ESTIMATION OF VARIANCE FROM RANKED SET SAMPLES

by

Arup Bose
Indian Statistical Institute
and
Purdue University

and

Nagaraj K. Neerchal
University of Maryland,
Baltimore County

Technical Report #96-11

This is a Preliminary Report

Department of Statistics
Purdue University
West Lafayette, IN USA

April 1996

*

ESTIMATION OF VARIANCE FROM RANKED SET SAMPLES

by

Arup Bose
Indian Statistical Institute
and
Purdue University

and

Nagaraj K. Neerchal
University of Maryland,
Baltimore County

This is a Preliminary Report

Abstract

We propose an unbiased estimator of the population variance when a ranked set sample is available, and compare it with other estimators.

Key Words and Phrases. Ranked set sampling, ordered statistic, mean squared error, relative precision, concomittants of ordered statistics.

AMS 1991 Subject Classification: Primary 62G05, Secondary 62F11

1. Introduction. Ranked set sampling (RSS) has established itself as a useful sampling procedure, specially in environmental studies. See Patil, Sinha and Tallie (1994) for a comprehensive review of the theory, methods and applications of this procedure. From the review, it is evident that most of the work has concentrated on evaluating this method and its variants in the context of estimating the *mean* of a population.

We focus on the problem of estimating the population *variance* σ^2. Stokes (1980) discussed this problem and provided an estimator. She compared her estimator with the usual variance estimator based on simple random sampling (SRS) through their mean squared error (MSE). She concluded that the gain in benefits by using the RSS variance estimator in place of the SRS estimator is rather small, specially when the population mean is unknown.

We propose an alternative *unbiased* estimator when replications are available. We then compare the performances of all three estimators by comparing their MSEs.

In practice, the rankings of the values may not be perfect. One such situation is when an auxiliary variable is used to rank the primary variable. We study the effect of this on our estimator by using a real data set.

2. Variance Estimators. Let $S_i = (X_{i1}, \ldots, X_{im})$ be random samples of size m from a population F, $1 \leq i \leq m$. Let $X_{(i)}$ be the ith *ordered* value in S_i. Then the ranked set sample is defined as the m independent values $X_{(i)}$, $1 \leq i \leq m$. If this process is replicated r times and $X_{(i)j}$ is the value of $X_{(i)}$ obtained on the jth replication then $X_{(i)j}, 1 \leq i \leq m$, $1 \leq j \leq r$ is the ranked set sample with mr independent values being quantified.

Stokes (1980) suggested the following estimator of the variance

$$\hat{\sigma}^2 = (mr-1)^{-1} \sum_{j=1}^{r} \sum_{i=1}^{m} (X_{(i)j} - \overline{X})^2 \ldots \ldots \tag{1}$$

where

$$\overline{X} = (mr)^{-1} \sum_{j=1}^{r} \sum_{i=1}^{m} X_{(i)j} \ldots \ldots \tag{2}$$

The bias and MSE of this estimator were established to be

$$\text{Bias}(\hat{\sigma}_1^2) = [m(mr-1)]^{-1} \sum_{i=1}^{m} (\mu_{(i)} - \mu)^2 \ldots \ldots \tag{3}$$

$$\text{where} \qquad \mu_{(i)} = E_F(X_{(i)}) \ \ldots \ \ldots \tag{4}$$

$$\mu = E_F(X_{11}) \ \ldots \ \ldots \tag{5}$$

$$\begin{aligned}
\text{MSE}(\hat{\sigma}_1^2) = {} & \frac{1}{mr}\mu_4 + \frac{(mr-3)(mr+1)}{(mr-1)^2\ m^2 r} \sum \tau_{(r)}^4 \\
& + \frac{4}{m^2 r(mr-1)} \sum \tau_{(r)} E(X_{(r)} - \mu)^3 \\
& + \frac{6 - 2(mr)^2}{m^2\ r(mr-1)^2} \sum \tau_{(r)}^2 E(X_{(r)} - \mu)^2 \\
& - \frac{(mr)^2 - 2mr + 3}{(mr-1)^2 m^2 r} \sum \sigma_{(r)}^4 \\
& + \frac{2}{m^2 r(mr-1)^2} \left(\sum \sigma_{(r)}^2 \right)^2 \\
& + \frac{1}{(mr-1)^2 r^2} \left(\sum \tau_{(r)}^2 \right)^2 \ldots \ldots
\end{aligned} \tag{6}$$

