B A D A N I A O P E R A C Y J N E I D E C Y Z J E

Nr 3–4 2005

Barbara KOWALCZYK*

ON DIFFERENT ESTIMATORS OF A POPULATION MEAN BASED ON RANKED SETS

A review of results concerning the problem of sampling based on ranked sets is presented. From an infinite or finite population n independent samples of n elements each are drawn. The samples are ranked and then n elements are chosen to be measured.

Keywords: *ranked sets, order statistics, sampling*

1. Introduction

A method of sampling based on ranked sets is an efficient alternative to simple random sampling which uses measurements on selected subsets of the primary sample. It can be applied in many studies where the exact measurement of an element is very difficult (in terms of money, time, labour and organization) but the variable of interest, although not easily measurable, can be relatively easily ranked (order) at no cost or very little additional cost. The ranking can be done on the basis of visual inspection, prior information, earlier sampling episodes or other rough methods not requiring actual measurement. If there is a related variable which is readily observable and can be easily ranked, and is correlated with the variable of interest, the ranking can also be done on the basis of this concomitant variable.

The standard example illustrating the matter under discussion is the following. If interest lies in estimating the mean height of trees, then measuring the height of the sampled trees could pose a problem, but it would be relatively easy to rank small sets of trees on the basis of visual inspection of their heights. And the cost of ranking is insignificant compared with the cost of measuring.

* Warsaw School of Economics, Institute of Econometrics, Al. Niepodległości 162, 02-554 Warsaw, bkowal@sgh.waw.pl

2. Standard ranked set sampling

2.1. Sampling method

The first step of ranked set sampling (RSS) procedure is to draw from an infinite population n random samples with n elements in each sample. Let $X_{11}, X_{12},..., X_{1n}$; $X_{21}, X_{22},..., X_{2n}$; ...; $X_{n1}, X_{n2},..., X_{nn}$ be independent random variables all having the same cumulative distribution function $F(x)$ with mean μ_x and variance σ_x^2. The second step is to rank each element within each set with respect to the variable of interest X. But ranking should be done without actual measurements. Let $X_{i(1:n)}, X_{i(2:n)},..., X_{i(n:n)}$ denote the corresponding order statistics of $X_{i1}, X_{i2},..., X_{in}$. After ranking the units appear as follows: $X_{1(1:n)}, X_{1(2:n)},..., X_{1(n:n)}$; $X_{2(1:n)}, X_{2(2:n)},..., X_{2(n:n)}$; ...; $X_{n(1:n)}, X_{n(2:n)}$, ..., $X_{n(n:n)}$. So now we have n ordered samples. From the first sample we choose for the actual measurement the element with the smallest rank $X_{1(1:n)}$. From the second sample we choose the element with the second smallest rank $X_{2(2:n)}$. We continue the procedure until the element with the largest rank from the n-th sample is chosen $X_{n(n:n)}$. This procedure yields a total number of n elements chosen to be measured, one from each sample. The chosen n elements constitute a ranked set sample. The mean of the ranked set sample is denoted by $\overline{X}_{[n]}$, where

$$\overline{X}_{[n]} = \frac{1}{n}\sum_{i=1}^{n} X_{i(i:n)}. \tag{1}$$

RSS was first suggested by McIntyre [4]. The appropriate statistical theory was delivered by Takahasi and Wakimoto [9].

2.2. Efficiency of the estimator

Let us denote the usual order statistics of a simple random sample $X_1, X_2,..., X_n$ by $X_{1:n}, X_{2:n},..., X_{n:n}$. It has to be noted that $X_{i(i:n)}$, $i = 1, \ldots, n$ are obviously independent as contrasted with the usual order statistics $X_{i:n}$, $i = 1, \ldots, n$ which are correlated. Both variables $X_{i(i:n)}$ and $X_{i:n}$ have the same cumulative distribution function.

Throughout the paper we assume that the elements can be ordered at no cost or insignificant additional cost. So, we take into account the cost of measurements only and compare the variance of $\overline{X}_{[n]}$ with the variance of the usual sample mean $\overline{X}_n$.

