SPECIAL CASE COMPARISON BETWEEN RSSMC AND RIDOUT'S RANKED SET SAMPLING: INCOMPLETE ORDER CASES

[1]PANLOP NAKAKATE, [2]SUKUMAN SARIKAVANIJ

[1]Department of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok, 10140, Thailand
[2]Department of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok, 10140, Thailand
E-mail: [1]p.nakakate@gmail.com, [2]sukuman@hotmail.com

Abstract- This study showed an analysis of a bias size and a variance size, when incomplete order samples were selected in Ridout's ranked set sampling and imbalanced samples were selected on bivariate cases by using a ranked set sampling of multiple characteristics. Each case was compared by using a mean square error (MSE) and then the suggested addition rules on modified Ridout ranked set sampling were discussed.

Index Terms- Bivariate, Median ranked set sampling, Ranked set sampling, Ridout ranked set sampling

I. INTRODUCTION

A simple random sampling is popular to estimate a mean of an interesting population group because it can reduce expenses in collecting samples. If each unit of a sample has high expense, it needs to fix a small sample size but it can estimate a mean precisely. McIntyre proposed a ranked set sampling (RSS) ,which could reduce a size of sample than a simple random sampling (SRS) but still effective, if interesting variables could be ordered easily without measuring, [1] This method uses the benefit of knowing sample orders. In the other word, we know that samples come from which order statistics. This reduces the variance between groups of sample from different orders, so the variance of estimators also decreased. A sample size proportion of RSS and SRS having the same effectiveness can be calculated from the relative precision(RP). Due to a sample reduction, RSS has been studied widely.

Besides RSS using well with univariate cases, it is also suitable for bivariate cases which both variables having high correlation. However it is problematic with both variables, which have low correlation, because it collects samples by considering the first variable order only and the paired sample of the first variable. [2] Then Ridout presented a sampling method for a case of bivariate, which both variables have low correlation. He created the sampling steps that gather all order statistics of sample according to n. The steps of this sampling method give a higher chance to receive all order statistics of sample than McIntyre's method.

In addition, P. Sontisamran and A. Pongpullponsak modified a RSS method for using with univariate cases which have two ranking variables. (Two concomitant variables are correlated to interesting variables.) This method has the process of selecting samples like Ridout's but it has different objective from gathering all order statistics of samples. It gathers samples, which have orders as median instead, by modifying an indicator variable, which is used for selecting samples. The sample from this method will has orders as same as median or nearest to median. They called the method as ranked set sampling of multiple characteristics (RSSMC) and brought it to construct the control chart. [3] The caution of this method is both ranking variables must have the same weight and distribution of both ranking variables and interesting variables must be symmetric. In general, if the weights of both variables are inequivalent, it will use the regression to find weights in ordering. If the weights of both variables are close to each other, P.Sontisamran and A. Pongpullponsak's method will be more convenient in calculating. When we consider this method can be adapted for using with bivariate cases, which each variable must have a ranking variable for ordering. Although RSSMC and Ridout's will be better than McIntyre's in normal but both cannot guarantee that will receive wanted sample orders. P. Nakakate, A. Pongpullponsan and S. Sarikavanij modified the sampling method of Ridout and presented it as modified Ridout RSS [4] for adding a higher chance to receive wanted sample orders. These wanted sample orders are from all sample orders and median sample orders in case of Ridout and RSSMC, respectively. Nevertheless the modified Ridout RSS method will have a higher chance in selecting samples, it cannot guarantee to get all sample orders. As a result, it might not receive samples from all orders or sample orders being the same as median called imbalance order samples. Thus they revealed the basic solution by using variables to measure the balance of both left and right sides. But this method did not suggest the solution when the balance of both sides cannot be arranged. So there is an interesting issue in studying further from Ridout RSS and RSSMC, which are initial concepts of the method as mentioned for modifying conditions of selecting samples and for convenience in applying other cases.

