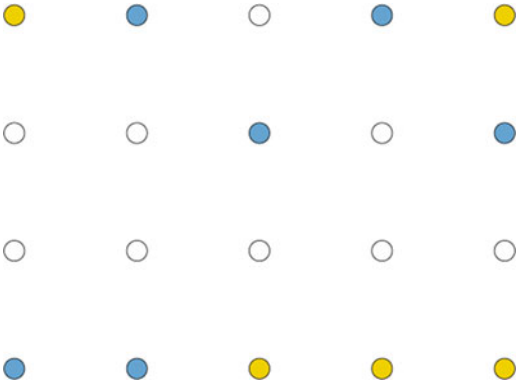
5

Capture–Recapture Experiments



He still couldn't be sure that
he hadn't landed in a trap.
—**Ian Rankin**, ***Resurrection Men***.—

Roadmap

This chapter deals with a very special case of survey models. Surveys are used in many settings to evaluate some features of a given population, including its main characteristic, the *size* of the population. In the case of capture–recapture surveys, individuals are observed and identified either once or several times and the repeated observations can be used to draw inference on the population size and its dynamic characteristics. Along with the original model, we will also introduce extensions that can be seen as a first entry into hidden Markov chain models, detailed further in Chap. 6. In particular, we cover the generic *Arnason–Schwarz* model that is customarily used by biologists for open populations.

On the methodological side, we provide an introduction to the accept–reject method, which is the central simulation technique behind most standard random generators and relates to the Metropolis–Hastings methodology in many ways.

J.-M. Marin and C.P. Robert, *Bayesian Essentials with R*, Springer Texts in Statistics, DOI 10.1007/978-1-4614-8687-9_5,

5.1 Inference in a Finite Population

In this chapter, we consider the problem of estimating the unknown size, N, of a population, based on a *survey*, that is, on a partial observation of this population. To be able to evaluate a population size without going through the enumeration of all its members is obviously very appealing, both timewise and moneywise, especially when sampling those members has a perturbing effect on them.[1]

A primary type of survey (which we do not study in this chapter) is based on knowledge of the structure of the population. For instance, in a political survey about voting intentions, we build a sample of 1,000 individuals, say, such that the main sociological groups (farmers, civil servants, senior citizens, etc.) are represented in proportion in the sample. In that situation, there is no statistical inference, so to speak, except about the variability of the responses, which are in the simplest cases binomial variables.

Obviously, such surveys require primary knowledge of the population, which can be obtained either by a (costly) census, like those that states run every 5 or 10 years, or by a preliminary exploratory survey that aims at uncovering these hidden structures. This secondary type of survey is the purpose of this chapter, under the name of *capture–recapture* (or *capture–mark–recapture*) experiments, where a few individuals sampled at random from the population of interest bring some information about the characteristics of this population and in particular about its size.

The capture–recapture models were first used in biology and ecology to estimate the size of animal populations, such as herds of caribous (e.g., for culling) or of whales (e.g., for the International Whaling Commission to determine fishing quotas), cod populations, and the number of different species in a particular area. 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, where the estimation of the size of populations at risk is always delicate (e.g., homeless people, prostitutes, illegal migrants, drug addicts, etc.);
- official statistics for reducing the cost of a census[2] or improving its efficiency on delicate or rare subcategories (as in the U.S. census undercount procedure and the new French census);
- finance (e.g., in credit scoring, defaulting companies, etc.) and marketing (consumer habits, telemarketing, etc.);

[1] In the most extreme cases, sampling an individual may lead to its destruction, as for instance in forestry when estimating the volume of trees or in meat production when estimating the content of fat in meat.

[2] Even though a census is formally a deterministic process since it aims at the complete enumeration of a given population, it inevitably involves many random components at the selection, collection, and processing levels (Särndal et al., 2003).

– fraud detection (e.g., phone, credit card, etc.) and document authentication (historical documents, forgery, etc.); and
– software debugging, to determine an evaluation of the number of bugs in a computer program.

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, in a computer program, the total number N of bugs is unknown but one can record the number n_1 of bugs detected in a given perusal. Similarly, 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 throughout this chapter is called **eurodip** and is related to a population of birds called *European dippers (Cinclus cinclus)*. These birds are closely dependent on streams, feeding on underwater invertebrates, and their nests are always close to water. The capture–recapture data on the European dipper contained in **eurodip** 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. Birds were at least 1 year old when initially banded. In **eurodip**, 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. For instance, the three lines from **eurodip**

```
1 0 0 0 0 0 0
1 3 0 0 0 0 0
0 2 2 2 1 2 2
```

indicate that the first dipper was only captured the first year in zone 1 and that the second dipper was captured in years 1981 and 1982 and moved from zone 1 to zone 3 between those years. The third dipper was captured every year but 1981 and moved between zones 1 and 2 during the remaining year.

