Sankhyā: The Indian Journal of Statistics
1997, Volume 59, Series B, Pt. 1, pp. 66-75

ON THE ESTIMATION OF SIZE OF A FINITE POPULATION

By S. SENGUPTA
Calcutta University, Calcutta
and
MOLOY DE*
Indian Statistical Institute, Calcutta

SUMMARY. The problem considered is that of unbiased estimation of the size of a finite population under capture-mark-release-recapture (CMRR) sampling procedure. The existing results are supplemented with various other results and the CMRR procedure is compared with the Negative Binomial and the Negative Hypergeometric sampling schemes in terms of the ASN and the variance of the UMVUE of the population size.

1. Introduction

The problem of estimating the population size N of a finite population is known to be of great importance. Well known problems of this kind are the estimation of the total number of fish in a lake, the estimation of the total number of wild animals in a forest etc.. Several authors had already considered the problem in the past and had suggested different methods of sampling with associated estimation procedures (see Boswell et al (1988), Seber(1982) and the references therein). The basic procedure is to initially catch, mark and release k population units into the target population and then to recatch units randomly from the population in one or more samples.

For unbiased estimation of N, a simple procedure (to be called procedure I) is to recatch and release units one by one until $m(\leq k)$ of the k initially marked units are recaptured. If S_m denotes the number of trials required, then S_m follows a Negative Binomial distribution with success probability $\frac{k}{N}$ and the

Paper received. December 1996; revised April 1997.
AMS (1980) *subject classification.* Primary 62D05; secondary 62L12
Keywords and phrases. CMRR sampling,Comparison of ASN, Comparison of variance, Estimation of population size, Negative Binomial sampling,Negative Hypergeometric Sampling, UMVU Estimators,Variance estimator.
*Work is supported by a Senior Research Fellowship of the Council of Scientific and Industrial Research.

uniformly minimum variance unbiased estimator (UMVUE) of N is obtained from the well known results on Negative Binomial distribution (see Johnson and Kotz (1969), page 126) as $\hat{N}_I = \frac{k\,S_m}{m}$ with variance $V(\hat{N}_I) = \frac{N(N-k)}{m}$. Also the expected number of trials for the procedure is $ASN(I) = E(S_m) = \frac{mN}{k}$.

If units are sampled one by one without being released into the population until $m(\leq k)$ of the k initially marked units are recaptured (to be called procedure II), then S_m, the number of trials required, follows a Negative Hypergeometric distribution and in this case the UMVUE of N is given by $\hat{N}_{II} = \frac{(k+1)S_m}{m} - 1$ with

$$V(\hat{N}_{II}) = \frac{(N+1)(N-k)(k+1-m)}{m(k+2)} \quad \text{and} \quad ASN(II) = E(S_m) = \frac{m(N+1)}{k+1}$$

(see Johnson and Kotz (1969),page 157).

A simple modification of the procedure I (to be called procedure III) is also suggested in the literature as follows: initially k population units are marked and released into the target population and then units are sampled at random, marked and released one by one until m marked units are recaptured. The procedure is a special case of a more general procedure suggested in Goodman(1953) and is termed as capture-mark-release-recapture (CMRR) sampling scheme. Using more general methods, Goodman(1953) obtained the UMVUE of N for this procedure as the quotient of two determinants and gave some simplified expressions for $k = 1$. Darroch (1958) had shown that Goodman's estimator, for $k = 1$, can also be expressed as the ratio of two Stirling numbers or differences of zero. Hossain (1995) had considered the special case of $k = m = 1$ in which case the UMVUE of N is $\binom{S_1+1}{2}$, where S_1 is the number of trials required.

The purpose of this paper is to supplement these studies with various other results and to compare procedure III with procedures I and II in terms of the ASN and the variance of the UMVUE of N. It is demonstrated in section 4 that the procedure III is always better than the procedure I and also appears to be better than the procedure II when N is considerably large. The supplementary results are discussed in sections 2 and 3.