II. THEORETICAL BACKGROUND

2.1 Ranked set sampling

Ranked set sampling was proposed by McIntyre in 1952. It is a sampling which takes benefit from

knowing sample orders of each unit. It has steps as follows:
-Set a sample of size n
-Set n = mxr
-Randomize a sample of size mxm
-Divide into m equal sets
-Select samples which are the 1st order from the 1st set, the 2nd order from the 2nd set,… , and the nth from the nth set. It will get a sample set of size m.
-Redo for r cycles, it will get a sample group of size n from
this process will be the representative of each order. McIntyre claimed that RSS is more effective than SRS without mathematical proof.
2.1.1 Estimator of RSS
A. Wolfe revealed that an arithmetic mean of ranked set sampling is unbiased. It can be written in summation form of integral and set permutation in binomial form. [5]

If we use properties as $\sum_{i=1}^{n} f_{i:n}(x) = nf(x)$, we might write shorter as follows:

$$E\left(\bar{X}_{RSS}\right) = E\left(\frac{X_{1:n}+\cdots+X_{n:n}}{n}\right) = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} x f_{i:n}(x)\,dx,$$

$$= \frac{1}{n}\int x\sum_{i=1}^{n} f_{i:n}(x)\,dx = \frac{1}{n}\int_{-\infty}^{\infty} xnf(x)\,dx = \mu.$$

2.1.2 Variance of RSS

$$Var\left(\bar{X}_{RSS}\right) = Var\left(\frac{X_{1:n}+\cdots+X_{n:n}}{n}\right),$$

$$= \frac{1}{n^2}\left(\sum_{i=1}^{n}\int_{-\infty}^{\infty} x^2 f_{i:n}(x)\,dx - \sum_{i=1}^{n}\mu_{i:n}^2\right),$$

$$= \frac{1}{n}\int_{-\infty}^{\infty} x^2 f(x)\,dx - \sum_{i=1}^{n}\mu_{i:n}^2,$$

and $Var\left(\bar{X}\right) = \frac{1}{n}\int_{-\infty}^{\infty} x^2 f(x)\,dx - n\mu^2,$

since $\sum_{i=1}^{n}\mu_{i:n} = \mu$, then $\sum_{i=1}^{n}\mu_{i:n}^2 \geq n\mu^2$.

Therefore $Var\left(\bar{X}_{RSS}\right) \leq Var\left(\bar{X}\right)$ and their differences are $\sum_{i=1}^{n}\mu_{i:n}^2 - n\mu^2$.

2.2 Median ranked set sampling
RSS will be the most effective precision, if all units of a sample are ordered correctly. Muttlak presented median RSS by selecting orders from a median position instead of all orders [6]. When a median order is used, correctly ordering depends on a median order only. If there are populations with asymmetric distribution, we will get a sample for calculating a median estimator. If there are populations with symmetric distribution, we will use the sample for calculating both of median and mean by using median RSS process. This process is almost the same as RSS process but it selects samples from a median position of all sets.

2.2.1 Estimator of median ranked set
A sample arithmetic mean from median RSS is an unbiased estimator of μ if and only if a population distribution is symmetric.

$$E\left(\bar{X}_{Med}\right) = E\left(\frac{X_{(Med:n)1}}{n}+\cdots+\frac{X_{(Med:n)n}}{n}\right)$$

If a population distribution is symmetric,

$E\left(\bar{X}_{Med}\right) = \text{median} = \mu$.

2.2.2 Variance of median ranked set estimator
A variance of median ranked set estimator is divided into two cases as follows:

$$Var\left(\bar{X}_{Med}\right) = \begin{cases} \frac{1}{n}Var\left(X_{\frac{n+1}{2}}\right), & \text{if } n \text{ is odd,} \\ \frac{1}{2n}\left(Var\left(X_{\frac{n}{2}}\right)+Var\left(X_{\frac{n}{2}+1}\right)\right), & n \text{ is even.} \end{cases}$$

2.3 Bivariate ranked set sampling
McIntyre's concept is used for RSS in collecting data from a bivariate population. This concept uses an order from a sample variable in ordering and its result is this variable having properties as RSS. But the properties of the other variable are collected from ordering of the first variable based on correlation of both variables. If the correlation is very high, it will be similar to RSS. If the correlation is very low, it will be similar to SRS. Thus, Ridout (2003) proposed a RSS sampling method from the variables, which has low correlation. This method is also used with variables having high correlation.

2.4 Ridout ranked set sampling
Ridout (2003) presented the sample selecting method for bivariate RSS. This method adds higher chance to gather all samples. It can be concluded concisely as follows:
-Set the 1st variable and the 2nd variable as i =1 and i =2, respectively.
-Set $n_{i[j]}$ as a size of sample which is at the jth order of the ith variable.

