{N}\frac{\mu_i^2}{m_i r_i}\sum_{t=1}^{r_i}\left[\left(Y_{i[t]} - \overline{Y}_i\right)^2 + \beta^2\left(\overline{X}_{i(t)} - \overline{X}_i\right)^2 - 2\beta\left(Y_{i[t]} - \overline{Y}_i\right)\left(\overline{X}_{i(t)} - \overline{X}\right)\right]
\end{aligned} \tag{16.30}$$

In the next section, we examine an application of two-stage sampling using a real data set.

16.5 NUMERICAL ILLUSTRATION WITH REAL DATA

For the purpose of numerical illustration, as in Salinas et al. (2018), we study a population consisting of the faculty from nine departments (as listed in Table 16.1) of the College of Arts and Sciences at Texas A&M University—Kingsville to investigate the performance of the proposed new calibrated estimator in two-stage RSS.

We assumed the study variable as the annual salary of a faculty member and the auxiliary variable as their experience at Texas A&M University—Kingsville, with the first year as the start date, irrespective of age or previous experience at other institutes. We cleaned the data set by including only those faculty members with an annual salary greater than $10,000, irrespective of experience. In this numerical illustration, we have $N = 9$. Let $\overline{Y}_i$ and $\overline{X}_i$ denote the average salary and average experience, respectively, of a faculty member in the ith department. Let S^2_{iy} and S^2_{ix} denote the population variances for the salary and experience, respectively, within the ith department. Let M_i be the total number of faculty members in the ith department. Let ρ_{ixy} denote the value of the correlation coefficient between salary and experience in the i-th department. A brief description of the parameters of the population in each of the above nine departments is given in Tables 16.2(a) and 16.2(b).

Thus $S^2_{by} = 132791075.1$, $S^2_{bx} = 11.20$, $S_{bxy} = 34307.62$ and $\rho_{bxy} = 0.88956$.

A SAS code (see Appendix A) was written to investigate the percent relative efficiency values. The percent relative efficiency of the RSS over the simple random sampling is defined as

$$RE = \frac{\text{Min. } V(\bar{y}_{lr})}{V(\bar{y}_{\text{RSS}(c)})} \times 100\%. \tag{16.31}$$

Following Singh, Tailor, and Singh (2014), realized (RD) ratios of the judgment-based ranked values to that of true population mean were defined for the study and auxiliary variables as

$$RD_{1i}[t] = \frac{\overline{Y}_{i[t]}}{\overline{Y}_i} \tag{16.32}$$

and

$$RD_{2i}(t) = \frac{\overline{X}_{i(t)}}{\overline{X}_i} \tag{16.33}$$

for $t = 1, 2, 3, \ldots, r_i$ in each cycle within the ith FSU.

Table 16.1 Departments as FSUs

1	Arts & Communications
2	Biological & Health Science
3	Chemistry
4	Language & Literature
5	History & Political Science
6	Music
7	Mathematics
8	Physics & Geosciences
9	Psychology & Sociology

Table 16.2a Descriptive Parameters at the Departmental Level

Dept.	M_i	$\overline{Y}_i$	$\overline{X}_i$	S^2_{iy}	S^2_{ix}	ρ_{ixy}
1	17	51659	12.12	410109956	72.74	0.31533
2	12	62809	12.75	349923262	110.93	0.75701
3	11	63299	10.91	279143130	71.49	0.81145
4	22	49296	9.47	351684087	83.51	0.46757
5	17	56109	9.64	465988718	71.19	0.89101
6	18	54650	10.40	240097753	67.83	0.34254
7	20	60144	15.11	359926960	107.16	0.45578
8	12	51917	10.92	606389811	99.90	0.44375
9	17	56635	11.82	556387242	198.65	0.64247

Table 16.2b Descriptive Parameters at the Departmental and Overall Levels

Dept	u_i	$u_i\overline{Y}_i$	$(u_i\overline{Y}_i-\overline{Y})^2$	$u_i\overline{X}_i$	$(u_i\overline{X}_i-\overline{X})^2$	$(u_i\overline{Y}_i-\overline{Y})(u_i\overline{X}_i-\overline{X})$
1	1.0479	54135.80	2593357.28	12.70	1.549	-2004.27
2	0.7397	46461.45	86206391.78	9.43	4.101	18801.60
3	0.6781	42921.92	164461827.25	7.40	16.472	52048.96
4	1.3562	66853.48	123371839.41	12.84	1.922	15398.81
5	1.0479	58799.16	9320599.89	10.10	1.834	-4134.68
6	1.1096	60639.04	23939974.42	11.54	0.007	407.18
7	1.2329	74150.14	338705199.13	18.63	51.441	131997.89
8	0.7397	38404.36	300739262.55	8.08	11.416	58592.84
9	1.0479	59350.38	12990149.02	12.39	0.865	3352.63
Sum	9.0000	501715.72	1062328600.72	103.11	89.61	274460.94

In this simulation study we considered several values of

$$RD_{1i}[t] = H_1 + 0.08e_t \tag{16.34}$$

and

$$RD_{2i}[t] = H_2 + 0.08e_t \tag{16.35}$$

where $e_t \sim N(0,1)$.

Then different values of $H_1 = \{0.85,\ 1.00,\ 1.15\}$ and $H_2 = \{0.75,\ 1.00,\ 1.25\}$ are investigated for different situations. The choice of H_1 is made so that the judgment ranking could be 85% of the original true mean value, could be perfect ranking, or could be 15% higher. More variation in judgment ranking is not considered, since judgment ranking will introduce measurement errors in the study variable, Y. The value of H_2 is given a wider range from 0.75 to 1.25, with a step of 0.25,

Table 16.3 Percent *RE* of the Two-Stage RSS

n	m	H_1	H_2	***RE* (%)**
3	30	0.9	0.8	104.7
3	30	0.9	1.0	114.8
3	30	0.9	1.3	205.5
3	30	1.0	0.8	112.9
3	30	1.0	1.0	103.5
3	30	1.0	1.3	130.1
3	30	1.0	0.8	123.0
3	30	1.0	1.0	104.3
3	30	1.0	1.3	117.1
3	30	1.2	0.8	238.8
3	30	1.2	1.0	120.6
3	30	1.2	1.3	105.8
5	30	0.9	0.8	101.8
5	30	0.9	1.0	117.7
5	30	0.9	1.3	209.6
5	30	1.0	0.8	113.6
5	30	1.0	1.0	103.8
5	30	1.0	1.3	131.1
5	30	1.0	0.8	120.8
5	30	1.0	1.0	103.0
5	30	1.0	1.3	128.1
5	30	1.2	0.8	200.3
5	30	1.2	1.0	113.6
5	30	1.2	1.3	102.5
3	60	0.9	0.8	104.7
3	60	0.9	1.0	122.5
3	60	0.9	1.3	227.9
3	60	1.0	0.8	114.6
3	60	1.0	1.0	105.1
3	60	1.0	1.3	140.2
3	60	1.0	0.8	113.7
3	60	1.0	1.0	105.3
3	60	1.0	1.3	122.3
3	60	1.2	0.8	209.7
3	60	1.2	1.0	127.2
3	60	1.2	1.3	104.0
5	60	0.9	0.8	104.6
5	60	0.9	1.0	129.7
5	60	0.9	1.3	260.6
5	60	1.0	0.8	115.4
5	60	1.0	1.0	105.9
5	60	1.0	1.3	135.4
5	60	1.0	0.8	116.3
5	60	1.0	1.0	103.4
5	60	1.0	1.3	122.5
5	60	1.2	0.8	195.1
5	60	1.2	1.0	113.4
5	60	1.2	1.3	105.7

since it is not in the hands of the investigator to control the value of the auxiliary variable, X. Recall that judgment ranking is made only for the study variable. We used proportional allocation to select SSUs from the FSUs with $m_i = mM_i/(N\overline{M})$ to select a total sample of the required size m as reflected in Table 16.3.

In the numerical comparisons, we consider two first-stage sample sizes of $n = 3, 5$ and total second-stage sample sizes of $m = 30$, 60 with proportional allocation across all nine FSUs. For the three values of H_1 and H_2 considered, it has been observed that the value of percent RE ranges from 101.80% to 260.59%, with a median value of 110.85%, a mean value of 131.91%, and a standard deviation of 41.15%. For $n = 3$, the minimum value of RE is 103.49%, maximum value is 238.76%, median value is 115.95%, mean value is 132.43%, with a standard deviation of 41.72%. For $n = 5$, the minimum value of RE is 101.80%, maximum value is 260.59%, median value is 115.85%, mean value is 131.41%, with a standard deviation of 41.47%.

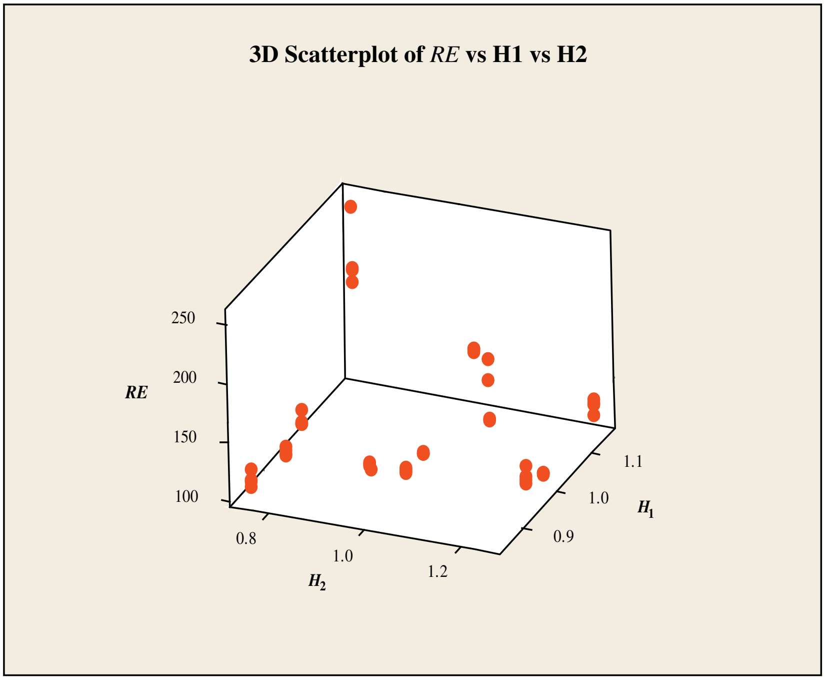


FIGURE 16.1

Graphical presentation of *RE* values versus judgment ranking.

A graphical presentation of the values of percent relative efficiency as a function of H_1 and H_2 is shown in Fig. 16.1. Overall we conclude that use of RSS while selecting SSUs could be beneficial over the use of simple random sampling.

ACKNOWLEDGMENTS

The authors are thankful to Prof. Dr. Carlos N. Bouza-Herrera for his valuable comments/suggestions on the original version of this chapter. Regular help from the Editorial Project Manager, Susan Ikeda, is also duly acknowledged. This chapter is an extension of the first author’s thesis work for the completion of an MS in Statistical Analytics, Computing and Modeling.

REFERENCES

Al-Omari, A.I., Bouza, C.N., 2014. Review of ranked set sampling: modifications and applications. Rev. Investig. Oper. 35 (3), 215–240.

Koyuncu, N., 2017. Calibration estimator of population mean under stratified ranked set sampling design. Commun. Stat.: Theory Methods (Available online.

Salinas, V.I., Sedory, S.A., Singh, S., 2018. Calibrated estimators in two-stage sampling. Commun. Stat.: Theory Methods (Avaliable online) .

Särndal, C.E., Swensson, B., Wretman, J.H., 1992. Model Assisted Survey Sampling. Springer-Verlag, New York.

Singh, H.P., Tailor, R., Singh, S., 2014. General procedure for estimating the population mean using ranked set sampling. J. Stat. Comput. Simul. 84 (5), 931−945.

Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S., Asok, C., 1984. Sampling Theory with Applications. Indian Society of Agricultural Statistics, New Delhi & Iowa State University Press, Ames, USA.

APPENDIX A

```
*SAS CODES USED IN THE NUMERICAL ILLUSTRATION;
DATA DATA1;
INPUT DEPT MI YIM XIM SIY2 SIX2 RHOIXY;
CARDS;
1  17 51659  12.12   410109956     72.74   0.31533
2  12 62809  12.75   349923262   110.93   0.75701
3  11 63299  10.91   279143130     71.49   0.81145
4  22 49296    9.47   351684087     83.51   0.46757
5  17 56109    9.64   465988718     71.19   0.89101
6  18 54650  10.40   240097753     67.83   0.34254
7  20 60144  15.11   359926960   107.16   0.45578
8  12 51917  10.92   606389811     99.90   0.44375
9  17 56635  11.82   556387242   198.65   0.64247
;
%MACRO VERO(III, H1_IN, H2_IN, MS_IN, NS_IN);
DATA DATA2;
SET DATA1;
N = 9;
NS = &NS_IN;
f = NS/N;
M_BAR = 16.2222222;
ui = MI/M_BAR;
ms = &MS_IN;
msi = ms*MI/(N*M_BAR);
SIGYI2 = (MI-1)*SIY2/MI;
SIGXI2 = (MI-1)*SIX2/MI;
SYb2 = 132791075.1;
SXb2 = 11.20;
SXYb = 34307.62;
RHOXYb = 0.88956;
BET = SXYb/SXb2;
COVXYI = RHOIXY * SQRT(SIY2*SIX2);
TERM_1 = ((1-f)/ns)*SYb2*(1-RHOXYb**2);
TERM_2I = (ui**2/(NS*N))*(SIGYI2 + BET**2*SIGXI2 - 2 * BET *COVXYI)/msi;
PROC MEANS DATA = DATA2 NOPRINT;
VAR TERM_2I;
OUTPUT OUT = DATA3 SUM = SUM_TERM_2I;
DATA DATA4;
SET DATA3;
KEEP SUM_TERM_2I;
DATA DATA5;
SET DATA2;
IF _N_ = 1 THEN SET DATA4;
VAR_LR = TERM_1 + SUM_TERM_2I;
```

```
KEEP VAR_LR BET YIM XIM;
DATA VAR_LR;
SET DATA5;
IF _N_=1;
RUN;
DATA DATA6;
SET DATA1;
IF _N_=1 THEN SET VAR_LR;
DATA DATA7;
SET DATA5;
SET DATA1;
ri = 3;
H1 = &H1_IN;
H2 = &H2_IN;
DO I = 1 TO ri;
R1T = H1 + 0.08*RAND("NORMAL");
R2T = H2 + 0.08*RAND("NORMAL");
OUTPUT;
END;
KEEP DEPT R1T R2T BET H1 H2 YIM XIM ;
DATA DATA8;
SET DATA7;
V1 = YIM**2*(R1T-1)**2 + BET**2*XIM**2*(R2T-1)**2-2*BET* YIM*XIM*(R1T-1)
*(R2T-1);
PROC SORT DATA = DATA8;
BY DEPT;
PROC MEANS DATA = DATA8 NOPRINT;
VAR V1;
BY DEPT;
OUTPUT OUT = DATA9 MEAN = MEAN_VI;
DATA DATA10;
SET DATA9;
DROP _TYPE_  _FREQ_;
PROC SORT DATA=DATA2;
BY DEPT;
PROC SORT DATA = DATA10;
BY DEPT;
DATA DATA11;
MERGE DATA2 DATA10;
BY DEPT;
MINUS_TERM = (ui**2/msi)*MEAN_VI/(NS*N);
PROC MEANS DATA = DATA11 NOPRINT;
VAR MINUS_TERM;
OUTPUT OUT = DATA12 SUM = SUM_MINUS_TERM;
DATA DATA13;
SET DATA12;
```

```
DROP _TYPE_ _FREQ_;
DATA DATA14;
MERGE DATA13 DATA5;
IF _N_ =1;
DATA DATA15;
SET DATA14;
VARP = VAR_LR-SUM_MINUS_TERM;
RE = VAR_LR*100/VARP;
KEEP RE;
DATA DATA16;
SET DATA2;
IF _N_=1 THEN SET DATA15;
DATA DATA17;
SET DATA16;
IF _N_ = 1 THEN SET DATA7;
KEEP NS MS H1 H2 RE;
DATA DATA18&III;
SET DATA17;
IF _N_=1;
DATA DATA19;
SET DATA181 DATA182 DATA183 DATA184 DATA185 DATA186 DATA187
DATA188 DATA189 DATA1810 DATA1811 DATA1812 DATA1813 DATA1814
DATA1815 DATA1816 DATA1817 DATA1818 DATA1819 DATA1820 DATA1821
DATA1822 DATA1823 DATA1824 DATA1825 DATA1826 DATA1827 DATA1828
DATA1829 DATA1830 DATA1831 DATA1832 DATA1833 DATA1834 DATA1835
DATA1836 DATA1837 DATA1838 DATA1839 DATA1840 DATA1841 DATA1842
DATA1843 DATA1844 DATA1845 DATA1846 DATA1847 DATA1848;
PROC PRINT DATA = DATA19;
VAR MS NS H1 H2 RE;
RUN;
%MEND VERO(III, H1_IN, H2_IN, MS_IN, NS_IN);
%VERO(1, 0.85, 0.75, 30, 3);
%VERO(2, 0.85, 1.00, 30, 3);
%VERO(3, 0.85, 1.25, 30, 3);
%VERO(4, 0.95, 0.75, 30, 3);
%VERO(5, 0.95, 1.00, 30, 3);
%VERO(6, 0.95, 1.25, 30, 3);
%VERO(7, 1.00, 0.75, 30, 3);
%VERO(8, 1.00, 1.00, 30, 3);
%VERO(9, 1.00, 1.25, 30, 3);
%VERO(10, 1.15, 0.75, 30, 3);
%VERO(11, 1.15, 1.00, 30, 3);
%VERO(12, 1.15, 1.25, 30, 3);
%VERO(13, 0.85, 0.75, 30, 5);
%VERO(14, 0.85, 1.00, 30, 5);
%VERO(15, 0.85, 1.25, 30, 5);
```

```
%VERO(16, 0.95, 0.75, 30, 5);
%VERO(17, 0.95, 1.00, 30, 5);
%VERO(18, 0.95, 1.25, 30, 5);
%VERO(19, 1.00, 0.75, 30, 5);
%VERO(20, 1.00, 1.00, 30, 5);
%VERO(21, 1.00, 1.25, 30, 5);
%VERO(22, 1.15, 0.75, 30, 5);
%VERO(23, 1.15, 1.00, 30, 5);
%VERO(24, 1.15, 1.25, 30, 5);
%VERO(25, 0.85, 0.75, 60, 3);
%VERO(26, 0.85, 1.00, 60, 3);
%VERO(27, 0.85, 1.25, 60, 3);
%VERO(28, 0.95, 0.75, 60, 3);
%VERO(29, 0.95, 1.00, 60, 3);
%VERO(30, 0.95, 1.25, 60, 3);
%VERO(31, 1.00, 0.75, 60, 3);
%VERO(32, 1.00, 1.00, 60, 3);
%VERO(33, 1.00, 1.25, 60, 3);
%VERO(34, 1.15, 0.75, 60, 3);
%VERO(35, 1.15, 1.00, 60, 3);
%VERO(36, 1.15, 1.25, 60, 3);
%VERO(37, 0.85, 0.75, 60, 5);
%VERO(38, 0.85, 1.00, 60, 5);
%VERO(39, 0.85, 1.25, 60, 5);
%VERO(40, 0.95, 0.75, 60, 5);
%VERO(41, 0.95, 1.00, 60, 5);
%VERO(42, 0.95, 1.25, 60, 5);
%VERO(43, 1.00, 0.75, 60, 5);
%VERO(44, 1.00, 1.00, 60, 5);
%VERO(45, 1.00, 1.25, 60, 5);
%VERO(46, 1.15, 0.75, 60, 5);
%VERO(47, 1.15, 1.00, 60, 5);
%VERO(48, 1.15, 1.25, 60, 5);
RUN;
PROC EXPORT DATA=DATA19 LABEL OUTFILE = 'C:\SASDATAFILES\RSS_2.XLS'
DBMS=EXCEL REPLACE;
RUN;
```

CHAPTER

ESTIMATION OF POPULATION MEAN USING INFORMATION ON AUXILIARY ATTRIBUTE: A REVIEW 17

Rajesh Singh[1], **Prabhakar Mishra**[1] **and Carlos N. Bouza-Herrera**[2]

[1]*Department of Statistics, Banaras Hindu University, Varanasi, Uttar Pradesh, India* [2]*Faculty of Mathematics and Computation, University of Havana, Havana, Cuba*

17.1 INTRODUCTION

In the sampling literature, auxiliary information is commonly used to improve estimates. Many authors have suggested estimators based on auxiliary information. However in many practical situations, instead of the existence of auxiliary variables there exist some auxiliary attributes, e.g., ϕ, which are highly correlated with the study variable y, such as:

(i) Sex (ϕ) and height of persons (y);
(ii) Amount of milk produced (y) and a particular breed of cow (ϕ);
(iii) Amount of yield of wheat crop and a particular variety of wheat (ϕ).

