Ann. Inst. Statist. Math.
Vol. 47, No. 4, 621–636 (1995)

FINITE POPULATION CORRECTIONS FOR RANKED SET SAMPLING*

G. P. PATIL, A. K. SINHA AND C. TAILLIE

Center for Statistical Ecology and Environmental Statistics, Department of Statistics, Pennsylvania State University, University Park, PA 16802-2112, U.S.A.

(Received April 7, 1993; revised March 22, 1995)

Abstract. Ranked set sampling (RSS) for estimating a population mean μ is studied when sampling is without replacement from a completely general finite population $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)'$. Explicit expressions are obtained for the variance of the RSS estimator $\hat{\mu}_{\mathrm{RSS}}$ and for its precision relative to that of simple random sampling without replacement. The critical term in these expressions involves a quantity $\gamma = (\boldsymbol{x} - \mu)'\boldsymbol{\Gamma}(\boldsymbol{x} - \mu)$ where $\boldsymbol{\Gamma}$ is an $N \times N$ matrix whose entries are functions of the population size N and the set-size m, but where $\boldsymbol{\Gamma}$ does not depend on the population values $\boldsymbol{x}$. A computer program is given to calculate $\boldsymbol{\Gamma}$ for arbitrary N and m. When the population follows a linear (resp., quadratic) trend, then γ is a polynomial in N of degree $2m + 2$ (resp., $2m + 4$). The coefficients of these polynomials are evaluated to yield explicit expressions for the variance and the relative precision of $\hat{\mu}_{\mathrm{RSS}}$ for these populations. Unlike the case of sampling from an infinite population, here the relative precision depends upon the number of replications of the set size m.

Key words and phrases: Linear range, observational economy, order statistics from finite populations, quadratic range, relative savings, sampling efficiency, sampling from finite populations, sampling without replacement.

1. Introduction

The method of ranked set sampling (RSS) was introduced by McIntyre (1952) as a cost-efficient alternative to simple random sampling for those situations where outside information is available allowing one to rank small sets of sampling units according to the character of interest without actually quantifying the units. McIntyre was concerned with estimating agricultural yields where the ranking could be done on the basis of visual inspection. One of the strengths of the

* Prepared with partial support from the Statistical Analysis and Computing Branch, Environmental Statistics and Information Division, Office of Policy, Planning, and Evaluation, United States Environmental Protection Agency, Washington, DC under a Cooperative Agreement Number CR-821531. The contents have not been subjected to Agency review and therefore do not necessarily reflect the views of the Agency and no official endorsement should be inferred.

method, however, is that its implementation and performance require only that ranking be possible but they do not depend in any way on how the ranking is accomplished.

A *basic cycle* of the method involves the random selection of m^2 units from the population. These units are randomly partitioned into m subsets, each containing m sampling units. The members of every subset are ranked according to the character of interest. Then the lowest ranked member is quantified from the first set, the second lowest ranked member is quantified from the second set, and so on until the highest ranked member of the last set is quantified. This yields m quantifications from among the m^2 selected units. Since m is usually taken as small in order to facilitate the ranking, this may not be enough measurements for reasonable inference and the basic cycle is repeated r times to give $n = mr$ quantifications out of m^2r selected units. The arithmetic mean of these n measurements is the RSS estimator $\hat{\mu}_{\text{RSS}}$ of the population mean μ. The integers m and r are design parameters known as the *set-size* and the *replication factor* (or the *number of cycles*), respectively.

Performance of the RSS estimator is generally benchmarked against that of the simple random sampling (SRS) estimator $\hat{\mu}_{\text{SRS}}$ with the same number of quantifications. For this purpose, one may employ either the *relative precision*,

$$\text{RP} = \frac{\text{var}(\hat{\mu}_{\text{SRS}})}{\text{var}(\hat{\mu}_{\text{RSS}})},$$

or the *relative savings*,

$$\text{RS} = 1 - 1/\text{RP}.$$

There was little followup on McIntyre's (1952) proposal until the late 1960s when Halls and Dell (1966) published a field evaluation and Takahasi and Wakimoto (1968) developed the statistical theory for the RSS method. When sampling is from a continuous population and the ranking is perfect, Takahasi and Wakimoto proved that $\hat{\mu}_{\text{RSS}}$ is unbiased for μ and is at least as efficient as $\hat{\mu}_{\text{SRS}}$. They also obtained the variance of the RSS estimator as

$$\text{var}(\hat{\mu}_{\text{RSS}}) = \frac{1}{mr}\left[\sigma^2 - \frac{1}{m}\sum_{i=1}^{m}(\mu_{(i:m)} - \mu)^2\right], \tag{1.1}$$

where σ^2 is the population variance and $\mu_{(i:m)}$ is the expected i-th out of m order statistic from the population. From (1.1), Takahasi and Wakimoto established the bounds

$$1 \le \text{RP} \le \frac{m+1}{2}, \tag{1.2}$$

or,

$$0 \le \text{RS} \le \frac{m-1}{m+1}, \tag{1.3}$$

where the upper bounds are sharp and are achieved exactly when the population follows a uniform distribution. The upper bound in (1.3) indicates that ranked set sampling can result in very substantial savings when compared with simple random sampling. Specifically, the method can result in savings in the number of quantifications by as much as 33, 50, 60, 67 percent when $m = 2, 3, 4, 5$, respectively. The savings achieved in practice are somewhat less due to ranking errors and population skewness.