where

$$\mu_4 = E_F\ (X_{11} - \mu)^4 \ \ldots \ \ldots \tag{7}$$

$$\sigma_{(r)}^2 = V(X_{(r)}) \ldots \ \ldots \tag{8}$$

$$\tau_{(r)} = \mu_{(r)} - \mu \ldots \ \ldots \tag{9}$$

To compare this estimator with the usual variance estimator s^2, based on mr i.i.d. observations on F, note that

$$\lim_{r \to \infty} (mr)\ \text{MSE}\ (\hat{\sigma}_1^2) = \mu_4 - \frac{1}{m} \sum_{i=1}^{m} (\sigma_{(i)}^2 + \mu_{(i)}^2)^2 \ldots \ \ldots \tag{10}$$

$$\lim_{r \to \infty} (mr)\ \text{MSE}\ (s^2) = \mu_4 - \sigma^4 \ldots \ \ldots \tag{11}$$

Thus the relative precision (RP) of $\hat{\sigma}_1^2$ compared to s^2, satisfies

$$\lim_{s \to \infty} \text{RP}\ (s^2, \hat{\sigma}_1^2) = \lim_{r \to \infty} \frac{\text{MSE}\ (s^2)}{\text{MSE}\ (\hat{\sigma}_1^2)} = \frac{\mu_4 - \sigma^4}{\mu_4\ - m^{-1} \sum \left(\sigma_{(i)}^2 + \tau_{(i)}^2 \right)^2} \geq 1 \ldots \ \ldots \tag{12}$$

Hence $\hat{\sigma}_1^2$ is "better" than s^2. However, the estimator $\hat{\sigma}_1^2$ does not exploit the replications to achieve unbiasedness. To obtain an estimator which is unbiased note that

$$E\ m^{-1}\sum_{i=1}^{m} X_{(i)j} = \mu \ldots\ \ldots \tag{13}$$

$$E\ m^{-1}\sum_{i=1}^{m} X_{(i)j}^2 = \sigma^2 + \mu^2 \ldots\ \ldots \tag{14}$$

To write down our estimator, we use the standard convention of $(\cdot)$ denoting that the corresponding index has been summed and of a bar denoting an average. Our estimator $\hat{\sigma}_2^2$ is defined as

$$\hat{\sigma}_2^2 = \frac{1}{mr}\sum_{j=1}^{r}\sum_{i=1}^{m} X_{(i)j}^2 - \binom{r}{2}^{-1} \sum_{1\le j<j'\le r} \overline{X}_{(\cdot)j}\,\overline{X}_{(\cdot)j'} \ \ldots\ \ldots \tag{15}$$

The basic properties of $\hat{\sigma}_2^2$ are given by

Result 1. The estimator $\hat{\sigma}_2^2$ satisfies

(a) $\hat{\sigma}_2^2 \ge 0$

(b) $E\ \hat{\sigma}_2^2 = \sigma^2$

(c) Letting μ_2', μ_3', μ_4' denote the second, third and fourth moments of F,

$$\begin{aligned} V(\hat{\sigma}_2^2) = {} & \frac{1}{mr}\mu_4' - \frac{1}{m^2 r}\sum_{i=1}^{m}\left(\sigma_{(i)}^2 + \mu_{(i)}^2\right)^2 \\ & + \binom{r}{2}^{-1}\left[2(r-2)\mu^2\frac{1}{m^2}\sum_{i=1}^{m}\sigma_{(i)}^2\right. \\ & \left. + \left(\frac{\mu_2'}{m} + \frac{1}{m^2}\sum_{i\ne i'}\mu_{(i)}\mu_{(i')}\right)^2 - \mu^4\right] \\ & - \frac{4}{r}\mu\left[\frac{\mu_3'}{m} - \frac{1}{m^2}\sum_{i=1}^{m}\mu_{(i)}\left(\sigma_{(i)}^2 + \mu_{(i)}^2\right)\right] \ldots\ \ldots \end{aligned} \tag{16}$$

