APPLICATIONES MATHEMATICAE
37,2 (2010), pp. 237–245

S. Sengupta (Kolkata)

# ESTIMATION OF THE SIZE OF A CLOSED POPULATION

*Abstract.* The problem considered is that of estimation of the size ($N$) of a closed population under three sampling schemes admitting unbiased estimation of $N$. It is proved that for each of these schemes, the uniformly minimum variance unbiased estimator (UMVUE) of $N$ is inadmissible under square error loss function. For the first scheme, the UMVUE is also the maximum likelihood estimator (MLE) of $N$. For the second scheme and a special case of the third, it is shown respectively that an MLE and an estimator which differs from an MLE by at most one have uniformly smaller mean square errors than the respective UMVUE's.

**1. Introduction.** Capture-recapture sampling is widely used in ecology and wild life studies to estimate the size ($N$) of a closed population and associated demographic parameters such as survival rates. The basic procedure is to initially catch, mark and release a sample of units into the target population and then, assuming a thorough mixing of the marked units with the population, to recatch marked and unmarked units randomly from the population in one or more subsequent samples. Capture rates and marked to unmarked ratios are the basis for parametric inference. We refer to Boswell et al. (1988) and Seber (1982) for a comprehensive review of such sampling and associated estimation procedures.

For unbiased estimation of $N$, a simple procedure (to be called *procedure* 1) is to recatch and release units one by one until $m\,(\leq k)$ of $k$ initially marked units are recaptured. If $S_m$ is the number of trials required, then $S_m$ follows a negative binomial distribution with success probability $k/N$ and the uniformly minimum variance unbiased estimator (UMVUE) of $N$ is

2010 *Mathematics Subject Classification*: 62D05, 62F10.
*Key words and phrases*: capture-recapture sampling, inadmissibility, estimation of population size, maximum likelihood estimator, uniformly minimum variance unbiased estimator.

DOI: 10.4064/am37-2-8

obtained from the well known results on negative binomial distribution (see Johnson and Kotz, 1969, p. 126) as $\hat{N}_1^* = (k/m)S_m$.

If units are sampled one by one without being replaced into the population until $m\,(\leq k)$ of $k$ initially marked units are recaptured (to be called *procedure* 2), then $S_m$, the number of trials required, follows a negative hypergeometric distribution and in this case the UMVUE of $N$ is given by

$$\hat{N}_2^* = \frac{k+1}{m}\,S_m - 1$$

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

Another modification of the procedure 1 (to be called *procedure* 3) suggested in the literature is as follows. Initially $k$ population units are marked and released into the target population and then units are sampled at random and released one by one after marking an unmarked sampled unit until $m$ marked units are recaptured. The procedure is a special case of a more general procedure suggested in Goodman (1953) and is termed *capture-mark-release-recapture* (CMRR) sampling scheme. Uniformly minimum variance unbiased estimation of $N$ for CMRR sampling had been discussed in Goodman (1953), Darroch (1958) and Sengupta and De (1997) and maximum likelihood estimation had been studied in Darroch (1958) and Samuel (1968) among others.

In this paper we consider the problem of estimation of $N$ under the above three sampling schemes for the special case of $m = 1$. We prove that for each of these procedures, the UMVUE of $N$ is inadmissible under square error loss function. For procedure 1, the UMVUE is also the maximum likelihood estimator (MLE) of $N$. For procedure 2 and for procedure 3 with $k = 1$, it is shown respectively that an MLE and an estimator which differs from an MLE by at most one have uniformly smaller mean square errors than the respective UMVUE's.