Because of this potential for observational economy, the RSS method has received growing attention both from statisticians and substantive scientists. See Patil *et al.* (1994) for an historical review of the theory, methods, and applications of ranked set sampling. However, these researches have been mostly concerned with sampling from infinite (continuous) populations. To our knowledge, the only exception is a paper in Japanese by Takahasi and Futatsuya (1988) giving a formula for the variance of $\hat{\mu}_{\mathrm{RSS}}$ when sampling is from a finite population. Unfortunately, the Takahasi-Futatsuya formula includes a general covariance term that depends on the structure of the population and is difficult to evaluate. In fact, Takahasi and Futatsuya obtain an explicit expression only for the combination of $m = 2$ and a discrete uniform population.

The present paper derives explicit expressions for $\mathrm{var}(\hat{\mu}_{\mathrm{RSS}})$ and for the corresponding relative savings when sampling is from an arbitrary finite population $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)'$. We show that the dependence on the population structure $\boldsymbol{x}$ is according to a bilinear function $\gamma = (\boldsymbol{x} - \mu)'\boldsymbol{\Gamma}(\boldsymbol{x} - \mu)$ of the centered population vector, $\boldsymbol{x} - \mu$, whose array of coefficients $\boldsymbol{\Gamma}$ can be written down explicitly in terms of binomial coefficients. Further, $\boldsymbol{\Gamma}$ involves only N and m and does not depend upon either r or $\boldsymbol{x}$. Upon asking, we can provide a computer program, in the GAUSS language, that evaluates $\boldsymbol{\Gamma}$ for arbitrary m and N (subject to memory limitations).

Next, we specialize to populations following either a linear or a quadratic trend in their range. In each case, γ is a polynomial function of N having degree $2m+2$ (linear trend) or $2m+4$ (quadratic trend). Using the GAUSS program to evaluate γ allows us to determine the coefficients of these polynomial functions and to write down $\mathrm{var}(\hat{\mu}_{\mathrm{RSS}})$ and RS explicitly as functions of (N, m, r) for the two population structures considered.

A final matter examined in the paper is how the relative performance of ranked set sampling depends upon the replication factor r. In the conventional case of sampling from an infinite population, the successive cycles are true replications, and RS (or RP) does not depend upon r. However, RS does depend upon r when sampling is without replacement from a finite population since the different cycles are not independent. In fact, we show that

$$\mathrm{RS} = \frac{1 - f/r}{1 - f} \times \mathrm{RS}^{(1)}, \tag{1.4}$$

where $f = n/N$ is the quantification (sampling) fraction and $\mathrm{RS}^{(1)}$ is the relative savings for a single cycle. This convenient formula means that only a single cycle has to be considered in theoretical investigations.

The present paper is concerned with derivations of the foregoing results and with numerical computations of RS for finite populations with linear and quadratic ranges. A companion paper (Patil *et al.* (1993)) discusses the implications of these developments, provides the results of some numerical computations, and compares the performance of RSS with that of systematic sampling and stratified random sampling from finite populations with a linear trend in their range.

First, we make a few remarks concerning the interplay between ranked set sampling and order statistics, and how that interplay is affected by finiteness of the population.

2. Order statistics for different sampling methods

If $X_1, X_2, \ldots, X_m$ is a random sample from an *infinite* population, then the SRS estimator of μ is

$$\hat{\mu}_{\text{SRS}} = \bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i. \tag{2.1}$$

Letting $X_{1:m}, X_{2:m}, \ldots, X_{m:m}$ be the order statistics of $X_1, X_2, \ldots, X_m$, the estimator (2.1) can be expressed as

$$\hat{\mu}_{\text{SRS}} = \frac{1}{m}\sum_{i=1}^{m} X_{i:m}. \tag{2.2}$$

Now consider ranked set sampling with only one cycle ($r = 1$) from the same infinite population and write $X_{(i:m)}$ for the quantification of the i-th ranked unit from the i-th set. Parentheses are used in the subscript to indicate that the $X_{(i:m)}$ are order statistics from *disjoint* sets, whereas the $X_{i:m}$ result from ordering a single set of m units. Although $X_{(i:m)}$ has the same marginal distribution as $X_{i:m}$, the different $X_{(i:m)}$ are independent while the $X_{i:m}$ are positively correlated. The RSS estimator is

$$\hat{\mu}_{\text{RSS}} = \frac{1}{m}\sum_{i=1}^{m} X_{(i:m)}. \tag{2.3}$$

Comparing equations (2.3) and (2.2), the estimator $\hat{\mu}_{\text{SRS}}$ is seen to have the larger variance because of the positive correlation among the $X_{i:m}$. This accounts for the superiority of RSS over simple random sampling and shows that the essence of McIntyre's method consists in obtaining direct *independent* measurements of the order statistics.

