Biometrika (1988), **75**, 1, *pp.* 178–80
Printed in Great Britain

A note on the exact maximum likelihood estimation of the size of a finite and closed population

By JOSÉ GALVÃO LEITE, JORGE OISHI
AND CARLOS ALBERTO DE BRAGANÇA PEREIRA
Instituto de Matemática e Estatística, Universidade de São Paulo, C.P. 20570, São Paulo, SP 01498, Brazil

Summary

Using data obtained by the general capture-recapture sequential sampling process, an analytical expression for the maximum likelihood estimate of the population size is introduced. It is shown that the bounded likelihood functions have at most two maxima. For the simple one-to-one case the estimate is unique.

Some key words: Capture-recapture sequential sampling process; Maximum likelihood estimate; Sufficient statistic.

1. Introduction

The objective of this note is to present a closed analytical expression for the maximum likelihood estimate to the size, N, of a finite and closed population when the data are obtained by capture-recapture sequential sampling. Inferences about N based on data obtained in special cases have been considered by many authors; see, for instance, Seber (1982, Ch. 3, 4; 1986) for a complete reference list.

Consider a population of size N that changes neither in size nor in form; that is, the population is closed. From this population, $k>1$, random samples without replacement are sequentially selected from the population. Each of these samples is returned to the population before the next is selected. For the jth $(j=1,\ldots,k)$ sample, the scientist records the sample size $m_j \geq 1$ and the number U_j of units selected for the first time; that is, units that were not selected in samples 1 to $j-1$. The statistic $T_k = U_1+\ldots+U_k$ is the number of distinct units selected in the whole sampling process. Leite & Pereira (1987) show that this statistic is sufficient and that the smallest factor of the likelihood function that depends on the value of N, the likelihood kernel, is

$$K(N,t)=N!\left\{(N-t)!\prod_{j=1}^{k}\binom{N}{m_j}\right\}^{-1} I_t(N),$$

where t is the observed value of T_k and $I_t(.)$ is the indicator function of $N_t=\{n\geq t\}$. The probability distribution of T_k (Leite & Pereira, 1987) is

$$\operatorname{pr}\{T_k=t \mid N\}=K(N,t)t!\sum_{i=0}^{t}(-1)^{t-i}\binom{t}{i}\binom{i}{m_1}\ldots\binom{i}{m_k} I^*(t),$$

where $m=\max\{m_1,\ldots,m_k\}$, $s=m_1+\ldots+m_k$ and $I^*(t)$ is the indicator function of

$$\{m \leq t \leq \min(N,s)\}.$$

2. Main results

If $t=m$, its smallest possible value, $K(N,t)$ is a decreasing function of N so that the maximum likelihood estimate is $\hat{N}=t$. At the other extreme, when $t=s$, $K(N,t)$ is an increasing function of N, so that $\hat{N}=\infty$.

Let now $m<t<s$, and consider the function

$$f_t(x)=(1-xt)^{-1}\prod_{j=1}^{k}(1-xm_j)$$

defined in $0\leqslant x\leqslant t^{-1}$. This function is continuous in $[0, t^{-1})$, is equal to unity at $x=0$, goes to ∞ as x increases to t^{-1}, and, if $n\geqslant t$,

$$f_t\left(\frac{1}{n+1}\right)=\frac{K(n+1, t)}{K(n, t)}.$$

The behaviour of f_t is described next.

LEMMA. *For $m<t<s$, the equation $f_t(x)=1$ has a unique positive solution x_0 in the open interval $(0, t^{-1})$. Also, $f_t(x)<1$ if $0<x<x_0$ and $f_t(x)>1$ if $x_0<x<t^{-1}$.*

Proof. Note that $f_t=g/h_t$, where g and h_t are defined as

$$g(x)=\prod_{j=1}^{k}(1-xm_j),\quad h_t(x)=1-xt.$$

The first and second derivatives of g are, respectively, negative and positive. Hence, g is a decreasing convex continuous function. Also, h_t is a decreasing linear function, $g(0)=h_t(0)=1$, $g(t^{-1})>0$ and $h_t(t^{-1})=0$. The derivative of (h_t-g) evaluated at $x=0$ is $m_1+\ldots+m_k-t>0$. Consequently, there is a unique point, x_0, in the open interval $(0, t^{-1})$, such that $g(x_0)=h_t(x_0)$, $g(x)<h_t(x)$ if $0<x<x_0$, and $g(x)>h_t(x)$ if $x_0<x<t^{-1}$. □

THEOREM. *A maximum likelihood estimate of N exists and is defined as*

$$\hat{N}=\begin{cases} t & (t=m),\\ t+n_t-1 & (m<t<s),\\ \infty & (t=s),\end{cases}$$

where $n_t=\min\{n;\ (t+n-m_1)\ldots(t+n-m_j)<n(t+n)^{k-1}\}$. Also this estimate is unique except when

$$\prod_{j=1}^{k}(t+n_t-m_j-1)=(n_t-1)(t+n_t-1)^{k-1}. \tag{1}$$

In this case the only two possible estimates are $(t+n_t-1)$ and $(t+n_t-2)$. If $m_1=\ldots=m_k=1$ it is always unique.