2. Estimaton of Population Size under Procedure III

Let us consider the procedure III and let, for $j = 1, \ldots, m$, S_j denote the number of trials required to recapture j marked units and $S_j^* = S_j - j$. It is easy to verify that $P[S_1^* = s_1^*, \cdots, S_m^* = s_m^*] =$

$$(1 - \frac{k}{N})(1 - \frac{k+1}{N}) \cdots (1 - \frac{s_m^* + k - 1}{N}) \frac{s_1^* + k}{N} \cdots \frac{s_m^* + k}{N},$$

$$0 \leq s_1^* \leq \cdots \leq s_m^* \leq N - k, \qquad \ldots(2.1)$$

whence the probability distribution of S_m^* is obtained as

$$P[S_m^* = s_m^*] = \binom{N-k}{s_m^*} \frac{s_m^*+k}{N^{m+s_m^*}} g_{m-1}(k, s_m^*+k), 0 \le s_m^* \le N-k, \quad \ldots(2.2)$$

where, for non-negative integers $a, b(\ge a), c$, $g_c(a,b)$ is defined as

$$g_c(a,b) = (b-a)! \sum \cdots \sum_{a \le i_1 \le \cdots \le i_c \le b} i_1 \cdots i_c \quad \ldots(2.3)$$

with $g_0(a,b) = (b-a)!$. The function $g_c(a,b)$ can also be expressed as

$$g_c(a,b) = \Delta^{b-a} a^{b-a+c} = \Delta^{b-a} x^{b-a+c}|_{x=a}$$

$$= b^{b-a+c} - \binom{b-a}{1}(b-1)^{b-a+c}$$

$$+ \binom{b-a}{2}(b-2)^{b-a+c} - \cdots + (-1)^{b-a} a^{b-a+c} \quad \ldots(2.4)$$

This may be proved by induction on b for fixed a noting that (2.4) is trivially true for $b = a$ whatever be c.

A useful identity follows from (2.2) viz.

$$\sum_{s_m^*=0}^{N-k} \binom{N-k}{s_m^*} \frac{s_m^*+k}{N^{s_m^*}} g_{m-1}(k, s_m^*+k) = N^m \quad \ldots(2.5)$$

From (2.1) and (2.2) it can also be seen by standard methods that S_m or equivalently S_m^* is a complete sufficient statistic for the parameter space $\{N \ge k\}$. It follows, therefore, that the unbiased estimator of N based on S_m^* is also the UMVUE which is obtained in the following theorem.

THEOREM 2.1. *The UMVUE of N for the suggested procedure is given by* $\hat{N}_{III} = \dfrac{g_m(k, S_m^*+k)}{g_{m-1}(k, S_m^*+k)} = \dfrac{\Delta^{S_m^*} k^{S_m^*+m}}{\Delta^{S_m^*} k^{S_m^*+m-1}}$.

PROOF. It is enough to prove that $\hat{N}_{III}$ is an unbiased estimator of N and this follows immediately from (2.2) and (2.5). □

In particular when $k = 1$, the UMVUE of N, as obtained in Goodman(1953) and Darroch(1958), is

$$\frac{g_m(1, S_m^*+1)}{g_{m-1}(1, S_m^*+1)} = \frac{\Delta^{S_m^*} 1^{S_m^*+m}}{\Delta^{S_m^*} 1^{S_m^*+m-1}} = \frac{\Delta^{S_m^*+1} O^{S_m^*+m+1}}{\Delta^{S_m^*+1} O^{S_m^*+m}}$$

using the identity $\frac{1}{\nu}\Delta^{\nu}O^{n} = \Delta^{\nu-1}1^{n-1}$. If, further, $m = 1$, the UMVUE of N is

$$\frac{g_1(1, S_1^* + 1)}{g_0(1, S_1^* + 1)} = \sum_{i=1}^{S_1^*+1} i = \binom{S_1^* + 2}{2} = \binom{S_1 + 2}{2},$$

as given in Hossain(1995).

