Hacettepe Journal of Mathematics and Statistics
Volume 47 (3) (2018), 755 – 761

Distribution function estimation using concomitant-based ranked set sampling

Ehsan Zamanzade*[†] and M. Mahdizadeh[‡]

Abstract

Ranked set sampling (RSS) is a data collection method designed to exploit auxiliary ranking information. In this paper, a new estimator of distribution function is proposed when RSS is done by using a concomitant variable. It is shown by simulation study that the alternative estimator can be considerably more efficient than the standard one, especially when the rankings are perfect.

Keywords: Relative efficiency, Concomitant variable, Judgment ranking

2000 AMS Classification: 62F10; 62D05.

Received : 10.02.2016 *Accepted* : 28.08.2016 *Doi* : 10.15672/HJMS.201814420708

1. Introduction

There are situations where the actual quantification of the variable of interest, say Y, is difficult (expensive, time-consuming or destructive), but sampling units can be easily ordered without actual measurement. Ranking is usually done by expert opinion, concomitant variable, or a combination of them. Ranked set sampling (RSS) is sampling plan which can be effectively employed in the above situations. It often results in improved statistical inference, for many population attributes, over simple random sampling (SRS).

The RSS was introduced by McIntyre [6] in the context of estimating pasture yields. He noticed that measuring a plot yield requires harvesting its crops, but an expert can simply sort the adjacent plots in terms of their pasture yields by eye inspection. Motivated by McIntyre [6]'s work, RSS has found many applications in other contexts, including forestry, medicine, biometrics, environmental monitoring and entomology. For a book-length treatment of RSS and its applications, see Chen [1].

*Department of Statistics, University of Isfahan, Isfahan, 81746-73441, Iran, Email: e.zamanzade@sci.ui.ac.ir; ehsanzamanzadeh@yahoo.com

[†]Corresponding Author.

[‡]Department of Statistics, Hakim Sabzevari University, P.O. Box 397, Sabzevar, Iran, Email: mahdizadeh.m@live.com

There has been much research in RSS since its introduction. Takahasi and Wakimoto [12] were the first who proved that sample mean in RSS, $\bar{Y}_{RSS}$, is unbiased for the population mean and more efficient than sample mean in SRS, $\bar{Y}_{SRS}$. Furthermore, they showed that the efficiency of $\bar{Y}_{RSS}$ to $\bar{Y}_{SRS}$ is maximized when the population distribution is uniform. Stokes and Sager [11] considered the problem of estimating the cumulative distribution function (CDF), and proved that RSS CDF estimator is more efficient than its counterpart in SRS regardless of the ranking quality. Stokes [10], MacEachern et al. [7] and Zamanzade and Vock [16] proposed some variance estimators based on a ranked set sample. As the ranking process in RSS is performed without obtaining precise values of the sample units, it may not to be accurate (perfect). Frey [3] and Li and Balakrishnan [5] proposed some nonparametric tests for assessing perfect ranking assumption which were followed by Vock and Balakrishnan [13], Zamanzade et al. [14] and Zamanzade et al. [15]. The problem of estimating the population mean and variance when RSS is applied by measuring a concomitant variable were discussed by Frey [4], Zamanzade and Mohammadi [17] and Zamanzade and Vock [16]. In this work, we plan to develop a more efficient CDF estimator when judgment ranking is performed using a concomitant variable.

In Section 2, the concomitant-based RSS is described, and our CDF estimator is presented. In Section 3, the proposed estimator is compared with the standard one in RSS. In Section 4, the new method is illustrated using a real data example. Final conclusions are given in Section 5.

2. The CDF estimation in concomitant-based RSS

Let Y and X be the variable of the interest and the concomitant variable, respectively. We assume that the exact measurement of the variable of interest Y is expensive, destructive or time-consuming, but actual quantification of the concomitant variable X can be easily obtained. The concomitant-based RSS can be described as follows:

(1) Draw a simple random sample of size k^2 from (Y, X), and partition them into k samples of size k each.

(2) In each sample of size k, obtain the exact quantification of the concomitant variable X, and then sort the sample according to the X values. It is assumed that the researcher is aware about the sign of correlation coefficient between X and Y.

(3) Actually measure the Y value of the ith ($i = 1, \dots, k$) ordered unit in the ith sample.

(4) Repeat the steps (1)-(3), n times (cycles) to obtain a ranked set sample of size nk.