The proof follows directly from the Lemma. The uniqueness in the one-by-one case is not in agreement with Samuel (1968). To see this, for $1<t<k$, note that (1) simplifies to

$$(t+n_t-2)^k=(n_t-1)(t+n_t-1)^{k-1}$$

and, defining x to be the integer $t+n_t-1$, we can write $(x-1)^k=(x-t)x^{k-1}$. This last equation is equivalent to

$$(t-k)x^{k-2}+\sum_{i=1}^{k}\binom{k}{i}(-1)^i x^{k-i-1}=x^{-1}(-1)^{k-1}.$$

Since $x>t>1$ is an integer, the left-hand side of this equation must also be an integer but the right-hand side cannot be an integer. Hence the equation has no integer solution.

3. COMMENTS

Most recent publications on inference about N are restricted to the case on $k=2$. See for instance Isaki (1986) and Pollock, Hines & Nichols (1985). However, with the simple expression of the maximum likelihood estimator obtained from N, some of their analysis can be extended to the general case of $k\geq 2$.

Before using $\hat{N}$ one needs to observe the following limitations.

(i) The parameter space N_t changes with the observed value t of T_k.

(ii) the random variables U_i, for $i=1,\dots,k$, that form the data are not independently and identically distributed. In fact they are not even exchangeable.

(iii) The estimator defined from $\hat{N}$ has no finite moments.

(iv) When either event $\{T_k=m\}$ or $\{T_k=s\}$ occurs the value of $\hat{N}$ is not related to the value of k.

Facts (i), (ii) and (iii) restrict the use of standard statistical procedures. The use of Bayesian procedures may be the way to counter these problems since they rely on the observed data rather than on the distributional properties of T_k. To the best of our knowledge Freeman (1972) is the only available reference for the Bayesian estimation of N.

The strongest limitation on $\hat{N}$ is, in our opinion, introduced by (iv). To make the number of samples, k, relevant when $T_k=m$ or $T_k=s$, one needs to use prior knowledge.

ACKNOWLEDGEMENTS

The authors are grateful to Luis Pericchi for his comments on earlier drafts and the referee for useful suggestions.

REFERENCES

FREEMAN, P. R. (1972). Sequential estimation of the size of a population. *Biometrika* **59**, 9–17.

ISAKI, C. T. (1986). Bias of the dual system estimator and some alternatives. *Comm. Statist.* A **15**, 1435–50.

LEITE, J. G. & PEREIRA, C. A. DE B. (1987). An urn model for the multi-sample capture/recapture sequential tagging process. *Sequential Anal.* **6** (2), 179–86.

POLLOCK, K. H., HINES, J. E. & NICHOLS, J. D. (1985). Goodness-of-fit tests for open capture-recapture models. *Biometrics* **41**, 399–410.

SAMUEL, E. (1968). Sequential maximum likelihood estimation of the size of a population. *Ann. Statist.* **39**, 1057–68.

SEBER, G. A. F. (1982). *The Estimation of Animal Abundance*, 2nd ed. London: Griffin.

SEBER, G. A. F. (1986). A review of estimating animal abundance. *Biometrics* **42**, 267–92.

[*Received September* 1986. *Revised April* 1987]

References

Sequential Estimation of the Size of a Population

P. R. Freeman

Biometrika, Vol. 59, No. 1. (Apr., 1972), pp. 9-17.

Stable URL:

http://links.jstor.org/sici?sici=0006-3444%28197204%2959%3A1%3C9%3ASEOTSO%3E2.0.CO%3B2-L

Goodness-of-Fit Tests for Open Capture-Recapture Models

Kenneth H. Pollock; James E. Hines; James D. Nichols

Biometrics, Vol. 41, No. 2. (Jun., 1985), pp. 399-410.

Stable URL:

http://links.jstor.org/sici?sici=0006-341X%28198506%2941%3A2%3C399%3AGTFOCM%3E2.0.CO%3B2-F

A Review of Estimating Animal Abundance

G. A. F. Seber

Biometrics, Vol. 42, No. 2. (Jun., 1986), pp. 267-292.

Stable URL:

http://links.jstor.org/sici?sici=0006-341X%28198606%2942%3A2%3C267%3AAROEAA%3E2.0.CO%3B2-K