Theorem 1. The mean of a ranked set sample $\overline{X}_{[n]}$ is an unbiased estimator of the population mean μ_X and its variance is given by:

$$\mathrm{Var}(\overline{X}_{[n]}) = \frac{1}{n}\left(\sigma_X^2 - \frac{1}{n}\sum_{i=1}^{n}(EX_{i:n} - \mu_X)^2 \right) \tag{2}$$

Proof can be found in [9].

Corollary 1. The mean of a ranked set sample $\overline{X}_{[n]}$ is more efficient than the usual sample mean $\overline{X}_n$ under simple random sampling SRS, i.e., when both estimators are constructed on the basis of the same number *n* of actual measurements, then

$$\mathrm{Var}\,(\overline{X}_{[n]}) \le \mathrm{Var}\,(\overline{X}_n)\,.$$

Efficiency of the estimator $\overline{X}_{[n]}$ compared with the usual sample mean $\overline{X}_n$ is given by:

$$\frac{\mathrm{Var}\,(\overline{X}_{[n]})}{\mathrm{Var}\,(\overline{X}_n)} = 1 - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{EX_{i:n} - \mu_X}{\sigma_X} \right)^2 \tag{3}$$

The corollary can be easily obtained by comparing the variance given in (2) with $\mathrm{Var}\,(\overline{X}_n) = \dfrac{\sigma_X^2}{n}$.

2.3. Several cycles of RSS procedure

Ranking without actual measurement is in many practical situations easier when there are not too many elements to compare. So *n* is generally chosen to be rather small. To provide enough quantifications for inference the entire process is repeated *r* times until the random variable *X* has been measured *nr* times, where *nr* is desired sample size. These *nr* elements $X_{1(1:n)j}, X_{2(2:n)j}, \ldots, X_{n(n:n)j}$, j = 1, 2, ..., *r* form the ranked set sample based on *r* cycles. So $X_{i(i:n)j}$ denotes the *i*-th order statistics from the *i*-th sample in the *j*-th cycle. The mean of a ranked set sample based on *r* cycles is denoted by $\overline{X}_{[n]r}$, where

$$\overline{X}_{[n]r} = \frac{1}{nr}\sum_{j=1}^{r}\sum_{i=1}^{n} X_{i(i:n)j} \,. \tag{4}$$

Several cycles $(r > 1)$ of RSS procedure are due to practical demands only and do not improve efficiency of the estimation, which can be seen from the theorem given below.

Theorem 2. The mean $\overline{X}_{[n]r}$ is an unbiased estimator of the population mean μ_X and its variance is given by:

$$\mathrm{Var}\,(\overline{X}_{[n]r}) = \frac{1}{nr}\left(\sigma_X^2 - \frac{1}{n}\sum_{i=1}^{n}(EX_{i:n} - \mu_X)^2\right). \tag{5}$$

Under given sample size $m = nr$ the variance $\mathrm{Var}\,(\overline{X}_{[n]r})$ is a decreasing function of n and takes the smallest value for $r = 1$.

Proof can be found in [9].

Corollary 2. Efficiency of the estimator $\overline{X}_{[n]r}$ compared with the usual sample mean $\overline{X}_{nr}$ of size nr is given by:

$$\frac{\mathrm{Var}\,(\overline{X}_{[n]r})}{\mathrm{Var}\,(\overline{X}_{nr})} = 1 - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{EX_{i:n} - \mu_X}{\sigma_X}\right)^2 . \tag{6}$$

3. Ranking by a concomitant variable

3.1. Sampling method

In many practical situations ranking by visual inspection or prior information is rather difficult or even impossible. So the ranking may be accomplished by means of some concomitant variable Y that is relatively easily measured and is correlated with the variable of interest X. To carry out the ranking n bivariate simple random samples of size n are drawn from an infinite population. From the first sample of size n, the X associated with the smallest ordered Y is measured. From the second sample of size n the X associated with the second smallest Y is measured. We continue this way until the X associated with the largest Y from the n-th sample is chosen for measurement.

The whole cycle is repeated *r* times, so the total number of elements to be measured is *nr*.