It follows similarly that the UMVUE of N^2 is $\frac{g_{m+1}(k, S_m^* + k)}{g_{m-1}(k, S_m^* + k)}$ and, hence, the UMVUE of $V(\hat{N}_{III})$ is

$$\begin{aligned}
\hat{V}(\hat{N}_{III}) &= \hat{N}_{III}^2 - \frac{g_{m+1}(k, S_m^* + k)}{g_{m-1}(k, S_m^* + k)} \\
&= \frac{g_m^2(k, S_m^* + k) - g_{m-1}(k, S_m^* + k) g_{m+1}(k, S_m^* + k)}{g_{m-1}^2(k, S_m^* + k)} \\
&= \frac{(\Delta^{S_m^*} k^{S_m^*+m})^2 - \Delta^{S_m^*} k^{S_m^*+m-1} \Delta^{S_m^*} k^{S_m^*+m+1}}{(\Delta^{S_m^*} k^{S_m^*+m-1})^2}.
\end{aligned}$$

It can be proved by induction on b for fixed a that

$$g_c^2(a, b) \geq g_{c-1}(a, b)\, g_{c+1}(a, b) \qquad \ldots(2.6)$$

noting that (2.6) is trivially true for $b = a$ whatever be c (see Hardy *et al.* (1952), page 52). The inequality (2.6) ensures that $\hat{V}(\hat{N}_{III})$ is uniformly non-negative.

3. The Special Case of $m = 1$

In particular when m=1, the probability distribution of S_1^* simplifies to

$$P[S_1^* = s_1^*] = (1 - \frac{k}{N})(1 - \frac{k+1}{N}) \cdots (1 - \frac{s_1^* + k - 1}{N}) \frac{s_1^* + k}{N},$$

$$0 \leq s_1^* \leq N - k \qquad \ldots(3.1)$$

and the estimators reduce to

$$\begin{aligned}
\hat{N}_{III} &= \sum_{k}^{S_1^*+k} i = \binom{S_1^*+k+1}{2} - \binom{k}{2}, \\
\hat{V}(\hat{N}_{III}) &= (\sum_{k}^{S_1^*+k} i)^2 - \sum \sum_{k \le i_1 \le i_2 \le S_1^*+k} i_1 i_2 \\
&= \sum \sum_{k \le i_1 \le i_2 \le S_1^*+k} i_1 i_2 - \sum_{k}^{S_1^*+k} i^2 \\
&= \sum_{k}^{S_1^*+k} i\{\binom{i+1}{2} - \binom{k}{2}\} - \sum_{k}^{S_1^*+k} i^2 \\
&= \sum_{k}^{S_1^*+k} i\{\binom{i}{2} - \binom{k}{2}\} \\
&= \frac{(S_1^*+k-1)(S_1^*+k)(S_1^*+k+1)(3S_1^*+3k+2)}{24} \\
&\quad - \binom{k}{2}\binom{S_1^*+k+1}{2} + \frac{(k-1)k(k+1)(3k-2)}{24}.
\end{aligned}$$

For $m = 1$, we derive a closed expression for $V(\hat{N}_{III})$ in theorem 3.1 stated below. The derivation is based on the following lemmas.

LEMMA 3.1. *For every i=0,1,...,N-k,*

$$\sum_{s_j^*=i}^{N-k} P[S_1^* = s_1^*] = \frac{N}{i+k} P[S_1^* = i].$$

PROOF. It is easy to verify the result from (3.1). □

LEMMA 3.2. *For any given f,* $E \sum_{i=k}^{S_1^*+k} i\, f(i) = N E\, f(S_1^* + k)$.

PROOF.

$$\begin{aligned} E\sum_{i=k}^{S_1^*+k} if(i) &= \sum_{s_1^*=0}^{N-k} P[S_1^* = s_1^*] \sum_{i=k}^{s_1^*+k} if(i) \\ &= \sum_{i=0}^{N-k} (i+k)f(i+k) \sum_{s_1^*=i}^{N-k} P[S_1^* = s_1^*] \\ &= N \sum_{i=0}^{N-k} f(i+k)P[S_1^* = i] \quad \text{(by lemma 3.1)} \\ &= N\ E\ f(S_1^* + k). \end{aligned}$$

□

LEMMA 3.3.

$$\begin{aligned} E(S_1^* + k) &= N\ E(\frac{1}{S_1^* + k}) + (k-1) \\ &= \frac{(N-k)!}{N^{N-k}}\{1 + \frac{N}{1!} + \frac{N^2}{2!} + \cdots + \frac{N^{N-k}}{(N-k)!}\} + (k-1). \end{aligned}$$