What happens when the population is finite and sampling is without replacement? First, the observations X_i in equation (2.1) are not independent and, in fact, are negatively correlated. The negative correlation reduces the variance of $\hat{\mu}_{\text{SRS}}$ and is the reason that sampling from finite populations is more efficient than sampling from equally diffuse infinite populations, as indicated by the usual finite population correction. There is a similar effect for ranked set sampling. Even

though the different sets are disjoint, they are not statistically independent and the various $X_{(i:m)}$ in (2.3) are negatively correlated, which has the beneficial effect of reducing the variance of $\hat{\mu}_{\text{RSS}}$ for finite populations. Unfortunately, the statistical analysis of $\hat{\mu}_{\text{RSS}}$ also becomes more involved since we have to determine the joint distribution of $X_{(i:m)}$ and $X_{(j:m)}$ and not just the marginal distributions of these variates. The determination of these joint and marginal distributions is taken up in the next section. In the case of RSS with r cycles, we will write $X_{(i:m)k}$ for the quantification of the i-th ranked unit in the i-th set of the k-th cycle for $k = 1, 2, \ldots, r$.

3. Order statistics from finite populations

Let $\Omega = \{x_1, x_2, \ldots, x_N\}$ be a finite population with mean μ and variance σ^2. Without loss of generality, we can suppose that $x_1 \leq x_2 \leq \cdots \leq x_N$ and we write $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)'$. Let a set of size m be selected at random and without replacement from Ω and define the event

$$\{i \Rightarrow s\}$$

to mean that the i-th ranked unit in the subset is the s-th ranked unit in the population. Also, write

$$A_i^s = \Pr\{i \Rightarrow s\}, \tag{3.1}$$

and let $\boldsymbol{A}_i$ denote the N-dimensional column vector having A_i^s as its s-th component. If $X_{(i:m)}$ is the quantification of the i-th ranked unit from the set, then

$$\begin{aligned} E[X_{(i:m)}] &\equiv \mu_{(i:m)} \\ &= \sum_{s=1}^{N} x_s \Pr(X_{(i:m)} = x_s) \\ &= \sum_{s=1}^{N} x_s \Pr(\{i \Rightarrow s\}) \\ &= \sum_{s=1}^{N} x_s A_i^s \\ &= \boldsymbol{A}_i' \boldsymbol{x}. \end{aligned} \tag{3.2}$$

In other words, the vector $\boldsymbol{A}_i$ defines the probability distribution of the order statistic $X_{(i:m)}$. Similarly,

$$\begin{aligned} \operatorname{var}(X_{(i:m)}) &\equiv \sigma^2_{(i:m)} \\ &= \boldsymbol{A}_i' \boldsymbol{x}^2 - (\boldsymbol{A}_i' \boldsymbol{x})^2, \end{aligned} \tag{3.3}$$

where $\boldsymbol{x}^2$ is the component-wise square of $\boldsymbol{x}$.

Next we study the joint distribution of the order statistics from disjoint sets. To this end, let two disjoint sets, each of size m, be drawn without replacement from Ω and write

$$\{i \Rightarrow s, j \Rightarrow t\}$$

for the event that the i-th ranked unit from set 1 has rank s in the population and the j-th ranked unit from set 2 has rank t in the population. Define

$$B_{ij}^{st} = \Pr(\{i \Rightarrow s, j \Rightarrow t\}) \tag{3.4}$$

and let $\boldsymbol{B}_{ij}$ be the $N \times N$ matrix with B_{ij}^{st} as its (s,t)-th component. Notice that $\boldsymbol{B}_{ij} = \boldsymbol{B}_{ji}'$ since

$$B_{ij}^{st} = B_{ji}^{ts}. \tag{3.5}$$

Let $X_{(i:m)1}$ and $X_{(j:m)2}$ be the quantifications of the i-th and the j-th ranked units from set 1 and set 2, respectively. Then,

$$\begin{aligned} E[X_{(i:m)1}X_{(j:m)2}] &= \sum_{s,t=1}^{N} x_s x_t \Pr(X_{(i:m)1} = x_s, X_{(j:m)2} = x_t) \\ &= \sum_{s,t=1}^{N} x_s x_t B_{ij}^{st} \\ &= \boldsymbol{x}'\boldsymbol{B}_{ij}\boldsymbol{x}. \end{aligned}$$

Consequently,

$$\begin{aligned} \operatorname{cov}(X_{(i:m)1}, X_{(j:m)2}) &\equiv C_{ij} \\ &= \boldsymbol{x}'(\boldsymbol{B}_{ij} - \boldsymbol{A}_i\boldsymbol{A}_j')\boldsymbol{x}. \end{aligned} \tag{3.6}$$

The equations (3.2), (3.3), and (3.6) express the first two joint moments of the order statistics in terms of the matrices $\boldsymbol{A}_i$ and $\boldsymbol{B}_{ij}$. We next develop explicit expressions for these matrices in terms of binomial and multinomial coefficients. As usual, we agree that the binomial coefficient $\binom{p}{q}$ and the multinomial coefficient $\binom{p}{q_1,q_2}$ vanish unless $0 \le q, q_1, q_2, q_1 + q_2 \le p$.