Consider a sample of size n drawn by simple random sampling without replacement (SRSWOR) from a population of size N. Let y_i and ϕ_i denote the observations on variable y and ϕ, respectively, for the ith unit ($i = 1,2,\ldots\ldots,N$). Let

$\phi_i = 1$, if ith unit of population possesses attribute $\phi = 0$, otherwise.

Let $A = \sum_{i=1}^{N} \phi_i$ and $a = \sum_{i=1}^{n} \phi_i$ denote the total number of units in the population and sample, respectively, possessing attributes. Let $P = A/N$ and $p = a/n$ denote the proportion of units in the population and sample, respectively, possessing attribute ϕ.

17.2 ESTIMATION OF POPULATION MEAN USING SINGLE AUXILIARY ATTRIBUTE INFORMATION

Taking into consideration the point biserial correlation coefficient between auxiliary attribute ϕ and the study variable y, Naik and Gupta (1996) defined the ratio estimator for population mean $\overline{Y}\left(= \frac{1}{N}\sum_{i=1}^{N} y_i\right)$ of the study variable y as follows

Ranked Set Sampling. DOI: https://doi.org/10.1016/B978-0-12-815044-3.00017-4

$$t_1 = \bar{y}(P \frown P) \tag{17.1}$$

where $\bar{y}\left(=\frac{1}{n}\sum_{i=1}^{n} y_i\right)$ is the sample mean of the study variable y.

The mean square error (MSE) of the ratio estimator t_1, up to the first order of approximation is given by

$$\mathrm{MSE}(t_1) = \left(\frac{1-f}{n}\right)\bar{Y}^2\left[C_y^2 + C_p^2\left(1-2k_p\right)\right] \tag{17.2}$$

where,

$$f = \frac{n}{N}, \quad C_y^2 = \frac{S_y^2}{\bar{Y}^2}, \quad S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2,$$
$$C_p^2 = \frac{S_\phi^2}{P^2}, \quad S_\phi^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\phi_i - P\right)^2, \quad k_p = \frac{\rho_{pb}C_y}{C_p}$$
$$\rho_{pb} = \frac{S_{y\phi}}{S_y S_\phi} \quad S_{y\phi} = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)\left(\phi_i - P\right)$$

It is well known that under SRSWOR, the variance of the usual unbiased estimator is

$$\mathrm{Var}(\bar{y}) = \frac{1-f}{n}\bar{Y}^2 C_y^2 \tag{17.3}$$

Jhajj, Sharma and Grover (2006) defined a general class of estimator as $t_2 = g(\bar{y}, v)$, where $v = \hat{P}/P$ and $g(\bar{y}, v)$ is a parametric function of $\bar{y}$ and v such that $g\left(\bar{Y}, 1\right) = \bar{Y}$, $\forall \bar{Y}$ and the function $g(\bar{y}, v)$ satisfies certain regularity conditions. The optimum MSE of the estimator t_2 is given by

$$\mathrm{MSE}(t_2)_{\mathrm{opt}} \cong \left(\frac{1-f}{n}\right)S_y^2\left(1-\rho_{pb}^2\right) \tag{17.4}$$

The above expression is equal to variance of the linear regression estimator

$$t_3 = \bar{y} + b\left(P - \hat{P}\right) \tag{17.5}$$

where b is the sample regression coefficient whose population regression coefficient is given by

$$\beta = \rho_{pb}S_y/S_\phi.$$

Shabbir and Gupta (2007) suggested the ratio type estimator for the population mean $\bar{Y}$ as

$$t_4 = \bar{y}\left[d_1 + d_2\left(P - \hat{P}\right)\right]\left(P/\hat{P}\right) \tag{17.6}$$

where (d_1, d_2) are suitably chosen constants whose sum need not be unity.

For the optimum values of d_1 and d_2 as

$$d_1^* = \frac{1}{1 + \left(\frac{1-f}{n}\right)C_y^2\left(1-\rho_{pb}^2\right)},$$

and

$$d_2^* = \frac{\left(\rho_{pb}C_y - C_p\right)}{\left[1 + \theta C_y^2\left(1-\rho_{pb}^2\right)\right]PC_p},$$

the minimum MSE of the estimator t_4 is given by

$$\text{min. MSE}(t_4) = \frac{\left(\frac{1-f}{n}\right)S_y^2\left(1-\rho_{pb}^2\right)}{1+\left(\frac{1-\mathrm{f}}{\mathrm{n}}\right)C_y^2\left(1-\rho_{pb}^2\right)} \tag{17.7}$$

Singh, Cauhan, Sawan, and Smarandache (2007) introduced the following ratio and product type exponential estimators of $\overline{Y}$

$$t_5 = \overline{y}\exp\left(\frac{P-\hat{P}}{P+\hat{P}}\right) \tag{17.8}$$

$$t_6 = \overline{y}\exp\left(\frac{\hat{P}-P}{\hat{P}+P}\right) \tag{17.9}$$

Singh, Cauhan, Sawan, and Smarandache (2007) further defined the following class of exponential estimators of $\overline{Y}$

$$t_7 = \overline{\mathrm{y}}\left[\alpha\exp\left(\frac{\mathrm{P}-\hat{\mathrm{P}}}{\mathrm{P}+\hat{\mathrm{P}}}\right)+(1-\alpha)\exp\left(\frac{\hat{P}-P}{\hat{P}+P}\right)\right] \tag{17.10}$$

where α is a suitably chosen constant.

The MSEs of the estimators t_5, t_6, and t_7, up to the first order of approximation, are respectively given by

$$\text{MSE}(t_5) = \left(\frac{1-f}{n}\right)\overline{\mathrm{Y}}^2\left[C_y^2+\frac{C_p^2}{4}-\rho_{pb}C_yC_p\right], \tag{17.11}$$

$$\text{MSE}(t_6) = \left(\frac{1-f}{n}\right)\overline{Y}^2\left[C_y^2+\frac{C_p^2}{P}+\rho_{pb}C_yC_p\right]. \tag{17.12}$$

$$\text{min. MSE}(t_7) = \frac{1-f}{n}\overline{Y}^2C_y^2\left(1-\rho_{pb}^2\right). \tag{17.13}$$

Singh, Chauhan, Sawan, and Smarandache (2008) suggested an estimator t_8 as

$$t_8 = \left[\overline{y}+b\left(P-\hat{P}\right)\right]\frac{P}{\hat{P}} \tag{17.14}$$

where b is the sample regression coefficient.

Singh, Chauhan, Sawan, and Smarandache (2008) also suggested the following estimator t_9 as

$$t_9 = \frac{\overline{y}+b\left(P-\hat{P}\right)}{\mathrm{m}_1\hat{P}+m_2}(\mathrm{m}_1P+m_2) \tag{17.15}$$

where $m_1(\neq 0)$, m_2 are either real numbers or the functions of the known parameters of the attribute (see Singh and Kumar, 2011).

Abd-Elfattah, El-Sherpieny, Mohamed, and Abdou (2010) proposed an estimator t_{10} as

$$t_{10} = m_1\frac{\overline{y}+b\left(P-\hat{P}\right)}{\hat{P}}+m_2\frac{\overline{y}+b\left(P-\hat{P}\right)}{\hat{P}+B_2(\phi)}(P+B_2(\phi)) \tag{17.16}$$

where m_1 and m_2 are weights that satisfy the condition $m_1+m_2=1$ and $B_2(\phi)$ is the population coefficient of kurtosis of auxiliary attribute.

Adapting Rao's (1991) idea, Grover and Kaur (2011) defined the following estimator t_{11}

$$t_{11} = \alpha\bar{y} + \beta\left(P - \hat{P}\right) \tag{17.17}$$

where α and β are suitably chosen constants. The optimum MSE, up to first order of approximation, of this estimator is given by

$$\text{min. MSE}(t_{11}) = \frac{\left(\frac{1-f}{n}\right)\overline{Y}^2 C_y^2\left(1 - \rho_{pb}^2\right)}{1 + \left(\frac{1-\mathrm{f}}{\mathrm{n}}\right)\left(1 - \rho_{pb}^2\right)} \tag{17.18}$$

Grover and Kaur (2011) suggested the following exponential type estimator of $\overline{Y}$

$$t_{12} = \left[\alpha\bar{y} + \beta\left(P - \hat{P}\right)\right]\exp\left(\frac{P - \hat{P}}{P + \hat{P}}\right) \tag{17.19}$$

where α and β are any constants and their values are suitably chosen. The optimum values of α and β are respectively

$$\alpha_{\text{opt}} = \frac{-C_p^2\left[2 - \frac{\left(\frac{1-f}{n}\right)M_2}{2} + \left(\frac{1-f}{n}\right)\left(\frac{C_p^2}{2} - \rho_{pb}C_yC_p\right)\right]}{2\left[\frac{1-f}{n}M_2^2 - C_p^2\left(1 + \frac{1-f}{n}M_1\right)\right]}$$

and

$$\beta_{\text{opt}} = \frac{\overline{Y}\left[M_2\left\{2 + \frac{\theta M_2}{2} + \frac{\theta}{2}\left(\frac{C_p^2}{2} - \rho_{pb}C_yC_p\right)\right\} - (1 + \theta M_1)C_p^2\right]}{2P\left[fM_2^2 - C_p^2(1 + \theta M_1)\right]}$$

where $M_1 = C_p^2 + C_y^2 - 2\rho_{pb}C_yC_p$ and $M_2 = C_p^2 - \rho_{pb}C_yC_p$.

On substituting these optimum values of α and β, we get the minimum MSE of the estimator t_{12} as

$$\text{min. MSE}(t_{12}) = \frac{\theta\overline{Y}^2 C_y^2\left(1 - \rho_{pb}^2\right)}{1 + \theta\mathrm{C}_\mathrm{y}^2\left(1 - \rho_{pb}^2\right)} - \frac{\theta^2\overline{Y}^2 C_p^2\left[4C_y^2\left(1 - \rho_{pb}^2\right) + \frac{C_p^2}{4}\right]}{16\left[1 + \mathrm{f}\mathrm{C}_\mathrm{y}^2\left(1 - \rho_{pb}^2\right)\right]}. \tag{17.20}$$

Koyuncu (2012) suggested an estimator t_{13} as

$$t_{13} = \left[w_1\bar{y} + w_2\left(P - \hat{P}\right)\right]\left(\frac{\eta\mathrm{P} + \lambda}{\eta\hat{\mathrm{P}} + \lambda}\right) \tag{17.21}$$

where η and λ are either real numbers or functions of the known parameter associated with an auxiliary attribute.

Koyuncu (2012) also proposed an improved estimator t_{14} as

$$t_{14} = \left[w_1\bar{y} + w_2\left(\frac{\hat{P}}{P}\right)^{\gamma}\right]\exp\left(\frac{\eta\left(P - \hat{P}\right)}{\eta\left(\mathrm{P} + \hat{\mathrm{P}}\right) + 2\lambda}\right) \tag{17.22}$$

where γ is a suitable real number, and w_1 and w_2 are suitable weights.

Singh and Solanki (2012) suggested estimator t_{15} as

$$t_{15} = \bar{y}\left[d_1 + d_2\left(P - \hat{P}\right)\right]\left(\frac{\psi P + \delta\eta}{\psi\hat{P} + \delta\eta}\right)^{\alpha} \tag{17.23}$$

where ψ and η are either real numbers or function of known parameters of the auxiliary attribute. The scalar α takes values -1 and $+1$, δ is an integer which takes values $+1$ and -1 for designing the estimators such that $(\psi\hat{P} + \delta\eta)$ and $(\psi P + \delta\eta)$ are nonnegative and (d_1, d_2) are suitably chosen constants such that the sum of the constants (d_1, d_2) need not be unity. It was shown that the proposed estimator performs better than many existing estimators.

Sharma, Verma, Sanaullah, and Singh (2013) studied some exponential ratio-product type estimators using information on auxiliary attributes. They studied the properties of the estimators under second order of approximation.

Singh, Kumar, and Singh (2013) suggested a family of ratio estimators for estimating population mean $\overline{Y}$ as

$$t_{16} = \alpha_1\bar{y} + \alpha_2\bar{y}\left(\frac{m_1 P + m_2}{m_1\hat{P} + m_2}\right)^{\alpha} \tag{17.24}$$

where m_1 and m_2 are either the real number or the functions of the parameters of the attribute and α_1 and α_2 are real constants to be determined.

Singh, Kumar, and Singh (2013) suggested another family of estimators as

$$t_{17} = \left(w_1\bar{y} + w_2\left(P - \hat{P}\right)\right)\left(\frac{aP + b}{a\hat{P} + b}\right)^{\alpha}\exp\left\{\frac{(aP + b) - \left(a\hat{P} + b\right)}{(aP + b) + \left(a\hat{P} + b\right)}\right\}^{\alpha} \tag{17.25}$$

where w_1 and w_2 are constants whose sum is not necessarily equal to one.

Sharma, Singh, and Kim (2013) proposed the following four estimators for estimating $\overline{Y}$ as

$$t_{18} = (1 - \alpha)\bar{y} + \alpha\bar{y}\frac{P}{\hat{P}} \tag{17.26}$$

where α is any real constant.

$$t_{19} = \bar{y}\left(\frac{P}{\beta P + (1 - \beta)\hat{P}}\right)^{g} \tag{17.27}$$

where g and β are any real constants.

$$t_{20} = \bar{y}\left(2 - \left(\frac{\hat{P}}{P}\right)^{w}\right) \tag{17.28}$$

where w is a constant.

$$t_{21} = \bar{y}\left(2 - \left\{\left(\frac{\hat{P}}{P}\right)^{\lambda}\exp\left(\frac{\delta\left(\hat{P} - P\right)}{\left(\hat{P} + P\right)}\right)\right\}\right) \tag{17.29}$$

where λ is a constant.

Sharma, Singh, and Kim (2013) studied the properties of these four estimators under a second order of approximation.

Barak and Barak (2013) proposed the following three unbiased estimators

$$t_{22} = \bar{y} + \left(\frac{P}{\hat{P}}\right) - 1 \tag{17.30}$$

$$t_{23} = \bar{y} - e^{(\hat{P}-P)} + 1 \tag{17.31}$$

and

$$t_{24} = \bar{y} - e^{(P-\hat{P})} + 1 \tag{17.32}$$

Yadav and Adewara (2013) suggested the following exponential estimators for estimating $\overline{Y}$

$$t_{25} = k\bar{y}\exp\left(\frac{P-\hat{P}}{P+\hat{P}}\right) \tag{17.33}$$

$$t_{26} = k\bar{y}\exp\left(\frac{\hat{P}-P}{\hat{P}+P}\right) \tag{17.34}$$

where k is any constant.

$$t_{27} = \bar{y}\exp\left(\frac{\hat{P}^{*}-P}{\hat{P}^{*}+P}\right) \tag{17.35}$$

where $\hat{P}^{*} = (1+g)P - g\hat{P}$ and $g = \frac{n}{N-n}$.

$$t_{28} = \bar{y}\left[\alpha\exp\left(\frac{P-\hat{P}}{P+\hat{P}}\right) + (1-\alpha)\exp\left(\frac{\hat{P}^{*}-P}{\hat{P}^{*}+P}\right)\right] \tag{17.36}$$

where α is a real constant.

Malik and Singh (2015) proposed a class of estimators for the population mean $\overline{Y}$ in double sampling as

$$t_{29} = \bar{y}\left[g_1 + g_2\left(\hat{P}' - \hat{P}\right)\right]\left\{\frac{w\hat{P}'+\eta}{w\hat{P}+\eta}\right\}^{\alpha}\exp\left\{\frac{\left(w\hat{P}'+\eta\right)-\left(w\hat{P}+\eta\right)}{\left(w\hat{P}'+\eta\right)+\left(w\hat{P}+\eta\right)}\right\}^{\beta} \tag{17.37}$$

where g_1 and g_2 are suitably chosen constants whose sum is not necessarily equal to unity and (w, η) are either real numbers or function of known parameters of the auxiliary attribute.

Saini and Kumar (2015) proposed a new exponential type product estimator as

$$t_{30} = \bar{y} - k\left(t_{30}^{*} - 1\right) \tag{17.38}$$

where k is any constant and

$$t_{30}^{*} = \exp\left[\frac{NP - n\hat{P}}{N-n} - P\right]$$

They have shown that their proposed estimator t_{30} is always more efficient than the exponential type product estimator (Bahl and Tuteja, 1991) and product estimator (Naik and Gupta, 1996).

17.3 ESTIMATION OF POPULATION MEAN USING TWO (OR MORE) AUXILIARY ATTRIBUTE INFORMATION

The regression estimator of $\overline{Y}$ based on two auxiliary attributes, is given by

$$t = \bar{y} + b_1\left(P_1 - \hat{P}_1\right) + b_2\left(P_2 - \hat{P}_2\right), \tag{17.39}$$

where $b_j = \frac{S_{y\varphi_j}}{S_{\varphi_j}^2}$ for $j = 1,2$.

Malik and Singh (2013a) suggested an improved estimator of $\overline{Y}$ using two auxiliary attributes and using point biserial and Phi-correlation given by

$$t=\bar{y}\exp\left(\frac{P_1-\hat{P}_1}{P_1+\hat{P}_1}\right)^{\gamma_1}\exp\left(\frac{P_2-\hat{P}_2}{P_2+\hat{P}_2}\right)^{\gamma_2}+b_1\left(P_1-\hat{P}_1\right)+b_2\left(P_2-\hat{P}_2\right) \tag{17.40}$$

where γ_1 and γ_2 are two unknown constants.

Malik and Singh (2013b) proposed following three estimators using two auxiliary attributes

$$t=\bar{y}\left(\frac{P_1}{\hat{P}_1}\right)^{\alpha_1}\left(\frac{\mathrm{P}_2}{\hat{\mathrm{P}}_2}\right)^{\alpha_2} \tag{17.41}$$

$$t=\bar{y}\exp\left(\frac{P_1-\hat{P}_1}{P_1+\hat{P}_1}\right)^{\beta_1}\exp\left(\frac{\hat{P}_2-P_2}{\hat{P}_2+P_2}\right)^{\beta_2} \tag{17.42}$$

and

$$t=w_0\bar{y}+w_1\bar{y}\left(\frac{P_1}{\hat{P}_1}\right)^{\alpha_1}\left(\frac{P_2}{\hat{P}_2}\right)^{\alpha_2}+w_2\bar{y}\exp\left(\frac{P_1-\hat{P}_1}{P_1+\hat{P}_1}\right)^{\beta_1}\exp\left(\frac{\hat{P}_2-P_2}{\hat{P}_2+P_2}\right)^{\beta_2} \tag{17.43}$$

where α_1, α_2, β_1, and β_2 are real constants and $w_i(i=0,1,2)$ are suitably chosen constants.

