ON ESTIMATING THE SIZE OF A POPULATION AND ITS INVERSE BY CAPTURE MARK METHOD

By P. K. PATHAK*

Indian Statistical Institute

SUMMARY. The problem of estimating the size of a population and its inverse is considered here on the basis of several independent simple random (without replacement) samples selected from the population. New estimates of the population size and its reciprocal are given. The estimate of the reciprocal of the population size is the unbiased minimum variance estimate. The estimate of the population size is unbiased and has minimum variance only if the total sample size is not less than the population size. Variance estimates of these estimates are also given.

1. Introduction

The problem of estimating the size of a population is known to be of great importance in biological and other related problems, e.g., well-known problems of this kind are the estimation of the total number of fish in a lake and the estimation of the total number of wild animals in a forest, etc. Several authors have already considered this problem in the past and have devised methods of sampling (see references). In this paper simple random sampling (without replacement) at several stages has been considered for this purpose. Czen Pin and Dzan Dzo (1961) have also considered this method and termed it as capture-mark method. Here, this method is also referred to as capture-mark method. Bailey (1951) mentions that in certain ecological problems one is more interested in estimating the reciprocal of the population size rather than the population size itself; the problem of estimating the reciprocal of the population size is therefore also considered here.

To begin with, the following lemma is given which will be found useful later.

Lemma 1.1 : *Let $A_1, \ldots, A_m$ be m events defined on a probability space. Let $A = \bigcup_{i=1}^{m} A_i$ and $B_i = (A - A_i)$ $i = 1, \ldots, m$. Then*

$$P\left[\bigcap_{i=1}^{m} A_i\right] = P(A) - \Sigma' P(B_1) + \Sigma' P(B_1 \cap B_2) \ldots \qquad \ldots \quad (1.1)$$

where the summation Σ' is taken over all combinations of B's chosen from $B_1, \ldots, B_m$.

The proof is omitted.

---

*At present with University of Illinois on leave from Indian Statistical Institute.

2. Capture-mark method

The capture-mark method of sampling from a finite population of size N is as follows: k simple random (without replacement) subsamples of size $n_1, n_2, \ldots n_k$ respectively are drawn independently of each other and the number, M, of distinct population units is noted. The random variable M is used to get estimates of the population size and its reciprocal.

To get the probability distribution of M, it can be verified on letting $A_i = \{i\text{-th population unit is selected in the sample}\}$ $(i = 1, \ldots, m)$ in Lemma 1.1 that the probability of getting any preassigned m distinct units, in this sampling scheme, is given by*

$$P_1 = \frac{\prod_{i=1}^{k}\binom{m}{n_i} - \binom{m}{1}\prod_{i=1}^{k}\binom{m-1}{n_i} + \ldots (-)^{m-\max n_i}\binom{m}{m-\max n_i}\prod_{i=1}^{k}\binom{\max n_i}{n_i}}{\prod_{i=1}^{k}\binom{N}{n_i}}. \quad \ldots \quad (2.1)$$

The probability distribution of M is immediately obtained from (2.1) as given below

$$P[M = m] = \frac{\binom{N}{m}\left\{\prod_{i=1}^{k}\binom{m}{n_i} - \binom{m}{1}\prod_{i=1}^{k}\binom{m-1}{n_i} + \ldots\right\}}{\prod_{i=1}^{k}\binom{N}{n_i}}. \quad \ldots \quad (2.2)$$

A useful equality follows from (2.2)

$$\sum_{m=\max n_i}^{\min(n, N)} \binom{N}{m}\left\{\prod_{i=1}^{k}\binom{m}{n_i} - \binom{m}{1}\prod_{i=1}^{k}\binom{m-1}{n_i} + \ldots\right\} = \prod_{i=1}^{k}\binom{N}{n_i}. \quad \ldots \quad (2.3)$$

It is worthwhile to mention here that Czen Pin and Dzan Dzo (1961) have proved that if $n = \sum_{i=1}^{k} n_i \to \infty$ in such a manner that $n^2 \Big/ \left(\sum_{i>j} n_i n_j\right)$ remains bounded, then

$$P[M = m] = e^{-\lambda}\frac{\lambda^{n-m}}{(n-m)!}\left[1 + 0\left(\frac{n}{N}\right)\right] \quad \ldots \quad (2.4)$$

where $\lambda = \left(\sum_{i>j} n_i n_j\right)/N$.

