Biometrika (2022), **xx**, x, *pp*. 1–5

Printed in Great Britain

Estimating the size of a population through repeated sampling: a new view on capture-recapture procedures

BY G. KAUERMANN, N. SAPARGALI

Department of Statistics, Ludwig-Maximilians-Universität München, 80539 Munich, Germany

goeran.kauermann@stat.uni-muenchen.de nurzhan.sapargali@stat.uni-muenchen.de

SUMMARY

There should be a single paragraph summary which should not contain formulae or symbols, followed by some key words in alphabetical order. Typically there are 3–8 key words, which should contain nouns and be singular rather than plural. The summary contains bibliographic references only if they are essential. It should indicate results rather than describe the contents of the paper: for example, 'A simulation study is performed' should be replaced by a more informative phrase such as 'In a simulation our estimator had smaller mean square error than its main competitors.'

Some key words: Capture-recapture estimator; Inclusion probability; Population size.

1. INTRODUCTION

2. MODEL AND LIKELIHOOD

Our starting point is a basic capture-recapture model with equal capture probabilities. Consider a closed population $\mathcal{P} = \{1, \ldots, N\}$ from which we independently draw $T > 1$ samples of size n using some sampling scheme. Define S_t to be the set of individuals included into sample t and

$\pi_i = \text{pr}(i \in S_t)$ to be the inclusion probability of i. Let $X_{i(T)}$ denote the frequency count of i being included into T samples

$$X_{i(T)} = \sum_{t=1}^{T} \mathbb{1}\{i \in S_t\};$$

If $\pi_i = n/N$ for all i, then $X_{1(T)}, \ldots, X_{N(T)}$ are modeled as identically and independently distributed binomial variables

$$\text{pr}(X_{i(T)} = x_i; T, N, n) = \binom{T}{x_i} \left(\frac{n}{N}\right)^{x_i} \left(1 - \frac{n}{N}\right)^{T - x_i}$$

However, since we only observe $X_{i(T)} > 0$, the above distribution must be truncated at zero. For convenience purposes, it is also useful to partition $\mathcal{P}$ into $D = \bigcup_{t=1}^{T} S_t = \{i \in \mathcal{P} : X_{i(T)} > 0\}$ and $U = \mathcal{P} \setminus D = \{i \in \mathcal{P} : X_{i(T)} = 0\}$. Accordingly, population size can be expressed as a sum of cardinalities of D and U, i.e. $N = N_D + N_U$, where the first term is known and the latter term is to be estimated. The resulting likelihood of the data is

$$\mathcal{L}(N_U) = \prod_{i \in D} \frac{\text{pr}(X_{i(T)} = x_i; T, N, n)}{1 - \text{pr}(X_{i(T)} = 0; T, N, n)} = \prod_{i \in D} \binom{T}{x_i} \frac{n^{x_i} (N_D + N_U - n)^{T - x_i}}{(N_D + N_U)^T - (N_D + N_U - n)^T}$$

Setting $T = 2$ and recognizing that $\sum_{i \in D} x_i = nT$ yields the following maximum likelihood estimator of N_U

$$\widehat{N}_U = \frac{(n - N_D)^2}{2n - N_D} \tag{1}$$

Thus, the population size is estimated by

$$\widehat{N} = N_D + \widehat{N}_U = \frac{n^2}{2n - N_D}$$

The above expression can be rewritten in a more familiar manner, if we introduce the set of recaptured individuals $R = S_1 \cap S_2 = \{i \in \mathcal{P} : X_{i(2)} = 2\}$. The size of $D = S_1 \cup S_2$ is then the sum of sample sizes at each draw minus the size of R

$$\widehat{N} = \frac{n^2}{2n - 2n + N_R} = \frac{n^2}{N_R} \tag{2}$$

Equation (2) is the special case of the Lincoln-Peterson estimator with fixed n for each sample draw (Pollock et al., 1990).