Singh and Malik (2013) suggested a class of estimators of the form

$$t=\sum_{i=0}^{3}w_i t_i \quad (\in H) \tag{17.44}$$

such that $\sum_{i=0}^{3}w_i=1$ and $w_i\in\mathfrak{R}$, where $t_0=\bar{y}$, $t_1=\bar{y}\left(\frac{P_1}{\hat{P}_1}\right)$, $t_2=\bar{y}\left(\frac{\hat{P}_2}{P_2}\right)$ and $t_3=\bar{y}\left(\frac{P_1}{\hat{P}_1}\right)\left(\frac{\hat{P}_2}{P_2}\right)$

Following Olkin (1958), Verma, Singh, and Florentin (2013) proposed an estimator

$$t=\bar{y}\left[w_1\frac{P_1}{\hat{P}_1}+w_2\frac{P_2}{\hat{P}_2}\right] \tag{17.45}$$

where w_1 and w_2 are constants, such that $w_1+w_2=1$.

Verma, Singh, and Florentin (2013) proposed another estimator t as

$$t=\left[k\bar{y}+k_1\left(P_1-\hat{P}_1\right)\right]\exp\left[\frac{P_2-\hat{P}_2}{P_2+\hat{P}_2}\right] \tag{17.46}$$

where k and k_1 are constants.

They also proposed the following estimator

$$t=\bar{y}+k_2\left(P_1-\hat{P}_1\right)+k_3\left(P_2-\hat{P}_2\right) \tag{17.47}$$

where k_2 and k_3 are constants.

Haq and Shabbir (2014) proposed some improved estimators using two auxiliary attributes. They proposed a chain-ratio-product type estimator of $\overline{Y}$ as

$$t* = \frac{\bar{y}}{4}\left(\frac{P_1}{\hat{P}_1}+\frac{\hat{P}_1}{P_1}\right)\left(\frac{P_2}{\hat{P}_2}+\frac{\hat{P}_2}{P_2}\right). \tag{17.48}$$

Following Rao (1991), Haq and Shabbir (2014) proposed a difference-type estimator of $\overline{Y}$ as

$$t = k_1\bar{y} + k_2\left(P_1 - \hat{P}_1\right) + k_3\left(P_2 - \hat{P}_2\right) \tag{17.49}$$

where k_1, k_2 and k_3 are real constants.

Haq and Shabbir (2014) also proposed an improved chain-ratio-product-difference type estimator of $\overline{Y}$ as

$$t = w_1 t* + w_2\left(P_1 - \hat{P}_1\right) + w_3\left(P_2 - \hat{P}_2\right) \tag{17.50}$$

where w_1, w_2 and w_3 are suitably chosen constants to be determined.

Haq and Shabbir (2014) have shown that the estimators proposed by them are better than other estimators considered in the paper in SRS and also in a two-phase scheme.

Singh, Malik, Adewara, and Florentin (2014) proposed some multivariate ratio estimators with known population proportion of two auxiliary characteristics for finite population. Following Olkin (1958), they proposed an estimator as

$$t_{ap} = \sum_{i=1}^{k} w_i r_i P_i \tag{17.51}$$

where

i. $w_i{'}s$ are weights such that $\sum_{i=1}^{k} w_i = 1$;

ii. $P_i{'}s$ are the proportion of the auxiliary attribute and assumed to be known; and

iii. $r_i = \frac{\bar{y}}{\hat{P}_i}$, $(i = 1,2,.\ldots\ldots,k)$, $\bar{y}$ is the sample mean of the study variable y and $\hat{P}_i$ is the proportion of auxiliary attributes P_i based on a SRS of size n drawn WOR from a population of size N.

Following Naik and Gupta (1996) and Singh, Cauhan, Sawan, and Smarandache (2007), they proposed another estimator t_s as

$$t_s = \prod_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{r_i} P_i \tag{17.52}$$

Singh, Malik, Adewara, and Florentin (2014) also proposed two alternative estimators based on geometric mean and harmonic mean, respectively, as

$$t_{gp} = \prod_{i=1}^{k} (r_i P_i)^{w_i} \tag{17.53}$$

and

$$t_{hp} = \left(\sum_{i=1}^{k} \frac{w_i}{r_i P_i}\right)^{-1} \tag{17.54}$$

such that $\sum_{i=1}^{k} w_i = 1$.

They have shown that the MSEs of estimators based on geometric, harmonic mean, and Verma, Singh and Florentin (2013) type estimator are the same. However, the bias of the ratio-type estimator based on harmonic mean is least.

Kungu and Odongo (2014) proposed a generalized estimator for estimating population mean of study variable y with the use of multiauxiliary attributes, given by

$$t_{rp} = \bar{y}\left(\frac{P_1}{p_1}\right)^{\alpha_1}\left(\frac{P_2}{p_2}\right)^{\alpha_2} - \left(\frac{P_k}{p_k}\right)^{\alpha_k}\left(\frac{p_{k+1}}{P_{k+1}}\right)^{\beta_{k+1}}\left(\frac{p_{k+2}}{P_{k+2}}\right)^{\beta_{k+2}} - \left(\frac{p_q}{P_q}\right)^{\beta_q} \tag{17.55}$$

where $\alpha's$ and $\beta's$ are arbitrary constants.

Sharma, Verma, and Singh (2015) proposed an improved family of estimators for estimating $\overline{Y}$ when information on two auxiliary attributes is available, as:

$$t_N = \bar{y}\left[w_1\left(\frac{p_1}{P_1}\right)^{\delta}\exp\left\{\frac{\eta_1(P_1 - p_1)}{\eta_1(P_1 + p_1) + 2\lambda_1}\right\} + w_2\left(\frac{p_2}{P_2}\right)^{\beta}\exp\left\{\frac{\eta_2(P_2 - p_2)}{\eta_2(P_2 + p_2) + 2\lambda_2}\right\}\right] \tag{17.56}$$

where δ and β are constants that can takes values (0, 1, -1) for designing different estimators.

η_1, λ_1, η_2, and λ_2 are either real numbers or the function of the known parameters. w_1 and w_2 are suitably chosen constants to be determined such that the MSE of the class of estimator t_N is minimum.

Saghir and Shabbir (2012) proposed an exponential ratio type estimator in stratified sampling as:

$$t_{ss} = \bar{y}^{*}_{st}\exp\left(\frac{P_1 - p_{1st}}{P_1 + (a-1)p_{1st}}\right)\exp\left(\frac{P_2 - p_{2st}}{P_2 + (b-1)p_{2st}}\right) \tag{17.57}$$

Malik and Singh (2013c) proposed an estimator in stratified sampling as:

$$t_{ms} = \bar{y}^{*}_{st}\exp\left(\frac{P_1 - p_{1st}}{P_1 + p_{1st}}\right)^{\alpha_1}\exp\left(\frac{P_2 - p_{2st}}{P_2 + p_{2st}}\right)^{\alpha_2} + b_1(P_1 - p_{1st}) + b_2(P_2 - p_{2st}) \tag{17.58}$$

where α_1 and α_2 are real constants.

17.4 CONCLUSION

In this chapter we have reviewed the work of the authors on the use of auxiliary attributes in construction of improved estimators for estimating unknown population mean. We have incorporated the work carried out using single and two auxiliary attributes. We hope this work will be helpful for researchers who are working in the construction of improved estimators using auxiliary information.

ACKNOWLEDGMENT

The authors are thankful to the learned referees for giving suggestions which helped in improving the presentation of this chapter.

REFERENCES

Abd-Elfattah, A.M., El-Sherpieny, E.A., Mohamed, S.M., Abdou, O.F., 2010. Improvement in estimating the population mean in simple random sampling using information on auxiliary attribute. Appl. Math. Comput. 215, 4198–4202.

Bahl, S., Tuteja, R.K., 1991. Ratio and product type exponential estimators. J. Inf. Optim. Sci. 12 (1), 159–164.

Barak, M.S., Barak, A.S., 2013. Some unbiased estimators for estimating the population mean in simple random sampling using information on auxiliary attribute. Int. J. Sci. Res. 4 (1), 2773–2776.

Grover, L.K., Kaur, P., 2011. An improved exponential estimator of finite population mean in simple random sampling using an auxiliary attribute. Appl. Math. Comput. 218 (7), 3093–3099.

Haq, A., Shabbir, J., 2014. An improved estimator of finite population mean when using two auxiliary attribute. Appl. Math. Comput. 241, 14–24.

Jhajj, H.S., Sharma, M.K., Grover, L.K., 2006. A family of estimators of population mean using information on auxiliary attribute. Pak. J. Stat. 22 (1), 43–50.

Koyuncu, N., 2012. Efficient estimators of population mean using auxiliary attributes. Appl. Math. Comput 218 (22), 10900–10905.

Kungu, J., Odongo, L., 2014. Ratio-cum-product estimator using multiple auxiliary attributes in single phase sampling. OJS 4, 239–245.

Malik, S., Singh, R., 2013a. A family of estimators of population mean using information on point bi-serial and phi correlation coefficient. Int. J. Stat. Econ. 10 (1), 75–89.

Malik, S., Singh, R., 2013b. An improved estimator using two auxiliary attributes. Appl. Math. Comput. 219, 10983–10986.

Malik, S., Singh, R., 2013c. Dual to ratio cum product estimators of finite population mean using auxiliary attribute(s) in stratified random sampling. World Appl. Sci. J. 28 (9), 1193–1198.

Malik, S., Singh, R., 2015. Estimation of population mean using information on auxiliary attribute in two-phase sampling. Appl. Math. Comput. 261, 114–118.

Naik, V.D., Gupta, P.C., 1996. A note on estimation of mean with known population proportion of an auxiliary character. J. Indian Soc. Agric. Stat. 48 (2), 151–158.

Olkin, I., 1958. Multivariate ratio estimation for finite populations. Biometrika 45, 154–165.

Rao, T.J., 1991. On certain methods of improving ratio and regression estimators. Commun. Stat.: Theory Methods 20 (10), 3325–3340.

Saghir, A., Shabbir, J., 2012. Estimation of finite population mean in stratified random sampling using auxiliary attribute(s) under non response. Pak. J. Stat. Oper. Res. 8 (1), 73–82.

Saini, M., Kumar, A., 2015. Exponential type product estimator for finite population mean with information on auxiliary attribute. Appl. Appl. Math. 10 (1), 106–113.

Shabbir, J., Gupta, S., 2007. On estimating the finite population mean with known population proportion of an auxiliary variable. Pak. J. Stat. 23 (1), 1–9.

Sharma, P., Singh, R., Kim, J.M., 2013. Study of some improved ratio type estimators using information on auxiliary attributes under second order approximation. J. Sci. Res. 57, 138–146.

Sharma, P., Verma, H., Sanaullah, A., Singh, R., 2013. Some exponential ratio-product type estimators using information on auxiliary attributes under second order approximation. Int. J. Stat. Econ. 12 (3), 58–66.

Sharma, P., Verma, H.K., Singh, R., 2015. An efficient class of estimators using two auxiliary attributes. Chil. J. Stat. 6 (2), 59–68.

Singh, H.P., Solanki, R.S., 2012. Improved estimation of population mean in simple random sampling using information on auxiliary attribute. Appl. Math. Comput. 218, 7798–7812.

Singh, R., Kumar, M., 2011. A note on transformations on auxiliary variable in survey sampling. Mod. Assist. Stat. Appl. 6 (1), 17–19.

Singh, R., Malik, S., 2013. A family of estimators of population mean using information on two auxiliary attribute. World Appl. Sci. J. 23 (7), 950–955.

Singh, R., Cauhan, P., Sawan, N., Smarandache, F., 2007. Auxiliary Information and A Priori Values in Construction of Improved Estimators. Renaissance High Press, USA.

Singh, R., Chauhan, P., Sawan, N., Smarandache, F., 2008. Ratio estimators in simple random sampling using information on auxiliary attribute. Pak. J. Stat. Oper. Res. 4 (1), 47–53.

Singh, R., Kumar, M., Singh, H.P., 2013. On estimation of population mean using information on auxiliary attribute. Pak. J. Stat. Oper. Res. 9 (4), 361–369.

Singh, R., Malik, S., Adewara, A.A., Florentin, S., 2014. Multivariate ratio estimation with known population proportion of two auxiliary characters for finite population. In: Smarandache, F. (Ed.), Collected Papers Vol V. Europa Nova, Brussels, pp. 231–238.

Verma, H., Singh, R., Florentin, S., 2013. Some Improved Estimators of Population Mean Using Information on Two Auxiliary Attributes. Educational Publishing & Journal of Matter Regularity, Beijing.

Yadav, S.K., Adewara, A.A., 2013. On improved estimation of population mean using qualitative auxiliary information. Math. Theory Model. 3 (11), 42–50.

CHAPTER

RATIO AND PRODUCT TYPE EXPONENTIAL ESTIMATORS FOR POPULATION MEAN USING RANKED SET SAMPLING 18

Gajendra K. Vishwakarma[1], **Sayed Mohammed Zeeshan**[1] **and Carlos N. Bouza-Herrera**[2]

[1]*Department of Applied Mathematics, Indian Institute of Technology (ISM), Dhanbad, Jharkhand, India* [2]*Faculty of Mathematics and Computation, University of Havana, Havana, Cuba*

18.1 INTRODUCTION

Ranked set sampling (RSS) is a method of sampling which provides more structure to the collected sample items and increases the amount of information present in the sample. The method of RSS was first envisaged by McIntyre (1952) as a cost-efficient substitute to simple random sampling (SRS) for those circumstances where measurements are inconvenient or expensive to obtain but (judgment) ranking of units according to the variable of interests, say, Y, is comparatively easy and cheap. It is known that the estimate of the population mean using RSS is more efficient than that obtained using SRS. McIntyre (1952) and Takahasi and Wakimoto (1968) considered perfect ranking of the elements, that is, there are no errors in ranking the elements. Yet, in most circumstances, the ranking may not be done perfectly. Dell and Clutter (1972) demonstrated that the mean using the RSS is an unbiased estimator of the population mean, whether or not there are errors in ranking. Stokes (1977) considered the case where the ranking is done on the basis of a concomitant (auxiliary) variable X instead of judgment. We would expect the variable of interest will be highly correlated with the concomitant (auxiliary) variable. Stokes (1980) showed that the estimator of the variance based on RSS data is an asymptotically unbiased estimator of the population variance. Samawi and Muttlak (1996) deal the problem of estimating the population ratio of the two variables Y and X using the RSS procedure. In addition, RSS has been investigated by many researchers, such as Al-Saleh and Al-Omari (2002), Wolfe (2004), Mandowara and Mehta (2013), and Al-Omari and Bouza (2015).

RSS has many statistical applications in agriculture, biology, environmental science, medical science, etc. Let m random samples of size m bivariate units each and rank the bivariate units within each sample with respect to the auxiliary variate X. Next, select the ith smallest auxiliary variate X from the ith sample for $i = 1, 2, 3, \ldots m$ for actual measurement of the associated variate of interest Y with it. In this way, a total number of m measured bivariate units are obtained, one from each sample. The cycle may be repeated r times to get a sample of size $n = rm$ bivariate units. These $n = rm$ units build the RSS data. Note that we assume that the ranking of the variate X will

Ranked Set Sampling. DOI: https://doi.org/10.1016/B978-0-12-815044-3.00018-6

be perfect, while the ranking of the variate Y will be with errors, or at worst of a random order if the correlation between $Y$and$X$ is close to zero. Also, note that in RSS, rm^2 elements are identified, but only rm of them are quantified. So, comparing this sample with a simple random sample of size rm is reasonable. For more details about RSS, see Kaur et al. (1995).

We assume that ranking on the auxiliary variate, X, is perfect. The associated variate, Y, is then with an error unless the relation between X and Y is perfect. Let us denote $(X_{j(i)}, Y_{j[i]})$ as the pair of the ith order statistics of X and the associated element Y in the jth cycle. Then the ranked set sample is

$(X_{1(1),}Y_{1[1]})\ldots,(X_{1(m),}Y_{1[m]}),(X_{2(1),}Y_{2[1]}),\ldots,(X_{2(m),}Y_{2[m]}),.\,.\,., (X_{r(1),}Y_{r[1]}),\ldots,(X_{r(m),}Y_{r[m]}),$

To obtain biases and mean squared error, we consider

$$\left.\begin{aligned}
&T_{y(i)}=\left(\mu_{y(i)}-\mu_y\right), T_{x(i)}=\left(\mu_{x(i)}-\mu_x\right), T_{xy(i)}=\left(\mu_{x(i)}-\mu_x\right)\left(\mu_{y(i)}-\mu_y\right),\\
&\sigma^2_{y(i)}=E\left(Y_{(i)}-\mu_i\right)^2, \sigma^2_{x(i)}=E\left(X_{(i)}-\mu_i\right)^2,\\
&\sigma_{xy}=E\left(Y_{(i)}-\mu_y\right)\left(X_{(i)}-\mu_x\right),
\end{aligned}\right\}\tag{18.1}$$

and

$$\left.\begin{aligned}
&\sum_{i=1}^{n}T_{x(i)}=0,\quad \sum\nolimits_{i=1}^{n}T_{y(i)}=0,\\
&\sum_{i=1}^{n}\sigma^2_{x(i)}=n\sigma^2_x-\sum\nolimits_{i=1}^{n}T^2_{x(i)},\quad \sum\nolimits_{i=1}^{n}\sigma^2_{y(i)}=n\sigma^2_y-\sum\nolimits_{i=1}^{n}T^2_{y(i)},\\
&\sum\nolimits_{i=1}^{n}\sigma_{xy(i)}=n\sigma_{xy}-\sum\nolimits_{i=1}^{n}T_{xy(i)}.
\end{aligned}\right\}\tag{18.2}$$

The sample mean of each variate based on RSS data and using the results obtained in Dell and Clutter (1972) can be defined as follows:

$$\left.\begin{aligned}
\overline{X}_{(n)}&=\frac{1}{mr}\sum\nolimits_{j=1}^{r}\sum\nolimits_{i=1}^{m}X_{r(m),}\\
\overline{Y}_{[n]}&=\frac{1}{mr}\sum\nolimits_{j=1}^{r}\sum\nolimits_{i=1}^{m}Y_{r[m]}
\end{aligned}\right\}\tag{18.3}$$

with variance

$$\left.\begin{aligned}
\mathrm{Var}\left(\overline{X}\right)&=\frac{\sigma^2_x}{m}-\frac{1}{rm^2}\sum\nolimits_{i=1}^{m}T^2_{x(i)}\\
\mathrm{Var}\left(\overline{Y}\right)&=\frac{\sigma^2_y}{m}-\frac{1}{rm^2}\sum\nolimits_{i=1}^{m}T^2_{y[i]}
\end{aligned}\right\},\tag{18.4}$$

and covariance

$$\mathrm{Cov}\left(\overline{X},\overline{Y}\right)=\frac{\sigma_{xy}}{m}-\frac{1}{rm^2}\sum\nolimits_{i=1}^{m}T_{xy[i]}\tag{18.5}$$

Note that $\mu_{x(i)}$ and $\mu_{y(i)}$ depend on order statistics from some specific distributions and these values can be found in Arnold et al. (1992).