PROOF. By lemma 3.2,

$$E(S_1^* + k) = E \sum_{i=k}^{S_1^*+k} \binom{i}{i} + (k \quad 1) = N\ E(\frac{1}{S_1^* + k}) + (k-1)$$

Also, by (3.1),

$$\begin{aligned} N\ E(\tfrac{1}{S_1^*+k}) &= \sum_{s_1^*=0}^{N-k} (1 - \frac{k}{N}) \cdots (1 - \frac{s_1^* + k - 1}{N}) \\ &= \frac{(N-k)!}{N^{N-k}}\{1 + \frac{N}{1!} + \cdots + \frac{N^{N-k}}{(N-k)!}\} \end{aligned}$$

Hence, follows the proof. □

THEOREM 3.1. *For m=1,* $V(\hat{N}_{III}) = N^2 - N\ E(S_1^* + k)$

$$= N^2 - \frac{(N-k)!}{N^{N-k-1}}\{1 + \frac{N}{1!} + \cdots + \frac{N^{N-k}}{(N-k)!}\} - N(k-1).$$

PROOF. For $m = 1$, $V(\hat{N}_{III}) = E\hat{V}(\hat{N}_{III}) = E\sum_{i=k}^{S_1^*+k} i\{\binom{i}{2} - \binom{k}{2}\}$

$$\begin{aligned} &= N\ E[\binom{S_1^* + k}{2} - \binom{k}{2}], \text{ by lemma 3.2} \\ &= N\ E(\hat{N}_{III}) - N\ E(S_1^* + k) = N^2 - N\ E(S_1^* + k), \end{aligned}$$

whence the theorem follows by lemma 3.3 . □

4. Comparisons with Negative Binomial and Negative Hypergeometric Sampling Schemes

We study below the performance of the procedure III in terms of the ASN i.e. $E(S_m)$ and the variance of the estimator of N as compared to the procedures I and II.

We first consider the special case of $m = 1$ and prove the following theorem.

THEOREM 4.1. *For* $m = 1$,

(a) *ASN(III)≤ASN(II)≤ASN(I)*,
(b) $V(\hat{N}_{II}) \leq V(\hat{N}_{III}) \leq V(\hat{N}_I)$.

PROOF. (a) By (3.1) and lemma 3.1 and 3.3,

$$\begin{aligned} E(S_1) &= E(S_1^* + k) - (k-1) = N\, E\left(\frac{1}{S_1^* + k}\right) \\ &\leq N\Big[k^{-1}P(S_1^* = 0) + (k+1)^{-1}\sum_{s_1^*=1}^{N-k} P(S_1^* = s_1^*)\Big] \\ &= N[k^{-1}P(S_1^* = 0) + N(k+1)^{-2}P(S_1^* = 1)] = \frac{N+1}{k+1}, \end{aligned}$$

with equality iff $N = k$ or $k + 1$. This proves that, for $m = 1$, ASN(III) $\leq$ ASN(II). Also by simple comparison of ASN(I) and ASN(II), it follows that ASN(II) $\leq$ ASN(I) with equality iff N=k. Hence follows the proof of (a). □

PROOF. (b) By theorem 3.1, we have, for $m = 1$, $V(\hat{N}_{III}) = N^2 - N\, E(S_1^* + k)$ which is clearly $\leq N(N-k)$ with equality iff $N = k$. This proves that, for $m = 1$, $V(\hat{N}_{III}) \leq V(\hat{N}_I)$.

Also by (a), $E(S_1^* + k) = E(S_1) + (k-1) \leq \frac{N+1}{k+1} + (k-1)$ implying that, for $m = 1$, $V(\hat{N}_{III}) \geq \frac{Nk(N-k)}{k+1}$. Hence, for $m = 1$,

$$V(\hat{N}_{III}) - V(\hat{N}_{II}) = V(\hat{N}_{III}) - \frac{(N+1)k(N-k)}{k+2} \geq \frac{k(N-k)(N-k-1)}{(k+1)(k+2)} \geq 0$$

with equality iff $N = k$ or $k + 1$. This proves the theorem. □

Thus, for $m = 1$, the procedure III is generally better than the procedure I both in terms of ASN and the variance of the estimator of N. However, $V(\hat{N}_{III})$ is generally larger than $V(\hat{N}_{II})$, although ASN(III) is smaller than ASN(II). In table 4.1 given below we try to compare numerically the procedures II and III

for some selected values of k in terms of $ASN \times V(\hat{N})$ which is the inverse of efficiency per unit sample.

