Estimation of the Population Mean using Random Selection in Ranked Set Samples

I. Rahimov and H. A. Muttlak

Department of Mathematical Sciences,
King Fahd University of Petroleum & Minerals,
Dhahran 31261, Saudi Arabia

Abstract. The use of ranked set sample (RSS) to estimate the population mean is a well known for its advantages over usual method using simple random sample (SRS). In this paper we generalize the random selection in ranked set sampling proposed by Li et al. (1999) to come up with estimator of the population mean. It will be shown that this estimator is more practical and more efficient in some cases.

Key words: normal distribution, ranked set sampling, simple random sampling.

1. Introduction

The ranked set sampling (RSS) has attracted number of researchers and investigators as an efficient sampling method, particularly in the area of environmental and ecological investigations. McIntyre (1952) proposed the ranked set sampling (RSS) method without the statistical theory to support his suggestion. Takahasi and Wakimoto (1968) supplied mathematical theory to support McIntyre's suggestion. Dell and Clutter (1972) showed that errors in ranking reduce the efficiency of the RSS mean relative to the simple random sampling (SRS) mean.

Li et al. (1999) introduced the notion of random selection of m sets out of k sets, $m < k$, where k is the set size in the usual RSS method. They studied the properties of the estimators of the population mean and variance based on the new randomly selected sample.

For classified and extensively reviewed work in the area of RSS from 1952 to 1994 see Patil et al (1994).

The RSS method can be summarized as follows: from a population of interest, n random sets each of size *n* are selected. The members of each random set are ranked with respect to the variable of interest by a cost free method e.g. eye inspection. From the first ordered set, the smallest unit is selected for measurement. From the second ordered set, the second smallest unit is selected for measurement. This continues until the largest element from the last ordered set is measured. This process may be repeated r times (i.e. *m* cycles or replications) to yield a sample of size mn.

In this paper we investigate the estimation of the population mean under the random selection of ranked set sample proposed by Li et al. (1999) with random number of rows selected. The newly proposed estimator of the population mean is shown to be more efficient under some conditions.

In the usual RSS method we need m set of size m units each, but some time we may not have the recourses or the time to measure all the m units one from each set, the newly suggested method can be utilize to reduce the number of units to k instead of m where k<m. In the study done by Muttlak and Al-Sabah (2001) we can see that it is very easy to implement the new method. If we use the RSS method with the line intercept sampling method see Muttlak and McDonald (1992), we may not have n^2 units to perform the usual RSS method, in this case the newly suggested method can be used to resolve this problem.

2. Estimating the Population Mean

We consider the following family of random variables

$X_{11}, X_{12}, ..., X_{1n}; X_{21}, X_{22}, ..., X_{2n}; ....; X_{i1}, X_{i2}, ..., X_{in};...; X_{n1}, X_{n2}, ..., X_{nn}$ be n^2 independent and identically distributed random variables with cdf F (x), pdf f (x), mean μ and variance σ^2. If we order each set of n random variables, we will get

$$X_{1(1)}, X_{1(2)}, ..., X_{1(n)}$$

$$X_{2(1)}, X_{2(2)}, ..., X_{2(n)}$$

……………………..

$$X_{i(1)}, X_{i(2)}, ..., X_{i(n)}$$

……………………..

$$X_{n(1)}, X_{n(2)}, ..., X_{n(n)}$$

Let ν be integer valued random variable taking values from $\Lambda = \{1, 2, 3, ..., n\}$ independent of the family $\{X_{ij},\ i,\ j\ =\ 1,\ 2,\ ...,\ n\}$.

For a given value of ν we select ν indices $(k_1, k_2, ..., k_\nu)$ from Λ at random without replacement independently of the value ν. Further we select $X_{1(k_1)}$ from the first row i.e. k_1th order statistics, $X_{2(k_2)}$ from the second row and so on $X_{\nu(k_\nu)}$ from the νth row of the table of ordered sets for quantification. Our inference for μ will be based on the random size sample $\{X_{1(k_1)},\ X_{2(k_2)},\ ...,\ X_{\nu(k_\nu)}\}$ of order statistics. We propose

$$\overline{y}_\nu = \frac{1}{\nu}\sum_{i=1}^{\nu} X_{i(k_i)} ,$$

as an estimator of the population mean μ.