Assume now that π_i varies across population units meaning that (2) is no longer applicable. Let us consider the problem from the Bayesian perspective and assume a beta prior for inclusion probabilities with hyperparameters $\alpha > 0$ and $\beta > 0$

$$f_{\pi_i}(\upsilon; \alpha, \beta) = \frac{\upsilon^{\alpha-1}(1-\upsilon)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}, \quad \mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)};$$

where $\Gamma(x)$ is the gamma function. From Bayes' theorem, we have

$$f_{\pi_i|X_{i(T)}}(\upsilon \mid x_i) = \frac{\mathrm{pr}(X_{i(T)} = x_i \mid \upsilon; T) f_{\pi_i}(\upsilon; \alpha, \beta)}{\mathrm{pr}(X_{i(T)} = x_i)}$$

The marginal likelihood can be found in a straightforward manner

$$\mathrm{pr}(X_{i(T)} = x_i) = \int_0^1 \mathrm{pr}(X_{i(T)} = x_i | \upsilon; T) f_{\pi_i}(\upsilon; \alpha, \beta) d\upsilon = \binom{T}{x_i} \frac{\mathrm{B}(\alpha + x_i, \beta + T - x_i)}{\mathrm{B}(\alpha, \beta)}$$

Conditioning on $X_{i(T)} > 0$ results in the following truncated distribution

$$\mathrm{pr}(X_{i(T)} = x_i \mid X_{i(T)} > 0; \alpha, \beta, T) = \frac{\mathrm{pr}(X_{i(T)} = x_i)}{1 - \mathrm{pr}(X_{i(T)} = 0)} = \binom{T}{x_i} \frac{\mathrm{B}(\alpha + x_i, \beta + T - x_i)}{\mathrm{B}(\alpha, \beta) - \mathrm{B}(\alpha, \beta + T)}$$

Following the empirical Bayes approach, we estimate hyperparameters α and β by maximizing the marginal likelihood.

$$\begin{aligned}
\mathcal{L}(\alpha, \beta) &= \prod_{i \in D} \binom{T}{x_i} \frac{\mathrm{B}(\alpha + x_i, \beta + T - x_i)}{\mathrm{B}(\alpha, \beta) - \mathrm{B}(\alpha, \beta + T)} \\
&= \prod_{i \in D} \binom{T}{x_i} \frac{\Gamma(\alpha + x_i)\Gamma(\beta + T - x_i)\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha + \beta + T) - \Gamma(\alpha)\Gamma(\beta + T)\Gamma(\alpha + \beta)}
\end{aligned}$$

Using the recursive relation of the gamma function $\Gamma(x) = (x-1)\Gamma(x-1)$ and the fact that T and x_i are natural numbers, we can rewrite the marginal likelihood as

$$\begin{aligned}\mathcal{L}(\alpha,\beta) &= \prod_{i\in D}\binom{T}{x_i}\frac{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta)\prod_{j=1}^{x_i}(\alpha+x_i-j)\prod_{j=1}^{T-x_i}(\beta+T-x_i-j)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta)\{\prod_{j=1}^{T}(\alpha+\beta+T-j)-\prod_{j=1}^{T}(\beta+T-j)\}} \\ &= \prod_{i\in D}\binom{T}{x_i}\frac{\prod_{j=1}^{x_i}(\alpha+x_i-j)\prod_{j=1}^{T-x_i}(\beta+T-x_i-j)}{\prod_{j=1}^{T}(\alpha+\beta+T-j)-\prod_{j=1}^{T}(\beta+T-j)} \end{aligned} \tag{3}$$

For inclusion probabilities π_i, it holds

$$E(\sum_{i\in\mathcal{P}}\pi_i) = N\frac{\alpha}{\alpha+\beta} = n \tag{4}$$

The marginal likelihood (3) is to be maximized subject to (4). Note that (4) can also be viewed as centering our prior at n/N. In other words, the hyperparameters determine only the variance of the prior. From the constraint (4), we have $\beta = Nn^{-1}\alpha - \alpha$. Thus, the likelihood can be reparametrized in terms of N_U and α

$$\mathcal{L}(\alpha, N_U) = \prod_{i\in D}\binom{T}{x_i}\frac{\prod_{j=1}^{x_i}(\alpha+x_i-j)\prod_{j=1}^{T-x_i}(Nn^{-1}\alpha-\alpha+T-x_i-j)}{\prod_{j=1}^{T}(Nn^{-1}\alpha+T-j)-\prod_{j=1}^{T}(Nn^{-1}\alpha-\alpha+T-j)} \tag{5}$$

where $N = N_D + N_U$.