18.2 SOME EXISTING ESTIMATORS FOR THE POPULATION MEAN

For estimating the population mean $\overline{Y}$, the usual ratio and product estimators for $\overline{Y}$, respectively, as

$$\hat{\overline{Y}}_R = \overline{y}\frac{\overline{X}}{\overline{x}}, \tag{18.6}$$

$$\hat{\overline{Y}}_P = \overline{y}\frac{\overline{x}}{\overline{X}}, \tag{18.7}$$

and their MSEs up to the first degree of approximation are

$$\text{MSE}\left(\hat{\overline{Y}}_R\right) = \frac{\overline{Y}^2}{n}\left[\left(C_x^2 + C_y^2 - 2\rho C_x C_y\right)\right], \tag{18.8}$$

$$\text{MSE}\left(\hat{\overline{Y}}_P\right) = \frac{\overline{Y}^2}{n}\left[\left(C_x^2 + C_y^2 + 2\rho C_x C_y\right)\right], \tag{18.9}$$

Samawi and Muttlak (1996) approached ratio and product estimators under RSS as

$$\hat{\overline{Y}}_R^{\text{rss}} = \overline{y}_{[n]}\frac{\overline{X}}{\overline{x}_{(n)}}, \tag{18.10}$$

$$\hat{\overline{Y}}_P^{\text{rss}} = \overline{y}_{[n]}\frac{\overline{x}_{(n)}}{\overline{X}}, \tag{18.11}$$

and drived their MSEs to the first degree approximation as

$$\text{MSE}(\widehat{\overline{Y}}_R^{\text{rss}}) = \frac{\overline{Y}^2}{m}\left[\left(C_x^2 + C_y^2 - 2\rho C_x C_y\right) - \frac{1}{rm}\left(\frac{\sum_{i=1}^m T_{x(i)}^2}{\mu_x^2} + \frac{\sum_{i=1}^m T_{y[i]}^2}{\mu_y^2} - 2\frac{\sum_{i=1}^m T_{xy[i]}}{\mu_x \mu_y}\right)\right] \tag{18.12}$$

$$\text{MSE}(\widehat{\overline{Y}}_P^{\text{rss}}) = \frac{\overline{Y}^2}{m}\left[\left(C_x^2 + C_y^2 + 2\rho C_x C_y\right) - \frac{1}{rm}\left(\frac{\sum_{i=1}^m T_{x(i)}^2}{\mu_x^2} + \frac{\sum_{i=1}^m T_{y[i]}^2}{\mu_y^2} + 2\frac{\sum_{i=1}^m T_{xy[i]}}{\mu_x \mu_y}\right)\right] \tag{18.13}$$

For estimating the population mean $\overline{Y}$, Bahl and Tuteja (1991) give the ratio and product type exponential estimators as

$$\hat{\overline{Y}}_{Re} = \overline{y}\exp\left[\frac{\overline{X} - \overline{x}}{\overline{X} + \overline{x}}\right], \tag{18.14}$$

$$\hat{\overline{Y}}_{Pe} = \overline{y}\exp\left[\frac{\overline{x} - \overline{X}}{\overline{x} + \overline{X}}\right], \tag{18.15}$$

and derived their MSEs to the first-degree approximation as

$$\text{MSE}\left(\hat{\overline{Y}}_{Re}\right) = \frac{\overline{Y}^2}{n}\left[\left(\frac{C_x^2}{4} + C_y^2 - \rho C_x C_y\right)\right], \tag{18.16}$$

$$\text{MSE}\left(\hat{\overline{Y}}_{Pe}\right) = \frac{\overline{Y}^2}{n}\left[\left(\frac{C_x^2}{4} + C_y^2 + \rho C_x C_y\right)\right], \tag{18.17}$$

18.3 PROPOSED ESTIMATORS FOR POPULATION MEAN

We define the following ratio and product type exponential estimators for $\overline{Y}$under RSS, respectively, as

$$\hat{\overline{Y}}^{\text{rss}}_{Re} = \overline{y}_{[n]}\exp\left[\frac{\overline{X} - \overline{x}_{(n)}}{\overline{X} + \overline{x}_{(n)}}\right], \tag{18.18}$$

$$\hat{\overline{Y}}^{rss}_{Pe} = \overline{y}_{[n]}\exp\left[\frac{x_{(n)} - \overline{X}}{x_{(n)} + \overline{X}}\right], \tag{18.19}$$

Here we have ranked the auxiliary variate and, thus, there is an induced rank in study variate. The induced rank on the study variate will be perfect if the correlation between the variate is perfect, otherwise it will be worse if there is no correlation (the worst case will not affect our problem since it has already been proven by Dell and Clutter (1972)). Therefore, the MSE of $\hat{\overline{Y}}^{\text{rss}}_{Re}$ and $\hat{\overline{Y}}^{\text{rss}}_{Pe}$ using bivariate Taylor series expansion is given as

$$\text{MSE}\left(\hat{\overline{Y}}^{\text{rss}}_{Re}\right) = \frac{\overline{Y}^2}{m}\left[\left(\frac{C_x^2}{4} + C_y^2 + \rho C_x C_y\right) - \frac{1}{mr}\left(\frac{\sum_{i=1}^m T^2_{x(i)}}{4\overline{X}^2} + \frac{\sum_{i=1}^m T^2_{y[i]}}{\overline{Y}^2} + \frac{\sum_{i=1}^m T_{xy[i]}}{\overline{XY}}\right)\right] \tag{18.20}$$

$$\text{MSE}\left(\hat{\overline{Y}}^{\text{rss}}_{Pe}\right) = \frac{\overline{Y}^2}{m}\left[\left(\frac{C_x^2}{4} + C_y^2 + \rho C_x C_y\right) - \frac{1}{mr}\left(\frac{\sum_{i=1}^m T^2_{x(i)}}{4\overline{X}^2} + \frac{\sum_{i=1}^m T^2_{y[i]}}{\overline{Y}^2} + \frac{\sum_{i=1}^m T_{xy[i]}}{\overline{XY}}\right)\right] \tag{18.21}$$

Preposition: Let $W_{x(i)} = \frac{\mu_{x(i)} - \mu_i}{\mu_i}$ and $W_{y[i]} = \frac{\mu_{y[]} - \mu_i}{\mu_i}$ and also using the result from Dell and Clutter (1972) the above equation may be written as

$$\begin{aligned}\text{MSE}\left(\hat{\overline{Y}}^{\text{rss}}_{\text{Re}}\right) &= \frac{\overline{Y}^2}{m}\left[\left(\frac{C_x^2}{4} + C_y^2 - \rho C_x C_y\right) - \frac{1}{mr}\left(\sum_{i=1}^{m}\frac{{W_{x(i)}}^2}{4} + \sum_{i=1}^{m}{W_{y[i]}}^2 - 2\sum_{i=1}^{m}\frac{W_{x(i)}}{2}W_{y[i]}\right)\right]\\ &= \frac{\overline{Y}^2}{m}\left[\left(\frac{C_x^2}{4} + C_y^2 - \rho C_x C_y\right) - \frac{1}{mr}\sum_{i=1}^{m}\left(\tfrac{W_{x(i)}}{2} - W_{y[i]}\right)^2\right]\\ &= \text{MSE}\left(\hat{\overline{Y}}_{\text{Re}}\right) - \frac{\overline{Y}^2}{m^2 r}\sum_{i=1}^{m}\left(\tfrac{W_{x(i)}}{2} - W_{y[i]}\right)^2\end{aligned}$$

It is clear that $\sum_{i=1}^m \left(\frac{W_{x(i)}}{2} - W_{y[i]}\right)^2$ is greater than zero. Hence

$$\text{MSE}\left(\hat{\overline{Y}}^{\text{rss}}_{Re}\right) \le \text{MSE}\left(\hat{\overline{Y}}_{Re}\right). \tag{18.22}$$

Also, it can be proved in similar ways that

$$\text{MSE}\left(\hat{\overline{Y}}^{\text{rss}}_{Pe}\right) \le \text{MSE}(\hat{\overline{Y}}_{Pe}). \tag{18.23}$$

18.3.1 GENERALIZED EXPONENTIAL ESTIMATORS USING RSS

We propose a ratio-cum-product type exponential estimators using RSS as

$$\hat{\overline{Y}}^{\text{rss}}_{G} = \overline{y}_{[n]}\exp\left[\frac{\left(\frac{\overline{X}}{\overline{x}_{(n)}}\right)^{\alpha} - 1}{\left(\frac{\overline{X}}{\overline{x}_{(n)}}\right)^{\alpha} + 1}\right], \tag{18.24}$$

where α is some suitable real number whose values make the minimum MSE of $\hat{\overline{Y}}_G^{\text{rss}}$. It can also be noticed that for $\alpha = 1$ and $\alpha = -1$ the above equation becomes Bahl and Tuteja (1991) usual ratio and product exponential estimators, respectively.

Again using Taylor series expansion we get the MSE of $\hat{\overline{Y}}_G^{\text{rss}}$ as

$$\text{MSE}\left(\hat{\overline{Y}}_G^{rss}\right) = \frac{\overline{Y}^2}{m}\left[\left(\frac{\alpha^2 C_x^2}{4} - \rho\alpha C_x C_y + C_y^2\right) - \frac{1}{mr}\left(\frac{\sum_{i=1}^{m} \alpha^2 T_{x(i)}^2}{4\overline{X}^2} + \frac{\sum_{i=1}^{m} \alpha T_{xy[i]}}{\overline{XY}} + \frac{\sum_{i=1}^{m} T_{y[i]}^2}{\overline{Y}^2}\right)\right] \tag{18.25}$$

In order to get the minimum MSE we differentiate the above Eq. (18.25) by α and equate it with 0. Hence we get optimum value of α as

$$\alpha_{\text{opt}} = 2\left(\frac{\rho C_x C_y - \frac{\sum_{i=1}^{m} T_{xy[i]}}{mr\overline{XY}}}{C_x^2 - \frac{\sum_{i=1}^{m} T_{x(i)}^2}{mr\overline{X}^2}}\right). \tag{18.26}$$

Using the above result we get the minimum MSE of $\hat{\overline{Y}}_G^{\text{rss}}$ as

$$\text{MSE}\left(\hat{\overline{Y}}_G^{rss}\right)_{\min} = \frac{\overline{Y}^2}{n}\left[C_Y^2 - \frac{\sum_{i=1}^{n} T_{y(i)}^2}{mr\overline{Y}^2} - \frac{\rho C_Y C_X - \frac{\sum_{i=1}^{n} T_{xy[i]}}{mr\overline{XY}}}{C_X^2 - \frac{\sum_{i=1}^{n} T_{x(i)}^2}{mr\overline{X}^2}}\right]. \tag{18.27}$$

18.4 A SIMULATION STUDY

To illustrate how one can gain an insight into the application or properties of the proposed estimator, a computer simulation was conducted. Bivariate random observations were generated from a bivariate normal distribution with parameters μ_y, μ_x, σ_x, σ_y and correlation coefficient ρ. The sampling method explained above is used to pick RSS data with sets of size m and after r repeated cycles to get an RSS of size mr. A sample of size mr bivariate units is randomly chosen from the population (we refer to these data as SRS data). The simulation was performed with $m = 3, 4, 5$ and with $r = 3$ and 6 (i.e., with total sample sizes of 9, 12, 15, 18, 24, and 30) for the RSS and SRS data sets. Here, we have ranked the auxiliary variate X which induces ranking in study variate Y (ranking on Y will be perfect if $\rho = 1$ or will be with errors in ranking if $\rho < 1$). Using R software we have conducted 5,000 replications for estimates of the means and mean square errors. The results of these simulations are summarized by the percentage relative efficiencies of the estimators using the formula.

$$\text{PRE}\left[*, \hat{\overline{Y}}_R^{\text{rss}}\right] = \frac{\text{MSE}\left(\hat{\overline{Y}}_R^{\text{rss}}\right)}{\text{MSE}(*)} \times 100 \tag{18.28}$$

$$\text{PRE}\left[*, \hat{\overline{Y}}_P^{\text{rss}}\right] = \frac{\text{MSE}\left(\hat{\overline{Y}}_P^{\text{rss}}\right)}{\text{MSE}(*)} \times 100 \tag{18.29}$$

where, $* = \hat{\overline{Y}}_{Re}^{\text{rss}}, \hat{\overline{Y}}_{Pe}^{\text{rss}}, \hat{\overline{Y}}_G^{\text{rss}}$.

Table 18.1 Percentage Relative Efficiencies (PREs) of Different Estimators of $\bar{Y}$ With Respect to $\bar{Y}_R$

r	*m*	$\rho = 0.5$			$\rho = 0.6$			$\rho = 0.7$		
		$\hat{\bar{Y}}_R^{rss}$	$\hat{\bar{Y}}_{Re}^{rss}$	$\hat{\bar{Y}}_G^{rss}$	$\hat{\bar{Y}}_R^{rss}$	$\hat{\bar{Y}}_{Re}^{rss}$	$\hat{\bar{Y}}_G^{rss}$	$\hat{\bar{Y}}_R^{rss}$	$\hat{\bar{Y}}_{Re}^{rss}$	$\hat{\bar{Y}}_G^{rss}$
3	3	100.00	180.96	478.69	100.00	176.60	195.60	100.00	86.48	132.17
6		100.00	195.98	197.16	100.00	69.32	127.10	100.00	300.47	360.13
3	4	100.00	88.08	102.37	100.00	159.51	292.034	100.00	134.18	575.96
6		100.00	137.81	148.54	100.00	79.89	177.04	100.00	425.84	133949.50
3	5	100.00	227.19	228.51	100.00	271.15	3401.16	100.00	95.60	209.80
6		100.00	459.04	2552.20	100.00	358.13	611.95	100.00	80.21	231.18
r	*m*	$\rho = 0.8$			$\rho = 0.9$			$\rho = 0.99$		
		$\hat{\bar{Y}}_R^{rss}$	$\hat{\bar{Y}}_{Re}^{rss}$	$\hat{\bar{Y}}_G^{rss}$	$\hat{\bar{Y}}_R^{rss}$	$\hat{\bar{Y}}_{Re}^{rss}$	$\hat{\bar{Y}}_G^{rss}$	$\hat{\bar{Y}}_R^{rss}$	$\hat{\bar{Y}}_{Re}^{rss}$	$\hat{\bar{Y}}_G^{rss}$
3	3	100.00	111.31	587.32	100.00	60.10	319.82	100.00	16.90	130.38
6		100.00	64.72	102.16	100.00	634.43	763.76	100.00	7.21	163.74
3	4	100.00	135.82	716.66	100.00	323.09	26611.89	100.00	106.97	878.19
6		100.00	74.83	147.67	100.00	48.06	230.65	100.00	211.98	25648.44
3	5	100.00	542.55	887.39	100.00	165.64	771.64	100.00	21.49	135.16
6		100.00	472.05	664.57	100.00	60.40	106.33	100.00	10.58	181.36

Table 18.2 Percentage Relative Efficiencies (PREs) of Different Estimators of $\bar{Y}$ With Respect to $\bar{Y}_P$

r	*m*	$\rho = -0.5$			$\rho = -0.6$			$\rho = -0.7$		
		$\hat{\bar{Y}}_P^{rss}$	$\hat{\bar{Y}}_{Pe}^{rss}$	$\hat{\bar{Y}}_G^{rss}$	$\hat{\bar{Y}}_P^{rss}$	$\hat{\bar{Y}}_{Pe}^{rss}$	$\hat{\bar{Y}}_G^{rss}$	$\hat{\bar{Y}}_P^{rss}$	$\hat{\bar{Y}}_{Pe}^{rss}$	$\hat{\bar{Y}}_G^{rss}$
3	3	100.00	100.70	115.48	100.00	165.76	617.15	100.00	198.68	2148.31
6		100.00	244.90	291.79	100.00	142.33	143.22	100.00	247.09	2241.68
3	4	100.00	96.36	101.24	100.00	68.88	100.66	100.00	449.56	533.31
6		100.00	367.50	39936.55	100.00	148.73	335.87	100.00	305.86	6376.77
3	5	100.00	265.84	739.76	100.00	82.69	100.41	100.00	261.39	1898.54
6		100.00	156.45	302.43	100.00	84.13	100.08	100.00	140.24	432.75
r	*m*	$\rho = -0.8$			$\rho = -0.9$			$\rho = -0.99$		
		$\hat{\bar{Y}}_P^{rss}$	$\hat{\bar{Y}}_{Pe}^{rss}$	$\hat{\bar{Y}}_G^{rss}$	$\hat{\bar{Y}}_P^{rss}$	$\hat{\bar{Y}}_{Pe}^{rss}$	$\hat{\bar{Y}}_G^{rss}$	$\hat{\bar{Y}}_P^{rss}$	$\hat{\bar{Y}}_{Pe}^{rss}$	$\hat{\bar{Y}}_G^{rss}$
3	3	100.00	225.97	2637.77	100.00	184.83	6134.04	100.00	85.47	6979.82
6		100.00	104.58	364.84	100.00	123.55	775.12	100.00	199.44	34980.65
3	4	100.00	293.40	16504.85	100.00	407.98	49398.26	100.00	175.07	9795.42
6		100.00	151.47	669.86	100.00	111.68	410.73	100.00	151.95	13768.70
3	5	100.00	304.81	15635.19	100.00	124.38	901.03	100.00	172.62	24835.13
6		100.00	165.10	998.60	100.00	214.16	6096.85	100.00	176.63	17588.92

18.5 CONCLUSIONS

It is observed from Table 18.1 that the PREs of the proposed ratio type exponential estimator using rank set sampling, $\hat{\bar{Y}}_{Re}^{\mathrm{rss}}$ and the proposed generalized exponential estimators using rank set sampling $\hat{\bar{Y}}_{G}^{\mathrm{rss}}$ are more efficient compared to the existing Samawi and Muttlak (1996) ratio estimator $\hat{\bar{Y}}_{R}^{\mathrm{rss}}$. Also from Table 18.2, it can be observed the PREs of the proposed product type exponential estimator using RSS, $\hat{\bar{Y}}_{Pe}^{\mathrm{rss}}$ and the proposed generalized exponential estimators using RSS, $\hat{\bar{Y}}_{G}^{\mathrm{rss}}$ are more efficient compared to the existing product estimator $\hat{\bar{Y}}_{Pe}^{\mathrm{rss}}$.

Finally, from Tables 18.1 and 18.2 we can conclude that the proposed estimators $\hat{\bar{Y}}_{Re}^{\mathrm{rss}}$, $\hat{\bar{Y}}_{Pe}^{\mathrm{rss}}$, and $\hat{\bar{Y}}_{\mathrm{G}}^{\mathrm{rss}}$ are more appropriate estimators than the existing popular estimators $\hat{\bar{Y}}_{R}$, $\hat{\bar{Y}}_{P}$, $\hat{\bar{Y}}_{\mathrm{R}}^{\mathrm{rss}}$, and $\hat{\bar{Y}}_{P}^{\mathrm{rss}}$ has appreciable efficiency.

REFERENCES

Al-Saleh, M.F., Al-Omari, A.I., 2002. Multistage ranked set sampling. J. Stat. Plan. Inference 102 (2), 273–286.

Al-Omari, A.I., Bouza, C.N., 2015. Ratio estimators of the population mean with missing values using ranked set sampling. Environmetrics 26.2, 67–76.

Arnold, B.C., Balakrishnan, N., Nagaraja, H.N., 1992. A first course in order statistics. SIAM 54.

Bahl, S., Tuteja, R.K., 1991. Ratio and product type exponential estimator. Inf. Optim. Sci. 12, 159–163.

Dell, T.R., Clutter, J.L., 1972. Ranked set sampling theory with order statistics background. Biometrics 28, 545–555.

Kaur, A., Patil, G.P., Sinha, A.K., Taillie, C., 1995. Ranked set sampling: an annotated bibliography. Environ. Ecol. Stat. 2, 25–54.

Mandowara, V.L., Mehta, N., 2013. Efficient generalized ratio-product type estimators for finite population mean with ranked set sampling. Aust. J. Stat. 42, 137–148.

McIntyre, G.A., 1952. A method for unbiased selective sampling, using ranked sets. Crop Pasture Sci. 3, 385–390.