Properties of $\overline{y}_\nu$

Let $\mu_{i:n}$ and $\sigma_{i:n}^2$ be expectation and variance of ith-order statistic in a simple random sample (SRS) of size n. In what follows we will consider the properties of $\overline{y}_\nu$.

Since $P(k_i = j) = \frac{1}{n}, j = 1, 2, ..., n$, we have

$$E(X_{i(k_i)}) = \sum_{j=1}^{n} E(X_{i(j)})\frac{1}{n} = \mu .$$

Thus

$$E(\overline{y}_\nu) = E[E[\frac{1}{\nu}\sum_{i=1}^{\nu} X_{i(k_i)}|\nu]] = E[\frac{1}{\nu}\sum_{i=1}^{\nu} E(X_{i(k_i)})] = \mu .$$

To find variance of $\overline{y}_\nu$ we use the following relation

$$Var(\overline{y}_\nu) = E[Var[\overline{y}_\nu|\nu]] + Var[E(\overline{y}_\nu|\nu)]. \quad (1)$$

Since $E[\overline{y}_\nu|\nu] = \frac{1}{\nu}\sum_{i=1}^{\nu} E(X_{i(k_i)}) = \mu$, the second term in equation (1) equal zero.

According to theorem 2.1 in Li et al. (1999)

$$Var(\overline{y}_r) = \frac{Q}{nr}\frac{n-r}{n-1} + \frac{1}{nr}\sum_{i=1}^{n}\sigma_{i:n}^2 , \quad (2)$$

for any fixed r, where $Q = \sum_{i=1}^{n}(\mu_{i:n} - \mu)^2$. Letting $\sigma_{[n]}^2 = \frac{1}{n}\sum_{i=1}^{n}\sigma_{i:n}^2$, we obtain that

$$Var(\overline{y}_\nu) = [\frac{Q}{n-1} + \sigma^2_{[n]}]E[\frac{1}{\nu}] - \frac{Q}{n(n-1)}. \quad (3)$$

3. Variable Size Samples

Let now ν be a binomial random variable truncated at zero with parameters n and $\frac{1}{2}$, i.e. $P(\nu = i) = P(\xi = i | \xi \geq 1)$, where ξ follows binomial distribution $b(n, \frac{1}{2})$. Then it is easy to find that

$$P(\nu = i) = \binom{n}{i}(2^n - 1)^{-1}, i = 1, 2, ..., n. \quad (4)$$

The distribution of ν given above corresponds to a scheme of sampling introduced by Wright (1988), called a variable size simple random sampling (VSSRS). The sampling plan is such that each of the $2^n - 1$ nonempty subsets of Λ has an equal probability of selection. Since for any fixed r

$$P(k_1 = i_1, k_2 = i_2, ..., k_r = i_r) = \left[r! \binom{n}{r} \right]^{-1},$$

using Lemma 2.1 by Li et al. (1999) we have

$$P(k_1 = i_1, k_2 = i_2, ..., k_\nu = i_\nu) = (2^n - 1)^{-1} \sum_{t=1}^{n} \frac{1}{r!}. \quad (5)$$

The following result gives a comparison of the variance of $\overline{y}_\nu$ with the variance of fixed sample size estimator $\overline{y}_r$.

Theorem 1.

One. If $r \geq \frac{n 2^{n-1}}{2^n - 1}$, then $Var(\overline{y}_\nu) \geq Var(\overline{y}_r)$;

Two. If $r \leq \frac{n 2^{n-1} - 2^n + 1}{2^n - 1}$, then $Var(\overline{y}_\nu) \leq Var(\overline{y}_r)$.

Proof a. Using the Cauchy- Schwarz inequality, we obtain

$$1 = E[\frac{1}{\nu}\nu] \leq E[\frac{1}{\nu}]E[\nu]. \quad (6)$$

If we take into account inequality (6) in equation (3) we have

$$Var(\overline{y}_\nu) \geq [\frac{Q}{n-1} + \sigma^2_{[n]}]\frac{1}{E(\nu)} - \frac{Q}{n(n-1)}. \quad (7)$$

It is easy to see that equation (2) can be written as

$$Var(\bar{y}_r) = [\frac{Q}{n-1} + \sigma^2_{[n]}]\frac{1}{r} - \frac{Q}{n(n-1)}. \quad (8)$$

From equations (7) and (8) we obtain that $Var(\bar{y}_\nu) \geq Var(\bar{y}_r)$, if $r \geq E(\nu)$. On other hand, since $E(\nu) = E[\xi | \xi > 0]$, where ξ follows $b(n, \frac{1}{2})$, we obtain

$$E(\nu) = \frac{n\, 2^{n-1}}{2^n - 1} \quad (9)$$

This is complete the proof of part a.