Table 4.1 ASN $\times V(\hat{N})$ FOR PROCEDURES II AND III FOR $m = 1$

k	N	Procedure II	Procedure III
1	20	1396.5000	1556.9930
	25	2704.0000	2834.6630
	30	4644.8330	4607.6520
	35	7344.0000	6931.8530
	40	10926.5000	9858.6880
2	10	161.3333	178.6543
	15	554.6667	581.3116
	20	1323.0000	1308.9050
	25	2591.3330	2430.1390
	30	4484.6670	4005.8040
5	10	72.0238	74.9496
	15	304.7619	313.0205
	20	787.5000	786.0479
	25	1609.5240	1556.6080
	30	2860.1190	2680.9850

The computational study indicates that for N considerably large the procedure III is better than procedure II, but for smaller values of N the procedure II is better. We now consider the case of general $m(\leq k)$ and extend partially Theorem 4.1.

THEOREM 4.2. *For* $m \leq k$,

(a) $ASN(III) \leq ASN(II) \leq ASN(I)$,
(b) $V(\hat{N}_{III}) \leq V(\hat{N}_I)$.

PROOF. (a). For the procedure III,

$$E(S_m) = E(S_1) + E(S_2 - S_1) + \cdots + E(S_m - S_{m-1}).$$

Now, for $j = 2, \ldots, m$, the conditional probability distribution of $S_j - S_{j-1}$ given S_{j-1} is same as the unconditional probability distribution of S_1 with k replaced by $S^*_{j-1} + k$ (see Section 3). Hence by part (a) of theorem 4.1,

$$E(S_j - S_{j-1}) = EE(S_j - S_{j-1}|S_{j-1}) \leq E\left(\frac{N+1}{k + S^*_{j-1} + 1}\right) \leq \frac{N+1}{k+1}$$

for every $j = 1, \ldots, m$ with $S_0 = 0$. This implies that ASN(III) $\leq \frac{(N+1)m}{k+1}$ =ASN(II). Also, since ASN(II) $\leq$ ASN(I), the part (a) of the theorem follows. □

PROOF. (b) We first prove that $V(\hat{N}^*_{III}) \leq V(\hat{N}_I)$, where $\hat{N}^*_{III}$ is an unbiased estimator of N defined as $\hat{N}^*_{III} = \frac{1}{m}\sum_{j=1}^{m} \hat{N}(j)$, where, for $j =$

$1, \cdots, m$, $\hat{N}(j) = \sum_{S^*_{j-1}+k}^{S^*_j+k} i$ with $S_0 = 0$. Since the conditional probability distribution of $S^*_j - S^*_{j-1}$ given S^*_{j-1} is of the form (3.1) with k replaced by $S^*_{j-1}+k$, as for $m = 1$, $\hat{N}(j)$ is a conditionally unbiased and, hence, an unconditionally unbiased estimator of N for each j which implies the unbiasedness of $\hat{N}^*_{III}$. It follows by similar arguments that $\hat{N}(j), j = 1, \cdots, m$ are uncorrelated so that

$$V(\hat{N}^*_{III}) = \frac{1}{m^2} \sum_{j=1}^{m} V(\hat{N}(j)), \text{ where, for } j = 1, \cdots, m,$$

$$V(\hat{N}(j)) = E\, V(\hat{N}(j)|S^*_{j-1}) = N^2 - N\, E(S^*_j + k),$$

$V(\hat{N}(j)|S^*_{j-1})$ being given by Theorem 3.1 with S^*_1 replaced by $S^*_j - S^*_{j-1}$ and k replaced by $S^*_{j-1} + k$.