-Set $V_i = \text{var}[n_{i[j]}]$ and $V = \sum_{i=1}^{p} V_i$.

-Randomize a sample of size nxn and divide into n sets with equal numbers by following the normal RSS process.
-Order units of sample inside sets that each set is independent.
- Select n units of sample which make $V_i = \text{var}[n_{i[j]}]$ smallest.
We will get samples of all order from Ridout's method or one order might disappear that is similar to have one position of error ranking.

$\sum_{i=1}^{n} \frac{X_{R_i}}{n}$ is an unbiased estimator from Ridout RSS when X_{R_i} is a random variable of the measuring i^{th} unit of sample which is taken from selecting of Ridout RSS.

2.5 RSSMC from A. Pongpullponsak and P.Sontisamran

A.Pongpullponsak and P.Sontisamran modified the Ridout's method for constructing the control chart conveniently with selecting samples from median RSS instead. (Ridout's method will select all sample orders.)

If we have two concomitant variables that are correlated with only one interesting variable, it can be used in some cases. There are two necessary conditions; the former is the distribution of an interesting population and concomitant variables which have symmetry and the latter both concomitant variables must have equal effect to an interesting variable. Furthermore, regression will be used in ordering in general cases which have unequal weights (see [7]). The selecting process can be concluded as follows:

-Set new $n_{[ij]}$ as value of concomitant variable of the i^{th} variable at the j^{th} order

-V_i = var[n_{ij}] is still the same.

-Follow the Ridout process will get a median value or an approximate value instead of getting a sample of all orders.

We proposed that this sample selecting method can be applied with a bivariate case by using with two interesting variables and each variable having a concomitant variable in ordering. When we use the same process in selecting samples, we will get a sample which has order at median position of both variables.

This method is convenient for using with the control chart since it must calculate the mean and variance. For application with bivariate, it can be can be illustrated by the example below:

Suppose $X_1$and X_2 are interesting random variables and Y_1, Y_2 are high correlation variables with $X_1$and X_2, respectively.

-Randomize a sample of size nxn and divide them into n sets, which each set has equal numbers.

-Measure Y_1,Y_2 from all units of sample for ordering X_1 and X_2.

-Find a value of V of each set and select the sample which has the lowest value of V and keep doing until n sets. This will get a sample of size n.

If both of X and Y have symmetric distribution, each Y_i , which is closest to $\overline{Y}_i$, will be also close to a median value of Y_i . We considered each variable independently. In addition, X_i and Y_i have high correlation. So, samples have approximate or equal values as median of X_i.

When we consider both Y_1 and Y_2, the sample, which has the smallest value V, also has the most approximate value as medians of both X and Y.

The estimator $\sum_{i=1}^{n} \frac{X_{RSSMC_i}}{n}$ is an unbiased estimator from RSSMC when X_{RSSMC_i} is a random variable of a measured value from the i^{th} unit of sample which is selected from RSSMC.

2.6 Modified Ridout ranked set sampling

This method was modified from Ridout's sampling method by considering each option of Ridout to get the best sample. The calculation of this method is more complicated than Ridout's but it give us a higher chance to get samples of all orders. It adds median ranked set to be one option if a population has a symmetric distribution, but the symmetric distribution of population can be a limitation in using also. This method has similar process to sample selecting method of Ridout but Ridout's select a unit of sample from each set, meanwhile the modified Ridout method selects n unit of sample from each cycle altogether or each n sets when each set has size n. If it is impossible to select samples from all order, the modified Ridout method uses variables to check the symmetric sample selection of both left and right sides for getting a sample which can approximate mean better.

However, this method did not describe how to select samples if we get a set of size nxn which cannot choose a set which has all order samples and a set which has non zero variables. This case will be the unclear if it is remained as Ridout's method.

III. RESULT AND DISCUSSION

A mean estimator of original Ridout's method, and the adapted RSSMC with bivariate were analysed from expected values and variance of estimators of each case. A case can be selected required sample orders will be a comparison between balanced RSS and Median RSS. In contrast, a case cannot be selected required sample orders must be analysed individually.