Samawi, H.M., Muttlak, H.A., 1996. Estimation of ratio using rank set sampling. Biom. J. 38, 753–764.

Stokes, L., 1977. Ranked set sampling with concomitant variables. Commun. Stat.: Theory Methods 6, 1207–1211.

Stokes, L., 1980. Estimation of variance using judgment ordered ranked set samples. Biometrics 36, 35–42.

Takahasi, K., Wakimoto, K., 1968. On unbiased estimates of the population mean based on the sample stratified by means of ordering. Ann. Inst. Stat. Math. 20, 1–31.

Wolfe, D.A., 2004. Ranked set sampling: an approach to more efficient data collection. Stat. Sci. 19, 636–643.

CHAPTER

19 EXTROPY ESTIMATION IN RANKED SET SAMPLING WITH ITS APPLICATION IN TESTING UNIFORMITY

Ehsan Zamanzade[1] **and Mahdi Mahdizadeh**[2]

[1]Department of Statistics, University of Isfahan, Isfahan, Iran [2]Department of Statistics, Hakim Sabzevari University, Sabzevari, Iran

19.1 INTRODUCTION

There are situations in which obtaining exact values of sample units is difficult/expensive but ranking the sample units in a set of small size without referring to their precise values is easy/cheap. In such situations, ranked set sampling (RSS) serves as an efficient alternative to simple random sampling (SRS). RSS was firstly introduced by McIntyre (1952) when he realized that it is hard and time-consuming to obtain exact measurements of the mean pasture yield because it requires harvesting the corps, but an expert can fairly rank some adjacent plots using eye inspection. Although RSS was first motivated by an agricultural problem, it soon found applications in other fields, including forestry (Halls and Dell, 1966), environmental monitoring (Kvam, 2003), medicine (Chen et al., 2005; Zamanzade and Mahdizadeh, 2017a), biometrics (Mahdizadeh and Zamanzade, 2017a), reliability (Mahdizadeh and Zamanzade, 2017b), and educational studies (Wang et al., 2016).

To draw a ranked set sample, one first determines the set size H and a vector of in-stratum sample sizes $\mathbf{m} = (m_1, \ldots, m_H)$ such that $n = \sum_{h=1}^{H} m_h$ is the total sample size. We then draw a simple random sample of size nH from the population of interest and randomly partitions them into n sets each of size H. Each set of size H is then ranked from smallest to largest. The ranking process in this step is done using any cheap method which does not require referring to exact measurements of the sample units. From the first m_1 sets of size H, the sample units with smallest judgment rank are selected for actual measurements. From the next m_2 sets of size H, the sample units with judgment rank 2 are selected for quantification. This process is continued until the sample units with judgment rank H are selected for quantification from the last m_H sets of size H. The resulting ranked set sample is called unbalanced as the numbers of different judgment order statistics are not equal. A ranked set sample is called balanced if $m_1 = \ldots = m_H = m$, and the value of m in this case is called the cycle size.

Ranked Set Sampling. DOI: https://doi.org/10.1016/B978-0-12-815044-3.00019-8

A ranked set sample, in its general form, is denoted by $\left\{X_{[i]j}{:}i=1,\ldots,H;\ j=1,\ldots,m_i\right\}$, where $X_{[i]j}$ is the jth measured unit with judgment rank i. The term "*judgment rank*" and the subscript [.] are used to indicate that the ranking process is done without observing actual values of the units in the set and thus it may be inaccurate and contains errors (imperfect ranking). If the ranking is perfect, then subscript [.] is replaced with (.), and the resulting ranked set sample is denoted by $\left\{X_{(i)j}{:}i=1,\ldots,H;\ j=1,\ldots,m_i\right\}$. In this case, the distribution of $X_{(i)j}$ is the same as the distribution of the ith order statistic from a sample of size H. Throughout this chapter, we assume that the ranking process is consistent, which means that the same ranking process is applied to all sets of size H. Under a consistent ranking process, it can be simply shown that the following identity holds

$$F(t)=\frac{1}{H}\sum_{h=1}^{H}F_{[h]},$$

where $F_{[h]}$ is the cumulative distribution function (CDF) of a sample unit with judgment rank h.

19.2 EXTROPY ESTIMATION USING A RANKED SET SAMPLE

Let X be the variable of interest which is continuous with probability density function (pdf) f and cumulative distribution function (CDF) F. As a measure of uncertainty, entropy of the random variable of X is defined by Shannon (1948) as

$$H(f)=-\int_{-\infty}^{+\infty}\log(f(x))f(x)dx.$$

Due to numerous applications of entropy in statistics, information theory, and engineering, the problem of nonparametric estimation of $H(f)$ has received considerable attention. Vasicek (1976) was the first to propose estimating $H(f)$ using spacings of order statistics. His estimator is based on the fact that the entropy of a continuous random variable X with CDF F can be expressed as

$$H(f)=\int_{0}^{1}\log\left(\frac{d}{dp}F^{-1}(p)\right)dp.$$

He proposed estimating the entropy by using the empirical distribution function and applying a difference operator instead of a differential operator. Let $X_1,\ldots,X_n$ be a simple random sample of size n from the population of interest, with ordered values $X_{(1)}<\ldots<X_{(n)}$. Then Vasicek (1976)'s entropy estimator is given by

$$H_V^{\mathrm{srs}}=\frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{n}{2w}\left(X_{(i+w)}-X_{(i-w)}\right)\right\},$$

where $w\left(\le n/2\right)$ is an integer number called windows size, and $X_{(i)}=X_{(1)}$ for $i<1$, and $X_{(i)}=X_{(n)}$ for $i>n$.

Ebrahimi et al. (1994) improved Vasicek (1976)'s entropy estimator by assigning different weights to the observations at the boundaries. Their corrected entropy estimator is given by

$$H_E^{\mathrm{srs}}=\frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{n}{c_iw}\left(X_{(i+w)}-X_{(i-w)}\right)\right\},$$

where

$$c_i = \begin{cases} 1+\dfrac{i-1}{m} & i \leq m \\ 2 & m+1 \leq i \leq n-m, \\ 1+\dfrac{n-i}{m} & n-m+1 \leq i \leq n \end{cases}$$

W is the window size defined as before, and $X_{(i)} = X_{(1)}$ for $i < 1$ and $X_{(i)} = X_{(n)}$ for $i > n$.

As a complement dual of entropy, Lad et al. (2015) introduced a new measure which is called extropy, as follows:

$$J(X) = -\frac{1}{2}\int_{-\infty}^{+\infty} f^2(x)dx.$$

Lad et al. (2015) also investigated several interesting properties of extropy and resolved a fundamental question of Shannon's entropy measure. Qui (2017) provided some characteristic results, monotone properties as well as a lower bound for extropy of order statistics and record values. Qui and Jia (2018) used extropy for testing uniformity and showed that the resulting test has a good performance in comparison with its competitors in the literature including those tests based on entropy due to Zamanzade (2015).

By following the lines of Vasicek (1976) and Ebrahimi et al. (1994), Qui and Jia (2018) developed two estimators for extropy. Let $X_1, \ldots, X_n$ be a simple random sample of size n from the population of interest, with ordered values $X_{(1)} < \ldots < X_{(n)}$. Then the Qui and Jia (2018)'s extropy estimators are given by

$$J_{Q1}^{\text{srs}} = -\frac{1}{2n}\sum_{i=1}^{n}\frac{2w/n}{X_{(i+w)} - X_{(i-w)}}, J_{Q2}^{\text{srs}} = -\frac{1}{2n}\sum_{i=1}^{n}\frac{2c_i/n}{X_{(i+w)} - X_{(i-w)}},$$

where

$$c_i = \begin{cases} 1+\dfrac{i-1}{m} & i \leq m \\ 2 & m+1 \leq i \leq n-m, \\ 1+\dfrac{n-i}{m} & n-m+1 \leq i \leq n \end{cases}$$

W is the window size defined as before, and $X_{(i)} = X_{(1)}$ for $i < 1$ and $X_{(i)} = X_{(n)}$ for $i > n$.

Let $\left\{X_{[i]j}: i = 1, \ldots, H;\ j = 1, \ldots, m\right\}$ be a balanced ranked set sample of size $n = mH$ from the population of interest, with the corresponding ordered value $Z_1 < \ldots < Z_n$. Mahdizadeh and Arghami (2009) modified Vasicek's (1976) entropy estimator to be used in balanced RSS. Their proposed estimator has the form

$$H_V^{\text{rss}} = \frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{n}{2w}(Z_{i+w} - Z_{i-w})\right\},$$

where $Z_i = Z_1$ for $i < 1$, and $Z_i = Z_n$ for $i > n$.

Zamanzade and Mahdizadeh (2017b) developed some entropy estimators in balanced RSS using entropy estimators proposed by Ebrahimi et al. (1994). The new estimator is given by

$$H_E^{\text{rss}} = \frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{n}{c_i w}(Z_{i+w} - Z_{i-w})\right\},$$

where

$$c_i = \begin{cases} 1 + \dfrac{i-1}{m} & i \le m \\ 2 & m+1 \le i \le n-m \\ 1 + \dfrac{n-i}{m} & n-m+1 \le i \le n \end{cases},$$

$Z_i = Z_1$ for $i < 1$, and $Z_i = Z_n$ for $i > n$.

By following the lines of Mahdizadeh and Arghami (2009) and Mahdizadeh and Zamanzade (2017b), we can develop extropy estimators for RSS as follows:

$$J_{Q1}^{\text{rss}} = -\frac{1}{2n}\sum_{i=1}^{n}\frac{2w/n}{Z_{i+w}-Z_{i-w}}, J_{Q2}^{\text{rss}} = -\frac{1}{2n}\sum_{i=1}^{n}\frac{c_i w/n}{Z_{i+w}-Z_{i-w}},$$

where c_i is as defined before, $Z_i = Z_1$ for $i < 1$, and $Z_i = Z_n$ for $i > n$.

We conducted a simulation study to compare different extropy estimators in balanced RSS and SRS designs in terms of root of mean square error (RME). In doing so, we generated 100,000 samples of sizes $n = 10, 20, 30, 50$ from standard normal, standard uniform, and standard exponential distributions. The values of set size H are taken to be 2 and 5, and the value of window size w is selected according to Grzegorzewski and Wieczorkowski's (1999) heuristic formula, i.e., $w = \left[\sqrt{n} + 0.5\right]$, where $[x]$ is the integer part of x.

The imperfect rankings model that we utilize is the fraction-of-random-rankings model developed by Frey et al. (2007). Under this model, the distribution of ith judgment order statistic is a mixture of true ith order statistic and a random draw from the parent distribution, i.e.:

$$F_{[i]} = \lambda F_{(i)} + (1-\lambda)F,$$

where the parameter $\lambda \in [0,1]$ determines the quality of the ranking. The values of λ in this simulation study are selected from the set $\lambda \in \{0.5, 0.8, 1\}$, which corresponds to moderate, good, and perfect ranking, respectively.

Tables 19.1–19.3 show the estimated RMSEs and biases of the extropy estimators. Table 19.3 presents the results when the parent distribution is standard normal. It can be seen that the RSS estimators outperform their SRS counterparts. In both SRS and RSS schemes, J_{Q2} always works better than J_{Q1}. The performance of any extropy estimator improves if the total sample size (n), the set size (H), or the value of (λ) increases, provided that other factors are fixed.

The simulation results for standard exponential and standard uniform distributions are presented in Tables 19.2 and 19.3, respectively. The general trends are similar to those mentioned for Table 19.1.

19.3 EXTROPY-BASED TESTS OF UNIFORMITY IN RSS

In this section, we evaluate the performance of extropy-based test of uniformity in RSS and compare it with its SRS counterpart using Monte Carlo simulation. Testing uniformity is a very important problem from a practical point of view, because goodness-of-fit test can be expressed as a problem of testing uniformity. This follows from the probability integral transform theorem which

Table 19.1 Estimated RMSE and Bias of Different Extropy Estimators When Parent Distribution is Standard Normal Distribution With $J(f) = -0.141$

		RSS ($\lambda = 1$)				RSS ($\lambda = 0.8$)			
		J_{Q1}^{rss}		J_{Q2}^{rss}		J_{Q1}^{rss}		J_{Q2}^{rss}	
n	***H***	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**
10	2	0.13	− 0.10	0.07	− 0.04	0.13	− 0.10	0.07	− 0.04
	5	0.11	− 0.09	0.06	− 0.03	0.12	− 0.09	0.06	− 0.03
20	2	0.05	− 0.04	0.03	− 0.02	0.06	− 0.04	0.04	− 0.02
	5	0.05	−0.04	0.03	−0.01	0.05	−0.04	0.03	−0.01
30	2	0.04	−0.02	0.03	−0.01	0.04	−0.02	0.02	−0.01
	5	0.03	−0.02	0.02	−0.01	0.03	−0.02	0.02	−0.01
50	2	0.02	−0.01	0.02	0.00	0.02	−0.01	0.02	0.00
	5	0.02	−0.01	0.01	0.00	0.02	−0.01	0.02	0.00

		RSS ($\lambda = 0.5$)				SRS			
		J_{Q1}^{rss}		J_{Q2}^{rss}		J_{Q1}^{srs}		J_{Q2}^{srs}	
n	***H***	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**
10	2	0.13	−0.10	0.07	− 0.04	0.13	− 0.10	0.07	− 0.04
	5	0.12	− 0.10	0.07	− 0.04	0.13	− 0.10	0.07	− 0.04
20	2	0.06	− 0.04	0.03	− 0.02	0.06	− 0.04	0.04	− 0.02
	5	0.05	− 0.04	0.03	− 0.02	0.06	− 0.04	0.04	− 0.02
30	2	0.04	− 0.03	0.02	− 0.01	0.04	− 0.02	0.02	− 0.01
	5	0.04	− 0.02	0.02	− 0.01	0.04	− 0.02	0.02	− 0.01
50	2	0.02	− 0.01	0.02	0.00	0.02	− 0.01	0.02	0.00
	5	0.02	− 0.01	0.02	0.00	0.02	− 0.01	0.02	0.00

states that if the variable of interest X follows a continuous distribution with cumulative distribution function F, then $Y = F(X)$ follows a standard uniform distribution.

Qui and Jia (2018) showed that the standard uniform distribution maximizes the extropy $J(f)$ among all continuous distributions that possess a density function f and have a given support on (0,1). Based on this property, they then proposed the following test statistic for testing uniformity

$$T^{\text{srs}} = -J_{Q2}^{\text{srs}},$$

and they proposed the reject the null hypothesis of uniformity of large enough values of T^{srs}.

By following the lines of Qui and Jia (2018), one can also perform an extropy-based test of uniformity based on a ranked set sample using below test statistic

$$T^{\text{rss}} = -J_{Q2}^{\text{rss}},$$

and rejects the null hypothesis of uniformity of large enough values of T^{rss}.

Table 19.2 Estimated RMSE and Bias of Different Extropy Estimators When Parent Distribution is Standard Exponential Distribution With $J(f) = -0.25$

		RSS ($\lambda = 1$)				RSS ($\lambda = 0.8$)			
		J_{Q1}^{rss}		J_{Q2}^{rss}		J_{Q1}^{rss}		J_{Q2}^{rss}	
n	H	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias
10	2	0.24	0.15	0.12	− 0.04	0.27	− 0.16	0.14	− 0.04
	5	0.17	− 0.12	0.08	− 0.02	0.20	− 0.13	0.10	− 0.03
20	2	0.13	− 0.09	0.07	− 0.03	0.14	− 0.09	0.08	− 0.03
	5	0.11	− 0.08	0.05	− 0.02	0.12	− 0.08	0.07	− 0.02
30	2	0.09	− 0.07	0.06	− 0.02	0.10	− 0.07	0.06	− 0.02
	5	0.08	− 0.06	0.04	− 0.01	0.09	− 0.06	0.05	− 0.01
50	2	0.07	− 0.05	0.04	− 0.01	0.07	− 0.05	0.04	− 0.01
	5	0.05	− 0.04	0.03	− 0.01	0.06	− 0.04	0.04	− 0.01

		RSS ($\lambda = 0.5$)				SRS			
		J_{Q1}^{rss}		J_{Q2}^{rss}		J_{Q1}^{srs}		J_{Q2}^{srs}	
n	H	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias
10	2	0.27	− 0.16	0.15	− 0.05	0.26	− 0.16	0.15	− 0.05
	5	0.24	− 0.15	0.13	− 0.04	0.26	− 0.16	0.15	− 0.05
20	2	0.14	− 0.09	0.08	− 0.03	0.15	− 0.09	0.09	− 0.03
	5	0.13	− 0.09	0.08	− 0.02	0.15	− 0.09	0.09	− 0.03
30	2	0.10	− 0.07	0.06	− 0.02	0.11	− 0.07	0.07	− 0.02
	5	0.10	− 0.07	0.06	− 0.02	0.11	− 0.07	0.07	− 0.02
50	2	0.07	− 0.05	0.05	− 0.01	0.07	− 0.05	0.05	− 0.01
	5	0.07	0.05	0.04	− 0.01	0.07	− 0.05	0.05	− 0.01

Remark 1. We have not considered the test of uniformity based on J_{Q1}^{rss} in our comparison set, because we have observed that J_{Q2}^{rss} is uniformly better than J_{Q1}^{rss}.

In order to compare the power of different tests of uniformity, the following alternative distributions are considered

$$A_k{:}F(x) = 1 - (1-x)^k,\ 0 \le x \le 1, \qquad \text{(for } k = 1.5, 2)$$

$$B_k{:}F(x) = \begin{cases} 2x^k, & 0 \le x \le 0.5, \\ 1 - 2(1-x)^k, & 0.5 \le x \le 1, \end{cases} \qquad \text{(for } k = 1.5, 2, 3)$$

$$C_k{:}F(x) = \begin{cases} 0.5 - 2(0.5-x)^k, & 0 \le x \le 0.5, \\ 0.5 + 2(x-0.5)^k, & 0.5 \le x \le 1, \end{cases} \qquad \text{(for } k = 1.5, 2)$$

Table 19.3 Estimated RMSE and Bias of Different Extropy Estimators When Parent Distribution is Standard Uniform Distribution With $J(f) = -0.5$

		RSS ($\lambda = 1$)				**RSS ($\lambda = 0.8$)**			
		J_{Q1}^{rss}		J_{Q2}^{rss}		J_{Q1}^{rss}		J_{Q2}^{rss}	
n	***H***	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**
10	2	0.43	− 0.36	0.19	− 0.12	0.46	− 0.37	0.21	− 0.14
	5	0.37	− 0.31	0.15	− 0.09	0.40	− 0.34	0.17	− 0.11
20	2	0.24	− 0.21	0.11	− 0.08	0.24	− 0.22	0.11	− 0.08
	5	0.21	− 0.19	0.08	− 0.06	0.22	− 0.20	0.10	− 0.07
30	2	0.17	− 0.16	0.07	− 0.06	0.17	− 0.16	0.08	− 0.06
	5	0.16	− 0.15	0.06	− 0.05	0.16	− 0.15	0.07	− 0.05
50	2	0.12	− 0.11	0.05	− 0.04	0.12	− 0.11	0.05	− 0.04
	5	0.11	− 0.11	0.04	− 0.03	0.11	− 0.11	0.05	− 0.04
		RSS ($\lambda = 0.5$)				**SRS**			
		J_{Q1}^{rss}		J_{Q2}^{rss}		J_{Q1}^{srs}		J_{Q2}^{srs}	
n	***H***	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**
10	2	0.46	− 0.38	0.21	− 0.14	0.46	− 0.39	0.22	− 0.15
	5	0.44	− 0.37	0.20	− 0.13	0.46	− 0.39	0.22	− 0.15
20	2	0.25	− 0.22	0.12	− 0.09	0.25	− 0.22	0.12	− 0.09
	5	0.24	− 0.21	0.11	− 0.08	0.25	− 0.22	0.12	− 0.09
30	2	0.18	− 0.16	0.08	− 0.06	0.18	− 0.16	0.08	− 0.07
	5	0.17	− 0.16	0.08	− 0.06	0.18	− 0.16	0.08	− 0.07
50	2	0.12	− 0.12	0.05	− 0.04	0.12	− 0.12	0.05	− 0.04
	5	0.12	− 0.11	0.05	− 0.04	0.12	− 0.12	0.05	− 0.04

One can simply verify that as compared with uniform distribution, under alternative A, values closer to zero are more probable, whereas under alternative B, values near to 0.5 and under alternative C, values close to 0 and 1 are more probable.