In conclusion, we hope that the introduction above was motivating enough to convince the reader that population sampling models are deeply relevant in statistical practice. Besides, these models also provide an interesting application of Bayesian modeling and in particular they allow for the inclusion of often available prior information.

5.2 Sampling Models

5.2.1 The Binomial Capture Model

We start with the simplest model of all, namely the independent observation or *capture*[3] of n^+ individuals from a population of size N. For instance, a trap is positioned on a rabbit track for five hours and n^+ rabbits are found in the trap. While the population size $N \in \mathbb{N}$ is the parameter of interest, there exists a nuisance parameter, namely 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.) For this model,

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

and the corresponding likelihood is

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

Obviously, with a single observation n^+, we cannot say much on (N,p), but the posterior distribution is still well-defined. For instance, if we use the vague prior

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

the posterior distribution of N is

$$\begin{aligned}\pi(N|n^+) &\propto \frac{N!}{(N-n^+)!n^+!}\,N^{-1}\mathbb{I}_{N\geq n^+}\mathbb{I}_{\mathbb{N}^*}(N)\int_0^1 p^{n^+}(1-p)^{N-n^+}\mathrm{d}p\\ &\propto \frac{(N-1)!}{(N-n^+)!}\,\frac{(N-n^+)!}{(N+1)!}\,\mathbb{I}_{N\geq n^+\vee 1}\\ &= \frac{1}{N(N+1)}\,\mathbb{I}_{N\geq n^+\vee 1}\,, \end{aligned} \tag{5.1}$$

where $n^+ \vee 1 = \max(n^+,1)$. Note that this posterior distribution is defined even when $n^+ = 0$. If we use the (more informative) uniform prior

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

the posterior distribution of N is

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

[3] We use the original terminology of *capture* and *individuals*, even though the sampling mechanism may be far from genuine capture, as in whale sightseeing or software bug detection.

For illustrative purposes, consider the case of year 1981 in **eurodip** (which is the first column in the file):

```
> data(eurodip)
> year81=eurodip[,1]
> nplus=sum(year81>0)
[1] 22
```

where $n^+ = 22$ dippers were thus captured. By using the binomial capture model and the vague prior $\pi(N, p) \propto N^{-1}$, the number of dippers N can be estimated by the posterior median. (Note that the mean of (5.1) does not exist, no matter what n^+ is.)

```
> N=max(nplus,1)
> rangd=N:(10^4*N)
> post=1/(rangd*(rangd+1))
> 1/sum(post)
[1] 22.0022
> post=post/sum(post)
> min(rangd[cumsum(post)>.5])
[1] 43
```

For this year 1981, the estimate of N is therefore 43 dippers. (See Exercise 5.1 for theoretical justifications as to why the sum of the probabilities is equal to n^+ and why the median is exactly $2n^+ - 1$.) If we use the ecological information that there cannot be more than 400 dippers in this region, we can take the prior $\pi(N, p) \propto \mathbb{I}_{\{1,\ldots,400\}}(N)\mathbb{I}_{[0,1]}(p)$ and estimate the number of dippers N by its posterior expectation:

```
> pbino=function(nplus){
+    prob=c(rep(0,max(nplus,1)-1),1/(max(nplus,1):400+1))
+    prob/sum(prob)
+    }
> sum((1:400)*pbino(nplus))
[1] 130.5237
```

5.2.2 The Two-Stage Capture–Recapture Model

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

1. n_1 individuals from a population of size N are "captured", that is, sampled without replacement.
2. Those individuals are "marked", that is, identified by a numbered tag (for birds and fishes), a collar (for mammals), or another device (like the Social Security number for homeless people or a picture for whales), and they are then released into the population.