(d)

$$\lim_{r\to\infty}(mr)V(\hat{\sigma}_2^2) = \mu_4 - \frac{1}{m}\sum\left(\sigma_{(i)}^2 + \tau_{(i)}^2\right)^2 - 4\mu^2\frac{1}{m}\sum\sigma_{(i)}^2 \ldots \tag{17}$$

and the first, second and last term in the variance expression (16) may be replaced $(mr)^{-1}$ times the above expression.

References

Patil, G. P., Sinha, A. K. and Tallie, C. (1994). Ranked set sampling. In *Handbook of Statistics*, Vol. 12, 167-200. Elsevier Science B. V.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley, New York.

Stokes, S. L. (1980). Estimation of variance using judgement ordered ranked set samples. *Biometrics* 36, 35-42.

McIntyre, G. A. (1952). A method of unbiased selective sampling using ranked sets. *Austral. J. Agricultural Res.* 3, 385-390.

Appendix

Proof of Result 1. To prove (a) observe that

$$\hat{\sigma}_2^2 = \frac{1}{r}\sum_{j=1}^{r} \overline{X^2}_{(\cdot)j} - \frac{1}{r(r-1)}\left[\left(\sum_{j=1}^{m} \overline{X}_{(\cdot)j}\right)^2 - \sum_{j=1}^{m}\left(\overline{X}_{(\cdot)j}\right)^2\right]$$

$$= \frac{1}{r}\sum_{j=1}^{r}\left(\overline{X^2}_{(\cdot)j} - \left(\overline{X}_{(\cdot)j}\right)^2\right)$$

$$+\frac{r}{(r-1)}\left[\frac{1}{r}\sum_{j=1}^{r}\left(\overline{X}_{(\cdot)j}\right)^2 - \left(\frac{1}{r}\sum_{j=1}^{m}\overline{X}_{(\cdot)j}\right)^2\right]$$

By Cauchy-Schwartz inequality, each term in the first sum is nonnegative. The second term is nonnegative, again by Cauchy-Schwartz inequality.

(b) is an easy consequence of (13), (14) and the definition of the estimator. To prove (c), note that $\hat{\sigma}_2^2 = T_1 - T_2$

where

$$T_1 = \frac{1}{r}\sum_{j=1}^{r} \overline{X^2}_{(\cdot)j} \ \ldots \ \ldots \qquad (A1)$$

$$T_2 = \binom{r}{2}^{-1} \sum_{1\le j<j'\le r} \overline{X}_{(\cdot)j}\,\overline{X}_{(\cdot)j'} \ldots \ \ldots \qquad (A2)$$

$$
\begin{aligned}
V(T_1) &= \frac{1}{r}\, V(\overline{X^2}_{(\cdot)j}) \\
&= \frac{1}{r\, m^2} \sum_{i=1}^{m} V\left(X^2_{(i)1}\right) \\
&= \frac{1}{r\, m^2} \sum_{i=1}^{m} \left\{ E(X^4_{(i)}) - \left(EX^2_{(i)}\right)^2 \right\} \\
&= \frac{1}{r\, m}\, \mu_4' \;-\; \frac{1}{r\, m^2} \sum_{i=1}^{m} \left(\sigma^2_{(i)} \;+\; \mu^2_{(i)}\right)^2 \;\ldots
\end{aligned}
\tag{A3}
$$

Note that T_2 is a U-statistics with the kernel of order 2, $h(x, y) = xy$. Hence using Lemma A of Serfling (1980, page 183)

$$
V(T_2) = \binom{r}{2}^{-1} \left[2(r-2)\; V\{E(\overline{X}_{(\cdot)1}\; \overline{X}_{(\cdot)2} \mid \overline{X}_{(\cdot)1})\} \;+\; V(\overline{X}_{(\cdot)1}\; \overline{X}_{(\cdot)2})\right] \ldots \tag{A4}
$$

