Capture-recapture experiments

Jean-Michel Marin

Université de Montpellier
Institut Montpelliérain Alexander Grothendieck

We consider the problem of estimating the unknown size, N, of a population. The capture–recapture models were first used in biology and ecology to estimate the size of animal populations

While our illustrative dataset will be related to a biological problem, we stress that these capture–recapture models apply in a much wider range of domains, such as, for instance,

- sociology and demography;
- official statistics for reducing the cost of a census or improving its efficiency on delicate or rare subcategories;
- finance and marketing ;
- fraud detection and document authentication and
- software debugging.

In these different examples, the size N of the whole population is unknown but samples (with fixed or random sizes) can easily be extracted from the population.

For instance, the total number N of homeless people in a city like Philadelphia at a given time is not known but it is possible to count the number n_1 of homeless persons in a given shelter on a precise night, to record their ID, and to cross this sample with a sample of n_2 persons collected the night after in order to detect how many persons n_{12} were present in the shelter on both nights.

The dataset we consider is related to a population of birds called *European dippers (Cinclus cinclus)*. The capture–recapture data on the European dipper covers 7 years (1981–1987 inclusive) of observations in a zone of 200 km^2 in eastern France.

The data consist of markings and recaptures of breeding adults each year during the breeding period from early March to early June.

294 birds were captured (and eventually recaptured). The dataset is a matrix of 294 rows and seven columns.

Each row of seven digits corresponds to a capture–recapture story for a given dipper, 0 indicating an absence of capture that year and, in the case of a capture, 1, 2, or 3 representing the zone where the dipper is captured.

Plan

The Binomial Capture Model
The Two-Stage Capture–Recapture Model
The T-Stage Capture–Recapture Model
Open populations

The Binomial Capture Model

We start with the simplest model of all, namely the independent observation or capture of n^+ individuals from a population of size N.

While the population size $N \in \mathbb{N}$ is the parameter of interest, there exists a nuisance parameter: the probability $p \in [0, 1]$ with which each individual is captured.

This model assumes that catching the ith individual is independent of catching the jth individual and then that

$$n^+ \sim \mathcal{B}(N, p)\,.$$

The corresponding likelihood is

$$\ell(N, p|n^+) = \binom{N}{n^+} p^{n^+} (1-p)^{N-n^+} \mathbb{I}_{N \geq n^+} \,.$$

If we use the vague prior

$$\pi_1(N, p) \propto N^{-1} \mathbb{I}_{\mathbb{N}^*}(N) \mathbb{I}_{[0,1]}(p) \,,$$

the posterior distribution of N is

$$\begin{aligned}
\pi_1(N|n^+) &\propto \frac{N!}{(N-n^+)!n^+!} N^{-1} \mathbb{I}_{N \geq n^+} \mathbb{I}_{\mathbb{N}^*}(N) \int_0^1 p^{n^+} (1-p)^{N-n^+} \mathrm{d}p \\
&\propto \frac{(N-1)!}{(N-n^+)!} \frac{(N-n^+)!(n^+ + 1)!}{(N+1)!} \mathbb{I}_{N \geq n^+ \vee 1} \\
&\propto \frac{1}{N(N+1)} \mathbb{I}_{N \geq n^+ \vee 1} \,.
\end{aligned}$$

If we use the (more informative) uniform prior

$$\pi_2(N,p) \propto \mathbb{I}_{\{1,\ldots,S\}}(N)\mathbb{I}_{[0,1]}(p)\,,$$

the posterior distribution of N is

$$\pi_2(N|n^+) \propto \frac{1}{N+1}\mathbb{I}_{S\geq N\geq n^+\vee 1}\,.$$

For the european dipper and the year 1981, we have $n^+ = 22$.

The median of $\pi_1(N|n^+ = 22)$ is equal to 43 (the expectation does not exist).

If we use the ecological information that there cannot be more than $S = 400$ dippers in this region, $\mathbb{E}^{\pi_2}(N|n^+ = 22) \approx 131$.

The Two-Stage Capture–Recapture Model

A logical extension to the capture model is the *capture–mark–recapture* model, which considers two capture periods plus a marking stage:

i) n_1 individuals from a population of size N are "captured", that is, sampled without replacement.

ii) Those individuals are "marked" and they are released into the population.

iii) A second and similar sampling (once again without replacement) is conducted, with n_2 individuals captured.

iv) m_2 individuals out of the n_2's bear the identification mark and are thus characterized as having been captured in both experiments.