3.1 Analysing selecting incomplete order samples of Ridout

For analysing the expected value we need to define the following:

If we get all sample orders for a variable X_p,

when $p = \{1,2\}$ $E\left(X_{p_{R_i}}\right) = E\left(\bar{X}_{p_{RSS}}\right) = \mu.$

Set $X_{i_{k:n}}$ as an order that is not collected, $k \in \{1,2,\ldots,n\}$

Set $X_{i_{m:n}}$ as an order that is collected repeatedly, $m \in \{1,2,\ldots,n\} - \{k\}$,

3.1.1 Incomplete order samples case

For incomplete order of Ridout case, we can rewrite the mean estimator as:

$$E\left(X_{i_R}\right)=\frac{1}{n}E\left(X_{i_{(1n)1}}+\ldots X_{i_{(k-1n)k-1}}+X_{i_{(k+1n)k+1}}+\ldots X_{i_{(nn)n}}+X_{i_{(mn)k}}\right).$$

We use some manipulations as

$$=E\left(\bar{X}_{i_{RSS}}\right)+\frac{1}{n}E\left(X_{i_{mn}}-X_{i_{kn}}\right)=\mu+\frac{E\left(X_{i_{mn}}-X_{i_{kn}}\right)}{n}\neq\mu.$$

3.1.2 Variance of estimator of Ridout in case of incomplete orders

For this case, we can rewrite the variance estimator as:

$$Var\left(X_{i_R}\right)=\sum_{j=1}^{n}\frac{X_{i_{j:n}}}{n},$$

$$=Var\left(\frac{X_{i_{1:n}}+\cdots+X_{i_{k-1:n}}+X_{i_{k+1:n}}+\cdots X_{i_{n:n}}+X_{i_{m:n}}}{n}\right),$$

$$=Var\left(\bar{X}_{i_{RSS}}+\frac{X_{i_{m:n}}-X_{i_{k:m}}}{n}\right),$$

$$=Var\left(\bar{X}_{i_{RSS}}\right)+\frac{Var\left(X_{i_{m:n}}\right)-Var\left(X_{i_{k:m}}\right)}{n^2}.$$

The case of incomplete orders of Ridout depends on the only last cycle. If it is considered for one cycle, it has bias. When it is brought to perform estimation, it will be different from normal RSS (Getting all orders). In investigating the effectiveness of any estimator $\hat{\theta}$ we use a mean square error in considering by:

$$MSE(\hat{\theta})=bias(\hat{\theta})+Var(\hat{\theta}).$$

When n is large enough, the bias of incomplete selecting orders will not affect much to mean estimating. However, in practical way we do not use that big of size of a sample. MSE has incomplete selecting order depended on the missing k^{th} order and selecting repeated of m^{th} order.

3.2 Analysing selecting imbalanced order samples of RSSMC

For analysing the expected value we need to define the following:

Set n as odd $X_{med:j}$ .It is an $\left(\frac{n+1}{2}\right)^{th}$ order statistic for $j=\{1,\ldots,n\}$,

Set n as odd $X_{med:j}$ is an $\left(\frac{n}{2}\right)^{th}$ order statistics statistic for

$j=\left\{1,\ldots,\frac{n}{2}\right\}$,

and $X_{med:i}$ is an $\left(\frac{n}{2}+1\right)^{th}$ order statistics statistic for $i=\left\{\frac{n}{2}+1,\ldots,n\right\}$.

3.2.1 In case of symmetric distribution

It is separated for considering into 2 cases; considering especially for the i^{th} variable.

The first case is all selected sample being at a median position.

$$E\left(\bar{X}_{i_{MMRSS(0)}}\right)=E\left(\bar{X}_{i_{med}}\right)=E\left(\frac{1}{n}\left(X_{i_{med,1}}+\cdots+X_{i_{med,n}}\right)\right)$$

The second case is two sets which m is not a median but we can select at the k^{th} order when both of two variables having low correlation that is called imbalanced sample orders.