Under each alternative, we have generated 10,000 RSS and SRS samples of sizes 10, 20, 30, and 50. The value of set size in RSS is taken from $H \in \{2, 5\}$ and the quality of ranking is controlled by fraction of random ranking as described in Section 19.2 with $\lambda \in \{1, 0.8, 0.5\}$ and the value of window size (m) is selected from Grzegorzewski and Wieczorkowski's (1999) heuristic formula, i.e., $w = \left[\sqrt{n} + 0.5\right]$, where $[x]$ is the integer part of x.

The power estimates of extropy-based tests of uniformity at significant level $\alpha = 0.1$ are presented in Table 19.4.

We observe from Table 19.4 that the extropy-based test of uniformity in RSS outperforms its counterpart in SRS. It is of interest to note that the power of T_{Q2}^{rss} increases if sample size (n),set size (H), or the value of (λ) increases, provided that other factors are fixed. This is consistent with what we observed in the previous section.

Table 19.4 Power Estimates of Extropy-Based Tests of Uniformity for n = 10, 20, 30, 50, and $\alpha = 0.1$ in SRS and RSS Designs

		RSS ($k = 2$)			RSS ($k = 5$)			SRS
ALt		$\lambda = 1$	$\lambda = 0.8$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 0.8$	$\lambda = 0.5$	
A1.5	$n = 10$	0.23	0.22	0.21	0.27	0.23	0.23	0.21
A2		0.45	0.44	0.41	0.58	0.50	0.45	0.42
B1.5		0.26	0.25	0.23	0.33	0.27	0.27	0.23
B2		0.49	0.49	0.45	0.64	0.55	0.51	0.45
B3		0.85	0.85	0.83	0.95	0.90	0.86	0.82
C1.5		0.12	0.12	0.12	0.12	0.12	0.13	0.12
C2		0.19	0.20	0.19	0.20	0.20	0.20	0.19
A1.5	$n = 20$	0.36	0.34	0.33	0.44	0.38	0.36	0.33
A2		0.78	0.74	0.71	0.90	0.81	0.76	0.70
B1.5		0.37	0.35	0.32	0.46	0.39	0.38	0.32
B2		0.77	0.73	0.71	0.87	0.81	0.77	0.71
B3		0.99	0.99	0.99	1.00	1.00	0.99	0.99
C1.5		0.24	0.24	0.23	0.26	0.24	0.25	0.23
C2		0.53	0.51	0.49	0.58	0.53	0.53	0.50
A1.5	$n = 30$	0.50	0.48	0.47	0.57	0.52	0.48	0.44
A2		0.93	0.91	0.89	0.98	0.95	0.92	0.87
B1.5		0.49	0.48	0.46	0.57	0.53	0.50	0.44
B2		0.91	0.90	0.90	0.97	0.95	0.92	0.88
B3		1.00	1.00	1.00	1.00	1.00	1.00	1.00
C1.5		0.35	0.35	0.34	0.37	0.36	0.37	0.33
C2		0.75	0.74	0.73	0.80	0.78	0.76	0.72
A1.5	$n = 50$	0.72	0.70	0.69	0.84	0.78	0.68	0.67
A2		1.00	0.99	0.99	1.00	1.00	0.99	0.99
B1.5		0.70	0.69	0.68	0.81	0.76	0.70	0.67
B2		0.99	0.99	0.99	1.00	1.00	0.99	0.99
B3		1.00	1.00	1.00	1.00	1.00	1.00	1.00
C1.5		0.55	0.54	0.53	0.61	0.58	0.54	0.52
C2		0.95	0.95	0.94	0.98	0.97	0.95	0.94

REFERENCES

Chen, H., Stasny, E.A., Wolfe, D.A., 2005. Ranked set sampling for efficient estimation of a population proportion. Stat. Med. 24, 3319–3329.

Ebrahimi, N., Habibullah, M., Soofi, E., 1994. Two measures of sample entropy. Stat. Probab. Lett. 20, 225–234.

Frey, J., Ozturk, O., Deshpande, J.V., 2007. Nonparametric tests of perfect judgment rankings. J. Am. Stat. Assoc. 102 (478), 708–717.

Grzegorzewski, P., Wieczorkowski, R., 1999. Entropy-based goodness-of-fit test for exponentiality. Commun. Stat.: Theory Methods 28, 1183−1202.

Halls, L.K., Dell, T.R., 1966. Trial of ranked-set sampling for forage yields. For. Sci. 12, 22−26.

Kvam, P.H., 2003. Ranked set sampling based on binary water quality data with covariates. J. Agric. Biol. Environ. Stat. 8, 271−279.

Lad, F., Sanfilippo, G., Agro, G., 2015. Extropy: complementary dual of entropy. Stat. Sci. 30, 40−58.

Mahdizadeh, M., Arghami, N., 2009. Efficiency of ranked set sampling in entropy estimation and goodness-of-fit testing for the inverse gaussian law. J. Stat. Comput. Simul. 80, 761−774.

Mahdizadeh, M., Zamanzade, E., 2017a. Efficient body fat estimation using multistage pair ranked set sampling. To appear in Statistical Methods in Medical Research. Available from: https://doi.org/10.1177/0962280217720473.

Mahdizadeh, M., Zamanzade, E., 2017b. A new reliability measure in ranked set sampling. To appear in Statistical Papers. Available from: http://dx.doi.org/10.1007/s00362-016-0794-3.

McIntyre, G.A., 1952. A method for unbiased selective sampling using ranked set sampling. Aust. J. Agric. Res. 3, 385−390.

Qui, G., 2017. The extropy of order statistics and record values. Stat. Probab. Lett. 120, 52−60.

Qui, G., Jia, K., 2018. Extropy estimators with applications in testing uniformity, To appear in. Journal of Nonparametric Statistics.

Shannon, C.E., 1948. A mathematical theory of communications. Bell Syst. Tech. J 27, 623−656.

Vasicek, O., 1976. A test for normality based on sample entropy. J. R. Stat. Soc., Ser. B 38, 54−59.

Wang, X., Lim, J., Stokes, L., 2016. Using ranked set sampling with cluster randomized designs for improved inference on treatment effects. J. Am. Stat. Assoc. 111 (516), 1576−1590.

Zamanzade, E., 2015. Entropy testing uniformity based on new entropy estimators. J. Stat. Comput. Simul. 85 (16), 3191−3205.

Zamanzade, E., Mahdizadeh, M., 2017a. A more efficient proportion estimator in ranked set sampling. Statistics and Probability Letters 129, 28−33.

Zamanzade, E., Mahdizadeh, M., 2017b. Entropy estimation from ranked set samples with application to test of fit. Rev. Colomb. Estad. 40, 1−19.

FURTHER READING

Dell, T.R., Clutter, J.L., 1972. Ranked set sampling theory with order statistics background. Biometrics 28 (2), 545−555.

Takahasi, K., Wakimoto, K., 1968. On unbiased estimates of the population mean based on the sample stratified by means of ordering. Ann. Inst. Stat. Math. 20 (1), 1−31.

Zamanzade, E., Wang, X., 2018. Comput. Stat. Available from: https://doi.org/10.1007/s00180-018-0807-x.

CHAPTER

SELECTION AND ESTIMATION IN RANKED SET SAMPLING USING R 20

Antonio Arcos, Beatriz Cobo and María del Mar Rueda

Department of Statistics and Operational Research, University of Granada, Granada, Spain

20.1 INTRODUCTION

Ranked set sampling (RSS) is an alternative to simple random sampling that has been shown to outperform simple random sampling (SRS) in many situations. RSS, originally proposed by McIntyre (1952), has recently attracted a considerable amount of interest and research as an alternate data collection method to SRS.

McIntyre's study generated a rapidly expanding body of research literature in order to estimate parameters as means. Dell and Clutter (1972) proved that the sample mean based on the RSS is unbiased for the population mean regardless of the errors of ranking. Bouza (2002) estimated the mean in ranked set sampling with nonresponses and in 2009 proposed a procedure for estimating the mean of a sensitive quantitative character. Pelli and Perri (2017) improved mean estimation in ranked set sampling using the Rao regression-type estimator. Other authors were responsible for estimating the variance (Stokes, 1980a; MacEachern et al., 2002; Perron and Sinha, 2004) from a nonparametric point of view, distribution functions (Stokes and Sager, 1988), and correlation coefficients (Stokes, 1980b). There has been a growing literature in RSS methods in recent years; see, for example Wolfe (2010, 2012).

Applications of RSS have been limited mostly to ecological, agricultural, and environmental sampling. Case studies can be found in Halls and Dell (1966), Al-Saleh and Al-Shrafat (2001), and Murff and Sager (2006). Thorough reviews of the RSS literature can be found in Patil et al. (1999) and Patil (2002).

There are many statistical softwares for working with complex surveys, but there are few that have implemented modules to work with ranked set sampling. In this, as in other types of sampling, two aspects must be highlighted: (1) the process of selecting the sample and (2) the parameter estimation process. We are going to focus on the first, since the number of different estimators proposed for different parameters is so broad.

Therefore in this work we are going to analyze the little software available for the selection of RSS balanced samples (the basic method) and we provide pseudocode for certain modifications of the basic method.

We will emphasize language R (R Core Team, 2017) as it is the free software that is most commonly used in the scientific community nowadays.

Ranked Set Sampling. DOI: https://doi.org/10.1016/B978-0-12-815044-3.00020-4

20.2 NOTATION AND BASIC DEFINITIONS

We will use the same notation as in Chen et al. (2004). An initial simple random sample of k units from the population is selected and subjected to ordering on the attribute of interest via some ranking process. The item judged to be the smallest is included as the first item in our ranked set sample and is noted by $X_{[1]}$. We select the item judged to be the second smallest of the k units in a second random sample and include it in our ranked set sample for measurement of the attribute of interest. This second measured observation is denoted by $X_{[2]}$. This process is continued until we have selected the unit judgment ranked to be the largest of the k units in the kth random sample, denoted by $X_{[k]}$, for measurement and inclusion in our ranked set sample. The observations $X_{[1]}, X_{[2]}, \ldots, X_{[k]}$ represent a ranked set sample with set size k. In order to obtain a ranked set sample with a desired total number of measured observations $k \cdot m$, we repeat the entire *cycle* process m independent time, yielding the data $X_{[1]j}, X_{[2]j}, \ldots, X_{[k]j}$ for $j = 1, \ldots, m$. This is referred to as a balanced RSS, where *balanced* indicates that the same number of observations were taken at each of the judgment ranks.

This is the usual RSS procedure. An alternative method is allocating sample units into ranks in different proportions; thus obtaining an *unbalanced* ranked set sample.

20.3 USING R FOR RANKED SET SAMPLING

In this section we describe how to obtain samples by ranked set sampling using the software R and we provide a code.

In the R web, there is a package, called NSM3, that calculates the ranked set sampling. Concretely, compute the ranked set sampling given a set size and number of cycles based on a specified auxiliary variable. This function only considers the option of balanced RSS.

20.3.1 BALANCED RANKED SET SAMPLING

To create ranked sets we must partition the selected first-phase sample into sets of equal size. In order to plan an RSS design, we must therefore choose a set size that is typically small, around three or four, to minimize the ranking error. Call this set size k, where k is the number of sample units allocated to each set. Now proceed as follows:

- Step 1: randomly select k^2 units from the population.
- Step 2: allocate the k^2 selected units as randomly as possible into k sets, each of size k.
- Step 3: without yet knowing any values for the variable of interest, rank the units within each set based on a perception of relative values for this variable. This may be based on personal judgment or done with measurements of a covariate that is correlated with the variable of interest.
- Step 4: choose a sample for actual analysis by including the smallest ranked unit in the first set, then the second smallest ranked unit in the second set, continuing in this fashion until the largest ranked unit is selected in the last set.

- Step 5: repeat steps 1 through 4 for m cycles until the desired sample size, $n = mk$, is obtained for analysis.

R Code

`RSS` *computes the indices of a sample obtained for balanced RSS.*

USAGE

```
RSS(k,m,ranker)
```

ARGUMENTS

- k: set size;
- m: number of cycles;
- ranker: auxiliary variable used for judgment ranking.

VALUE

Returns a vector of the indices corresponding to the observations selected to be in the RSS.

FUNCTION CODE

```
library(NSM3)
RSS
function (k, m, ranker)
{
     N <- length(ranker)
     num.samples <- m * k
     SRS.index <- matrix(sample(1:N, num.samples * k), nrow = k)
     selected.rankers <- matrix(ranker[SRS.index], nrow = k)
     sorted <- apply(selected.rankers, 2, sort)
     sample.ranks <- apply(selected.rankers, 2, order)
     output <- 0
     for (i in 1:num.samples) {
          index <- floor((i - 1)/m) + 1
          output[i] <- SRS.index[sample.ranks[index, i], i]
     }
     return(sort(output))
}
<environment: namespace:NSM3>
```

20.3.2 UNBALANCED RANKED SET SAMPLING

We consider two situations.

20.3.2.1 Case 1

We are interested to obtain an URSS of size $n = k \cdot m + s, s \neq 0$. Thus we follow the same procedure as in the previous case RRS, and we include then the elements remaining. The procedure is explained below.

- Step 1: randomly select k^2 units from the population.
- Step 2: allocate the k^2 selected units as randomly as possible into k sets, each of size k.
- Step 3: without yet knowing any values for the variable of interest, rank the units within each set based on a perception of relative values for this variable.
- Step 4: choose a sample for actual analysis by including the smallest ranked unit in the first set, then the second smallest ranked unit in the second set, continuing in this fashion until the largest ranked unit is selected in the last set.
- Step 5: let the number of cycles $m = [n/k]$, that is to say, the integer is less than or equal to n/k. Repeat steps 1 through 4 for m cycles.
- Step 6: If $s > 0$, we choose integers $j_1, j_2, ..., j_s, (j_1 \neq j_2 \neq ... \neq j_s)$ from $1, 2, ..., k$. Then we select s independent SRSWR samples, each of size k. Observations from each of the samples are ranked with respect to their auxiliary variable. From the first sample, we select j_1th ranked observation, j_2th observation from the second sample and j_s ranked observation from the sth sample
- Step 7. The final sample contains the units obtained in steps 5 and 6.

R Code

`ranked1` *computes the indices of a sample obtained for balanced or unbalanced (Case1) RSS.*

USAGE

```
ranked1(n,k,ranker)
```

ARGUMENTS

- n: sample size;
- k: set size;
- ranker: vector which contains an auxiliary variable.

VALUE

Returns a vector which contains the sample indices for balanced or unbalanced (Case 1) RSS.

FUNCTION CODE

```
ranked1 = function(k,n,ranker){
    N = length(ranker)
    m = n%/%k
    s = n%%k
    SF = RSS(k,m,ranker)
    if(s! = 0){
         mas0 = sample(N,k*s, replace = T)
         x = ranker[mas0]
         gr0 = as.factor(rep(1:s,k))
         sf0data1 = data.frame(mas0,x,gr = gr0)
         sf0dataorder <- sf0data1[order(sf0data1$gr,sf0data1$x),]
         i = sample(k,s)
         SF <- c(SF, sf0dataorder[i,]$mas0)
         SF <- sort(SF)
    }
    return(SF)
}
```

20.3.2.2 Case 2

We are interested in an unbalanced sample in the following way: choose a sample for actual analysis by including the smallest ranked units in the first n_1 sets, then the second smallest ranked units in the n_2 second sets, continuing in this fashion until the largest ranked unit is selected in the last n_k sets. The final sample size is $n = n_1 + n_2 \cdots + n_k$.

R Code

`ranked2` *computes the indices of a sample obtained for unbalanced (Case2) RSS.*

USAGE

```
ranked2(k,ss,ranker)
```

ARGUMENTS

- k: set size;
- ss: vector with the sample size allocations;
- ranker: vector which contains an auxiliary variable.

VALUE

Returns a vector which contains the sample indices for unbalanced (Case 2) RSS.

FUNCTION CODE

```
ranked2 = function(k,ss,ranker){
   N = length(ranker)
   n = sum(ss)
   SRS.index <- matrix(sample(1:N, n*k), nrow = k)
   selected.rankers <- matrix(ranker[SRS.index], nrow = k)
```

```
    sorted <- apply(selected.rankers, 2, sort)
    sample.ranks <- apply(selected.rankers, 2, order)
    ssa<-cumsum(ss)
    output<-c()
    for (l in 1:ssa[1]) {
      SRS.index[,1:ssa[1]][,l][sample.ranks[,l]][1]->output1
      output<-c(output, output1)
     }
    for (t in 2:length(ss)) {
     for (ll in 1:(ss[t])) {
       SRS.index[,(ssa[t-1]+1):(ssa[t])][,ll]
       [sample.ranks[,(ssa[t-1]+1):(ssa[t])][,ll]][t]->outputt
       output<-c(output, outputt)
        }
    SF <- sort(output)
    }
  return(SF)
  }
```

20.3.3 THE MEDIAN RANKED SET SAMPLING METHOD

The ranked set sampling (RSS) method as suggested by McIntyre (1952) may be modified to come up with new sampling methods that can be made more efficient than the usual RSS method. It is known that there will be a loss in precision due to the errors in ranking the units. One modification to reduce the errors in ranking, namely median ranked set sampling (MRSS), is considered in this study; see Muttlak (1997) for details.

In the MRSS procedure, select k random samples of size k units from the population and rank the units within each sample with respect to a variable of interest. If the sample size k is odd, from each sample select for measurement the $\left((k+1)/2\right)$th smallest rank (the median of the sample). If the sample size is even, select for measurement from the first $k/2$ samples the $\left(k/2\right)$th smallest rank and from the second $k/2$ samples the $\left((k+2)/2\right)$th smallest rank. The cycle may be repeated m times to get the $n = k \cdot m$ units which form the MRSS sample.

R Code

`MRSS` *computes the indices of a sample obtained for RSS using the median method.*

USAGE

```
  MRSS(k, ranker)
```

ARGUMENTS

- k: set size;
- ranker: vector which contains an auxiliary variable.

VALUE

Returns a vector which contains the sample indices for MRSS,

FUNCTION CODE

```
MRSS = function (k,ranker){
N <- length(ranker)
num.samples <- k
SRS.index <- matrix(sample(1:N, num.samples * k), nrow = k)
selected.rankers <- matrix(ranker[SRS.index], nrow = k)
sorted <- apply(selected.rankers, 2, sort)
sample.ranks <- apply(selected.rankers, 2, order)
output<-c()
if (k%%2!=0) {
    (k+1)/2->med
    for (col in 1:k) {
        SRS.index[,1:k][,col][sample.ranks[,col]][med]->output1
        output<- c(output,output1)
        }
}
if (k%%2==0) {
        outputa<-c()
        outputb<-c()
        k/2->med
        for (col in 1:(k/2)) {
          SRS.index[,1:(k/2)][,col][sample.ranks[,col]][med]->output1
          outputa<- c(outputa,output1)
      }
      for (col in 1:(k/2)) {
          SRS.index[,((k/2)+1):k][,col][sample.ranks[,((k/2)+1):k][,col]][med+1]
->output2
          outputb<- c(outputb,output2)
      }
output<- c(outputa,outputb)
}
SF <- sort(output)
return(SF)
}
```

20.4 ESTIMATION USING RSS

While this may change as RSS methodology progresses, at this point in time standard software packages are sufficient to analyze RSS data once they have been collected.

