Models for Ranking Errors in Ranked Set Sampling

Aparna R. Gurao
Research Scholar, J.J.T. University, Rajasthan

Abstract - Ranked Set Sampling was proposed by McIntire (1951) in order to reduce measurement cost and still obtain an efficient estimator of the population mean. The purpose of this paper is to look in to some of the ranking error models proposed in the literature and evaluate their impact in the gain in precision of ranked set sampling over simple random sampling. This paper reviews some of the models suggested by researchers in the literature on ranked set sampling. Subsequent studies will be undertaken for developing new and hopefully better models for ranking errors in ranked set sampling.

Keywords: Ranked set sampling, ranking error, judgmental error, efficiency of estimator, ordered statistics

I. INTRODUCTION

Ranked Set Sampling can be viewed as an alternative to Simple Random Sampling for obtaining observations from a statistical population. In comparison to simple random sampling, ranked set sampling can provide improved estimators and more powerful statistical tests when observations are expensive, destructive, or time-taking and it is possible to compare the sampling units without making any measurement or quantification. McIntyre (1952) introduced ranked set sampling for estimating total crop yield. Dell and Clutter (1972) applied it to estimation of forage yield. Mode et al. (1999) used ranked set sampling for estimating stream habitat area. Ranked set sampling was used for estimating grazing effects on ryegrass and covers by Cobby et al. (1985), for estimating plutonium levels in soil by Gilbert (1995), and for gasoline sampling by Nussbaum and Sinha (1997). As matter of fact, there is a substantial amount of statistical research investigating benefits of ranked set sampling. Kaur et al. (1995) provide a partial review of parametric ranked set sampling, while Bohn (1996) provides a review of nonparametric ranked set sampling, that is, ranked set sampling for nonparametric procedures. Most of the comparisons between ranked set sampling and simple random sampling are based on the assumption that there is no error in comparing and judgmentally ranking sampling units. While this may be a desirable assumption, it is hardly realistic. Nevertheless, it is believed in many situations that even if ranking may not be prefect, it may still be beneficial because it may still be better than simple random sampling. In some sense, it may be argued that the worst case of ranking errors may simply lead to a random ranking of sampling units, making ranked set sampling equivalent to simple random sampling. In essence, ranked set sampling is not supposed to be worse than simple random sampling under any circumstances. Even then, ranking errors are serious and therefore must be investigated and understood so that their impact can be moderated, if not eliminated completely. Additionally, ranking errors must be investigated and analyzed because perfect ranking is a rather unrealistic assumption, that can hardly occur in natural situations.

The purpose of this paper is to look in to some of the ranking error models proposed in the literature and evaluate their impact in the gain in precision of ranked set sampling over simple random sampling. The paper begins with a discussion of some of the desirable properties of possible ranking error models. First and foremost, these models should be sufficiently flexible to cover a wide range of the degree of judgmental error in ranking. The model should capture, at one end, the case of random or blind ranking, as well as, at the other end, the case of prefect ranking, covering most of the reasonable cases of imperfect ranking. These models should also be applicable to different set sizes, so that they can be useful in a search for optimal set size under any specified ranking error model.

II. SOME RANKING ERROR MODELS IN THE LITERATURE

This section describes the notation for ranked set sampling used in this paper. It also contains some models of imperfect ranking in ranked set sampling from the literature.

The population of interest follows a probability distribution specified by the cumulative distribution function (CDF) F with a location parameter θ. A ranked set sample is obtained as follows. First, a random sample