Proof b. To prove part b we need the upper estimate for $E(\frac{1}{\nu})$. For any random variable ν having probability distribution given in equation (4), it is proved by Wright (1988) that

$$(E[\nu]-1)E(\frac{1}{\nu}) \leq 1 \quad (10)$$

If we use inequality (10) in equation (3) we obtain that

$$Var(\bar{y}_\nu) \leq [\frac{Q}{n-1} + \sigma^2_{[n]}]\frac{1}{E(\nu)-1} - \frac{Q}{n(n-1)}. \quad (11)$$

It follows from inequalities (9) and (11) that $Var(\bar{y}_\nu) \leq Var(\bar{y}_r)$, if $r \leq E(\nu) - 1$. Taking into account equation (9) we obtain the bound of for r in part b. This proves the theorem.

4. Best Linear Unbiased Estimator for the Normal Population Mean

In this section we consider the case when the initial population follows normal distribution with a mean μ and variance σ^2. Let $\mu_{(i:n)}$ and $V_{(i:n)}$ be the mean and variance of the ith-order statistic from a standard normal population when the sample size is n. Also ν as defined before, be an integer valued random variable taking values from the set Λ={1, 2, …, n} and independent of the family $\{X_{ij}, i, j = 1, 2, ..., n\}$.

We propose new estimator for μ based on best linear unbiased estimator proposed by Li et al. (1999) which can be defined as follows

$$\tilde{\mu}_{rss}(\nu) = \frac{M_1(\nu)T_1(\nu) - M_2(\nu)T_2(\nu)}{[M_1(\nu)]^2 - M_2(\nu)\sum_{i=1}^{\nu} 1/V_{(k_i:n)}}, \quad (12)$$

where $M_1(\nu)=\sum_{i=1}^{\nu}\mu_{(k_i:n)}/V_{(k_i:n)}$, $M_2(\nu)=\sum_{i=1}^{\nu}\mu^2{}_{(k_i:n)}/V_{(k_i:n)}$, $T_1(\nu)=\sum_{i=1}^{\nu}X_{i(k_i)}\mu_{(k_i:n)}/V_{(k_i:n)}$

and $T_2(\nu)=\sum_{i=1}^{\nu}X_{i(k_i)}/V_{(k_i:n)}$. It is shown by Li, et al. (1999) that $\tilde{\mu}_{rss}(j)$ for any fixed $1\le j\le n$ is unbiased estimator of μ and its variance equal to

$$Var(\tilde{\mu}_{rss}(j))=\frac{\sigma^2}{\binom{n}{j}}\sum_{\substack{k_1,\,\ldots,\,k_j=1\\ k_1<\ldots<k_j}}^{n}\frac{M_2(j)}{M_2(j)\sum_{i=1}^{j}\frac{1}{V_{(k_i:n)}}-M_1^2(j)} \tag{13}$$

Since ν and the family of random variables $\{X_{ij}\}$ are independent, we obtain by total probability arguments that

$$E[\tilde{\mu}_{rss}(\nu)]=E[E[\tilde{\mu}_{rss}(\nu)|\nu]]=\mu\,. \tag{14}$$

To find the variance we use the relation

$$Var[\tilde{\mu}_{rss}(\nu)]=E[Var[\tilde{\mu}_{rss}(\nu)|\nu]]+Var\big(E[\tilde{\mu}_{rss}(\nu)|\nu]\big). \tag{15}$$

Since, $E[\tilde{\mu}_{rss}(\nu)|\nu]=\mu$ the second term on the right side of (15) equals to zero. Consider the first term. Using equation (13) by total probability arguments we have the first term and, consequently, the variance is given as

$$Var(\tilde{\mu}_{rss}(\nu))=\sigma^2E\left\{\left[\binom{n}{\nu}\right]^{-1}\sum_{\substack{k_1,\,\ldots,\,k_\nu=1\\ k_1<\ldots<k_\nu}}^{n}\frac{M_2(\nu)}{M_2(\nu)\sum_{i=1}^{\nu}\frac{1}{V_{(k_i:n)}}-M_1^2(\nu)}\right\}. \tag{16}$$

5. Comparison with Sample Mean of Simple Random Sample (SRS)

Now we consider a comparison of our estimator $\overline{y}_\nu$ with SRS mean $\overline{X}$ in the case when the initial population is normal. Let

$$\alpha(n)=\frac{1}{n(n-1)}\sum_{i=1}^{n}\mu_{(i:n)}^2\ ,$$

where $\mu_{(i:n)}$ is the mean of the ith order statistics from a standard normal distribution.