The resulting ranked set sample is then denoted by

$$\left\{Y_{[i]ij} : i = 1, \dots, k;\ j = 1, \dots, n\right\},$$

where $Y_{[i]ij}$ is the ith judgement ordered unit in the jth cycle. We use the term *judgement order* and the subscript [.] in $Y_{[i]ij}$ to emphasize that the Y value of ith ordered unit in step (3), may not be the true ith order statistic of the variable of interest Y, in the ith sample. This is so because the ranking process in step (2) is performed according to the concomitant variable X. Obviously, the quality of the ranking depends on the correlation between concomitant variable and the variable of the interest. If the variable of interest Y is a one-to-one function of the concomitant variable X, then ranking process is perfect. In this case, we use the subscript (.) for the ranked set sample units, which are denoted by

$$\left\{Y_{(i)ij} : i = 1, \dots, k;\ j = 1, \dots, n\right\}.$$

The standard CDF estimator in RSS is given by

$$(2.1) \qquad \hat{F}_{RSS}(t) = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n} I\left(Y_{[i]ij} \le t\right).$$

The properties of $\hat{F}_{RSS}(t)$ have been studied by Stokes and Sager [11]. They proved that this estimator is unbiased and has less variance than empirical distribution function (EDF) in SRS of the same size, regardless of the quality of ranking.
In concomitant-based RSS with set size k and cycle size n, we measure $M = nk^2$ units on the concomitant variable, and use them for ranking purpose, but we only measure $N = nk$ of units on the variable of interest. The main idea behind our proposed procedure is to improve accuracy of the CDF estimation by exploiting the information contained in all measurements on the concomitant variable.

Let $\left\{X_{(i_1)i_2j} : i_1, i_2 = 1, \ldots, k;\ j = 1, \ldots, n\right\}$ be the set of all concomitant variable values which are used to obtain the ranked set sample of

$$\left\{Y_{[i]ij} : i = 1, \ldots, k;\ j = 1, \ldots, n\right\},$$

where $X_{(i_1)i_2j}$ is the i_1th ordered unit in the i_2th sample of the jth cycle. Let $X_{(1)}, \ldots, X_{(M)}$ be the ordered values of the concomitant variable

$$\left\{X_{(i_1)i_2j} : i_1, i_2 = 1, \ldots, k;\ j = 1, \ldots, n\right\}.$$

Also, suppose $\left(Y_{[1]}^m, X_{(1)}^m\right), \ldots, \left(Y_{[N]}^m, X_{(N)}^m\right)$ are quantifications on the variable of interest along with the corresponding values of the concomitant variable, where $X_{(1)}^m < \cdots < X_{(N)}^m$.

In view of the identity $F_Y(t) = E\left(E\left(I\left(Y \le t\right)|X\right)\right) = E\left(F_{Y|X}(t)\right)$, the estimator of $E\left(F_{Y|X}(t)\right)$ can be considered as population CDF estimator. Since

$$Var\left(I\left(Y \le t\right)\right) = E\left(Var\left(I\left(Y \le t\right)|X\right)\right) + Var\left(F_{Y|X}(t)\right),$$

one would expect the estimator of $F_{Y|X}(t)$ to have smaller variance than that of $E\left(I\left(Y \le t\right)\right)$. The quantity $E\left(F_{Y|X}(t)\right)$ is estimated by taking the average over kN estimates of $F_{Y|X_{(i)}}(t)$, for $i = 1, \ldots, kN$.

Here, we assume that $F_{Y|X}(t)$ is non-increasing function of X, and hence the estimates of $F_{Y|X_{(i)}^m}(t)$ should be non-increasing in i, as well. However, the estimates of $F_{Y|X_{(i)}^m}(t)$ may not be non-increasing in i due to sampling noise. One can resolve this problem by using nonparametric isotonic regression. Let $\hat{F}_{Y_{[i]}^m}(t) = I\left(Y_{[i]}^m \le t\right)$. We find the values of $\hat{F}_{Y_{[i]}^m}^{iso}(t)$ such that $\sum_{i=1}^{N}\left(\hat{F}_{Y_{[i]}^m}(t) - F_{Y_{[i]}^m}^{iso}(t)\right)^2$ is minimized under the constraint

$$F_{Y_{[1]}^m}^{iso}(t) \ge \cdots \ge F_{Y_{[N]}^m}^{iso}(t).$$

The $\hat{F}_{Y_{[i]}^m}^{iso}(t)$ values can be found quite efficiently by using pool adjacent violator algorithm (PAVA) (see Robertson [9], Chapter 1). It can be shown that for $i = 1, \ldots, N$,

$$\hat{F}_{Y_{[i]}^m}^{iso}(t) = \min_{r \le i}\max_{s \ge i}\sum_{g=r}^{s}\frac{\hat{F}_{Y_{[g]}^m}(t)}{s - r + 1}.$$