From (2.2) it can be seen by induction over N that M is a complete sufficient statistic for the parameter space $\{N \geqslant \max_{1 \leqslant i \leqslant k} n_i\}$. It thus follows that if minimum variance unbiased estimates of the population size and its inverse exist, they must be functions of M.

In the next two sections, the problem of estimating the population size and its inverse is considered in reverse order for some simplicity in the exposition.

*$\binom{r}{k}$ is to be regarded as zero for $k > r$.

3. Estimation of the inverse of the population size

The following theorem gives the unbiased minimum variance estimate of $\frac{1}{N}$

Theorem 1 : *The unbiased minimum variance estimate of* $1/N$ *is given by*

$$t_{-1}(m) = \frac{\left[\frac{1}{m}\prod_{i=1}^{k}\binom{m}{n_i} - \frac{\binom{m}{1}}{(m-1)}\prod_{i=1}^{k}\binom{m-1}{n_i} + \frac{\binom{m}{2}}{(m-2)}\prod_{i=1}^{k}\binom{m-2}{n_i} - \cdots\right]}{\left[\prod_{i=1}^{k}\binom{m}{n_i} - \binom{m}{1}\prod_{i=1}^{k}\binom{m-1}{n_i} + \binom{m}{2}\prod_{i=1}^{k}\binom{m-2}{n_i} - \cdots\right]}. \qquad \ldots \quad (3.1)$$

Proof: Let u_{11} and u_{21} be the first sample units respectively of the first two subsamples. Since $P[u_{11} = u_{21}] = 1/N$, an unbiased estimate of $1/N$ is given by

$$t_{-1} = \begin{cases} 1 & \text{if } u_{11} = u_{21} \\ 0 & \text{otherwise.} \end{cases} \qquad \ldots \quad (3.2)$$

Further, since M is a complete sufficient statistic, the unbiased minimum variance estimate of $\frac{1}{N}$ is given by

$$t_{-1}(m) = E[t_{-1} \mid M = m] = \frac{P[u_{11} = u_{21} \cap M = m]}{P[M = m]}. \qquad \ldots \quad (3.3)$$

In order to be able to express (3.3) as a function of $m, n_1, \ldots, n_k$, let $u_{(1)}, \ldots, u_{(m)}$ be the m distinct population units selected in the sample. Then on letting $A_j = [u_{11} = u_{21} = u_{(i)}$ and $u_{(j)}$ is selected in the sample$]$ $j = 1, \ldots, m$ in Lemma 1.1, it can be seen that the probability of getting a sample with $u_{(1)} \ldots, u_{(m)}$ and $u_{11} = u_{21} = u_{(i)}$ is given by

$$P[u_{11} = u_{21} = u_{(i)} \cap u_{(1)}, \ldots, u_{(m)}]$$

$$= \frac{\binom{m-1}{n_1-1}\binom{m-1}{n_2-1}\prod_{i=3}^{k}\binom{m}{n_i} - \binom{m-1}{1}\binom{m-2}{n_1-1}\binom{m-2}{n_2-1}\prod_{i=3}^{k}\binom{m-1}{n_i}\cdots}{N.N.\binom{N-1}{n_1-1}\binom{N-1}{n_2-1}\prod_{i=3}^{k}\binom{N}{n_i}}$$

$$= \frac{\frac{1}{m^2}\prod_{i=1}^{k}\binom{m}{n_i} - \frac{\binom{m}{1}}{m(m-1)}\prod_{i=1}^{k}\binom{m-1}{n_i} + \cdots}{\prod_{i=1}^{k}\binom{N}{n_i}}. \qquad \ldots \quad (3.4)$$

It follows from (3.4) that

$$P[u_{11} = u_{21} \cap M = m]$$

$$= \frac{\binom{N}{m}\left[\frac{1}{m}\prod_{i=1}^{k}\binom{m}{n_i} - \frac{\binom{m}{1}}{(m-1)}\prod_{i=1}^{k}\binom{m-1}{n_i} + \frac{\binom{m}{2}}{(m-2)}\prod_{i=1}^{k}\binom{m-2}{n_i} - \cdots\right]}{\prod_{i=1}^{k}\binom{N}{n_i}}. \qquad \ldots \quad (3.5)$$

The theorem is proved on combining (2.2), (3.3) and (3.5).