The maximum likelihood estimates of N_U and α have no closed-form expressions and must be found numerically.

Setting $T = 2$ and $\alpha \to \infty$ will yield the estimator of N_U equivalent to (1), as the variance of the prior tends to 0 resulting in a degenerate distribution centered at n/N. This can be shown more formally by taking the limit of (5) with respect to $\alpha \to \infty$.

First, partition D into the set of recaptured individuals $R = S_1 \cap S_2 = \{i \in \mathcal{P} : X_{i(2)} = 2\}$ and the set of individuals captured only once $C = \{i \in \mathcal{P} : X_{i(2)} = 1\}$ with sizes N_R and N_C

respectively. Then, the likelihood simplifies to

$$\begin{aligned}
\mathcal{L}(\alpha, N_U) &= \prod_{i \in R \cup C} \binom{2}{x_i} \frac{\prod_{j=1}^{x_i}(\alpha + x_i - j) \prod_{j=1}^{2-x_i}(Nn^{-1}\alpha - \alpha + 2 - x_i - j)}{\prod_{j=1}^{2}(Nn^{-1}\alpha + 2 - j) - \prod_{j=1}^{2}(Nn^{-1}\alpha - \alpha + 2 - j)} \\
&= \frac{\prod_{i \in C} 2\alpha(Nn^{-1}\alpha - \alpha + 1) \prod_{i \in R} \alpha(\alpha + 1)}{(2Nn^{-1}\alpha^2 - \alpha^2 + \alpha)^{N_C + N_R}} \\
&= \frac{\alpha^{N_C + N_R}(\alpha + 1)^{N_R}\{2(Nn^{-1}\alpha - \alpha + 2)\}^{N_C}}{(2Nn^{-1}\alpha^2 - \alpha^2 + \alpha)^{N_C + N_R}} \\
&= \left\{ \frac{\alpha + 1}{\alpha(2Nn^{-1} - 1) + 1} \right\}^{N_R} \left\{ \frac{2\alpha(Nn^{-1} - 1) + 4}{\alpha(2Nn^{-1} - 1) + 1} \right\}^{N_C}
\end{aligned}$$

The limit of this expression as $\alpha \to \infty$ is

$$\begin{aligned}
&\lim_{\alpha \to \infty} \left[\left\{ \frac{\alpha + 1}{\alpha(2Nn^{-1} - 1) + 1} \right\}^{N_R} \left\{ \frac{2\alpha(Nn^{-1} - 1) + 4}{\alpha(2Nn^{-1} - 1) + 1} \right\}^{N_C} \right] \\
&= \lim_{\alpha \to \infty} \left[\left\{ \frac{1 + \alpha^{-1}}{2Nn^{-1} - 1 + \alpha^{-1}} \right\}^{N_R} \left\{ \frac{2(Nn^{-1} - 1) + 4\alpha^{-1}}{2Nn^{-1} - 1 + \alpha^{-1}} \right\}^{N_C} \right] \\
&= \frac{2^{N_C}(Nn^{-1} - 1)^{N_C}}{(2Nn^{-1} - 1)^{N_D}} \qquad (6)
\end{aligned}$$

since $N_R + N_C = N_D$. By taking the logarithm of (6) and maximizing with respect to N we have

$$\widehat{N} = \frac{n(2N_D - N_C)}{2(N_C - N_D)}$$

As $N_C = N_D - N_R$ and $N_D = 2n - N_R$, we arrive to

$$\widehat{N} = \frac{n^2}{N_R}$$

which is the Lincoln-Peterson estimator for fixed n as in (2). From it, the estimator (1) follows.

3. DISCUSSION

REFERENCES

POLLOCK, K. H., NICHOLS, J. D., BROWNIE, C. & HINES, J. E (1990). *Statistical inference for capture-recapture experiments*. *Wildlife Monographs* **107**, 3–97.