But the first variance equals

$$
V(\mu\; \overline{X}_{(\cdot)2}) \;=\; \mu^2\, \frac{1}{m^2} \sum_{i=1}^{m} \sigma^2_{(i)} \ldots \tag{A5}
$$

$$
V(\overline{X}_{(\cdot)1}\; \overline{X}_{(\cdot)2}) \;=\; E(\overline{X}_{(\cdot)1}\; \overline{X}_{(\cdot)2})^2 \;-\; (E\overline{X}_{(\cdot)1}\; \overline{X}_{(\cdot)2})^2 \;=\; \left[E(\overline{X}_{(\cdot)1})^2\right]^2 - \mu^4 \;\ldots \tag{A6}
$$

$$
\begin{aligned}
E(\overline{X}_{(\cdot)1})^2 &= E\left(\frac{1}{m} \sum X_{(i)1}\right)^2 \\
&= \frac{1}{m^2}\, E\left(\sum_{i=1}^{m} \sum_{i'=1}^{m} X_{(i)1}\; X_{(i')1}\right) \\
&= \frac{1}{m^2} \sum_{i=1}^{m} E\; X^2_{(i)1} + \frac{1}{m^2} \sum_{i \neq i'} \mu_{(i)}\; \mu_{(i')} \\
&= \frac{1}{m}\mu_2' + \frac{1}{m^2} \sum_{i \neq i'} \mu_{(i)}\; \mu_{(i')} \cdots
\end{aligned}
\tag{A7}
$$

It remains to find the covariance between T_1 and T_2. Note that

$$
\text{Cov}\left(\overline{X^2}_{(\cdot)1},\; \overline{X}_{(\cdot)1}\; \overline{X}_{(\cdot)2}\right) \;=\; \mu\left(E\; \overline{X^2}_{(\cdot)1}\; \overline{X}_{(\cdot)1}\right) \;-\; \left(E\; \overline{X^2}_{(\cdot)1}\right) \mu^2 \;\ldots \tag{A8}
$$

$$
\begin{aligned}
E\overline{X^2}_{(\cdot)1}\,\overline{X}_{(\cdot)1} &= \frac{1}{m^2}E\Big(\sum_{i=1}^{m} X_{(i)}^2\Big)\Big(\sum_{i=1}^{m} X_{(i)}\Big) \\
&= \frac{1}{m^2}\left[E\sum_{i=1}^{m} X_{(i)}^3 + E\Big(\sum_{i=1}^{m} X_{(i)}\Big)\sum_{\substack{j=1\\ j\neq i}}^{m} X_{(j)}^2\right] \\
&= \frac{1}{m}\,\mu_3' + \frac{1}{m^2}\sum_{i=1}^{m}\mu_{(i)}\Big(m\mu_2' - EX_{(i)}^2\Big) \\
&= \frac{1}{m}\,\mu_3' + \mu\,\mu_2' - \frac{1}{m^2}\sum_{i=1}^{m}\mu_{(i)}\Big(\sigma_{(i)}^2 + \mu_{(i)}^2\Big)\ldots\ \ldots
\end{aligned}
\tag{A9}
$$

Hence

$$
\mathrm{Cov}\left(\overline{X^2}_{(\cdot)1}, \overline{X}_{(\cdot)1}\,\overline{X}_{(\cdot)2}\right) = \mu\left(\frac{\mu_3'}{m} - \frac{1}{m^2}\sum_{i=1}^{m}\mu_{(i)}(\sigma_{(i)}^2 + \mu_{(i)}^2)\right)\ldots\ \ldots \tag{A10}
$$

Using equation (A3) - (A10), the expression (16) for the variance follows. To prove (d), observe that from (16), the limit equals

$$
\mu_4' - \frac{1}{m}\sum(\sigma_{(i)}^2 + \mu_{(i)}^2)^2 - 4\mu\left(\mu_3' - \frac{1}{m}\sum\mu_{(i)}(\sigma_{(i)}^2 + \mu_{(i)}^2)\right)\ldots\ \ldots \tag{A11}
$$

