

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 $\mathcal{P} = \{1, \dots, N\}$ 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 $X_{1(T)}, \dots, X_{N(T)}$ 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 $D = \cup_{t=1}^T S_t = \{i \in \mathcal{P} : X_{i(T)} > 0\}$ 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

$$\hat{N}_U = \frac{(n - N_D)^2}{2n - N_D}$$

Thus, the population size is estimated by

$$\hat{N} = N_D + \hat{N}_U = \frac{n^2}{2n - N_D} \quad (1)$$

Equation (1) 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 (1) is no longer applicable.

Let us consider the problem from the Bayesian perspective and assume a beta prior for inclusion

probabilities with hyperparameters α and β

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

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

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

The marginal likelihood can be found in a straightforward manner

$$\text{pr}(X_{i(T)} = x_i) = \int_0^1 \text{pr}(X_{i(T)} = x_i | v) f_{\pi_i}(v) dv = \binom{T}{x_i} \frac{B(\alpha + x_i, \beta + T - x_i)}{B(\alpha, \beta)}$$

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

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

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

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

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 re-write the marginal likelihood as

$$\begin{aligned} \mathcal{L}(\alpha, \beta) &= \prod_{i \in D} \binom{T}{x_i} \frac{\Gamma(\alpha) \Gamma(\beta) \Gamma(\alpha + \beta) \prod_{j=1}^{x_i} (\alpha + x_i - j) \prod_{j=1}^{T-x_i} (\beta + T - x_i - j)}{\Gamma(\alpha) \Gamma(\beta) \Gamma(\alpha + \beta) \{ \prod_{j=1}^T (\alpha + \beta + T - j) - \prod_{j=1}^T (\beta + T - j) \}} \\ &= \prod_{i \in D} \binom{T}{x_i} \frac{\prod_{j=1}^{x_i} (\alpha + x_i - j) \prod_{j=1}^{T-x_i} (\beta + T - x_i - j)}{\prod_{j=1}^T (\alpha + \beta + T - j) - \prod_{j=1}^T (\beta + T - j)} \end{aligned} \quad (2)$$

REFERENCES

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