THEOREM 3.1. *The components of* $\boldsymbol{A}_i$, $i = 1, 2, \ldots, m$, *are given by*

$$A_i^s = \frac{\binom{s-1}{i-1}\binom{N-s}{m-i}}{\binom{N}{m}}, \qquad s = 1, 2, \ldots, N.$$

PROOF. In order for the event $\{i \Rightarrow s\}$ to be true, exactly $i-1$ units must be selected from among the smallest $s-1$ members of the population and $m-i$ units must be selected from among the $N-s$ largest members of the population.

THEOREM 3.2. *If $s < t$, then*

$$B_{ij}^{st} = \sum_{\lambda=0}^{m-i} \frac{\binom{s-1}{i-1}\binom{t-s-1}{\lambda}\binom{N-t}{m-i-\lambda}\binom{t-1-i-\lambda}{j-1}\binom{N-t-m+i+\lambda}{m-j}}{\binom{N}{m,m}}.$$

If $s = t$, then $B_{ij}^{st} = 0$; if $s > t$, then $B_{ij}^{st} = B_{ji}^{ts}$.

PROOF. Similar to the argument in Theorem 3.1. Here λ is the number of units in set 1 whose value lies between x_s and x_t. In fact, λ must satisfy all of the following restrictions:

$$\begin{aligned} &0 \le \lambda \le t - s - 1, \\ &2m - i - j + t - N \le \lambda \le m - i, \\ &\lambda \le t - i - j. \end{aligned}$$

However, one can simply let $\lambda = 0, 1, \ldots, m - i$ and our convention regarding the vanishing of the binomial coefficients will do the rest.

Next, we observe the following results for the component-wise sums of the matrices $\boldsymbol{A}_i$ and $\boldsymbol{B}_{ij}$:

$$\begin{aligned} \sum_{i=1}^{m} A_i^s &= \Pr(\text{unit } s \text{ is in the selected set}) \\ &= m/N, \end{aligned} \tag{3.7}$$

$$\begin{aligned} \sum_{i=1}^{m}\sum_{j=1}^{m} B_{ij}^{st} &= \Pr(\text{unit } s \text{ is in set 1 and unit } t \text{ is in set 2}) \\ &= \frac{m^2}{N(N-1)}(1 - \delta_{st}) \end{aligned} \tag{3.8}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m}[B_{ij}^{st} - A_i^s A_j^t] = \frac{m^2}{N^2(N-1)} - \frac{m^2}{N(N-1)}\delta_{st}, \tag{3.9}$$

where δ_{st} is the Kronecker delta symbol.

4. Moments of the RSS estimator

Suppose that mr sets, each of size m, are selected randomly and without replacement from Ω. Let the lowest ranked unit be quantified in each of the first r sets:

$$X_{(1:m)1}, X_{(1:m)2}, X_{(1:m)3}, \ldots, X_{(1:m)r}.$$

In each of the next r sets, the second ranked unit is quantified to yield:

$$X_{(2:m)1}, X_{(2:m)2}, X_{(2:m)3}, \ldots, X_{(2:m)r}.$$

This process continues until the highest ranked unit is quantified in each of the last r sets:

$$X_{(m:m)1}, X_{(m:m)2}, X_{(m:m)3}, \ldots, X_{(m:m)r}.$$

The ranked set estimator of μ is the average of these quantifications:

$$\hat{\mu}_{\mathrm{RSS}} = \frac{1}{rm} \sum_{k=1}^{r} \sum_{i=1}^{m} X_{(i:m)k}. \tag{4.1}$$

Our next result establishes the unbiasedness of the RSS estimator for finite populations.

THEOREM 4.1. *The ranked set estimator of μ is unbiased.*

PROOF. An informal proof simply notes that the average of the $\mu_{(i:m)}$ over $i = 1, 2, \ldots, m$, is μ itself. A formal proof, in the context of finite populations, uses equation (3.7):

$$\begin{aligned}
E[\hat{\mu}_{\mathrm{RSS}}] &= \frac{1}{rm} \sum_{k=1}^{r} \sum_{i=1}^{m} E[X_{(i:m)k}] \\
&= \frac{1}{rm} \sum_{k=1}^{r} \sum_{i=1}^{m} \boldsymbol{A}_i' \boldsymbol{x} \\
&= \frac{1}{rm} \sum_{k=1}^{r} \sum_{i=1}^{m} \sum_{s=1}^{N} A_i^s x_s \\
&= \frac{1}{m} \sum_{s=1}^{N} \left(\sum_{i=1}^{m} A_i^s \right) x_s \\
&= \frac{1}{N} \sum_{s=1}^{N} x_s \\
&= \mu.
\end{aligned}$$