$$E\left(\bar{X}_{i_{MMRSS(1)}}\right)=E\left(\frac{X_{i_{(med,n)1}}}{n}+\cdots+\frac{X_{i_{(med,n)m-1}}}{n}+\right.$$

$$\left.\frac{X_{i_{(med,n)m+1}}}{n}+\cdots+\frac{X_{i_{(med,n)n}}}{n}+\frac{X_{i_{(k,n)m}}}{n}\right),$$

$$=\mu+E\left(\frac{X_{i_{(k,n)m}}}{n}-\frac{X_{i_{(med,n)m}}}{n}\right).$$

When k in RSSMC has value as med+1 or med-1which each sample has a chance to be selected. It is similar to Ridout variance. An estimator of RSSMC in case of selecting order which getting unwanted samples is;

$$Var\left(\bar{X}_{i_{RSSMC}}\right)=Var\left(\bar{X}_{i_{med}}\right)+\frac{Var\left(X_{i_{mk}}\right)-Var\left(X_{i_{med}}\right)}{n^2}$$

3.3 Analysing MSE from incomplete sample orders

Although expected values of both sampling methods do not have differences from RSS, it causes from the expected values calculated from repeated redoing for n times when n comes close to infinity. Bringing incomplete orders to estimate will have bias for each time.

This will make MSE values increasing. Having chance like this with Ridout's method depends on the last cycle and RSSMC depends on the position selected samples from every cycles. So the MSE value is based on the variance of the selected orders, which are not wanted, and the difference of mean between wanted orders and unwanted orders. Possible cases can be considered into two cases.

The first case of the incomplete order which missing for one order can be divided into small cases as follows:

-If m and k have different value more than one, it is clear that they make MSE value higher. So this case should be controlled by using the variable to check the balance of left and right from modified Ridout RSS.

- A case of repeated orders are different from missing order for one order on normal distribution or other symmetric distributions which are resemble to each other.

We can consider the paired of sample order m, k that minimize mean bias from Table 1 which is arranged from [8].

Table 1 Showing $E\left(X_{i:n}\right)-E\left(X_{j:n}\right)$ for j = i+1 with n = 3, …, 8.

	i=2,j=1	i=3,j=2	i=4,j=3	i=5,j=4	i=6,j=5	i=7,j=6	i=8,j=7
n=3	0.8463	0.8463					
n=4	0.7324	0.5940	0.7324				
n=5	0.6679	0.4950	0.4950	0.6679			
n=6	0.6255	0.6216	0.0403	0.6216	0.6255		
n=7	0.5948	0.4047	0.3527	0.3527	0.4047	0.5948	
n=8	0.5714	0.3794	0.3203	0.3050	0.3203	0.3794	0.5714

The differences of order means, which are close to median and the consecutive orders, will less than the differences of order means which is far from the median with the consecutive orders. So there are three most suitable options as follows:

The 1st option is an incomplete sample in a median position; a repeated sample in median+1 or median -1 position.

The 2nd option is an incomplete sample in median-1 position; a repeated sample in median position.

The 3rd option is an incomplete sample in median+1 position; a repeated sample in median position.

The RSSMC will always choose a primary order as a median which is the first option as mentioned previously. For the second and third option occur in Ridout's method including the first option. However, these three options and the case of incomplete samples at max position, repeated samples at max-1 have equal chance to be selected in Ridout RSS. On the other hand, the opportunities for getting incomplete order samples in any position are equal if a repeated order is different from it for one order.

The second case of more than two incomplete orders of symmetric distribution has two cases as follows:

-The gathered orders have left and right balance; unbiased mean estimators and small different variance from normal RSS.

-The gathered orders do not have left and right balance. This case is similar to the repeated orders that are different from the incomplete order only one order. The orders, which are not collected, will change to left and balance instead. So the best sample is still the original options which are near the median position.

The conclusion of both methods gets the all required orders which have resembled effectiveness. The possibility to get all required orders in RSSMC is less than Ridout's method. If we cannot gather all required orders, RSSMC has a higher chance to make MSE less.

3.4 Simulation

The simulation section showed the comparison between incomplete orders near to the median position order cases and incomplete orders far from the median position order cases on bivariate standard normal distribution in each m and k. When we have complete orders then m = k. For incomplete comparison, we considered only the right side due to the symmetric property. This comparison is divided into two cases that are set size = 4 and set size = 4 on the Table 2 and Table 3, respectively.