If we assume a *closed population*, a fixed population size N throughout the capture experiment, a constant capture probability p for all individuals, and complete independence between individuals and between captures:

$$n_1 \sim \mathcal{B}(N, p)\,,$$

$$m_2|n_1 \sim \mathcal{B}(n_1, p)\,,$$

$$n_2 - m_2|n_1, m_2 \sim \mathcal{B}(N - n_1, p)\,.$$

Let $n^c = n_1 + n_2$ and $n^+ = n_1 + (n_2 - m_2)$ denote the total number of captures over both periods and the total number of captured individuals.

The corresponding likelihood $\ell(N, p|n_1, n_2, m_2)$ is

$$\begin{aligned}
&\binom{N-n_1}{n_2-m_2} p^{n_2-m_2}(1-p)^{N-n_1-n_2+m_2}\mathbb{I}_{\{0,\ldots,N-n_1\}}(n_2-m_2) \\
&\qquad \times \binom{n_1}{m_2} p^{m_2}(1-p)^{n_1-m_2}\binom{N}{n_1} p^{n_1}(1-p)^{N-n_1}\mathbb{I}_{\{0,\ldots,N\}}(n_1) \\
&\qquad \propto \frac{N!}{(N-n_1-n_2+m_2)!} p^{n_1+n_2}(1-p)^{2N-n_1-n_2}\mathbb{I}_{N\geq n^+} \\
&\qquad \propto \binom{N}{n^+} p^{n^c}(1-p)^{2N-n^c}\mathbb{I}_{N\geq n^+}\,,
\end{aligned}$$

which shows that (n^c, n^+) is a sufficient statistic.

If we choose the prior $\pi(N,p) = \pi(N)\pi(p)$ such that $\pi(p)$ is a $\mathcal{U}([0,1])$ density, the conditional posterior distribution on p is such that

$$\pi(p|N,n_1,n_2,m_2) = \pi(p|N,n^c) \propto p^{n^c}(1-p)^{2N-n^c}\,.$$

That is,

$$p|N,n^c \sim \mathcal{B}e(n^c+1, 2N-n^c+1).$$

If $\pi_1(N) \propto \mathbb{I}_{\mathbb{N}^*}(N)$, we get

$$\pi_1(N|n_1,n_2,m_2) = \pi(N|n^c,n^+) \propto \binom{N}{n^+} B(n^c+1, 2N-n^c+1)\mathbb{I}_{N\geq n^+\vee 1}\,,$$

where $B(a,b)$ denotes the beta function. This distribution is called a *beta-Pascal* distribution, but it is not very tractable.

The intractability in the posterior distribution $\pi_1(N|n_1, n_2, m_2)$ is due to the infinite summation resulting from the unbounded support of N.

If we have information about an upper bound S on N and use the corresponding uniform prior,

$$\pi_2(N) \propto \mathbb{I}_{\{1,\ldots,S\}}(N)\,,$$

the posterior distribution of N is thus proportional to

$$\pi_2(N|n^+, n^c) \propto \binom{N}{n^+} \frac{\Gamma(2N - n^c + 1)}{\Gamma(2N + 2)}\, \mathbb{I}_{\{n^+ \vee 1,\ldots,S\}}(N)\,,$$

and, in this case, it is possible to calculate the posterior expectation of N with no approximation error.

For the first two years of the european dippers experiment, which correspond to the first two columns and the first 70 rows of the dataset, $n_1 = 22$, $n_2 = 60$, and $m_2 = 11$.

Hence, $n^c = 82$ and $n^+ = 71$

We get $\mathbb{E}^{\pi_2}(N|n^+ = 71, n^c = 82) \approx 165$.

A simpler model used in capture–recapture settings is the hypergeometric model, also called the *Darroch model*.

This model can be seen as a conditional version of the two-stage model when conditioning on both sample sizes n_1 and n_2:

$$m_2|n_1, n_2 \sim \mathcal{H}(N, n_2, n_1/N)\,. \tag{1}$$

If we choose the uniform prior $\mathcal{U}(\{1,\ldots,400\})$ on N, we get

$$\pi(N|m_2) \propto \binom{N-n_1}{n_2-m_2} \Big/ \binom{N}{n_2} \mathbb{I}_{\{n^+ \vee 1,\ldots,400\}}(N)\,.$$

The posterior expectation can be computed numerically by simple summations.