We next obtain the variance of $\hat{\mu}_{\mathrm{RSS}}$. As in equation (3.6), we let C_{ij} denote the covariance between $X_{(i:m)k}$ and $X_{(j:m)\ell}$. From (4.1), it follows that

$$\begin{aligned}
(rm)^2 \operatorname{var}(\hat{\mu}_{\mathrm{RSS}}) &= r\sigma^2_{(1:m)} + r\sigma^2_{(2:m)} + \cdots + r\sigma^2_{(m:m)} \\
&\quad + r(r-1)C_{11} + r^2 C_{12} + \cdots + r^2 C_{1m} \\
&\quad + r^2 C_{21} + r(r-1)C_{22} + \cdots + r^2 C_{2m} \\
&\quad \vdots \\
&\quad + r^2 C_{m1} + r^2 C_{m2} + \cdots + r(r-1)C_{mm} \\
&= r \sum_{i=1}^{m} \sigma^2_{(i:m)} + r^2 \sum_{i=1}^{m} \sum_{j=1}^{m} C_{ij} - r \sum_{i=1}^{m} C_{ii}.
\end{aligned} \tag{4.2}$$

In order to simplify this expression, we need the following result:

THEOREM 4.2. *The sum of the variances of the order statistics is*

$$\sigma^2_{(1:m)} + \sigma^2_{(2:m)} + \cdots + \sigma^2_{(m:m)} = m\sigma^2 - \sum_{i=1}^{m}(\mu_{(i:m)} - \mu)^2, \tag{4.3}$$

while the sum of their covariances is

$$\sum_{i=1}^{m}\sum_{j=1}^{m} C_{ij} = -\frac{m^2}{N-1}\sigma^2. \tag{4.4}$$

PROOF. Using the fact that $\boldsymbol{A}_i$ gives the distribution of $X_{(i:m)}$, we obtain

$$\begin{aligned}\sigma^2_{(i:m)} &= E[(X_{(i:m)} - \mu)^2] - (\mu_{(i:m)} - \mu)^2 \\ &= \sum_{s=1}^{N} A_i^s (x_s - \mu)^2 - (\mu_{(i:m)} - \mu)^2.\end{aligned}$$

The equation (4.3) now follows by summing over i and applying the summation formula (3.7). Equation (4.4) is immediate from equations (3.6) and (3.9).

Putting these results into (4.2), we obtain our first formula for the variance of the ranked set estimator:

$$\operatorname{var}(\hat{\mu}_{\mathrm{RSS}}) = \frac{1}{rm^2}\left\{\frac{m(N-1-mr)}{N-1}\sigma^2 - \sum_{i=1}^{m}(\mu_{(i:m)} - \mu)^2 - \sum_{i=1}^{m} C_{ii}\right\}. \tag{4.5}$$

This is essentially the formula of Takahasi and Futatsuya (1988), except that these authors have not given an explicit expression for C_{ii}. However, an alternative formula proves to be more convenient for actually calculating the variance. We observe that the last term in (4.5) must remain unchanged if the population is centered, i.e., if x_s is replaced by $x_s - \mu$ and $\mu_{(i:m)}$ is replaced by $\mu_{(i:m)} - \mu$. Now

$$\begin{aligned}C_{ii} &= \boldsymbol{x}'(\boldsymbol{B}_{ii} - \boldsymbol{A}_i\boldsymbol{A}_i')\boldsymbol{x} \\ &= \boldsymbol{x}'\boldsymbol{B}_{ii}\boldsymbol{x} - \mu^2_{(i:m)},\end{aligned}$$

which becomes, after centering,

$$C_{ii} = (\boldsymbol{x} - \mu)'\boldsymbol{B}_{ii}(\boldsymbol{x} - \mu) - (\mu_{(i:m)} - \mu)^2.$$

Putting this into (4.5), we see that the terms $\sum(\mu_{(i:m)} - \mu)^2$ cancel giving

$$\operatorname{var}(\hat{\mu}_{\mathrm{RSS}}) = \frac{1}{mr}\left\{\frac{N-1-mr}{N-1}\sigma^2 - \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{x} - \mu)'\boldsymbol{B}_{ii}(\boldsymbol{x} - \mu)\right\}. \tag{4.6}$$

Thus, it is only necessary to calculate the matrices $\boldsymbol{B}_{ii}$, $i = 1, 2, \ldots, m$, in order to obtain $\text{var}(\hat{\mu}_{\text{RSS}})$ for a general finite population. In fact, we note that only the matrix,

$$\sum_{i=1}^{m} \boldsymbol{B}_{ii} \equiv \frac{\boldsymbol{\Gamma}}{\binom{N}{m,m}}, \tag{4.7}$$

is needed. In terms of $\boldsymbol{\Gamma}$, equation (4.6) becomes

$$\begin{aligned} \text{var}(\hat{\mu}_{\text{RSS}}) &= \frac{1}{mr}\left\{\frac{N-1-mr}{N-1}\sigma^2 - \frac{1}{m\binom{N}{m,m}}\gamma\right\} \\ &= \frac{1}{mr}\left\{\frac{N-1-mr}{N-1}\sigma^2 - \frac{m!(m-1)!}{N(N-1)\cdots(N-2m+1)}\gamma\right\} \\ &= \frac{1}{mr}\left\{\frac{N-1-mr}{N-1}\sigma^2 - \bar{\gamma}\right\}, \end{aligned} \tag{4.8}$$

where

$$\gamma = (\boldsymbol{x} - \mu)'\boldsymbol{\Gamma}(\boldsymbol{x} - \mu) \tag{4.9}$$

$$\bar{\gamma} = \frac{m!(m-1)!}{N(N-1)\cdots(N-2m+1)}\gamma. \tag{4.10}$$

In view of Theorem 3.2, the matrix $\boldsymbol{\Gamma}$ is symmetric with zeros on the diagonal. We record the following expressions for $\boldsymbol{\Gamma}$ when $m = 2$ and $m = 3$.