**2. Inadmissibility of UMVUE.** For procedure $i$, $i = 1, 2, 3$, let $S$ be the number of trials required to recapture the first marked unit and let $P_{iN}(s)$ denote $P_N\ (S = s)$. It then follows that

$$\text{(2.1)}\quad P_{1N}(s) = \left(1 - \frac{k}{N}\right)^{s-1}\frac{k}{N}, \qquad s = 1, 2, \ldots,$$

$$\text{(2.2)}\quad P_{2N}(s) = \left(1 - \frac{k}{N}\right)\left(1 - \frac{k}{N-1}\right)\ldots\left(1 - \frac{k}{N-s+2}\right)\frac{k}{N-s+1}, \qquad s = 1, \ldots, N-k+1,$$

$$\text{(2.3)}\quad P_{3N}(s) = \left(1 - \frac{k}{N}\right)\left(1 - \frac{k+1}{N}\right)\ldots\left(1 - \frac{s+k-2}{N}\right)\frac{s+k-1}{N}, \qquad s = 1, \ldots, N-k+1.$$

The following lemmas are used in the derivation of the results that follow.

LEMMA 2.1.

$$\sum_{s=j}^{\infty} P_{1N}(s) = \frac{N}{k} P_{1N}(j), \tag{2.4}$$

$$\sum_{s=j}^{N-k+1} P_{2N}(s) = \frac{N-j+1}{k} P_{2N}(j), \tag{2.5}$$

$$\sum_{s=j}^{N-k+1} P_{3N}(s) = \frac{N}{j+k-1} P_{3N}(j). \tag{2.6}$$

*Proof.* It is easy to verify (2.4) from (2.1). Also note that (2.5) is trivially true for $j = N - k + 1$. Hence, assuming that it is true for $j + 1$, we have

$$\begin{aligned}\sum_{s=j}^{N-k+1} P_{2N}(s) &= \left(1-\frac{k}{N}\right)\cdots\left(1-\frac{k}{N-j+2}\right)\\ &\quad\times\left\{\frac{k}{N-j+1}+\frac{N-j}{k}\left(1-\frac{k}{N-j+1}\right)\frac{k}{N-j}\right\}\\ &= \left(1-\frac{k}{N}\right)\cdots\left(1-\frac{k}{N-j+2}\right) = \frac{N-j+1}{k} P_{2N}(j),\end{aligned}$$

which proves (2.5). By similar arguments, (2.6) follows.

LEMMA 2.2. *Let $E_{iN}$ denote the expectation under procedure $i$, $i = 1, 2, 3$, for given $N$. Then for any given real-valued function $f$,*

$$kE_{1N} \sum_{j=1}^{S} f(j) = NE_{1N}[f(S)], \tag{2.7}$$

$$kE_{2N} \sum_{j=1}^{S} f(j) = E_{2N}[(N-S+1)f(S)], \tag{2.8}$$

$$E_{3N} \sum_{j=1}^{S} (j+k-1)f(j) = NE_{3N}[f(S)]. \tag{2.9}$$

*Proof.* By (2.4) we have

$$\begin{aligned}kE_{1N} \sum_{j=1}^{S} f(j) &= k\sum_{s=1}^{\infty} P_{1N}(s) \sum_{j=1}^{s} f(j) = k\sum_{j=1}^{\infty} f(j) \sum_{s=j}^{\infty} P_{1N}(s)\\ &= N\sum_{j=1}^{\infty} f(j)P_{1N}(j) = NE_{1N}[f(S)],\end{aligned}$$

which proves (2.7). Similarly, (2.8) and (2.9) follow, respectively, from (2.5) and (2.6).

In particular, for $f(j) = 1$ for all $j$, (2.7)–(2.9) yield $E_{iN}[\hat{N}_i^*] = N$, $i = 1, 2, 3$, where

$$\hat{N}_1^* = kS, \tag{2.10}$$

$$\hat{N}_2^* = (k+1)S - 1, \tag{2.11}$$

$$\hat{N}_3^* = \sum_{j=k}^{S+k-1} j = \binom{S+k}{2} - \binom{k}{2}. \tag{2.12}$$

For $m = 1$, it is well known that $S$ is complete sufficient for the parameter space $\{N \geq k\}$. Hence, it follows that $\hat{N}_i^*$ is the UMVUE of $N$ under procedure $i$, $i = 1, 2, 3$.