Rounded posterior expectation of the dipper population size under a uniform prior $\mathcal{U}(\{1,\ldots,400\})$

m_2	0	3	4	7	8	9	10	11	12	15
$\mathbb{E}^{\pi}[N\|m_2]$	355	329	316	252	224	197	172	152	135	101

Rounded posterior expectation of the dipper population size under a uniform prior $\mathcal{U}(\{1,\ldots,S\})$ for $m_2 = 11$

S	100	150	200	250	300	350	400	450	500
$\mathbb{E}^{\pi}[N\|m_2]$	95	125	141	148	151	151	152	152	152

Getting back to the two-stage capture model with probability p of capture, the posterior distribution of (N, p) associated with the vague prior $\pi(N, p) \propto 1/N$ is proportional to

$$\frac{(N-1)!}{(N-n^+)!}\, p^{n^c} (1-p)^{2N-n^c}\,.$$

If $n^+ > 0$, both conditional posterior distributions are standard distributions since

$$\begin{aligned} p|n^c, N &\sim \mathcal{B}e(n^c+1, 2N-n^c+1) \\ N-n^+|n^+, p &\sim \mathcal{N}eg(n^+, 1-(1-p)^2)\,. \end{aligned}$$

Therefore, while the marginal posterior in N is difficult to manage, the joint distribution of (N, p) can be approximated by a Gibbs sampler, as follows:

Initialization: Generate $p^{(0)} \sim \mathcal{U}([0,1])$.

Iteration i ($i \geq 1$):

1. Generate $N^{(i)} - n^+ \sim \mathcal{N}eg(n^+, 1-(1-p^{(i-1)})^2)$.
2. Generate $p^{(i)} \sim \mathcal{B}e(n^c+1, 2N^{(i)} - n^c + 1)$.

The T-Stage Capture–Recapture Model

A further extension to the two-stage capture–recapture model is to consider instead a series of T consecutive captures.

If we denote by n_t the number of individuals captured at period t ($1 \leq t \leq T$) and by m_t the number of recaptured individuals ($m_1 = 0$), under the same assumptions as in the two-stage model, then $n_1 \sim \mathcal{B}(N, p)$ and, conditionally on the $j-1$ previous captures and recaptures ($2 \leq j \leq T$),

$$m_j \sim \mathcal{B}\left(\sum_{t=1}^{j-1}(n_t - m_t), p\right) \quad \text{and} \quad n_j - m_j \sim \mathcal{B}\left(N - \sum_{t=1}^{j-1}(n_t - m_t), p\right).$$

The likelihood $\ell(N, p|n_1, n_2, m_2 \ldots, n_T, m_T)$ is thus

$$\binom{N}{n_1} p^{n_1}(1-p)^{N-n_1} \prod_{j=2}^{T} \left[\binom{N - \sum_{t=1}^{j-1}(n_t - m_t)}{n_j - m_j} p^{n_j - m_j + m_j} \right.$$
$$\left. \times (1-p)^{N - \sum_{t=1}^{j-1}(n_t - m_t)} \binom{\sum_{t=1}^{j-1}(n_t - m_t)}{m_j} (1-p)^{\sum_{t=1}^{j-1}(n_t - m_t) - m_j} \right]$$
$$\propto \frac{N!}{(N-n^+)!} p^{n^c}(1-p)^{TN - n^c} \mathbb{I}_{N \geq n^+}$$

if we denote the sufficient statistics as

$$n^+ = \sum_{t=1}^{T}(n_t - m_t) \quad \text{and} \quad n^c = \sum_{t=1}^{T} n_t \,,$$

the total numbers of captured individuals and captures over the T periods, respectively.

For the uniform prior $\mathcal{U}(\{1,\ldots,S\})$ on N and $\mathcal{U}([0,1])$ on p, the posterior distribution of N is then proportional to

$$\pi(N|n^+,n^c) \propto \binom{N}{n^+} \frac{(TN-n^c)!}{(TN+1)!}\, \mathbb{I}_{\{n^+\vee 1,\ldots,S\}}(N).$$

In the european dippers example, for the whole set of observations, we have $T = 7$, $n^+ = 294$, and $n^c = 519$. Under the uniform prior with $S = 400$, we obtain

$$\mathbb{E}^\pi(N|n^+ = 294, n^c = 82) \approx 373$$

While this value seems dangerously close to the upper bound of 400 on N and thus leads us to suspect a strong influence of the upper bound S, the computation of the posterior expectation for $S = 2500$ gives

$$\mathbb{E}^\pi(N|n^+ = 294, n^c = 82) \approx 374$$

which shows the limited impact of this hyperparameter S.