THEOREM 4.3. *Let $m = 2$ and $1 \le s < t \le N$. Then*

$$\Gamma^{st} = \Gamma^{ts} = (N-t)(N-s-2) + (s-1)(t-3).$$

THEOREM 4.4. *Let $m = 3$ and $1 \le s < t \le N$. Then*

$$\begin{aligned} \Gamma^{st} &= \Gamma^{ts} \\ &= \frac{1}{4}(N-t)(N-t-1)(N-s-3)(N-s-4) \\ &\quad + (N-t)(t-4)(N-s-3)(s-1) + (N-t)(s-1)(N-5) \\ &\quad + \frac{1}{4}(s-1)(s-2)(t-4)(t-5). \end{aligned}$$

5. Relative precision of the RSS estimator

We now compare the performance of ranked set sampling with that of simple random sampling. When sampling is without replacement from a finite population, the variance of the SRS estimator is

$$\operatorname{var}(\hat{\mu}_{\mathrm{SRS}}) = \frac{N-mr}{N-1}\frac{\sigma^2}{mr}$$

when there are $n = mr$ quantifications. In conjunction with (4.8), this yields the following expression for the relative precision of RSS:

$$\begin{aligned} \mathrm{RP} &= \frac{\operatorname{var}(\hat{\mu}_{\mathrm{SRS}})}{\operatorname{var}(\hat{\mu}_{\mathrm{RSS}})} \\ &= \frac{1}{1-\dfrac{N-m}{N-mr}\left[\dfrac{1}{N-m}+\dfrac{N-1}{N-m}\dfrac{\bar{\gamma}}{\sigma^2}\right]} \\ &= \frac{1}{1-\dfrac{N-m}{N-mr}\mathrm{RS}^{(1)}}, \end{aligned} \tag{5.1}$$

where

$$\mathrm{RS}^{(1)} = \frac{1}{N-m}+\frac{N-1}{N-m}\frac{\bar{\gamma}}{\sigma^2}. \tag{5.2}$$

Here, it should be noted that, while RP depends on the replication factor r, the quantities γ, $\bar{\gamma}$, and $\mathrm{RS}^{(1)}$ are each independent of r. From the relative precision (5.1), we obtain the relative savings RS as

$$\mathrm{RS} = \frac{N-m}{N-mr}\mathrm{RS}^{(1)} \tag{5.3}$$

$$= \frac{1-f/r}{1-f}\times \mathrm{RS}^{(1)}. \tag{5.4}$$

In particular, putting $r = 1$ in this equation, we see that $\mathrm{RS}^{(1)}$ is the relative savings for a single cycle of the RSS procedure. We also observe that RS is a monotone increasing function of r for a given value of m.

6. Linear and quadratic range

Here we specialize to particular population structures and obtain explicit expressions for $\mathrm{RS}^{(1)}$. The pertinent formulae are equations (4.10) and (5.2) which are repeated here for convenience of the reader:

$$\frac{\bar{\gamma}}{\sigma^2} = \frac{m!(m-1)!}{N(N-1)\cdots(N-2m+1)}\frac{\gamma}{\sigma^2} \tag{6.1}$$

$$\mathrm{RS}^{(1)} = \frac{1}{N-m}+\frac{N-1}{N-m}\frac{\bar{\gamma}}{\sigma^2}. \tag{6.2}$$

For the populations considered, it turns out that γ and σ^2 are polynomials in N. Further, after scaling, the populations converge to continuous populations with finite, nonzero RS as $N \to \infty$. But this implies that

$$\deg_N \gamma = 2m + \deg_N \sigma^2.$$

Indeed, from equation (6.2) the ratio $\bar{\gamma}/\sigma^2$ has a finite, nonzero limit when $N \to \infty$. But then equation (6.1) gives the desired conclusion. For the specific populations considered, we find empirically that γ is divisible by

$$(N+1)N(N-1)(N-2)\cdots(N-2m+1). \tag{6.3}$$

Accordingly, we may write

$$\gamma = a_m(N+1)N(N-1)(N-2)\cdots(N-2m+1)P_m(N), \tag{6.4}$$

where a_m is a constant and $P_m(N)$ is a monic polynomial in N with

$$\deg_N P_m = \deg_N \sigma^2 - 1.$$

Equations (6.1) and (6.2) now imply that

$$\begin{aligned}\lim_{N\to\infty} \mathrm{RS}^{(1)} &= \lim_{N\to\infty} \frac{\bar{\gamma}}{\sigma^2} \\ &= m!(m-1)!a_m \lim_{N\to\infty} \frac{(N+1)P_m(N)}{\sigma^2}.\end{aligned}$$