For procedure $i$, $i = 1, 2, 3$, consider an alternative estimator $\hat{N}_i = \hat{N}_i(S)$ of $N$ defined as

$$\hat{N}_i = \hat{N}_i^* - X \tag{2.13}$$

where $X = X(S)$ satisfies the following conditions:

$$X(s) = 0 \quad \text{for } s = 1, \tag{2.14}$$

$$X(s+1) - X(s) = 0 \text{ or } 1 \quad \text{for } s \geq 1, \tag{2.15}$$

$$X(s) > 0 \quad \text{for at least one } s \geq 2. \tag{2.16}$$

In the theorem below we prove that any estimator of the form (2.13), though negatively biased, dominates the UMVUE under square error loss function for each of the three procedures. Thus, for $m = 1$ and for each $i = 1, 2, 3$, the UMVUE of $N$ for procedure $i$ is inadmissible under square error loss and we have a class of estimators dominating the UMVUE.

THEOREM 2.3. *For $i = 1, 2, 3$, $E_{iN}[\hat{N}_i - N]^2 \leq E_{iN}[\hat{N}_i^* - N]^2$ with strict inequality for all $N \geq N_0$, where for $i = 1$, $N_0 = k + 1$ and for $i = 2, 3$, $N_0 - k + 1$ is the smallest integer $s$ for which $X(s) > 0$.*

*Proof.* Note that

$$\begin{aligned} E_{iN}[\hat{N}_i - N]^2 &= E_{iN}[\hat{N}_i^* - X - N]^2 \\ &= E_{iN}[\hat{N}_i^* - N]^2 + E_{iN}[X]^2 - 2E_{iN}[X(\hat{N}_i^* - N)]. \end{aligned} \tag{2.17}$$

Now, by (2.7)–(2.9), we have

$$\begin{aligned} NE_{1N}[X] &= kE_{1N}\sum_{j=1}^{S} X(j) = kE_{1N}\Big[SX - \sum_{j=1}^{S-1}\{X(S) - X(j)\}\Big] \\ &= E_{1N}\Big[\hat{N}_1^* X - k\sum_{j=1}^{S-1}\{X(S) - X(j)\}\Big] \\ &= E_{1N}\Big[\hat{N}_1^* X - k\sum_{j=1}^{X} j\Big] = E_{1N}[\hat{N}_1^* X - kX(X+1)/2], \end{aligned} \tag{2.18}$$

$$\begin{aligned}(2.19)\quad NE_{2N}[X] &= E_{2N}[(S-1)X+(N-S+1)X]\\ &= E_{2N}\Big[(S-1)X+k\sum_{j=1}^{S}X(j)\Big]\\ &= E_{2N}\Big[((k+1)S-1)X-k\sum_{j=1}^{S-1}\{X(S)-X(j)\}\Big]\\ &= E_{2N}\Big[\hat{N}_2^*X-k\sum_{j=1}^{X}j\Big]=E_{2N}[\hat{N}_2^*X-kX(X+1)/2],\end{aligned}$$

$$\begin{aligned}(2.20)\quad NE_{3N}[X] &= E_{3N}\sum_{j=1}^{S}(j+k-1)X(j)\\ &= E_{3N}\Big[X\sum_{j=1}^{S}(j+k-1)-\sum_{j=1}^{S}(j+k-1)(X(S)-X(j))\Big]\\ &= E_{3N}\Big[\hat{N}_3^*X-\sum_{j=1}^{S}(j+k-1)(X(S)-X(j))\Big]\\ &\le E_{3N}\Big[\hat{N}_3^*X-k\sum_{j=1}^{S}(X(S)-X(j))\Big]\\ &= E_{3N}\Big[\hat{N}_3^*X-k\sum_{j=1}^{X}j\Big]=E_{3N}[\hat{N}_3^*X-kX(X+1)/2].\end{aligned}$$

Hence, (2.17)–(2.20) imply that

$$E_{iN}[\hat{N}_i-N]^2-E_{iN}[\hat{N}_i^*-N]^2\le E_{iN}[X^2-kX(X+1)]\le 0,\quad i=1,2,3,$$

with strict inequality for all $N\ge N_0$, and this completes the proof of the theorem.