of size m is selected from the population, where m is called the set size. The m observations are ranked from largest to smallest in some way without making any measurement on the sampling units in the sample. It is usually assumed that this ranking is done judgmentally by visual inspection or with help of a concomitant variable that does not require making any measurement. The sampling unit believed to be the largest, also called the first judgment order statistic, is selected for inclusion in the sample. It is denoted by $X_{[1]}$. Then a second simple random sample of size m is selected from the population and the sampling unit that is judged to be second largest is selected for inclusion in the sample. This sampling unit is called the second judgment order statistic and is denoted by $X_{[2]}$. This procedure is repeated until, at the m^{th} step, a simple random sample of size m is selected and the sampling unit that is judged to be the smallest is selected for inclusion in the sample. It is the m^{th} judgment order statistic and is denoted by $X_{[m]}$. This cycle of m selections is repeated r times to obtain a total of rm^2 sampling units in the sample. The sampling units in the sample are denoted using a subscript for the judgment order and a superscript for the cycle. In this way, the sampling units in the first cycle are $X^1_{[1]}, X^1_{[2]}, \dots, X^1_{[m]}$. In general, the sampling units selected in the i^{th} cycle are $X^i_{[1]}, X^i_{[2]}, \dots, X^i_{[m]}$ for $i = 1, 2, \dots, r$. The resulting sample, in this way, consists of the mr sampling units $X^i_{[j]}, j = 1, 2, \dots, m; i = 1, 2, \dots, r$ and is called a ranked set sample with set size m and r cycles.

Some attempts to model imperfect ranking mechanisms can be found in the literature. Dell and Clutter (1972) developed a model by adding a random error to the observations. Thus, $Y_i = X_i + \epsilon_i$, where ϵ_i is the error in the judgment ranking procedure, and sampling units are judgmentally ranked according to the variable Y instead of X. further, $X_{[j]}$ is the X value of the sampling unit corresponding to $Y_{(j)}$, the j^{th} judgment order statistic. Dell and Clutter considered the case where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$, while the variable X can be assumed to follow any distribution that may be reasonable in practice. The major problem with this model is that it is over simplified for ranked set sampling. Further, the authors mention that ranking errors can be influenced by the set size and the magnitudes of the sampling units themselves. However, none of these considerations can be taken into account through this model.

Stokes (1977) suggested ranking based on a concomitant variable. Her model can be used to determine the relative precision of ranked set sampling under imperfect ranking if a concomitant variable exists and the two variables for certain strict distributional assumptions. She assumes that a concomitant variable is available for the variable of interest and it can be measured accurately almost without any cost. She suggested that this concomitant variable should be used for ranking sampling units in ranked set sampling. For example, suppose that the variable of interest, X, has mean μ_X and variance σ_X^2, while the concomitant variable Y has mean μ_Y and variance σ_Y^2. Further, assume that $(X - \mu_X)/\sigma_X$ and $(Y - \mu_Y)/\sigma_Y$ follow a common distribution. This model has a good property that, when it fits a given situation well, it provides nice results on the efficiency of ranked set sampling relative to simple random sampling. The model can also incorporate the information obtained through the concomitant variable in the properties of the concomitant ranked set sampling estimator of the population mean of the variable of interest. The main problem with this model, however, is that its assumptions are very restrictive and most practical situations are unlikely to satisfy these assumptions.

Bohn and Wolfe (1994) developed a ranking error model based on the expected spacing. The expected spacing between two order statistics $X_{(i)}$ and $X_{(j)}$ is given by $|E[X_{(i)} - X_{(j)}]|$. Bohn and Wolfe use these expected spacing for defining the probabilities that the item that has actual rank i in the set is chosen as the j^{th} judgment order statistic. More precisely, let $p_{i,j}$ denote the probability that the truly rank i sampling unit is erroneously assigned rank j. In other words,

$$p_{t,j} = P(X_{[j]} = X_{(t)}) \quad (1)$$

The P matrix is then defined as the matrix having these elements. In other words,

$$P = \left((p_{t,j})\right) = \left((P(X_{[j]} = X_{(t)}))\right) \quad (2)$$

The case of perfect ranking is obtained by putting $p_{t,j} = 1$ and $p_{t,j} = 0$ for $t \neq j$. If the ranking is completely randomly done, then $p_{t,j} = \frac{1}{m}$ for all t and j. However, in general, no specific structure is required for the $p_{t,j}$'s. The only requirement is that, for every $j = 1,2,\dots,m$, we have $\sum_{t=1}^{m} p_{t,j} = 1$ because the j^{th} judgment order statistic has to be $X_{(t)}$ for some $t = 1,2,\dots,m$. The $p_{t,j}$'s are the parameters of the model and it is not possible to estimate these probabilities from a ranked set sample, because the true order statistic will never be known unless all sampling units in the set are subjected to measurement. Bohn and Wolfe take

$$p_{t,j} = \frac{a_j}{|E[X_{(t)} - X_{(j)}]|}, for\ t \neq j \quad (3)$$

and

$$p_{t,t} = c_t. \quad (4)$$

This specification of the $p_{t,j}$'s and the doubly stochastic restriction on the matrix P made up by the elements $p_{t,j}$ result in a complete specification of the model, with only one problem. This model heavily depends on the underlying probability distribution of the variable of interest X through its expected spacing. What can be derived is the following result related to the distribution of the judgment order statistic in terms of the true rank order.