Open Populations

Moving towards more realistic settings, we now consider the case of an *open population* model, where the population size does not remain fixed over the experiment but, on the contrary, there is a probability q for each individual to leave the population at each time.

We study here a model where only the individuals captured during the first capture experiment are marked and subsequent recaptures are registered.

For three successive capture experiments, we have

$$n_1 \sim \mathcal{B}(N, p)\,, \quad r_1|n_1 \sim \mathcal{B}(n_1, q)\,, \quad r_2|n_1, r_1 \sim \mathcal{B}(n_1 - r_1, q)\,,$$

for the distributions of the first capture population size and of the numbers of individuals who vanished between the first and second, and the second and third experiments, respectively, and

$$c_2|n_1, r_1 \sim \mathcal{B}(n_1 - r_1, p), \quad c_3|n_1, r_1, r_2 \sim \mathcal{B}(n_1 - r_1 - r_2, p)\,,$$

for the number of recaptured individuals during the second and the third experiments.

Here, only n_1, c_2, and c_3 are observed. The numbers of individuals removed at stages 1 and 2, r_1 and r_2, are not observed.

The likelihood $\ell(N, p, q, r_1, r_2 | n_1, c_2, c_3)$ is given by

$$\binom{N}{n_1} p^{n_1} (1-p)^{N-n_1} \binom{n_1}{r_1} q^{r_1} (1-q)^{n_1-r_1} \binom{n_1 - r_1}{c_2} p^{c_2} (1-p)^{n_1-r_1-c_2}$$
$$\times \binom{n_1 - r_1}{r_2} q^{r_2} (1-q)^{n_1-r_1-r_2} \binom{n_1 - r_1 - r_2}{c_3} p^{c_3} (1-p)^{n_1-r_1-r_2-c_3} .$$

If we use the prior $\pi(N,p,q) \propto \mathbb{I}_{\{0,\ldots,S\}}(N)\mathbb{I}_{[0,1]}(p)\mathbb{I}_{[0,1]}(q)$, the associated conditionals are

$$\begin{aligned}
\pi(p|N,q,\mathcal{D}^*) &\propto p^{n^+}(1-p)^{u_+}\mathbb{I}_{[0,1]}(p)\,,\\
\pi(q|N,p,\mathcal{D}^*) &\propto q^{r_1+r_2}(1-q)^{2n_1-2r_1-r_2}\mathbb{I}_{[0,1]}(q)\,,\\
\pi(N|p,q,\mathcal{D}^*) &\propto \frac{N(N-1)!}{(N-n_1)!}(1-p)^N\mathbb{I}_{S\geq N\geq n_1}\,,\\
\pi(r_1|p,q,n_1,c_2,c_3,r_2) &\propto \frac{(n_1-r_1)!\,q^{r_1}(1-q)^{-2r_1}(1-p)^{-2r_1}}{r_1!(n_1-r_1-r_2-c_3)!(n_1-c_2-r_1)!}\,,\\
\pi(r_2|p,q,n_1,c_2,c_3,r_1) &\propto \frac{q^{r_2}[(1-p)(1-q)]^{-r_2}}{r_2!(n_1-r_1-r_2-c_3)!}\,,
\end{aligned}$$

where $\mathcal{D}^* = (n_1,c_2,c_3,r_1,r_2)$ and

$$\begin{aligned}
&u_1 = N-n_1,\ u_2 = n_1-r_1-c_2,\ u_3 = n_1-r_1-r_2-c_3\,,\\
&n^+ = n_1+c_2+c_3,\ u_+ = u_1+u_2+u_3\,.
\end{aligned}$$

Therefore, the full conditionals are

$$\begin{aligned} p|N,q,\mathcal{D}^* &\sim \mathcal{B}e(n^+ + 1, u_+ + 1)\,, \\ q|N,p,\mathcal{D}^* &\sim \mathcal{B}e(r_1 + r_2 + 1, 2n_1 - 2r_1 - r_2 + 1)\,, \\ r_2|p,q,n_1,c_2,c_3,r_1 &\sim \mathcal{B}\left(n_1 - r_1 - c_3, \frac{q}{q + (1-q)(1-p)}\right)\,. \end{aligned}$$