For example, let *h(x)* be any function of *x*. For an RSS sample, the estimator

$$\hat{\mu}_{h\mathrm{RSS}} = \frac{1}{mk}\sum_{r=1}^{k}\sum_{i=1}^{m} h(X_{[r]i})$$

is unbiased for the expectation of $h(X)$ if the ranking mechanism in RSS is consistent (Chen et al., 2004).

The natural estimates of $V(\hat{\mu}_{h\mathrm{RSS}})$ using an RSS sample are given by

$$s^2_{\mathrm{RSS}} = \frac{1}{mk-1}\sum_{r=1}^{k}\sum_{i=1}^{m}\left(h(X_{[r]i}) - \hat{\mu}_{h\mathrm{RSS}}\right)^2$$

These estimators can easily be calculated from standard software packages.

20.5 EXAMPLES

20.5.1 RANKING WITH AN INEXPENSIVE QUANTITATIVE MEASUREMENT

When auxiliary information is available for the entire population of size *N* (an inexpensive quantitative measurement), the previous functions can be used for select units included in the RSS sample. The following lines show how to do it with the different types of RSS previously reported.

```
set.seed (1)
response<-rnorm(200,20,2)
auxiliary<-rnorm(200,10,1)
#Get the indices for a RSS with set size 3 and 2 cycles
RSS(2,3,auxiliary)
#66   74 133 147 172 183
#Balanced
ranked1(3,12,auxiliary)
#[1]   7  67  68  72  83  88 107 128 142 179 180 200
#Unbalanced Case 1
ranked1(3,13,auxiliary)
#[1]   2   4  10  12  48  51  56  70  99 115 149 173 200
#Using MRSS
MRSS(3,auxiliary)
#[1]  17  66 114
MRSS(4,auxiliary)
#[1]  12  54 147 164
#Unbalanced Case 2
ranked2(3,c(2,3,4,5),auxiliary)
#[1]  14  32  63  93  134 136 151 169 191
```

In all previous examples, the response observed can be easily show using, for example:

```
response[MRSS(3,auxiliary)]
#[1] 21.37948 18.86266 17.90403
```

20.5.2 RANKING WITH A PROFESSIONAL JUDGMENT

Suppose you want to determine the average production in an olive grove like the one shown in the figure:

There are $N = 2070$ olive trees. It is planned to select a sample of size $n = 30$, taking sets of size $m = 3$, which will be sorted by visual inspection. A labeling and object count computer program provides the indices $i = 1, \ldots, 2070$ that identifies the objects in the photograph. The following program lines select three simple random samples:

```
N = 2070
m = 3
num.samples <- m * m
index <- matrix(sample(1:N, num.samples), nrow = m, byrow = T)
index
          [,1] [,2] [,3]
[1,]   213 462 333
[2,]   1321 1865 334
[3,]   234 331 5
```

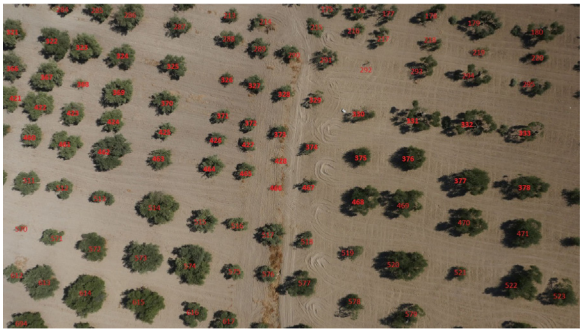
The user provides the order by visual inspection of the matrix

$$\begin{pmatrix} 213 & 462 & 333 \\ 1321 & 1865 & 334 \\ 234 & 331 & 5 \end{pmatrix}$$

so that it is the matrix

$$\begin{pmatrix} \mathbf{213} & 333 & 462 \\ 334 & \mathbf{1865} & 1321 \\ 234 & 5 & \mathbf{331} \end{pmatrix},$$

from the photographic information of the figure:



Finally, units 213, 331, and 1865 are selected for observation of the main variable. The previous process is repeated until all the variables of interest in the sample are observed.

20.6 ADDITIONAL SOFTWARE

To our knowledge, there are few programs that perform RSS. One of them is the one described briefly below.

Visual Sample Plan (VSP) is a software tool developed by Pacific Northwest National Laboratory (PNNL), initially conceived and sponsored through DOE-Office of Health, Safety and Security (HHS), that supports the development of a sampling plan and statistical data analysis. VSP has many sampling design and statistical analysis modules focused on soils, sediments, surface water, streams, groundwater, buildings, and others. Many statistical sampling designs are available, including ranked set sampling.

Either professional judgment or an inexpensive quantitative (screening) measurement of the variable of interest can be used to do the ranking when ranked set sampling is used. VSP calculates the number of samples and field ranking locations needed to estimate the mean using ranked set sampling and places the field ranking locations on the map using simple random sampling.

Ranked set sampling design for estimating a mean is implemented for a balanced or unbalanced design. It is possible to determine the number of samples, take into account the sample size, the set size, the relative precision, and the number of cycles to compute the total number of samples that should be collected. For ranked set sampling, VSP produces field sample markers on the map that have different shapes and colors. It is also possible to include cost-effectiveness parameters in the analysis.

ACKNOWLEDGMENTS

This work is partially supported by Ministerio de Economía y Competitividad (grant MTM2015-63609-R, Spain), by Consejería de Economía, Innovación, Ciencia y Empleo (grant SEJ2954, Junta de Andalucía, Spain) and Ministerio de Educación, Cultura y Deporte (grant FPU, Spain).

REFERENCES

Al-Saleh, M.F., Al-Shrafat, K., 2001. Estimation of average milk yield using ranked set sampling. Environmetrics 12, 395–399.

Bouza, C.N., 2002. Estimation of the mean in ranked set sampling with non responses. Metrika 56, 171–179.

Chen, Z., Bai, Z.D., Sinha, B.K., 2004. Ranked Set Sampling: Theory and Applications Lecture Notes in Statistics, vol. 176. Springer-Verlag, New York.

Dell, T.R., Clutter, J.L., 1972. Ranked set sampling theory with order statistics background. Biometrics 28, 545–555.

Halls, L.K., Dell, T.R., 1966. Trial of ranked-set sampling for forage yields. For. Sci. 12, 22–26.

MacEachern, S.N., Öztürk, O., Wolfe, D.A., Stark, G.V., 2002. A new ranked set sample estimator of variance. J. R. Stat. Soc. B 64, 177–188.

McIntyre, G.A., 1952. A method of unbiased selective sampling using ranked sets. J. Agric. Res. 3, 385–390.

Murff, E.J.T., Sager, T.W., 2006. The relative efficiency of ranked set sampling in ordinary least squares regression. Environ. Ecol. Stat. 13, 41–51.

Muttlak, H.A., 1997. Median ranked set sampling. J. Appl. Stat. Sci. 6, 245–255.

Patil, G., Sinha, A., Taillie, C., 1999. Ranked set sampling: a bibliography. Environ. Ecol. Stat. 6, 91–98.

Patil, G.P., 2002. Ranked set sampling, Encyclopedia of Environmetrics, 3. John Wiley & Sons, Ltd, Chichester, pp. 1684–1690.

Pelli, E., Perri, P.F., 2017. Improving mean estimation in ranked set sampling using the Rao regression-type estimator. Braz. J. Probab. Stat. Available from: http://imstat.org/bjps/papers/BJPS350.pdf.

Perron, F., Sinha, B.K., 2004. Estimation of variance based on a ranked set sample. J. Stat. Plan. Inference 120, 21–28.

R Core Team, 2017. R: A language and environment for statistical computing. R Found. Stat. Comput. Vienna, Austria. URL http://www.R-project.org/.

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

Stokes, S.L., 1980b. Inferences on the correlation coefficient in bivariate normal populations from ranked set samples. J. Am. Stat. Assoc. 75, 989–995.

Stokes, S.L., Sager, T.W., 1988. Characterization of a ranked set sample with application to estimating distribution functions. J. Am. Stat. Assoc. 83, 374–381.

Wolfe, D.A., 2010. Ranked set sampling. Wiley interdisciplinary reviews. Comput. Stat. 2, 460–466.

Wolfe, D.A., 2012. Ranked set sampling: its relevance and impact on statistical inference. ISRN Probab. Stat.

FURTHER READING

Bouza, C.N., 2009. Ranked set sampling and randomized response procedures for estimating the mean of a sensitive quantitative character. Metrika 70, 267–277.

CHAPTER

VARIANCE ESTIMATION OF PERSONS INFECTED WITH AIDS UNDER RANKED SET SAMPLING

21

Carlos N. Bouza-Herrera[1], Jose F. García[2], Gajendra K. Vishwakarma[3] and Sayed Mohammed Zeeshan[3]

[1]Faculty of Mathematics and Computation, University of Havana, Havana, Cuba [2]DACEA, Universidad Juárez Autónoma de Tabasco, Villahermosa, Tabasco, Mexico [3]Department of Applied Mathematics, Indian Institute of Technology (ISM), Dhanbad, Jharkhand, India

21.1 INTRODUCTION

Ranked set sampling is an alternative sample design, which generally provided gains in accuracy with respect to simple random sampling with replacement (SRSWR). It was proposed for estimating the yield of pastures by McIntyre (1952). He established this method to estimate the mean pasture yield using RSS and found its inferences more efficient than selecting the sample using a simple random sampling (SRS) design. The units may be ranked by means of a cheap procedure and then an order statistics is selected from each of the independent samples selected using SRS with replacement (SRSWR). It turned out that the use of ranked set sampling is highly beneficial and leads to estimators which are more precise than the usual sample mean per unit ones. The method is now referred to as the ranked set sampling (RSS) method in the literature. Takahasi and Wakimoto (1968) were the first to prove that the mean estimator from RSS is more efficient than that from SRS. This led to a lot of research that has been done by various authors including Dell and Clutter (1972), Stokes (1980), Patil et al. (1995), MacEachern et al. (2002), Chen et al. (2003), Perron and Sinha (2004), and Frey (2011).

In this chapter, we propose a model using RSS, instead of SRS with replacement (SRSWR), for studies of variance. The rest of this chapter is organized as follows: Section 21.2 develops the study of the one-way analysis of variance. Section 21.3 is devoted to the presentation of estimators of the variance. Section 21.4 is devoted to the development of numerical studies of the behavior of the analyzed models in testing hypothesis. We discuss the results obtained from the use of SRSWR and develop alternative RSS models in the next section. Samples of persons infected with the AIDS virus are analyzed and the behavior of the accuracy of the different alternative estimators are also discussed.

Ranked Set Sampling. DOI: https://doi.org/10.1016/B978-0-12-815044-3.00021-6

21.2 ESTIMATION OF THE TREATMENT EFFECTS IN A ONE-WAY LAYOUT IN RANKED SET SAMPLING

Consider the one-way layout

$$Y_{ij} = \mu_i + \mathring{a}_{ij} = \mu + \alpha_i + \mathring{a}_{ij}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n(i). \tag{21.1}$$

This issue is important in many applications and has been studied extensively. Let Y be the variable of interest. We select independent samples of size $n(i)$, $i = 1, \ldots, k$, using simple random sampling with replacement (SRSWR), for estimating the parameters of interest μ, $\alpha_i = (\mu_i - \mu)$, $i = 1, \ldots, k$. We assume that for any $i = 1, \ldots, k$ and $j = 1, \ldots, n(i)$, $E(\mathring{a}_{ij}) = 0$, $V(\mathring{a}_{ij}) = \sigma^2{}_i$ and $\mathrm{Cov}\left(\mathring{a}_{ij}\mathring{a}_{i'j'}\right) = 0$, if $i \neq i'$ and/or $j \neq j'$. The usual estimation of the effects α_i is

$$\alpha_i^* = \frac{\sum_{j=1}^{n(i)} y_{ij}}{n(i)} - \frac{\sum_{i=1}^{k}\sum_{j=1}^{n(i)} y_{ij}}{n} = \left(\bar{y}_i - \bar{y}\right); \quad \text{where } n = \sum_{i=1}^{k} n(i) \tag{21.2}$$

Its variance is given by

$$V(\alpha_i^*) = E\left(\bar{y}_i \pm \mu_i \pm \mu - \bar{y}\right)^2 = \frac{\sigma_i^2}{n(i)} + \frac{\sigma^2}{n} + \left(\mu_i - \mu\right)^2. \tag{21.3}$$

Muttlak (1998) proposed to use RSS. As usual, the model was based on the selection of $n(i)$ independent samples of size $n(i)$ using SRSWR and to rank each of them. That is we have hypothetically for each $i = 1, \ldots, k$ $s_i = \left\{\left(Y_{i11} \ldots, Y_{i1n(i)}\right)_1, \ldots, \left(Y_{in(i)1} \ldots, Y_{in(i)n(i)}\right)_{n(i)}\right\}$ and by ranking, we have the ranked samples $\left\{\left(Y_{i1(1)} \ldots, Y_{i(n(i))}\right), \ldots, \left(Y_{in(i)(1)} \ldots, Y_{in(i)(n(i))}\right)\right\}$. Y is measured in the statistic of order (SO) t in the tth sample. Then our set of results for treatment "i" is

$$s(i) = \left\{\left(Y_{i1(1)}\right), \left(Y_{i2(2)}\right), \ldots, \left(Y_{it(t)}\right) \ldots, \left(Y_{in(1)n(i)}\right)\right\} = \left\{Y_{i(1)}, Y_{i(2)}, \ldots, Y_{i(t)} \ldots, Y_{i(n(i))}\right\} \tag{21.4}$$

We deal with the linear model

$$Y_{i(j)t} = \mu_i + \varepsilon_{i(j)t} = \mu + \alpha_i + \varepsilon_{i(j)t,} i = 1, \ldots, k, \quad j = 1, \ldots, n(i)$$

$$\bar{y}_{(i)} = \frac{\sum_{j=1}^{n(i)} y_{i(j)}}{n(i)}, \quad \bar{y}_{\mathrm{RSS}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n(i)} y_{i(j)j}}{n}, \quad n = \sum_{i=1}^{k} n(i)$$

It is unbiased and

$$V(\bar{y}_{(i)}) = \frac{\sum_{j=1}^{n(i)} \sigma^2{}_{i(j)}}{n^2(i)}, \quad \sigma^2{}_{i(j)} = V(y_{i(j)})$$

It was hypothesized that, $\sigma_{i(j)}^2 = \sigma_{(i)}^2$, which is the counterpart of the hypothesis used in the development of the one-way layout ANOVA. $\sigma_i^2 = \sigma^2$, in the inferences based on SRSWR. Using the relation established by Takahasi and Wakimoto (1968) we can derive that for any $i = 1, \ldots, k$

$$V(\bar{y}_{(i)}) = \frac{\sigma^2{}_i}{n(i)} - \frac{\sum_{j=1}^{n(i)} \Delta^2_{i(j)}}{n^2(i)}, \quad \Delta_{i(j)} = \mu_{i(j)} - \mu_{(i)}, \quad \mu_{i(j)} = E(y_{i(j)}),$$

Let us look for the RSS counterpart of the results in Eqs. (21.2) and (21.3).

Proposition 2.1: $\alpha^*_{(i)} = \dfrac{\sum_{j=1}^{n(i)} y_{i(j)}}{n(i)} - \dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n(i)} y_{i(j)}}{n} = \left(\bar{y}_{(i)} - \bar{y}_{\mathrm{RSS}}\right)$ is unbiased and more accurate than α^*_i.

Proof: Due to the unbiasedness of the RSS estimators

$$E(\alpha^*_{(i)}) = E(\alpha^*_i) = E(\bar{y}_{(i)}) - E(\bar{y}_{\mathrm{RSS}}) = \mu_i - \mu$$

and $V(\alpha^*_{(i)}) = E\left(\bar{y}_{(i)} \pm \mu_i \pm \mu - \bar{y}_{\mathrm{RSS}}\right)^2 = \dfrac{\sum_{j=1}^{n(i)} \sigma^2_{i(j)}}{n^2(i)} + \dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n(i)} \sigma^2_{i(j)}}{n^2} + \left(\mu_i - \mu\right)^2$

We have that $\sigma^2_{i(j)} = \sigma^2_i - \Delta^2_{i(j)}$ then substituting in the above equation

$$V(\alpha^*_{(i)}) = \frac{\sigma^2_i}{n(i)} + \frac{\sum_{i=1}^{k} n(i)\sigma^2_i}{n^2} + \left(\mu_i - \mu\right)^2 - \psi(1)$$

where

$$\Psi(1) = \frac{\sum_{j=1}^{n(i)} \Delta^2_{i(j)}}{n^2(i)} + \frac{\sum_{i=1}^{k}\sum_{j=1}^{n(i)} \Delta^2_{i(j)}}{n^2} \geq 0$$

represents the gain in accuracy due to the use of RSS.

Remark 1: If $\forall i = 1, \ldots, k, \quad \sigma^2_i = \sigma^2 \Rightarrow \dfrac{\sigma^2}{n(i)} + \dfrac{\sigma^2}{n}$ and the usual relation is obtained.

21.3 ESTIMATION OF THE VARIANCE IN RSS

A basic relationship in RSS is

$$\sigma^2 = \frac{1}{k}\sum_{r=1}^{k} \sigma^2_{(r)} + (\mu_{(r)} - \mu_{(r')})^2; \text{ if } r \neq r' \tag{21.5}$$

Stokes (1980) suggested as an estimator of it, for one cycle,

$$\sigma^2_S = \frac{1}{(k-1)}\sum_{r=1}^{k} \left(Y_{(r)i} - \mu_{\mathrm{rss}}\right)^2 \text{ where, } \mu_{\mathrm{rss}} = \frac{1}{k}\sum_{r=1}^{k} Y_{(r)}$$

and its expectation is $E\left(\sigma^2_S\right) = \sigma^2 + \dfrac{1}{k(k-1)}\sum_{r=1}^{k} \left(\mu_{(r)} - \mu_{\mathrm{rss}}\right)^2$

Considering the structure of the one-way ANOVA the estimator proposed by Stokes (1980) is given in the next proposition.

Proposition 3.1: Stokes (1980); $\sigma_S^2 = \sigma^2 + \dfrac{1}{(nk-1)}\sum_{i=1}^{n}\sum_{r=1}^{k}\left(Y_{(r)i} - \mu_{\text{rss}}\right)^2$

where $\mu_{\text{rss}} = \dfrac{1}{nk}\sum_{i=1}^{n}\sum_{r=1}^{k} Y_{(r)i}$ estimates the RSS variance and if $n \to \infty$ then $E\left(\sigma_S^2\right) = \sigma^2$.

Stokes derived that this estimator overestimates σ^2 and its variance is

$$V\left(\sigma_S^2\right) = \frac{n}{(nk-1)^2}\left\{\left(\frac{nk-1}{nk}\right)^2\sum_{r=1}^{k}\mu_{4(r)} + 4\sum_{r=1}^{k}\Delta_{(r)}^2\sigma_{(r)}^2 + 4\left(\frac{nk-1}{nk}\right)\sum_{r=1}^{k}\Delta_{(r)}\mu_{3(r)}\right.$$
$$\left. + \frac{4n}{k^2n^2}\sum\sum_{r<r'}\sigma_{(r)}^2\sigma_{(r')}^2 - \frac{2(n-1)-(nk-1)^2}{k^2n^2}\sum_{r=1}^{k}\sigma_{(r)}^4\right\}$$

Note that this error depends on moments of the distribution of the order statistics then the variable's distribution must be known. Hence to derive an explicit formula is very complex.