**3. Comparison between MLE and UMVUE.** For $m=1$, the likelihood functions of $N$ given $S=s$ under procedure $i$ is $L_i(N)=P_{iN}(s)$, $i=1,2,3$, given by (2.1)–(2.3), where $N\ge k$ for procedure 1 and $N\ge s+k-1$ for procedures 2 and 3. We note that for $s=1$, $L_i(N)$ is decreasing in $N$ for each $i=1,2,3$ so that the maximum likelihood estimate is $k$ for $s=1$.

For $s>1$, it may be seen by direct differentiation that $\log L_1(N)$ is maximum for $N=kS$. Thus for procedure 1, the MLE is the same as the UMVUE and hence inadmissible under square error loss.

It may also be verified that for $k=1$, $L_2(N)$ is decreasing in $N$ for every $s\ge 1$, and for $k>1$, $L_2(N)$ is $>$, $=$ or $<$ $L_2(N-1)$ according as $N$ is

$<, =$ or $> kS$. Thus for procedure 2, an MLE of $N$ is $\hat{N}_{\mathrm{MLE2}} = kS$, which is unique for $k = 1$. Since $\hat{N}_{\mathrm{MLE2}}$ is of the form (2.13) with $X = S - 1$, it readily follows from Theorem 2.3 that for $N > k$, $\hat{N}_{\mathrm{MLE2}}$ has a smaller mean square error (mse) than the respective UMVUE $\hat{N}_2^*$, while for $N = k$, the two mse's are equal. Following the proof of Theorem 2.3 it can be shown that for $k > 1$, the mse of an alternative MLE $\hat{N}'_{\mathrm{MLE2}}$ is also smaller than that of $\hat{N}_2^*$ for $N > k + 1$ and is equal for $N = k, k + 1$, where $\hat{N}'_{\mathrm{MLE2}} = k$, for $S = 1$ and $\hat{N}'_{\mathrm{MLE2}} = kS - 1$ for $S > 1$ though in this case (2.15) is not satisfied for $s = 1$.

For procedure 3, MLE of $N$ cannot be obtained in an explicit form. Consider, for $s > 1$, the equation

$$\frac{d \log L_3(N)}{dN} = 0 \quad \text{or} \quad \sum_{j=k}^{s+k-2} \frac{j}{N - j} = 1 \tag{3.1}$$

and note that the LHS of (3.1) is decreasing in $N$ and tends to 0 as $N$ tends to $\infty$. Also for $N = s+k-1$, the LHS of (3.1) is $\geq s+k-2 \geq 1$. Hence, (3.1) has a unique solution, say $N_0 = N_0(s)$, and $\log L_3(N)$ is increasing in $N$ for $N < N_0$ and is decreasing in $N$ for $N > N_0$. Thus for procedure 3, MLE of $N$ is $\hat{N}_{\mathrm{MLE3}}$, where $\hat{N}_{\mathrm{MLE3}} = k$ for $s = 1$, while for $s > 1$, $\hat{N}_{\mathrm{MLE3}} = N_0$ if $N_0$ is an integer, and $\hat{N}_{\mathrm{MLE3}} = [N_0]$ or $[N_0] + 1$ according as $L_3([N_0])$ is $\geq$ or $\leq L_3([N_0]+1)$ if $N_0$ is not an integer. Here $[c]$ denotes the largest integer not exceeding $c$. Clearly, for $s = 2$, $\hat{N}_{\mathrm{MLE3}} = N_0 = 2k$.

In what follows we prove that, for $k = 1$, the estimator $\hat{N}_3$ defined as

$$\hat{N}_3 = \begin{cases} k & \text{for } s = 1, \\ [N_0] & \text{for } s > 1, \end{cases} \tag{3.2}$$

is of the form (2.13). The following lemma is useful for this purpose.