Corollary 1 : *In the particular case when $n_1 = n_2 = \ldots = n_k = 1$, the unbiased minimum variance estimate of $\frac{1}{N}$ is given by*

$$t_{-1}(m) = \frac{C_m(k-1)}{C_m(k)} \qquad \ldots \quad (3.6)$$

where $C_m(k) = m^k - \binom{m}{1}(m-1)^k + \ldots (-)^{m-1}\binom{m}{m-1}$ $C_m(k)$'s are called the differences of zeroes.

The author (1961) has tabulated the values of $\frac{C_m(k-1)}{C_m(k)}$ for all m and $k = 1$ to 50.

It can be shown in a similar manner that the best unbiased estimate of $V[t_{-1}(m)]$ is given by

$$v[t_{-1}(m)] = t_{-1}^2(m) - t_{-2}(m) \qquad \ldots \quad (3.7)$$

where

$$t_{-2}(m) = \frac{\left[\frac{1}{m^2}\prod_{i=1}^{k}\binom{m}{n_i} - \frac{\binom{m}{1}}{(m-1)^2}\prod_{i=1}^{k}\binom{m-1}{n_i} + \ldots\right]}{\left[\prod_{i=1}^{k}\binom{m}{n_i} - \prod_{i=1}^{k}\binom{m-1}{n_i} + \ldots\right]}.$$

4. Estimation of the population size

The search for an unbiased estimate of N leads to the following theorem.

Theorem 2 : *The unbiased minimum variance estimate of N exists if and only if the sample size $n = \sum_{i=1}^{k} n_i \geqslant N$ in which case the required estimate is given by*

$$t_1(m) = \frac{\left[m\prod_{i=1}^{k}\binom{m}{n_i} - (m-1)\binom{m}{1}\prod_{i=1}^{k}\binom{m-1}{n_i} + \ldots\right]}{\left[\prod_{i=1}^{k}\binom{m}{n_i} - \binom{m}{1}\prod_{i=1}^{k}\binom{m-1}{n_i} + \ldots\right]}. \qquad \ldots \quad (4.1)$$

Proof: Suppose that there exists such an estimate. Let it be $t_1(m)$. Then we have from the condition of unbiasedness

$$\sum_{m=\max n_i}^{\min(n,N)} t_1(m)\binom{N}{m}\frac{\left[\prod_{i=1}^{k}\binom{m}{n_i} - \binom{m}{1}\prod_{i=1}^{k}\binom{m-1}{n_i} + \ldots\right]}{\prod_{i=1}^{k}\binom{N}{n_i}} = N \qquad \ldots \quad (4.2)$$

for all $N \geqslant \max_i n_i$.

Putting successively $N = \max n_i, \max n_i+1, \ldots$, we get the only possible estimate, $t_1(m)$, as defined above. It is found on taking the expectation of $t_1(m)$, with the help of (2.3), that $t_1(m)$ is an unbiased estimate of N if and only if $n \geqslant N$; otherwise

$$E[t_1(m)] = \left[N - \frac{\binom{N}{n+1}\left[\prod_{i=0}^{k}\binom{n+1}{n_i} - \prod_{i=0}^{k}\binom{n}{n_i} + \ldots\right]}{\prod_{i=1}^{k}\binom{N}{n_i}} \right] \quad \ldots \quad (4.3)$$

where $n_0 = 1$.

This completes the proof.

The bias of $t_1(m)$ decreases as n increases and would be negligible if n is large. Moreover, if in practice some approximation for N is available in advance, a correction for the bias can be made.