MacEachern et al. (2002) proposed to use as an estimator

$$\sigma_M^2 = \sigma_{M1}^2 + \sigma_{M2}^2 \tag{21.6}$$

where

$$\sigma_{M1}^2 = \frac{1}{2n^2k^2}\sum_{r \neq r'}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(Y_{(r)i} - Y_{(r')j}\right)^2 \tag{21.7}$$

and

$$\sigma_{M2}^2 = \frac{1}{2n(n-1)^2k^2}\sum_{r=1}^{k}\sum_{j=1}^{n}\sum_{j=1}^{n}\left(Y_{(r)i} - Y_{(r')j}\right)^2 \tag{21.8}$$

It is unbiased. The next proposition gives its properties.

Proposition 3.2: MacEachern et al. (2002); $\sigma_M{}^2$ is unbiased and its variance, if $\mu_{(r)} < \infty$, is

$$V\left(\sigma_M^2\right) = A + B + C + D + F$$

where $A = \dfrac{1}{nk^2}\sum_{r=1}^{k}\mu_{4(r)}$, $B = \dfrac{4}{nk^2}\sum_{r=1}^{k}\mu_{3(r)}\Delta_{(r)}$, $C = \dfrac{4}{nk^2}\sum_{r=1}^{k}\sigma_{(r)}^2\Delta_{(r)}^2$,

$D = \dfrac{4}{n^2k^4}\sum_{r<r'}\sigma_{(r)}^2\sigma_{(r')}^2$, $F = \dfrac{k^2(n-1)-2}{n(n-1)k^4}\sum_{r=1}^{k}\sigma_{(r)}^2$

Using the mean square errors (MSEs) and one-way ANOVA decomposition ideas we have that

$$MST = MS1 - MS2$$

taking $\mu_{\text{rss}(r)} = \dfrac{1}{n}\sum_{i=1}^{n} Y_{(r)i}$, then

$$MS1 = \frac{1}{k-1}\sum_{r=1}^{k}\sum_{i=1}^{n}\left(Y_{(r)i} - \mu_{\text{rss}}\right)^2$$

$$MS2 = \frac{1}{k-1}\sum_{r=1}^{k}\sum_{i=1}^{n}\left(Y_{(r)i} - \mu_{\text{rss}(r)}\right)^2$$

Then the rank-residual MSE is:

$$\text{MSR} = \frac{1}{k(n-1)}\sum_{r=1}^{k}\sum_{i=1}^{n}\left(Y_{(r)i} - \mu_{\text{rss}(r)}\right)^2$$

The expectations of the MSEs are

$$E(\text{MST}) = \frac{1}{k}\sum_{r=1}^{k}\sigma^2_{(r)} + \frac{1}{n(k-1)}\sum_{r=1}^{k}\left(\mu_{(r)} - \mu\right)^2$$

$$E(\text{MSR}) = \frac{1}{k}\sum_{r=1}^{k}\sigma^2_{(r)}.$$

Then the variance of σ^2_M as

$$\sigma^2_M = \frac{(k-1)\text{MST} + (nk-k+1)\text{MSR}}{nk}$$

and its expectations given by

$$E\left(\sigma^2_M\right) = \frac{n+2}{nk}\sum_{r=1}^{k}\sigma^2_{(r)} + \frac{1}{n^2k}\sum_{r=1}^{k}\left(\mu_{(r)} - \mu\right)^2$$

The ordering made using an auxiliary variable X is equivalent to the use of SRS in the inferior of the scenarios. It is well known that RSS is equivalent to it, in terms of accuracy, in such cases. That is, for any $r\mu_{(r)} = -\mu$. Hence the statistic

$$V(n) = \frac{\text{MST}}{\text{MSR}} \tag{21.9}$$

under the hypothesis of random ranking must be close to 1. Therefore, we can evaluate the usefulness of the ranking of Y, produced by X, by analyzing $V(n)$, as in regression analysis, through the coefficient of determination. In this case, large values of $V(n)$ imply that the ranking is more different than the ranking produced by pure randomness. That is reasoning similar to the nonparametric evaluation of the goodness of regression fitting. Under the hypothesis of normality $V(n)$ is distributed $F(k,\ k(n\ -\ 1))$ and inferences can be developed using the classic parametric theory using F tests.

21.4 MONTE CARLO EVALUATION

21.4.1 NORMALITY-BASED TESTS

One thousand runs (samples) were generated using the uniform *(0,2)*, normal *(0,1)*, exponential *(1)*, gamma with density function $f(x) = x^4\exp(-x)/\Gamma(5)$; $x > 0$, U-shaped with density function $f(x) = 3x^2/2$; $x \in [0,\ 1]$ and the lognormal *(0,1)* distribution. The sample size parameters were $n \in \{2,\ 3,\ 4,\ 5\}$

Table 21.1 Percentage of Acceptance of the True Hypothesis H_0 Using One-Way ANOVA [1000 Runs (Samples) Generated and $\alpha = 0.05$]

Distribution	*n*	*k*	SRS	RSS	Distribution	*n*	*k*	SRS	RSS
Uniform (0,2)	2	2	0.76	0.72	*Normal (0,1)*	2	2	0.88	0.87
	2	3	0.74	0.71		2	3	0.87	0.87
	2	4	0.78	0.71		2	4	0.88	0.87
	2	5	0.78	0.77		2	5	0.91	0.88
	5	2	0.79	0.77		5	2	0.93	0.89
	5	3	0.79	0.76		5	3	0.93	0.92
	5	4	0.82	0.78		5	4	0.93	0.93
	5	5	0.81	0.78		5	5	0.94	0.92
Exponential(1)	2	2	0.54	0.67	*Gamma* $x^4 \exp(-x)/\Gamma(5)$ $x > 0$	2	2	0.66	0.69
	2	3	0.54	0.67		2	3	0.66	0.69
	2	4	0.66	0.72		2	4	0.66	0.73
	2	5	0.71	0.73		2	5	0.68	0.76
	5	2	0.70	0.72		5	2	0.75	0.79
	5	3	0.71	0.72		5	3	0.76	0.77
	5	4	0.72	0.74		5	4	0.77	0.84
	5	5	0.74	0.79		5	5	0.77	0.88
U-shaped $f(x) = 3x^2/2$ $x \in [0, 1]$	2	2	0.54	0.68	*Lognormal(0,1)*	2	2	0.86	0.83
	2	3	0.56	0.69		2	3	0.86	0.82
	2	4	0.56	0.69		2	4	0.88	0.85
	2	5	0.59	0.69		2	5	0.88	0.85
	5	2	0.61	0.74		5	2	0.89	0.87
	5	3	0.69	0.74		5	3	0.88	0.87
	5	4	0.69	0.82		5	4	0.89	0.90
	5	5	0.71	0.84		5	5	0.89	0.90

and $k \in \{2, 3, 4, 5\}$. The ANOVA was performed using the normal approximation and $\alpha = 0.05$. An estimation of the percentage of samples in which we accepted the true hypothesis H_0 was computed for SRS and RSS and the results are presented in Table 21.1.

Remark: When data are not normally distributed, the RSS-ANOVA had a better performance than the classic SRS procedure. This could be due to the convergence of linear rank statistics to normality.

Tables 21.2–21.7 present the results of *1000* runs (samples) generated, using different bivariate distributions, with $\rho \in \{0.00, 0.05, 0.75, 0.90, 0.95\}$. We computed the values of *V(mean)* using the following formula as

$$V(mean) = \frac{1}{1000} \sum_{1 \le h \le 100} V(n)_h \tag{21.10}$$

Table 21.2 Values of *V(Mean)* Under Different Values of ρ and a Joint uniform Distribution

Distribution	*n*	*K*	$\rho = 0$	$\rho = 0.05$	$\rho = 0.50$	$\rho = 0.75$	$\rho = 0.90$	$\rho = 0.95$
Uniform (0,2)	2	2	1.07	1.59	3.43	4.85	5.55	6.77
	2	3	1.06	1.59	3.43	4.88	5.59	6.77
	2	4	1.07	1.59	3.42	4.89	5.80	7.08
	2	5	1.07	1.59	3.4 4	4.94	5.85	7.28
	5	2	1.08	1.59	3.46	4.94	5.84	7.39
	5	3	1.08	1.59	3.4 5	4.94	5.85	7.42
	5	4	1.06	1.65	3.47	4.96	5.87	7.48
	5	5	1.08	1.65	3.48	4.98	5.89	7.55

Table 21.3 Values of *V(Mean)* Under Different Values of ρ and a Joint Normal Distribution

Distribution	*n*	*k*	$\rho = 0$	$\rho = 0.05$	$\rho = 0.50$	$\rho = 0.75$	$\rho = 0.90$	$\rho = 0.95$
Normal (0,1)	2	2	1.07	1.15	8.63	8.91*	8.88*	9.09*
	2	3	1.06	1.15	8.73	8.98*	8.98*	9.17*
	2	4	1.07	1.14	8.72*	8.18*	8.78*	9.28*
	2	5	1.07	1.11	8.72*	8.78*	8.79*	9.28*
	5	2	1.02	1.11	8.74*	8.78*	8.84*	9.39*
	5	3	1.02	1.11	8.65*	9.07*	8.95*	9.47*
	5	4	1.06	1.13	8.79*	9.16*	9.07*	9.58*
	5	5	1.02	1.13	8.81*	9.28*	9.59*	9.75*

Table 21.4 Values of *V(Mean)* Under Different Values of ρ and a Joint Exponential Distribution

Distribution	*n*	*k*	$\rho = 0$	$\rho = 0.05$	$\rho = 0.50$	$\rho = 0.75$	$\rho = 0.90$	$\rho = 0.95$
Exponential(1)	2	2	1.07	2.25	3.73	5.05	6.18	6.66
	2	3	1.06	2.25	3.73	5.18	6.18	6.67
	2	4	1.07	2.24	3.72	5.15	6.28	6.68
	2	5	1.07	2.22	3.7 7	5.35	6.39	6.68
	5	2	1.02	2.22	3.76	5.55	6.44	6.69
	5	3	1.02	2.22	3.7 5	5.57	6.55	6.77
	5	4	1.06	2.23	3.77	5.56	6.57	6.88
	5	5	1.02	2.23	3.78	5.58	6.59	6.85

Table 21.5 Values of *V(Mean)* Under Different Values of ρ and a Joint Gamma Distribution

Distribution	*n*	*k*	$\rho=0$	$\rho=0.05$	$\rho=0.50$	$\rho=0.75$	$\rho=0.90$	$\rho=0.95$
Gamma $x^4\exp(-x)/\Gamma(5)$ $x>0$	2	2	1.17	2.69	6.06	6.67	6.61	7.44
	2	3	1.16	2.69	6.03	6.66	6.67	7.76
	2	4	1.17	2.69	6.06	6.69	6.71	7.79
	2	5	1.17	2.69	6.00	6.19	6.76	7.73
	5	2	1.19	2.76	6.06	6.19	6.77	7.78
	5	3	1.19	2.75	6.00	6.61	6.75	7.87
	5	4	1.20	2.82	6.00	6.66	6.77	7.86
	5	5	1.21	2.79	6.08	6.33	6.77	7.87

Table 21.6 Values of *V(Mean)* Under Different Values of ρ and a Joint *U*-Shaped Distribution With Densities Function $f(x)=3x^2/2$ $x\in[0, 1]$

Distribution	*n*	*k*	$\rho=0$	$\rho=0.05$	$\rho=0.50$	$\rho=0.75$	$\rho=0.90$	$\rho=0.95$
U-shaped $f(x)=3x^2/2$ $x\in[0, 1]$	2	2	1.15	2.27	3.43	3.85	4.44	5.55
	2	3	1.16	227	3.43	3.88	4.49	5.57
	2	4	1.15	2.27	3.42	3.89	4.30	5.58
	2	5	1.15	2.27	3.4 4	3.93	4.34	5.58
	5	2	1.18	2.27	3.46	3.93	4.34	5.59
	5	3	1.18	2.27	3.4 5	3.93	4.35	5.42
	5	4	1.16	2.30	3.47	3.96	4.37	5.48
	5	5	1.08	2.30	3.48	3.98	4.39	5.55

Table 21.7 Values of *V(Mean)* Under Different Values of ρ and a Joint Lognormal Distribution

Distribution	*n*	*k*	$\rho=0$	$\rho=0.05$	$\rho=0.5$	$\rho=0.75$	$\rho=0.90$	$\rho=0.95$
Lognormal(0,1)	2	2	1.17	2.22	3.23	4.55	5.55	6.11
	2	3	1.11	2.22	3.23	4.55	5.59	6.10
	2	4	1.17	2.22	3.23	4.54	5.58	6.18
	2	5	1.17	2.22	3.23	4.44	5.75	6.18
	5	2	1.18	2.22	3.22	4.44	5.77	6.26
	5	3	1.18	2.24	3.28	4.44	5.75	6.22
	5	4	1.17	2.32	3.27	4.46	5.77	6.36
	5	5	1.18	2.33	3.28	4.45	5.50	6.40

Table 21.2 exhibits that for values of $\rho = 0$, the value of $V(mean)$ is the smallest and then, its values are increased seriously for $\rho \geq 0.50$.

The results in Table 21.3 give a better idea of the effect of the correlation in detecting the non-random ordering of Y. The values of $V(mean)$ which are significant are marked with an "*." For $\rho = 0.50$ the significance is accepted for the pairs {(2,4), (2,5), (5,2), (5,3), (5,4), (5,5)}. The non-randomness of the ranking is accepted in all the cases for $\rho \geq 0.75$.

Table 21.4 gives an idea that for the exponential distribution for highly correlated variables, the value of $V(mean)$ is expected to be larger than 3.

For the gamma with density function $x^4 exp(-x)/\Gamma(5)$ and lognormal distributions (see Tables 21.5 and 21.7), the values of $V(mean)$ are expected to be close to 2 for $\rho < 0.50$.

Table 21.6 sustains a similar result for $\rho > 0.75$ in the case U-shaped distribution with density function $f(x) = 3x^2/2\ x \in [0, 1]$.

21.4.2 ANALYSIS OF THE TIME TO DEATH OF HIV-INFECTED PERSONS

We have considered a database of the lifetime of a set of 231 persons infected with HIV clustered by the risk-group. It constituted the following population:

G1—Drug users;
G2—Bisexual-homosexual men;
G3—Bisexual-lesbian women;
G4—Heterosexual men;
G5—Heterosexual women;
G6—Contaminated by blood transfusions;
G7—Sons of HIV-infected women;
G8—Unknown.

We selected 1000 independent (runs) samples from the data set to estimate treatment effects and compared them with the effect calculated with the population data. The estimated variance of treatment effects for the models using SRS and RSS for each group (i = 1,. . .,8) computed as

Table 21.8 Efficiencies of the Estimates of the Treatment Effects in eight Groups of Persons Infected With HIV (Variable Time to Death in Years) Using SRS and RSS

Group	$\hat{V}(\alpha_i^*)$	$\hat{V}(\alpha_{(i)}^*)$
1	1.97	1.24
2	1.52	2.78
3	1.20	2.19
4	1.05	2.11
6	1.85	1.72
6	1.45	1.11
7	1.44	1.79
8	1.94	1.30

$$\hat{V}(\alpha_i^*) = \frac{1}{1000}\sum_{t=1}^{1000}(\alpha_{ij}^* - \alpha_i)^2 = E(i;\ \text{srs}) \tag{21.11}$$

$$\hat{V}(\alpha_{(i)}^*) = \frac{1}{1000}\sum_{t=1}^{1000}(\alpha_{(i)j}^* - \alpha_i)^2 = E(i;\ \text{rss}) \tag{21.12}$$

We computed the ratio of the efficiencies and the results are given in Table 21.8. Note that both are considerably larger for the use of SRS. These results illustrate the behavior of RSS as an alternative for estimating the treatment effects and variability. Due to the nature of the data non-normality was present, hence the use of ANOVA for fixing the existence of the significance of the observed differences did not make sense.

21.5 CONCLUSIONS

The use of RSS in ANOVA is at least as good as the SRS methodology. This result supports that RSS-designed experiments can be analyzed using one-way ANOVA. The estimation of the variance using RSS allows establishing the closeness of the ranking to the perfect ranking, assumed in the modeling. *V(n)* is a nonparametric statistic that can be used for analyzing the quality of the ranking. Further study is needed to establish rules for evaluating the relative precision of RSS as a function of the quality of the ranking.

REFERENCES

Chen, Z., Bai, Z., Sinha, B., 2003. Ranked Set Sampling: Theory and Applications, vol. 176. Springer Science & Business Media, New York.

Dell, T.R., Clutter, J.L., 1972. Ranked set sampling theory with order statistics background. Biometrics 28 (2), 545–555.

Frey, J., 2011. A note on ranked-set sampling using a covariate. J. Stat. Plan. Inference 141 (2), 809–816.

MacEachern, S.N., Öztürk, Ö., Wolfe, D.A., Stark, G.V., 2002. A new ranked set sample estimator of variance. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) 64 (2), 177–188.

McIntyre, G.A., 1952. A method for unbiased selective sampling, using ranked sets. Aust. J. Agric. Res. 3 (4), 385–390.

Muttlak, H.A., 1998. Median ranked set sampling with concomitant variables and a comparison with ranked set sampling and regression estimators. Environmetrics 9 (3), 255–267.

Patil, G.P., Sinha, A.K., Taillie, C., 1995. Finite population corrections for ranked set sampling. Ann. Inst. Stat. Math. 47 (4), 621–636.

Perron, F., Sinha, B.K., 2004. Estimation of variance based on a ranked set sample. J. Stat. Plan. Inference 120 (1-2), 21–28.

Stokes, S.L., 1980. Estimation of variance using judgment ordered ranked set sample. Biometrics 36, 35–42.

Takahasi, K., Wakimoto, K., 1968. On unbiased estimates of the population mean based on the sample stratified by means of ordering. Ann. Inst. Stat. Math. 20 (1), 1–31.

Index

Note: Page numbers followed by "*f*" and "*t*" refer to figures and tables, respectively.

RANKED SET SAMPLING

65 Years Improving the Accuracy in Data Gathering

Edited by
Carlos N. Bouza-Herrera, Amer Ibrahim Falah Al-Omari

Ranked Set Sampling is an advanced survey technique, which seeks to improve the likelihood that collected sample data provide a good representation of the population, and minimizes costs associated with obtaining them. The main focus of many agricultural, ecological, and environmental studies is to develop well designed, cost-effective, and efficient sampling designs, giving ranked set sampling (RSS) techniques a particular place in resolving the disciplinary problems of economists in application contexts, particularly experimental economics. RSS has been successfully used across regression analysis, consumer expenditure surveys, and alternative index estimation, but the technique is not as well understood or used as it might be. This book seeks to place RSS at the heart of economic study designs.

- Focused on how researchers should manipulate RSS techniques for specific applications
- Discusses how RSS performs in popular statistical models such as regression and hypothesis testing
- Includes discussion of open theoretical research problems
- Provides mathematical proofs, enabling researchers to develop new models

Dr. Carlos N. Bouza-Herrera is a professor of mathematics, economics, and computation at University of Havana, Havana, Cuba. Dr. Herrera has headed more than 60 research projects and authored more than 200 papers. His main area of investigation is in mathematical statistics. He has been author, coauthor, or editor of 19 books. He is a consulting editor of *Current Index to Statistics and International Abstracts of Operations Research*.

Dr. Amer Ibrahim Falah Al-Omari is a professor of statistics at the Department of Mathematics, and Dean of Academic Research at Al al-Bayt University, Mafraq, Jordan. He is interested in ranked set sampling, entropy, missing data, order statistics, acceptance sampling plans, and statistical inference. He has authored over 100 articles.

ELSEVIER

ACADEMIC PRESS
An imprint of Elsevier
elsevier.com/books-and-journals

ISBN 978-0-12-815044-3

9 780128 150443