However,

$$
\mu_4' = \mu_4 + \mu^4 + 4\mu\,\mu_3 + 6\,\sigma^2\mu^2\ldots\ \ldots \tag{A12}
$$

$$
\mu_3' = \mu_3 + 3\sigma^2\mu + \mu^3\ldots\ \ldots \tag{A13}
$$

Hence

$$
\mu_4' - 4\mu\mu_3' = \mu_4 - 3\mu^4 - 6\sigma^2\mu^2\ldots\ \ldots \tag{A14}
$$

Observing that $\mu_{(i)} = \tau_{(i)} + \mu$,

$$\begin{aligned}
&\sum \left\{ \left(\sigma^2_{(i)} + \mu^2_{(i)} \right)^2 - 4\mu \; \mu_{(i)} \left(\sigma^2_{(i)} + \mu^2_{(i)} \right) \right\} \\
&= \sum \left(\sigma^2_{(i)} + \mu^2_{(i)} \right) \left\{ \sigma^2_{(i)} + \tau^2_{(i)} + \mu^2 + 2\mu\tau_{(i)} - 4\mu(\mu + \tau_{(i)}) \right\} \\
&= \sum \left(\sigma^2_{(i)} + \tau^2_{(i)} + \mu^2 + 2\mu\tau_{(i)} \right) \left\{ \sigma^2_{(i)} + \tau^2_{(i)} - 3\mu^2 - 2\mu\tau_{(i)} \right\} \\
&= \sum \left(\sigma^2_{(i)} + \tau^2_{(i)} \right)^2 + \sum \left(\sigma^2_{(i)} + \tau^2_{(i)} \right) \left(\mu^2 + 2\mu\tau_{(i)} - 3\mu^2 - 2\mu\tau_{(i)} \right) \\
&+ \sum \left(\mu^2 + 2\mu\tau_{(i)} \right) \left(-3\mu^2 - 2\mu\tau_{(i)} \right) \\
&= \sum \left(\sigma^2_{(i)} + \tau^2_{(i)} \right)^2 - 2\mu^2 \sum \left(\sigma^2_{(i)} + \tau^2_{(i)} \right) - 3m\mu^4 - 4\mu^2 \sum \tau^2_{(i)} \ldots
\end{aligned} \tag{A15}$$

Note that

$$\begin{aligned}
\sum \left(\sigma^2_{(i)} + \tau^2_{(i)} \right) &= \sum \left\{ EX^2_{(i)} - \mu^2_{(i)} + \mu^2_{(i)} + \mu^2 - 2\mu\mu_{(i)} \right\} \\
&= \sum \left(EX^2_{(i)} - \mu^2 \right) \\
&= m\sigma^2 \ldots
\end{aligned} \tag{A16}$$

$$\begin{aligned}
\sum \tau^2_{(i)} &= \sum \left(\mu^2_{(i)} + \mu^2 - 2\mu\mu_{(i)} \right) \\
&= \sum \mu^2_{(i)} - \mu^2 \ldots
\end{aligned} \tag{A17}$$

Using (A13) and (A15) - (A17) the expression in (A11) equals

$$\mu_4 - 3\mu^4 - 6\,\sigma^2\mu^2 - \left\{ \frac{1}{m} \sum_{i=1}^{m} \left(\sigma^2_{(i)} + \tau^2_{(i)} \right)^2 - 2\mu^2\sigma^2 - 3\mu^4 - 4\mu^2 \left(\sum_{i=1}^{m} \frac{\mu^2_{(i)}}{m} - \mu^2 \right) \right\}$$

$$= \mu_4 - \frac{1}{m} \sum_{i=1}^{m} \left(\sigma^2_{(i)} + \tau^2_{(i)} \right)^2 - 4\mu^2 \left\{ \sigma^2 - \frac{\sum_{i=1}^{m} \mu^2_{(i)}}{m} + \mu^2 \right\}$$

$$= \mu_4 - \frac{1}{m} \sum_{i=1}^{m} (\sigma^2_{(i)} + \tau^2_{(i)})^2 - 4\frac{\mu^2}{m} \sum_{i=1}^{m} \sigma^2_{(i)}$$