In the particular case when $n_1 = n_2 = \ldots = n_k = 1$. $t_1(m)$ may be expressed in terms of the differences of zeroes as

$$t_1(m) = \frac{C_m(k+1)}{C_m(k)}. \quad \ldots \quad (4.4)$$

An estimate of $V[t_1(m)]$ (unbiased if $\sum_{i=1}^{k} n_i \geqslant N$) is given by

$$v[t_1(m)] = t_1^2(m) - t_2(m) \quad \ldots \quad (4.5)$$

where

$$t_2(m) = \frac{\left[m^2 \prod_{i=1}^{k}\binom{m}{n_i} - (m-1)^2\binom{m}{1}\prod_{i=1}^{k}\binom{m-1}{n_i} + \ldots\right]}{\left[\prod_{i=1}^{k}\binom{m}{n_i} - \binom{m}{1}\prod_{i=1}^{k}\binom{m-1}{n_i} + \ldots\right]}.$$

5. Concluding remark

The estimates suggested in the preceding sections are difficult to compute in practice except when $n_1 = n_2 = \ldots = n_k = 1$ and $k \leqslant 50$. However, if the subsampling fractions $\frac{n_i}{N}$ $(i = 1, \ldots, k)$ are negligible so that subsampling may be assumed with replacement, the above described estimates may be approximated by

$$t_{-1}(m) \doteq \frac{C_m(\Sigma n_i - 1)}{C_m(\Sigma n_i)} \quad \ldots \quad (4.6)$$

$$v[t_{-1}(m)] \doteq t_1^2(m) - \frac{C_m(\Sigma n_i - 2)}{C_m(\Sigma n_i)} \quad \ldots \quad (4.7)$$

$$t_1(m) \doteq \frac{C_m(\Sigma n_i + 1)}{C_m(\Sigma n_i)} \quad \ldots \quad (4.8)$$

$$v[t_1(m)] \doteq t_1^2(m) - \frac{C_m(\Sigma n_i + 2)}{C_m(\Sigma n_i)}. \quad \ldots \quad (4.9)$$

These approximations are good only for $\Sigma n_i \leqslant 50$ as tables of the ratio $\frac{C_m(r-1)}{C_m(r)}$ are not yet available in the literature beyond $r = 50$.

As a further approximation to these estimates, it is of interest to point out that if $n = \sum_{i=1}^{k} n_i$ is large and if $\frac{n^2}{\sum_{i>j} n_i n_j}$ is bounded so that (2.4) may be used to approximate the numerators and denominators of these estimates, the asymptotic expressions for these estimates are given by

$$t_{-1}(m) \doteq \frac{(n-m)}{\sum_{i>j} n_i n_j} \quad \ldots \quad (4.10)$$

$$v[t_{-1}(m)] \doteq t^2_{-1}(m) - \frac{(n-m)(n-m-1)}{\left(\sum_{i>j} n_i n_j\right)^2} \quad \ldots \quad (4.11)$$

$$t_1(m) \doteq \frac{\left(\sum_{i>j} n_i n_j\right)}{(n-m+1)} \quad \ldots \quad (4.12)$$

$$v(t_1(m)) \doteq t_1^2(m) - \frac{\left(\sum_{i>j} n_i n_j\right)^2}{(n-m+1)(n-m+2)}. \quad \ldots \quad (4.13)$$

The estimate (4.12) has been suggested for estimating N by Czen Pin and Dzan Dzo (1961). An interesting discussion on the bias and variance of (4.12) and on other related problems like confidence interval estimation of N may be found in their paper. It has been proved by them that (4.12) attains least variance when $n_1 = n_2 = \ldots = n_k = 1$.

References

Bailey, N. T. J. (1951): On estimating the size of mobile population from recapture data. *Biometrika*, 38, 293-306.

Chapman, D. G. (1952): Inverse, multiple and sequential simple censuses. *Biometrics*, 8, 286-306.

Craig, C. C. (1953): On the utilization of marked specimens in estimating populations of flying insects. *Biometrika*, 40, 170-176.

Czen Pin and Dzan Dzo—I (1961): On estimating the size of population by capture-mark method. *Zastosowania Mat.* 6, 51-63.

De Lury, D. B. (1947): On estimation of biological populations. *Biometrics*, 3, 145-167.

Haldane, J. B. S. (1945): On method of estimating frequencies. *Biometrika*, 33, 222-225.

Lewontin, R. C. and Timothy Prout (1956): Estimation of the number of different classes in a population. *Biometrics*, 12, 211-223.

Moran, P. A. P. (1951): A mathematical theory of animal trapping. *Biometrika*, 38, 307-311.

Pathak, P. K. (1961): *Some Contributions to the Theory of Sampling*, Ph.D. thesis submitted to the Indian Statistical Institute.

Paper received : October, 1963