Now, we estimate $F_{Y|X^m_{(i)}}(t)$ by using linear interpolation of closest known values of $\hat{F}^{iso}_{Y^m_{[i]}}(t)$ on either side, i.e.,

$$\hat{F}_{Y|x}(t) = \begin{cases} \hat{F}^{iso}_{Y^m_{[1]}}(t), & x < X^m_{(1)}, \\ \hat{F}^{iso}_{Y^m_{[i]}}(t) + \frac{\hat{F}^{iso}_{Y^m_{[i+1]}}(t) - \hat{F}^{iso}_{Y^m_{[i]}}(t)}{X^m_{(i+1)} - X^m_{(i)}} \left(x - X^m_{(i)}\right), & X^m_{(i)} \le x < X^m_{(i+1)} \quad (i = 1, \ldots, N), \\ \hat{F}^{iso}_{Y^m_{[N]}}(t), & x \ge X^m_{(N)}. \end{cases}$$

Finally, the proposed CDF estimator is $\hat{F}_N(t) = \frac{1}{kN} \sum_{i=1}^{kN} \hat{F}_{Y|X_{(i)}}(t)$.

3. Monte Carlo Comparisons

We conducted a simulation study to assess the performance of the proposed estimator in concomitant-based RSS. To this end, we used an imperfect ranking model introduced by Dell and Clutter [2]. It assumes that (Y, X) has a bivariate normal distribution with correlation coefficient ρ. The selected values of ρ are $\rho = 1$ for perfect ranking, $\rho = 0.8$ for imperfect ranking with fairly good accuracy, and $\rho = 0$ for random ranking. It is worth mentioning that we assume that the researcher is aware of the sign of correlation coefficient between the interest variable and the concomitant variable, and therefore the simulation results do not depend on the sign of ρ.

We first take $Y\ (\in \mathbb{R})$ as the variable of interest, so the relation between variable of interest and concomitant variable is linear. We then take $e^Y\ (\in \mathbb{R}^+)$ as the variable of interest, therefore the relation between variable of interest and concomitant variable is non-linear. Also, three configurations of the sample size and the set size considered are $(N, k) = (10, 5), (10, 10)$ and $(20, 5)$. This allows us to observe the effect of increasing the sample (set) size when the set (sample) size is fixed.

For each combination of ρ, N and k, 100,000 samples were generated in RSS scheme. From each sample, the estimators $\hat{F}_N(t)$ and $\hat{F}_{RSS}(t)$ were computed for both response variables (Y and e^Y). Finally, the efficiency of $\hat{F}_N(t)$ relative to $\hat{F}_{RSS}(t)$ is estimated by

$$RE(t) = \frac{MSE\left(\hat{F}_{RSS}(t)\right)}{MSE\left(\hat{F}_N(t)\right)},$$

where $MSE\left(\hat{F}_{RSS}(t)\right)$ and $MSE\left(\hat{F}_N(t)\right)$ are estimated mean squared errors for the two CDF estimators based on 100,000 replications. The simulation results are presented in Figures 1 and 2. In any plot, the top, middle and bottom curves are corresponding to $\rho = 1$, $\rho = 0.8$ and $\rho = 0$, respectively.

The results confirm the preference of the new estimator. The proposed estimator has considerably better performance than its empirical counterpart when the rankings are perfect, especially at the boundaries. Interestingly, $\hat{F}_N(t)$ is still the better than $\hat{F}_{RSS}(t)$ when $\rho = 0.8$, in most cases. However, when the rankings are completely random ($\rho = 0$), then the estimated relative efficiencies are less than one, but the efficiencies loss are not much in this case. Furthermore, the relative efficiency increases as sample size (N) increases and the rankings are fairly good