Table 2 The MSE comparisons in each m and k, when set size = 4, number of cycle = 1.

	m=k	m_1=3,k_1=4 m_2=2,k_2=1	m=3,k=2	m=3,k=4	m=4,k=3
$MSE\left(\hat{\theta}\right)$	0.1091	0.0941	0.1260	0.1380	0.1538
$\sqrt{Var\left(\hat{\theta}\right)}$	0.3304	0.3067	0.3253	0.3173	0.3412

From Table 2, case set size = 4. Median RSS (m_1=3, k_1=4, m_2=2, k_2=1) has lower MSE than a complete order RSS (m=k) due to the lower variance (unbiased property). In incomplete order case, when m or k is close to median position (m=3, k=2), RSS has lower MSE than when m or k is far from median position (m=4, k=3) because it has more bias and more estimator variance. The case of m=3 ,k=4 has lower MSE than m=4 ,k=3 . This showed that k affects to MSE more than m because variance of $X_{i_{kcn}}$ is added to the estimator variance.

Table 3 The MSE comparisons in each m and k, when set size =5, number of cycle=1.

	m=k	m_1=3,k_1=1 m_2=3,k_2=2 m_3=3,k_3=4 m_4=3,k_4=5	m=3,k=4
$MSE\left(\hat{\theta}\right)$	0.0750	0.0547	0.0835
$\sqrt{Var\left(\hat{\theta}\right)}$	0.2740	0.2337	0.2720
	m=4,k=3	m=4,k=5	m=5,k=4
$MSE\left(\hat{\theta}\right)$	0.0871	0.0991	0.0879
$\sqrt{Var\left(\hat{\theta}\right)}$	0.2751	0.2809	0.2653

From Table 3, case set size=5. Median RSS (m_1=3, k_1=4, m_2=2, k_2=1) also has lower MSE than the complete order RSS (m=k) like the case of set size =4. The case m=3,k=4 has smallest MSE that is similar to case set size =4.

CONCLUSION

Both of simulation result and expected value calculation (adapt from [7]) are based on the standard normal distribution. This showed the two consecutive orders nearest to the median position having smallest mean difference. The simulation result on symmetric distribution showed median RSS is better than McIntyre RSS. These results can be applied in modified Ridout RSS that we should give the first priority to median RSS. However, in independent bivariate case, complete median RSS has low chance to occur. If we apply modified Ridout's RSS with extreme ranked set sampling on symmetric population, the priority of this RSS will be lower than McIntyre RSS and median RSS. When the value of left and right balance variable is greater than zero and

incomplete order Ridout occurs, the first priority in this case should be m=median position or k =median position while m and k are consecutive orders. If we add these rules to the modified Ridout's RSS, the MSE of this method will be decreased and the RP will be increased. Therefore the sample size can be reduced for saving cost and the mean estimating precision is more effective with the same sample size.

REFERENCES

[1] G. A. McIntyre, "A Method for unbiased selective sampling, using ranked sets". The American Statistician, vol. 59, no. 3, pp. 230-232, Aug. 2005.

[2] M. S. Ridout, "On ranked set sampling for multiple characteristics". Environmental and Ecological Statistics, vol. 10, no, 2, pp. 255-262, Jun 2003.

[3] A. Pongpulponsak, and P. Sontisamran, "statistical quality control based on ranked set sampling on multiple characteristics", Chiang Mai J. Sci., vol 40, no. 3, pp. 485-498, Jul. 2013.

[4] P. Nakakate and A. Pongpullponsak, "Modified Ridout 's Ranked set sampling", 13-15 Jul., Paper present at International conference on applied statistics, Phuket, Thailand, Thammasat university, 2016.

[5] A. W. Douglas, "Ranked set sampling: An approach to more efficient data collection", Statistician Science, vol. 19 no.4, pp. 636-643, Nov. 2004.

[6] H. A. Muttlak, "Median ranked set sampling with size biased probability of selection", Biometrical Journal, vol. 40, no. 4, pp. 455-465, Aug. 1998.

[7] Z. Chen, "Adaptive ranked-set sampling with multiple concomitant variables: an effective way to observational economy", Bernoulli, vol. 8, No. 3, pp. 313-322, jun. 2002.

[8] Arnold, Barry C., N. Balakrishnan, and H. N. Nagaraja. "A First Course in Order Statistics", New York: Wiley, 1992. N. pag. Print., pp. 86-94.

★★★