$$F_{[j]}(x) = \sum_{t=1}^{m} p_{t,j} F_{(t)} \ for\ j = 1,2,\dots,m. \quad (5)$$

This relationship can be used to compute the relative precisions of different estimators based on the ranked set sample obtained under this imperfect ranking model.

The expected spacing model appeals the intuition because it is reasonable to claim that, if the expected difference between any two order statistics is large, then it is unlikely that the ranking in a set will result in one of these order statistics being incorrectly taken to be the other. It is not claimed that it cannot happen, but only that it is less likely to happen than if the expected difference between them is small. One of the major drawbacks of the expected spacing models is that while it explicitly uses the expected spacing, it does not take in to account the variance of the underlying distribution at all. Another drawback pointed out by Presnell and Bohn (1999) is the assumption of independence. One argument in support of the model, nevertheless, is that while this assumption is not true, it appears to a reasonable approximation for a majority of distributions that are likely to be encountered in practice. Finally, it is important to note that the expected spacing model depends on the underlying distribution of the variable of interest, and thus is not nonparametric. Also, it is difficult to extend to set sizes large than five due to computational complexities. Even the model derived by Bohn and Wolfe for $m = 4$ is hard to work with. This drawback makes it impossible to use this model for determining optimal set size, or to decide which order statistics to select for making measurement when the sample size is greater than four.

In spite of the difficulties in implementing the model of Bohn and Wolfe, it is mathematically interesting and hence is studied to some depth in the following section.

III. PROPERTIES OF THE $\boldsymbol{P}$ MATRIX

The relationship

$$F_{[j]}(x) = \sum_{i=1}^{m} p_{i,j} F_{(i)}(x) \qquad (6)$$

is used by Bohn and Wolfe for the expected spacing model, where $p_{i,j}$ is the probability that the i^{th} order statistic from the set of size m is judgment ranked as the j^{th} order statistic in the set. In other words,

$$p_{i,j} = P\left(X_{[j]} = X_{(i)}\right), \quad for\ i,j = 1, 2, \ldots, m. \qquad (7)$$

Let

$$P = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,m} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m,1} & p_{m,2} & \cdots & p_{m,m} \end{bmatrix} \qquad (8)$$

For the relationship (6) to hold, it is necessary to have

$$\sum_{i=1}^{m} p_{i,j} = 1 \ for\ each\ j = 1, 2, \ldots, m. \qquad (9)$$

This requirement is automatically satisfied due to the procedure used to obtain a ranked set sample. It corresponds to the fact that the j^{th} judgment order statistic, which is selected from set number j, must be equal to exactly one true order statistic and thus corresponds to the requirement that the columns of $\boldsymbol{P}$ have all unit sum. This is a column condition.

The corresponding row condition, namely,

$$\sum_{j=1}^{m} p_{i,j} = 1 \ for\ each\ i = 1, 2, \ldots, m. \qquad (10)$$

is not assumed by Bohn and Wolfe in their model. The matrix $\boldsymbol{P}$ is considered to apply for an entire cycle of ranked set sampling. For this interpretation of $\boldsymbol{P}$, condition 10 is not necessary. As a matter of fact, this condition would imply that the i^{th} order statistic could be selected only once in a cycle. This, however, cannot be assumed for ranked set sampling with imperfect ranking. For example, in the first set of m sampling units, we could correctly have $X_{[1]} = X_{(1)}$, but in the second set of m sampling units, the sampling unit judged to be rank 2 is actually the rank 1 unit, thus giving $X_{[2]} = X_{(1)}$. This in now way implies that the same sampling unit is selected twice, since the order statistics are from two independent sample of size m each from the distribution F. However, what is worth noting in this situation is that, since the first order statistic is selected twice, consequently some other order statistic will not be selected at all in the cycle.