Lemma 3.1. *For $k = 1$ and for every $s > 1$,*

$$k + s - 1 \leq N_0(s + 1) - N_0(s) \leq k + s. \tag{3.3}$$

The proof of the lemma is given in the Appendix.

Let now $[N_0] = N_0 - \nu(s)$ and note that $0 \leq \nu(s) < 1$ for every $s > 1$. Also, by (2.12), $\hat{N}_3^*(s + 1) - \hat{N}_3^*(s) = s + k$. Hence writing $X(s) = \hat{N}_3^*(s) - \hat{N}_3(s)$, by Lemma 3.1, it follows that for $k = 1$, $X(1) = 0$, $X(2) = 1$, and for $s > 1$,

$$\begin{aligned} -1 \leq (s + k - 1) - (N_0(s + 1) - N_0(s)) &< X(s + 1) - X(s) \\ &= s + k - (N_0(s + 1) - N_0(s)) + (\nu(s + 1) - \nu(s)) \\ &< s + k + 1 - (N_0(s + 1) - N_0(s)) \leq 2, \end{aligned}$$

which means that $X(s+1) - X(s)$ is 0 or 1 since it is an integer. This shows that for $k = 1$, $\hat{N}_3$ is of the form (2.13) and it follows by Theorem 2.3 that for $N > k$, $\hat{N}_3$, which differs from MLE by at most one, has a smaller mse than the respective UMVUE $\hat{N}_3^*$, while for $N = k$, the two mse's are equal. We remark that an alternative estimator

$$\hat{N}_3' = \begin{cases} k & \text{for } s = 1, \\ [N_0] + 1 & \text{for } s > 1, \end{cases}$$

again differs from MLE by at most one and is of the form (2.13) for $k = 1$, since, for $s > 1$, $\hat{N}_3'(s+1) - \hat{N}_3'(s) = \hat{N}_3(s+1) - \hat{N}_3(s)$ and for $s = 1, 2$, $X = \hat{N}_3^* - \hat{N}_3' = 0$. Numerical calculations reveal that for $k = 1$, an exact MLE $\hat{N}_{\mathrm{MLE3}}$ is also of the form (2.13). We further believe that Lemma 3.1 and hence the consequent results are true for general $k \geq 1$. However, we have not been able to prove this so far.

**Appendix: Proof of Lemma 3.1.** Not to obscure the essential steps of the reasoning we first prove some necessary results in the following lemmas.

Lemma A.1. *For $k = 1$ and $s\ (\neq 4) \geq 2$, $N_0(s) \geq s^2/2$.*

*Proof.* By the arguments used to obtain $\hat{N}_{\mathrm{MLE3}}$, it is enough to show that

$$\text{(A.1)} \qquad \sum_{j=1}^{s-1} \frac{j}{s^2/2 - j} \geq 1$$

Now for $s \geq 6$,

$$\sum_{j=1}^{s-1} \frac{j}{s^2/2 - j} = \frac{2}{s^2} \sum_{j=1}^{s-1} j \left(1 - \frac{2j}{s^2}\right)^{-1} \geq \frac{2}{s^2} \sum_{j=1}^{s-1} j \left(1 + \frac{2j}{s^2}\right)$$
$$= \frac{3s^3 + s(s-6) + 2}{3s^3} \geq 1.$$

Also (A.1) may be verified by direct calculations for $s = 2, 3, 5$, which completes the proof of the lemma.