The constant a_m can be found from

$$m!(m-1)!a_m = a_{\sigma^2} \lim_{N\to\infty} \mathrm{RS}^{(1)},$$

where a_{σ^2} is the leading coefficient of σ^2, considered as a polynomial in N. But the quantity,

$$\lim_{N\to\infty} \mathrm{RS}^{(1)}, \tag{6.5}$$

can be computed directly from the limiting (continuous) population using equation (1.1). Thus, we finally obtain

$$\begin{aligned}\frac{\bar{\gamma}}{\sigma^2} &= m!(m-1)!a_m \frac{(N+1)P_m(N)}{\sigma^2} \\ &= \left(\lim_{N\to\infty} \mathrm{RS}^{(1)}\right) \frac{(N+1)P_m(N)}{\sigma^2/a_{\sigma^2}},\end{aligned} \tag{6.6}$$

and all that remains is the determination of the polynomial $P_m(N)$, which will be accomplished numerically using the GAUSS program.

6.1 *Linear range*

The population is defined by $x_s = s$ for $s = 1, 2, \ldots, N$. The population variance is $\sigma^2 = (N^2 - 1)/12$ and the limit (6.5) is $(m-1)/(m+1)$ since the population converges to the uniform distribution for large N. It is clear from (4.9) that γ is a polynomial function of N. Numerical work using the GAUSS program showed that this polynomial was divisible by the quantity (6.3) for $m \leq 12$. Larger values of m have not been studied numerically and this factorization of γ has not been established analytically. Since σ^2 is a quadratic in N, the polynomial $P_m(N)$ is linear of form $N - C_m$ for suitable constants C_m. Equation (6.6) now gives

$$\frac{\bar{\gamma}}{\sigma^2} = \frac{m-1}{m+1}\frac{(N+1)(N-C_m)}{N^2-1} = \frac{m-1}{m+1}\frac{N-C_m}{N-1}, \tag{6.7}$$

and

$$\mathrm{RS}^{(1)} = \frac{1}{N-m}\left(1 + \frac{m-1}{m+1}(N - C_m)\right). \tag{6.8}$$

Using the GAUSS program, the first few values of C_m are determined to be $C_m = 11/5,\ 44/35,\ 61/63,\ 129/154$ for $m = 2, 3, 4, 5$, respectively. Table 1 provides numerical computations for $m = 2, 3, 4, 5$, $r = 1, 2, 3, 4$ and $N = 4(1)9, 12, 16(2)20, 25, 27, 32, 36, 45, 48, 50, 64, 75, 80, 100, 125, 150, 200$ and ∞.

6.2 *Quadratic range*

The population is defined by $x_s = s^2$ for $s = 1, 2, \ldots, N$ and the population variance is

$$\sigma^2 = (N^2 - 1)(2N + 1)(8N + 11)/180.$$

As $N \to \infty$, the population converges (after scaling) to a continuous distribution with pdf given by $1/(2\sqrt{x})$, $0 < x < 1$. For this distribution, the limit (6.5) is found to be

$$\frac{1}{4}\frac{(m-1)(4m+7)}{(m+1)(m+2)}.$$

Again, it is clear from (4.9) that γ is a polynomial function of N and numerical work using the GAUSS program showed that this polynomial was divisible by the quantity (6.3) for $m \leq 12$. Larger values of m have not been studied numerically and this factorization of γ has not been established analytically. Since σ^2 is of the fourth degree in N, the polynomial $P_m(N)$ is a cubic. Equation (6.6) now gives

$$\frac{\bar{\gamma}}{\sigma^2} = 4\frac{(m-1)(4m+7)}{(m+1)(m+2)}\frac{(N+1)P_m(N)}{(N^2-1)(2N+1)(8N+11)} = 4\frac{(m-1)(4m+7)}{(m+1)(m+2)}\frac{P_m(N)}{(N-1)(2N+1)(8N+11)}, \tag{6.9}$$

and

$$\mathrm{RS}^{(1)} = \frac{1}{N-m}\left(1 + 4\frac{(m-1)(4m+7)}{(m+1)(m+2)}\frac{P_m(N)}{(2N+1)(8N+11)}\right). \tag{6.10}$$

Table 1. Relative savings (in percent) of RSS compared with SRS under a linear range when the set size, $m = 2(1)5$, the number of cycles, $r = 1(1)4$, and the population size, $N \geq m^2 r$.

Population Size (N)	Relative Savings															
	Set Size (m)															
	2				3				4				5			
	Cycles (r)				Cycles (r)				Cycles (r)				Cycles (r)			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
4	80															
5	64															
6	57															
7	52															
8	49	73														
9	47	65			81											
12	43	53	71		71											
16	40	47	56	70	64				83							
18	39	45	52	63	62	78			80							
20	38	43	49	58	61	74			78							
25	37	41	45	51	58	68			73				85			
27	37	40	44	49	58	66	77		72				84			
32	36	39	42	46	56	63	71		70	82			81			
36	36	38	41	44	56	61	68	76	69	79			79			
45	35	37	39	41	54	59	63	69	67	74			76			
48	35	37	39	41	54	58	62	68	66	73	81		75			
50	35	37	38	40	54	58	62	67	66	72	80		75	84		
64	35	36	37	39	53	56	59	62	65	69	75	81	73	80		
75	35	36	37	38	53	55	57	60	64	68	72	77	72	78	84	
80	34	35	36	37	52	55	57	59	64	67	71	76	72	77	83	
100	34	35	36	36	52	54	55	57	63	66	69	72	71	75	79	84
125	34	35	35	36	51	53	54	56	62	64	67	69	70	73	76	80
150	34	34	35	35	51	52	53	55	62	64	65	67	69	72	74	77
200	34	34	34	35	51	52	52	53	61	63	64	65	69	70	72	74
Inf.	33	33	33	33	50	50	50	50	60	60	60	60	67	67	67	67