Let $X_{1[1:n]j}, X_{2[2:n]j}, \ldots, X_{n[n:n]j}$ be a ranked set sample selected on the basis of an ordered concomitant variable Y in the j-th cycle. The mean of a RSS constructed on the basis of a concomitant variable Y in r cycles is denoted by $\overline{X}_{Y[n]r}$, where

$$\overline{X}_{Y[n]r} = \frac{1}{nr}\sum_{j=1}^{r}\sum_{i=1}^{n} X_{i[i:n]j} \,. \tag{7}$$

3.2. Efficiency of the estimator

Theorem 3. Assume that the regression of X on Y is linear, that is,

$$E(X \mid Y) = \mu_X + \frac{\rho_{XY}\sigma_X}{\sigma_Y}(Y - \mu_Y) \tag{8}$$

and

$$\mathrm{Var}\,(X \mid Y) = \sigma_X^2(1 - \rho_{XY}^2)\,. \tag{9}$$

Thus, $\overline{X}_{Y[n]r}$ is an unbiased estimator of a population mean μ_X and its variance is given by

$$\mathrm{Var}(\overline{X}_{Y[n]r}) = \frac{\sigma_X^2}{nr}\left(1 - \frac{\rho_{XY}^2}{n}\sum_{i=1}^{n}\left(\frac{EY_{i:n} - \mu_Y}{\sigma_Y}\right)^2\right). \tag{10}$$

Proof can be found in [8].

Corollary 3. Under assumptions (8) and (9) efficiency of the estimator $\overline{X}_{Y[n]r}$ constructed on the basis of a concomitant variable Y compared with the usual simple random sample mean $\overline{X}_{nr}$ of the same number of actual measurements nr is given by:

$$\frac{\mathrm{Var}\,(\overline{X}_{Y[n]r})}{\mathrm{Var}\,(\overline{X}_{nr})} = 1 - \frac{\rho_{XY}^2}{n}\sum_{i=1}^{n}\left(\frac{EY_{i:n} - \mu_Y}{\sigma_Y}\right)^2 . \tag{11}$$

4. Errors in ranking

Accurate ranking (when not based on some concomitant variable as was the case in the previous section) is the most difficult part of implementation of RSS procedure. When elements are ordered by the "ranker's judgment", the quantified element from the *i*-th sample in the *j*-th cycle may not be necessarily the *i*-th order statistic in that sample but rather the *i*-th "judgement order statistic" and is written $X^*_{i(i:n)j}$ to distinguish it from the actual order statistic $X_{i(i:n)j}$. In other words, errors in ranking cause that the element that is placed in the position to be quantified may differ from the element that should have been placed. Let us notice that the case of errors in ranking is equivalent to the case of ranking on the basis of a concomitant variable X^* when this concomitant variable is the "ranker's judgement".

Errors in ranking can be described by the model:

$$X^* = X + \varepsilon, \tag{12}$$

where

$$X \text{ and } \varepsilon \text{ are independent and } \varepsilon \sim N(0, \sigma_\varepsilon^2). \tag{13}$$

X represents the study variable, X^* refers to what the ranker "sees", and ε denotes judgement error. In this case we have:

$$\operatorname{Var}(X^* \mid X) = \text{const}. \tag{14}$$

To use the theory given in section 3 we need the opposite condition

$$\operatorname{Var}(X \mid X^*) = \text{const}, \tag{15}$$

which is not true in general under assumptions (12)–(13). The condition (15) holds for normal case, i.e. when additional assumption is made that the study variable is also normally distributed

$$X \sim N(\mu_X, \sigma_X^2). \tag{16}$$

So in normal model given by (12), (13), (16) errors in ranking are simply a special case of ranking by a concomitant variable (compare section 3), where the concomitant variable is X^* what ranker "sees".

Errors in ranking were considered by Dell and Clutter in [2], where various simulation results were given for different distributions. Analogous normal model was considered theoretically by David and Levine in [1].