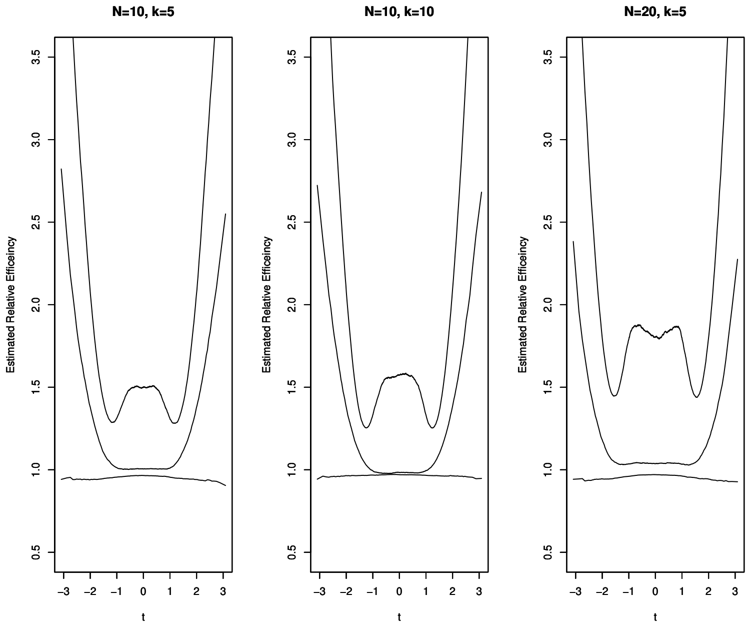

Figure 1. The estimated relative efficiencies in concomitant-based RSS scheme, when the relation between the target variable and concomitant variable is linear.

($\rho \geq 0.8$). These figures also suggest that a greater efficiency gain is obtained by increasing the cycle size (n) rather than set size (k).

4. Application to a real data set

In this section, we use a data set from Platt et al. [8] to illustrate the application of the new CDF estimator. It contains seven variables related to 396 conifer (pinus palustris) trees. We only consider two variables: X the diameter in centimeters at breast height, and Y the entire height in feet. The data set can be found in Appendix B of Chen [1].

We treat the tree data as the target population, where Y is the variable of interest, and X is concomitant variable. The correlation of coefficient between X and Y in the population is 0.91. The CDF of Y is given by

$$F_Y\left(t\right) = \frac{1}{396}\sum_{i=1}^{396} I\left(Y_i \leq t\right).$$

For the same choices of the sample size and the set size in Section 3, 100,000 samples were drawn from the population in concomitant-based RSS design. From each sample, the estimators $\hat{F}_N\left(t\right)$ and $\hat{F}_{RSS}\left(t\right)$ were computed at $t = F_Y^{-1}\left(p\right), p = 0.1, 0.25, 0.5, 0.75, 0.9$, where $F_Y^{-1}\left(.\right)$ is the quantile function associated with Y. Table 1 displays values of $RE\left(t\right)$ defined as in the previous section. It is observed

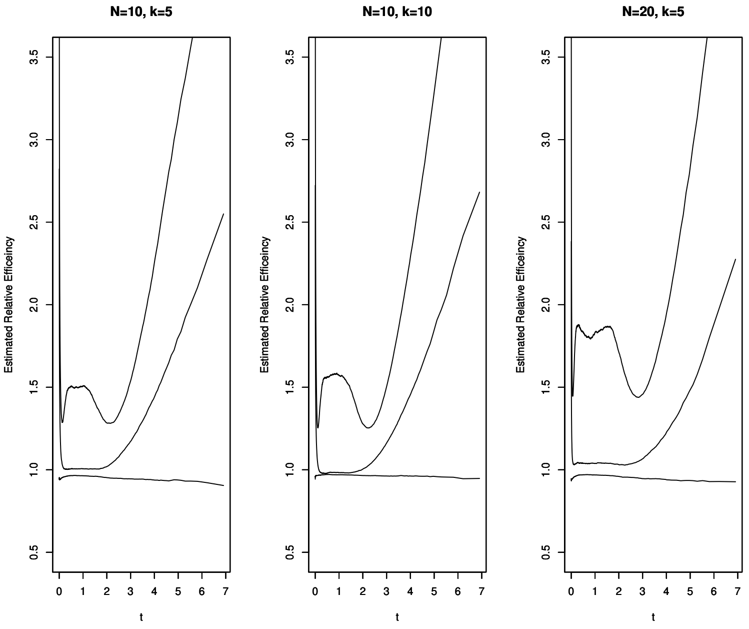

Figure 2. The estimated relative efficiencies in concomitant-based RSS scheme, when the relation between the target variable and concomitant variable is non-linear.