If it is further assumed that the underlying distribution of the population values is unimodal, it is reasonable to further assume that

$$p_{1,1} \geq p_{2,2} \geq \cdots \geq p_{\left[\frac{m}{2}\right],\left[\frac{m}{2}\right]}. \qquad (11)$$

$$p_{m,m} \geq p_{m-1,m-1} \geq \cdots \geq p_{\left[\frac{m}{2}\right],\left[\frac{m}{2}\right]}. \qquad (12)$$

This corresponds to the situation where it is easier to judgment rank extreme sampling units than central ones in such distributions. This is quite intuitive because values in the tails of unimodal distributions tend to be further apart compared to those that fall closer to the center or mode of the distribution.

It is also reasonable to assume that $p_{j,j} \geq p_{i,j}$ *for all* $i \neq j$ for every $j = 1, 2, \ldots, m$. That is, the most likely event is that of correct judgment ranking. This is reasonable because ranked set sampling would not be useful in situations where this assumption is not appropriate.

Furthermore, when the underlying distribution is symmetric, some other symmetries should also appear in the matrix P. For instance, $p_{1,1} = p_{m,m}, p_{2,2} = p_{(m-1),(m-1)}$, and so on. More generally, we should also expect $p_{i,j} = p_{(m+1-i),(m+1-j)}$. However, we cannot expect the P matrix to be totally symmetric. For example, the condition $p_{i,j} = p_{j,i}$ would mean that the probability of judgment ranking the i^{th} order statistic to be the j^{th} is the same as the probability of judgment ranking the j^{th} order statistic to be the i^{th}. There is not reason to believe that this can happen even when the underlying distribution is symmetric.

REFERENCES

[1] Bohn, L. L. (1996). A review of nonparametric ranked set sampling methodology. communications in Statistics, Theory and Methods, Vol. 25, pp. 2675 - 2685.

[2] Bohn, L. L. and Wolfe, D. A. (1994). The effect of imperfect judgment rankings on properties of procedures based on the ranked set samples analog of the Mann-Whitney-Wilcoxon statistic. Journal of the American Statistical Association, Vol. 89, pp. 168 - 176.

[3] Cobby, J. M., Ridout, M. S., Bassett, P. J., and Large, R. V. (1985). An investigation into the use of ranked set sampling on grass and grass clover swards. Grass and Forage Science, Vol. 40, pp. 257 - 263.

[4] Dell, T. R. and Clutter, J. L. (1972). Ranked set sampling theory with order statistics background. Biometrics, Vol. 28, pp. 545 - 553.

[5] Gilbert, R. O. (1995). DQO statistics bulletin: statistical methods for the data quality objective process. Technical Report PNL-SA-26377, The Pacific Northwest Laboratory.

[6] Kaur, A., Patil, G. P., Sinha, A., and Taillie, C. (1995). Ranked set sampling: an annotated bibliography. Environmental and Ecological Statistics, Vol. 2, pp. 25 - 54.

[7] McIntyre, G. A. (1952). A method for unbiased selective sampling, using ranked sets. Australian Journal of Agricultural Research, Vol. 3, pp. 385 - 390.

[8] Mode, N. A., Conquest, L. L., and Marker, D. A. (1999). Ranked set sampling for ecological research: accounting for the total costs of sampling. Environmetrics, Vol. 10, pp. 179 - 194.

[9] Nussbaum, B. D. and Sinha, B. D. (1997). Cost effective gasoline sampling using ranked set sampling. In: ASA Proceedings of the Section on Statistics and the Environment, pp. 83 - 87.

[10] Presnell, B. and Bohn, L. L. (1999). U-statistics and imperfect ranking in ranked set sampling. Journal of Nonparametric Statistics, Vol. 10, pp. 111 - 126.

[11] Stokes, S. L. (1977). Ranked set sampling with concomitant variables. Communications in Statistics, Theory and Methods, Vol. 6, pp. 1207 - 1211.