5. Extreme Ranked Set Sampling

5.1. Sampling method

Extreme ranked set sampling (ERSS) is a procedure analogous to ranked set sampling but based only on the lowest and the highest order statistics. It was introduced by [6]. ERSS involves random drawing from an infinite population n sets of n units each (n is an even number). From the first set of n elements the lowest ranked unit is measured. From the second set of n elements the largest ranked unit is measured. From the third set of n elements the lowest ranked unit is measured, and so on. From the last set the largest ranked unit is measured. This procedure yields a total number of n elements chosen to be measured, one from each sample. The chosen n elements constitute an extreme ranked set sample. The mean of the extreme ranked set sample is denoted by $\overline{X}_{E[n]}$, where

$$\overline{X}_{E[n]} = \frac{1}{n}\{X_{1(1:n)} + X_{2(n:n)} + X_{3(1:n)} + \ldots + X_{n(n:n)}\}. \tag{17}$$

As in previous cases the whole procedure can be repeated r times, so the mean of an extreme ranked set sample based on r cycles is defined as

$$\overline{X}_{E[n]r} = \frac{1}{nr}\sum_{j=1}^{r}\{X_{1(1:n)j} + X_{2(n:n)j} + X_{3(1:n)j} + \ldots + X_{n(n:n)j}\}. \tag{18}$$

ERSS procedure is definitely easier for implementation than the one based on all order statistics but at the same time is less efficient in many practical situations.

5.2. Efficiency of the estimator

Theorem 4. Expected value and variance of the estimator $\overline{X}_{E[n]r}$ are given by:

$$E\overline{X}_{E[n]r} = \frac{1}{2}(EX_{1:n} + EX_{n:n}), \tag{19}$$

$$\operatorname{Var}\overline{X}_{E[n]r} = \frac{1}{2nr}(\operatorname{Var}X_{1:n} + \operatorname{Var}X_{n:n}). \tag{20}$$

Theorem 5. Assume that the underlying distribution of X is symmetric. Then the mean of an extreme ranked set sample $\overline{X}_{E[n]r}$ is an unbiased estimator of a population mean μ and its variance is given by:

$$\text{Var}\,(\overline{X}_{E[n]r}) = \frac{\text{Var}\,X_{1:n}}{nr} \tag{21}$$

Proof can be found in [6].

Theorem 6. Assume that the underlying distribution of X is uniform $U(a,b)$. Then the mean of an extreme ranked set sample $\overline{X}_{E[n]r}$ is more efficient than the usual sample mean $\overline{X}_{nr}$, that is,

$$\text{Var}\,(\overline{X}_{E[n]r}) \le \text{Var}\,(\overline{X}_{nr}) \tag{22}$$

and more efficient than the mean of a ranked set sample $\overline{X}_{[n]r}$, that is,

$$\text{Var}\,(\overline{X}_{E[n]r}) \le \text{Var}\,(\overline{X}_{[n]r}) \tag{23}$$

Proof can be found in [6].

Remark 1. The mean of an extreme ranked set sample $\overline{X}_{E[n]r}$ is not an unbiased estimator of a population mean μ_X in general. So, extreme ranked set sampling is not a proper method of sampling in the case of non symmetric distributions. In [6], many simulations are conducted which confirm this result.

6. Ranked Set Sampling from a Finite Population

6.1. Sampling method

The first step of ranked set sampling procedure from a finite population is to draw n elements by simple random sampling without replacement (SRSWOR) from the given finite population of N elements. The drawing is repeated independently n times which yields n independent samples (sets) of size n. In each set separately distinct elements appear because within each set sampling is without replacement but some elements that appear in one sample may also appear in some other sample because different samples are drawn independently from the entire population of N elements.

The second step is to rank each sample without actual measurements. For the final sample the element with the smallest rank from the first sample is chosen, the element

with the second smallest rank from the second sample and so on until the element with the largest rank from the n-th sample is chosen.

Let X_k denote the value of a characteristic X for the k-th population element, $k = 1, 2, ..., N$. The X_k are treated in finite population theory as unknown but constant (non-random) values which are traditionally written in capital letters. Let x_{il}, $i = 1, 2, ..., n$, $l = 1, 2, ..., n$ denote the value of X for the unit drawn in the i-th sample and in the l-th draw. It is easily seen that x_{il} is a random variable which can take values $X_1, X_2, X_3, ..., X_N$, with probability $1/N$ each. Let $x_{i(1:n)}, x_{i(2:n)}, ..., x_{i(n:n)}$ denote the corresponding order statistics of $x_{i1}, x_{i2}, ..., x_{in}$.