Lemma A.2. *For $k = 1$ and for every $s > 1$, $N_0(s) \geq N_1(s)$ where*

$$N_1 = N_1(s) = \frac{s+k-1}{s+k} \sum_{j=k}^{s+k-2} j + (s+k-2).$$

*Proof.* For $k = 1$,

$$N_1(s) = \frac{s(s-1)}{2} + \frac{(s-1)(s+2)}{2(s+1)} \leq \frac{s^2}{2}.$$

Hence, for $s \neq 4$, the lemma follows from Lemma A.1. For $s = 4$, direct computation shows $\sum_{j=1}^{s-1} j/(N_1 - j) \geq 1$ and hence, as in the proof of Lemma A.1, $N_0(s) \geq N_1(s)$. Thus the lemma follows.

We now proceed to prove Lemma 3.1. As in the proof of Lemma A.1, it is enough to show that

$$\sum_{j=k}^{s+k-1} \frac{j}{N_0+k+s-1-j} \geq 1, \tag{A.2}$$

$$\sum_{j=k}^{s+k-1} \frac{j}{N_0+k+s-j} \leq 1, \tag{A.3}$$

where $N_0$ is the solution of (3.1). Now

$$\begin{aligned}
\sum_{j=1}^{s+k-1} \frac{j}{N_0+k+s-1-j} &= \sum_{j=k}^{s+k-2} \frac{j}{N_0+k+s-1-j} + \frac{s+k-1}{N_0} \\
&= 1 + \sum_{j=k}^{s+k-2} j\left\{\frac{1}{N_0+k+s-1-j} - \frac{1}{N_0-j}\right\} + \frac{s+k-1}{N_0} \\
&= 1-(k+s-1)\sum_{j=k}^{s+k-2} \frac{j}{(N_0+k+s-1-j)(N_0-j)} + \frac{s+k-1}{N_0} \\
&\geq 1 - \frac{s+k-1}{N_0+1}\sum_{j=k}^{s+k-2}\frac{j}{N_0-j} + \frac{s+k-1}{N_0} = 1 - \frac{s+k-1}{N_0+1} + \frac{s+k-1}{N_0} \\
&\geq 1,
\end{aligned}$$

and this proves (A.2). Similarly we have, using the A.M.-H.M. inequality,

$$\begin{aligned}
&\sum_{j=1}^{s+k-1} \frac{j}{N_0+k+s-j} \\
&\qquad = 1 + \frac{s+k-1}{N_0+1} - (k+s)\sum_{j=k}^{s+k-2} \frac{j}{(N_0+k+s-j)(N_0-j)} \\
&\qquad \leq 1 + \frac{s+k-1}{N_0+1} - (k+s)\left[\sum_{j=k}^{s+k-2} \frac{j(N_0+k+s-j)}{N_0-j}\right]^{-1} \\
&\qquad = 1 + \frac{s+k-1}{N_0+1} - \left[\sum_{j=k}^{s+k-2} \frac{j}{k+s} + 1\right]^{-1} \leq 1
\end{aligned}$$

since, by Lemma A.2,

$$\frac{N_0+1}{s+k-1} \geq \sum_{j=k}^{s+k-2} \frac{j}{k+s} + 1,$$

and this proves (A.3).

**Acknowledgements.** The author is grateful to the referee for his useful comments on an earlier draft of the paper.

## References

M. T. Boswell, K. P. Burnham, and G. P. Patil (1988), *Role and use of composite sampling and capture-recapture sampling in ecological studies*, in: Handbook of Statistics, Vol. 6, P. R. Krishnaiah and C. R. Rao (eds.), North-Holland, Amsterdam, 469–488.

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

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

N. L. Johnson and S. Kotz (1969), *Distribution in Statistics: Discrete Distribution*, Wiley, New York.

E. Samuel (1968), *Sequential maximum likelihood estimation of the size of a population*, Ann. Math. Statist. 39, 1057–1068.

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

S. Sengupta and M. De (1997), *On the estimation of size of a finite population*, Sankhyā B 59, 66–75.

S. Sengupta<br>
Department of Statistics<br>
Calcutta University<br>
35, Ballygunge Circular Road<br>
Kolkata 700019, India<br>
E-mail: samindras@yahoo.co.in

*Received on 12.10.2009;*
*revised version on 4.1.2010* (2021)