Using the GAUSS program, the first few $P_m(N)$ were determined to be as follows:

$$P_2(N) = (35N^3 - 19N^2 - 137N - 65)/35$$
$$P_3(N) = (532N^3 + 278N^2 - 1003N - 572)/532$$
$$P_4(N) = (3542N^3 + 3048N^2 - 4373N - 2887)/3542$$
$$P_5(N) = (10296N^3 + 10484N^2 - 9564N - 7101)/10296.$$

Table 2 provides numerical computations of the relative savings for $m = 2, 3, 4, 5$, $r = 1, 2, 3, 4$ and values of N considered earlier for Table 1.

Table 2. Relative savings (in percent) of RSS compared with RSS under a quadratic range when the set size, $m = 2(1)5$, the number of cycles, $r = 1(1)4$, and the population size, $N \geq m^2 r$.

Population Size (N)	Relative Savings															
	Set Size (m)															
	2				3				4				5			
	Cycles (r)				Cycles (r)				Cycles (r)				Cycles (r)			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
4	74															
5	60															
6	53															
7	49															
8	46	68														
9	44	61			77											
12	40	50	66		67											
16	37	44	52	65	61				80							
18	37	42	49	59	59	74			77							
20	36	41	46	54	58	70			74							
25	35	38	42	47	56	64			70				83			
27	35	38	41	46	55	63	73		69				81			
32	34	37	39	43	54	60	68		67	78			78			
36	34	36	38	41	53	58	65	73	66	75			76			
45	33	35	37	39	52	56	60	66	64	71			73			
48	33	35	36	38	51	55	59	64	64	70	78		73			
50	33	34	36	38	51	55	59	63	63	69	77		72	81		
64	33	34	35	36	50	53	56	59	62	66	72	77	70	77		
75	32	33	34	35	50	52	54	57	61	65	69	74	69	75	81	
80	32	33	34	35	50	52	54	56	61	64	68	72	69	74	80	
100	32	33	33	34	49	51	53	54	60	63	66	69	68	72	76	81
125	32	32	33	34	49	50	51	53	60	62	64	66	67	70	73	77
150	32	32	33	33	49	50	51	52	59	61	63	65	67	69	72	74
200	32	32	32	33	48	49	50	51	59	60	61	63	66	68	70	72
Inf.	31	31	31	31	47	47	47	47	57	57	57	57	64	64	64	64

6.3 *Comparisons of relative savings*

Tables 1 and 2 show the relative savings due to RSS compared with SRSWOR for finite populations with linear and quadratic range when $N \geq m^2 r$, i.e. when the sampling fraction $\frac{n}{N} = \frac{mr}{N} \leq \frac{1}{m}$. Note that the relative savings increase with the set size for both populations. However, for a given set size, the number of cycles, and the population size, the relative savings seem to be slightly higher for linear range over quadratic range. We also observe that the last row of Table 1 corresponding to $N \to \infty$ matches with the values obtained by Dell and Clutter (1972) for the continuous uniform distribution as expected. Both tables suggest

an optimal choice of m for fixed $n = mr$, and it is in its maximal value satisfying $n = mr$, where r is to be a positive integer. Finally, the following observation may be of some practical significance. Relative-savings-wise, near-uniform finite populations of size $N \geq 25$ are pretty close to their uniform continuous population counterpart, whereas for right-skew finite populations, it is $N \geq 50$.

REFERENCES

Dell, T. R. and Clutter, J. L. (1972). Ranked set sampling theory with order statistics background, *Biometrics*, **28**, 545–553,

Halls, L. S. and Dell, T. R. (1966). Trial of ranked set sampling for forage yields, *Forest Science*, **12**, 22–26,

McIntyre, G. A. (1952). A method for unbiased selective sampling using ranked sets, *Australian Journal of Agricultural Research*, **3**, 385–390,

Patil, G. P., Sinha, A. K. and Taillie, C. (1993). Ranked set sampling from a finite population in the presence of a trend on a site, *Journal of Applied Statistical Science*, **1**(1), 51–65.

Patil, G. P., Sinha, A. K. and Taillie, C. (1994). Ranked set sampling, *Handbook of Statistics, Vol. 12: Environmental Statistics* (eds. G. P. Patil and C. R. Rao), 167–200, North-Holland, Elsevier, New York.

Takahasi, K. and Futatsuya, M. (1988). Ranked set sampling from a finite population, *Proc. Inst. Statist. Math.*, **36**, 55–68 (Japanese with English summary).

Takahasi, K. and Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering, *Ann. Inst. Statist. Math.*, **20**, 1–31.