The mean of a ranked set sample is denoted by $\bar{x}_{[n]}$, where

$$\bar{x}_{[n]} = \frac{1}{n}\sum_{i=1}^{n} x_{i(i:n)} \,. \tag{24}$$

When the whole procedure is repeated in r cycles the mean is given by

$$\bar{x}_{[n]r} = \frac{1}{nr}\sum_{j=1}^{r}\sum_{i=1}^{n} x_{i(i:n)j} \,, \tag{25}$$

where $x_{i(i:n)j}$ denotes the i-th order statistics from the i-th sample in the j-th cycle.

Theorem 7. The mean $\bar{x}_{[n]r}$ of a ranked set sample from a finite population based on r cycles is an unbiased estimator of the population mean $\bar{X} = \frac{1}{N}\sum_{k=1}^{N} X_k$ and its variance is given by:

$$D^2(\bar{x}_{[n]r}) = \frac{1}{nr}\left\{\left(1-\frac{1}{N}\right)S^2 - \frac{1}{n}\sum_{i=1}^{n}(Ex_{i:n} - \bar{X})^2\right\}, \tag{26}$$

where

$$S^2 = \frac{1}{N-1}\sum_{k=1}^{N}(X_k - \bar{X})^2 \,.$$

Proof for one cycle can be found in Kowalczyk [3]. Generalization for r cycles is straightforward.

Theorem 8. A statistic of the form

$$x_{[n]r} = N\bar{x}_{[n]r} \tag{27}$$

is an unbiased estimator of the population total $X = \sum_{k=1}^{N} X_k$ and its variance is given by:

$$D^2(x_{[n]r}) = \frac{N^2}{nr}\left\{\left(1-\frac{1}{N}\right)S^2 - \frac{1}{n}\sum_{i=1}^{n}(Ex_{i:n} - \overline{X})^2\right\}. \tag{28}$$

Proof is easily obtained from theorem 7 as $\mathrm{Var}\,(x_{[n]r}) = N^2\mathrm{Var}\,(\overline{x}_{[n]r})$.

Corollary 4. The mean $\overline{x}_{[n]r}$ of a ranked set sample from a finite population based on *r* cycles is more efficient than the common sample mean $\overline{x}_{nr,\mathrm{SRS}}$ based on *nr* actual measurements under simple random sampling with replacement (SRS), that is,

$$\mathrm{Var}\,(\overline{x}_{[n]r}) \le \mathrm{Var}\,(\overline{x}_{nr,\mathrm{SRS}}) = \frac{\sigma^2}{nr}, \tag{29}$$

where

$$\sigma^2 = \frac{1}{N}\sum_{k=1}^{N}(X_k - \overline{X})^2 = \frac{N-1}{N}S^2. \tag{30}$$

Remark 2. Comparing analytically the mean $\overline{x}_{[n]r}$ of a ranked set sample from a finite population, the variance of which is given by [26] with the sample mean $\overline{x}_{nr,\mathrm{SRSWOR}}$ based on *nr* actual measurements under SRSWOR, the variance of which is given by

$$\mathrm{Var}\,(\overline{x}_{nr,\mathrm{SRSWOR}}) = \left(1-\frac{nr}{N}\right)\frac{S^2}{nr} \tag{31}$$

we get

$$\begin{aligned}&\mathrm{Var}\,(\overline{x}_{[n]r}) \le \mathrm{Var}\,(\overline{x}_{nr}) \Leftrightarrow \\ &\Leftrightarrow \frac{1}{N-1}\frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2 \le \frac{1}{n-1}\frac{1}{n}\sum_{i=1}^{n}(Ex_{i:n} - \overline{X})^2.\end{aligned} \tag{32}$$

When the population size *N* is large compared with *n* the condition (32) should be satisfied.

7. Simulation Study

Data for the simulation are taken from Särndal, Swensson and Wretman [7]. Population of $N = 281$ municipalities in Sweden is considered. Sweden is divided into 284 municipalities but three largest municipalities: Stockholm, Göteborg and

Malmö are excluded from the analysis. Two different variables are taken into account:

Y – 1985 population in thousands (concomitant variable – easily accessible),

X – Revenues from the 1985 municipal taxation in millions of kronor (study variable).