Table 1. Estimated efficiencies of $\hat{F}_N(t)$ relative to $\hat{F}_{RSS}(t)$ at pth quantiles, for $p \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$.

	p				
(N,k)	0.1	0.25	0.5	0.75	0.9
$(5,5)$	1.03	1.37	1.19	1.40	1.02
$(5,10)$	1.00	1.43	1.15	1.44	1.00
$(10,5)$	1.22	1.68	1.23	1.52	1.03

that the proposed estimator outperforms the standard one in concomitant-based RSS design.

5. Conclusion

The idea of ranking using a concomitant variable has been widely employed for statistical inference. Under this setup, we consider the problem of CDF estimation in RSS. The standard estimator does not make efficient use of the available information. An alternative estimator, incorporating the concomitant variable information, is proposed and its finite sample behavior is investigated in simulation. The results suggest that the new approach tends to be highly efficient in some situations, especially if the rankings are perfect. They also suggest that the higher relative efficiency is achieved by increasing the number of cycles rather than set size, as long as the quality of ranking is fairly good.

Acknowledgement

The authors are thankful to two anonymous referees for their valuable comments and suggestions which improved the earlier version of this paper.

References

[1] Chen, Z., Bai, Z., and Sinha, B.K. *Ranked set sampling: Theory and Applications*, Springer, New York, 2004.

[2] Dell, T.R., and Clutter, J.L. *Ranked set sampling theory with order statistics background*, Biometrics **28**(2), 545-555, 1972.

[3] Frey, J., Ozturk, O., and Deshpande, J.V., *Nonparametric tests for perfect judgment rankings*, Journal of the American Statistical Association **102**(478), 708-717, 2007.

[4] Frey, J. *A note on ranked-set sampling using a covariate*, Journal of Statistical Planning and Inference **141**(2), 809-816, 2011.

[5] Li,T., and Balakrishnan, N. *Some simple nonparametric methods to test for perfect ranking in ranked set sampling. Journal of Statistical Planning and Inference* **138**(5), 1325-1338, 2008.

[6] McIntyre, G.A. *A method for unbiased selective sampling using ranked set sampling*, Australian Journal of Agricultural Research, 3, 385-390, 1952.

[7] MacEachern, S.N., Ozturk, O., Wolfe, D.A. and Stark, G.V. *A new ranked set sample estimator of variance*, Journal of the Royal Statistical Society: Series B. **64**(2), 177-188, 2002.

[8] Platt, W.J., Evans, G.M., and Rathbun, S.L. *The population dynamics of a long-lived conifer (Pinus palustris)*, American Naturalist 131, 491–525, 1988.

[9] Robertson, T., Wright, F.T., and Dykstra, R.L. *Order Restricted Statistical Inference*, Wiley, New York, 1988.

[10] Stokes, S.L. *Estimation of variance using judgment ordered ranked set samples*, Biometrics **36**(1), 35-42, 1980.

[11] Stokes, S.L., and Sager, T.W. *Characterization of a Ranked-Set Sample with Application to Estimating Distribution Functions*, Journal of the American Statistical Association **83**(402), 374-381, 1988.

[12] Takahasi, K. and Wakimoto, K. *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, 1968.

[13] Vock, M., and Balakrishnan, N. A. *Jonckheere-Terpstra-type test for perfect ranking in balanced ranked set sampling*, Journal of Statistical Planning and Inference **141**(2), 624-630, 2011.

[14] Zamanzade, E. , Arghami, N.R., Vock, M. *Permutation-based tests of perfect ranking*, Statistics & Probability Letters **82**(12), 2213-2220, 2012.

[15] Zamanzade, E., Arghami, N.R., and Vock, M. *A parametric test of perfect ranking in balanced ranked set sampling*, Communications in Statistics: Theory and Methods **43**(21), 4589-4611, 2014.

[16] Zamanzade, E,. and Vock, M. *Variance estimation in ranked set sampling using a concomitant variable*, Statistics & Probability Letters **105**, 1-5, 2015.

[17] Zamanzade, E,. and Mohammadi, M. *Some modified mean estimators in ranked set sampling using a covariate*, Journal of Statistical Theory and Applications **15**(2), 142-152, 2016.