Theorem 2. $Var(\overline{y}_n)\le Var(\overline{X})$ if and only if the random sample size ν such that

$$E\left[\frac{1}{\nu}\right]\le\frac{n^{-1}+\alpha(n)}{1+\alpha(n)}. \tag{17}$$

Proof. It follows from Lemma 2.3 in Li et al. (1999) that $\sigma^2_{[n]} = \sigma^2 - Q/n$. Since in the case of normal population with mean μ and standard deviation σ, the mean of the ith order statistic is $\mu_{i:n} = \mu + \sigma\mu_{(i:n)}$, we obtain that

$$Q = \sigma^2 \sum_{i=1}^{n} \mu^2_{(i:n)} . \tag{18}$$

If we use these relations in equation (3), we obtain that

$$Var(\overline{y}_\nu) = \sigma^2 \{E[\frac{1}{\nu}] - \alpha(n)(1-[\frac{1}{\nu}])\} . \tag{19}$$

Since $Var(\overline{X}) = \frac{\sigma^2}{n}$ it follows from (18) that $Var(\overline{y}_n) \le Var(\overline{X})$, if

$$E[\frac{1}{\nu}] - \alpha(n)(1-E[\frac{1}{\nu}]) \le \frac{1}{n} .$$

The last inequality is equivalent to (17). Let now assume that (17) is not satisfied, i.e.

$$E\left[\frac{1}{\nu}\right] > \frac{n^{-1}+\alpha(n)}{1+\alpha(n)} .$$

We have from this inequality that

$$E[\frac{1}{\nu}] - \alpha(n)(1-E[\frac{1}{\nu}]) > \frac{1}{n}$$

which yields due to (19) that $Var(\overline{y}_n) > Var(\overline{X})$. The theorem is proved.

Corollary 1. If n ≥ 4 and

$$E[\frac{1}{\nu}] \le \frac{3n-2}{n(2n-1)} , \tag{20}$$

then $Var(\overline{y}_\nu) \le Var(\overline{X})$.

Proof. It shown in Li and Ni Chuiv (1997) for a standard normal population with n ≥ 4 that

$$\frac{1}{n} \sum_{i=1}^{n} \mu^2_{(i:n)} > \frac{1}{2} . \tag{21}$$

Since (a +x)/(b+x), a, b > 0 is increasing function when b > a, taking into account (21) in (17), we obtain that if (20) satisfied, then $Var(\overline{y}_\nu) \le Var(\overline{X})$.

Acknowledgments

The authors are indebted to King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia for excellent research facilities.

REFERENCES

1. Dell and Clutter (1972). Ranked set sampling theory with order statistic background, *Biometrics*. 28, pp. 545-555.
2. Li, D. and Chuiv, N. N. (1997). On the efficiency of ranked set sampling strategies in parametric estimation. *Calcutta Statistical Association Bulletin*, 47, 23-42
3. Li, D. Sinha, B. K. and Perron, F. (1999). Random selection in ranked set sampling and its applications, *Journal of Statistical Planning and Inference*, 76, 185-201.
4. McIntyre, G. A. (1952). A method of unbiased selective sampling, using ranked sets, *Australian Journal of Agricultural Research*, 3, 385-390.
5. Muttlak, H. A. and McDonald, L. L. (1992). Ranked set sampling and the line intercept method: A more efficient procedure, *The Biometrical Journal*, 34, 329-46
6. Muttlak, H. A. and Al-Sabah (2001). Statistical quality control using ranked set sampling, *Technical Report no. FT/2000*-15, Department of Mathematical Sciences, KFUPM.
7. Patil, G. P., Sinha, A. K. and Taillie, C. (1994). Ranked set sampling, A *Handbook of Statistics*, vol. 12, pp167-200.
8. Takahasi K. and Wakimoto K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering, *Annals of the Institute of Statistical Mathematics,* 21, 249-55.
9. Wright, T. (1988). On some properties of variable size simple random sampling and a limit theorem, *Commination in Statistics Theory and Method*, 17(9), 2997-3016.