Population parameters are the following: $\overline{X} = 187.06$, $CV_X = 1.067$ (coefficient of variation), $\rho_{XY} = 0.992$. Two different estimators of a population mean $\overline{X}$ are considered:

- $\overline{x}_{\text{SRSWOR}}$ sample mean under simple random sampling without replacement,
- $\overline{x}_{\text{RSS}}$ ranked set sample mean based on one cycle $(r = 1)$, when ranking is implemented on the basis of a concomitant variable Y – 1985 population.

In the case of both sampling schemes sample size is $n = 20$. To compare different methods of estimation sampling is repeated 10000 times.

Table 1

Simulation results

	$\overline{x}_{\text{SRSWOR}}$	$\overline{x}_{\text{RSS}}$
Mean of 10000 repetitions	187.52	196.93
Bias	0.47	–0.13
Bias in %	0.25	–0.07
MSE	1867.65	409.45
Root mean square error	43.22	20.23
Root mean square error in %	23.10	10.82

S o u r c e: own calculations.

As one can see from table 1 ranked set sampling implemented on the basis of a concomitant variable proved to be more efficient for estimating population mean than simple random sampling without replacement. Gain in efficiency is very high: $\frac{43.22 - 20.23}{43.22} \cdot 100\% = 53.19\%$ as far as the root mean square error is concerned.

References

[1] DAVID H.A., LEVINE D.N., *Ranked Set Sampling in the Presence of Judgement Error*, Biometrics, 1972, 28, 553–555.

[2] DELL T.R., CLUTTER, J.L., *Ranked Set Sampling Theory with Order Statistics Background*, Biometrics, 1972, 28, 545–555.

[3] KOWALCZYK B., *Ranked Set Sampling and Its Applications in Finite Population Studies*, Statistics in Transition, 2004, Vol. 6, No 7, 1031–1046.
[4] MCINTYRE G.A., *A Method of Unbiased Selective Sampling, Using Ranked Sets*, Australian J. Agricultural Research, 1952, 3, 385–390.
[5] PATIL G.P., SINHA A.K., TAILLIE C., *Ranked Set Sampling*, Handbook of Statistics, 1994, Vol. 12, 167–201.
[6] SAMAWI H.M., MOHMMAD S., ABU-DAYYEH W., *Estimating the Population Means Using Extreme Ranked Set Sampling*, Biometrical Journal, 1996, 38, 577–586.
[7] SÄRNDAL C.E., SWENSSON B., WRETMAN J., *Model Assisted Survey Sampling*, Springer-Verlag, 1992.
[8] STOKES S.L., *Ranked Set Sampling with Concomitant Variables*, Communications in Statistics, Theory and Methods, 1977, 6, 1207–1211.
[9] TAKAHASI K., WAKIMOTO, K., *On Unbiased Estimates of the Population Mean Based on the Sample Stratified Means of Ordering*, Annals of the Institute of Statistical Mathematics, 1968, 20, 1–31.

O estymatorach średniej opartych na zbiorach porangowanych

Przedstawiono przegląd wyników dotyczących estymacji wartości średniej w populacji, gdy próba jest oparta na zbiorach porangowanych. Próbkowanie takie polega na wylosowaniu *n* prób po *n* elementów w każdej próbie. Następnie każdemu elementowi w próbie nadaje się rangę (bez wykonywania dokładnego pomiaru), a do próby właściwej włącza się po jednym elemencie z każdego zbioru. Autorka przedstawia sytuacje, w których: rangowanie bez dokonywania dokładnego pomiaru jest bezbłędne, sytuację dopuszczającą błędy w rangowaniu, rangowanie na podstawie cechy stowarzyszonej, a także rangowanie ograniczające się tylko do ekstremalnych statystyk pozycyjnych. Rozważany jest zarówno przypadek populacji nieskończonej, jak i skończonej. Przeglądowe wyniki teoretyczne zobrazowano badaniem symulacyjnym, przeprowadzonym na populacji rzeczywistej.

Słowa kluczowe: *zbiory porangowane, statystyki pozycyjne, próbkowanie*