Table 4.2. $V(\hat{N})$ and ASN$\times V(\hat{N})$ FOR PROCEDURES II AND III FOR $m > 1$

m	k	N	$V(\hat{N}_{II})$	$V(\hat{N}_{III})$	ASN(II) $\times V(\hat{N}_{II})$	ASN(III) $\times V(\hat{N}_{III})$
2	3	30	167.4000	323.6262	2594.7000	2786.8200
		35	230.4000	453.4350	4147.2000	4269.7000
		40	303.4000	605.7967	6219.7000	6159.6600
		45	386.4000	780.8748	8887.2000	8492.2370
		50	479.4000	978.8585	12224.7000	11301.5200
2	5	20	90.0000	121.7214	630.0000	673.1599
		25	148.5714	205.5523	1287.6190	1321.7240
		30	221.4286	311.6173	2288.0950	2260.9680
		35	308.5714	440.1366	3702.8570	3531.1440
		40	410.0000	591.2849	5603.3330	5169.4930
3	7	30	132.0370	188.2415	1534.9310	1633.0330
		35	186.6667	269.7184	2520.0000	2594.6260
		40	250.5556	366.0504	3852.2920	3846.2890
		45	323.7037	477.3474	5583.8890	5416.8470
		50	406.1111	603.7205	7766.8750	7333.7120
4	8	45	212.7500	340.7283	4349.5560	4578.4400
		50	267.7500	432.6468	6069.0000	6218.9910
		55	329.0000	534.4419	8188.4440	8160.7680
		60	396.5000	645.5012	10749.5600	10412.5200

Now $V(\hat{N}(j)) \leq N^2 - N\, E(S^*_1 + k) = V(\hat{N}(1))$ for every $j = 1, ..., m$ and consequently, by part (b) of theorem 4.1, $V(\hat{N}^*_{III}) \leq \frac{V(\hat{N}(1))}{m} \leq \frac{N(N-k)}{m} = V(\hat{N}_I)$. Since $\hat{N}_{III}$ is the UMVUE of N for procedure III, the part (b) of the theorem follows. □

We also believe that $V(\hat{N}_{II}) \leq V(\hat{N}_{III})$, although a completely satisfactory proof of this claim, for $m > 1$, has so far eluded us. In table 4.2 we present some numerical computations in this regard for some selected values of m and k. In the same table we also compute the values of ASN$\times V(\hat{N})$ for procedures II and

III. The computational study again indicates that in terms of $\text{ASN} \times V(\hat{N})$ the procedure III is better than the procedure II when N is considerably large.

ACKNOWLEDGEMENT. The authors are highly thankful to Professor Bikas K Sinha of the Indian Statistical Institute, Calcutta for suggesting the problem and for his active interest throughout this investigation.

References

BOSWELL, M. T., BURNHAM, K.P. and PATIL, G.P. (1988). Role and use of composite sampling and capture-recapture sampling in ecological studies. In *Handbook of Statistics*, Vol.6, Eds.P. R. Krishnaiah and C. R. Rao, North Holland, Amsterdam, 469-488.

DARROCH, J.N. (1958). The multiple-recapture census: I Estimation of a closed population. *Biometrika*, **45**, 343-359.

GOODMAN, L.A. (1953). Sequential sampling tagging for population size problems. *Ann. Math. Statist.*, **24**, 56-69.

HARDY, G.H., LITTLEWOOD, J. and POLYA, G. (1952). *Inequalities*, Cambridge University Press, Cambridge.

HOSSAIN, M.F. (1995). Unknown population size estimation: an urn model approach. *Journal of Statistical Studies*, **15**, 89-94.

JOHNSON, N.L. and KOTZ, S. (1969). *Distribution in Statistics: Discrete Distributions*, John Wiley, New York.

SEBER, G.A.F. (1982). *The Estimation of Animal Abundance and Related Parameters*, 2nd edition, Macmillan, New York.

DEPARTMENT OF STATISTICS
CALCUTTA UNIVERSITY
35, BALLYGUNGE CIRCULAR ROAD
CALCUTTA-700 019
INDIA.

STAT-MATH UNIT
INDIAN STATISTICAL INSTITUTE
203, B. T. ROAD
CALCUTTA-700 035
INDIA.