3. A second and similar sampling (once again without replacement) is conducted, with n_2 individuals captured.
4. 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* (that is, 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, we end up with a product of binomial models,

$$n_1 \sim \mathscr{B}(N,p)\,, \quad m_2|n_1 \sim \mathscr{B}(n_1,p)\,,$$

and

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

If

$$n^c = n_1 + n_2 \quad \text{and} \quad n^+ = n_1 + (n_2 - m_2)$$

denote the total number of captures over both periods and the total number of captured individuals, respectively, 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 $\mathscr{U}([0,1])$ 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,

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

Unfortunately, the marginal posterior distribution of N is more complicated. For instance, if $\pi(N) = \mathbb{I}_{\mathbb{N}^*}(N)$, it satisfies

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

where $B(a,b)$ denotes the beta function This distribution is called a *beta-Pascal* distribution, but it is not very tractable. The same difficulty occurs if $\pi(N) = N^{-1}\mathbb{I}_{\mathbb{N}^*}(N)$.

The intractability in the posterior distribution $\pi(N|n_1,n_2,m_2)$ is due to the infinite summation resulting from the unbounded support of N. A feasible approximation is to replace the missing normalizing factor by a finite sum with a large enough bound on N, the bound being determined by a lack of perceivable impact on the sum. But the approximation errors due to the computations of terms such as $\binom{N}{n^+}$ or $B(n^c+1, 2N-n^c+1)$ can become a serious problem when n^+ is large. However,

```
> prob=lchoose((471570:10^7),471570)+lgamma(2*(471570:10^7)-
+ 582681+1)-lgamma(2*(471570:10^7)+2)
> range(prob)
[1] -7886469 -7659979
```

shows that relatively large populations are manageable.

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

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

the posterior distribution of N is thus proportional to

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

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

For the first 2 years of the **eurodip** 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 $n^+ = 71$. Therefore, within the frame of the two-stage capture–recapture model[4] and the uniform prior $\mathcal{U}(\{1,\ldots,400\}) \times \mathcal{U}([0,1])$ on (N,p), the posterior expectation of N is derived as follows:

```
> n1=sum(eurodip[,1]>0)
> n2=sum(eurodip[,2]>0)
> m2=sum((eurodip[,1]>0) & (eurodip[,2]>0))
> nc=n1+n2
> nplus=nc-m2
> pcapture=function(T,nplus,nc){
+ #T is the number of capture episodes
+   lprob=lchoose(max(nplus,1):400,nplus)+
      lgamma(T*max(nplus,1):400-nc+1)-
+     lgamma(T*max(nplus,1):400+2)
+   prob=c(rep(0,max(nplus,1)-1),exp(lprob-max(lprob)))
```

[4] This analysis is based on the assumption that all birds captured in the second year were already present in the population during the first year.

```
+   prob/sum(prob)
+   }
> sum((1:400)*pcapture(2,nplus,nc))
[1] 165.2637
```

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 since (see Exercise 5.3)

$$m_2|n_1, n_2 \sim \mathscr{H}(N, n_2, n_1/N)\,, \tag{5.2}$$

the hypergeometric distribution If we choose the uniform prior $\mathscr{U}(\{1, \ldots, 400\})$ on N, the posterior distribution of N is thus

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

and posterior expectations can be computed numerically by simple summations.

For the first 2 years of the **eurodip** dataset and $S = 400$, the posterior distribution of N for the Darroch model is given by

$$\pi(N|m_2) \propto (n-n_1)!(N-n_2)!/\{(n-n_1-n_2+m_2)!N!\}\, \mathbb{I}_{\{71,\ldots,400\}}(N)\,,$$

the normalization factor being the inverse of

$$\sum_{k=71}^{400} (k-n_1)!(k-n_2)!/\{(k-n_1-n_2+m_2)!k!\}\,.$$

We thus have a closed-form posterior distribution and the posterior expectation of N is given by

```
pdarroch=function(n1,n2,m2){

  prob=c(rep(0,max(n1+n2-m2,1)-1),
       choose(n1,m2)*choose(max((n1+n2-m2),1):400-n1,n2-m2)/
       choose(max((n1+n2-m2),1):400,n2))
  prob/sum(prob)
  }
> sum((1:400)*pdarroch(n1,n2,m2))
[1] 137.5962
```

```
Table~5.1 shows the evolution of this
posterior expectation for different values of m_2,
obtained by

> for (i in 6:16) print(round(sum(pdarroch(n1,n2,i)*1:400)))
[1] 277
[1] 252
[1] 224
[1] 197
[1] 172
[1] 152
[1] 135
[1] 122
[1] 111
[1] 101
[1] 94
```

The number of recaptures is thus highly influential on the estimate of N. In parallel, Table 5.2 shows the evolution of the posterior expectation for different values of S (taken equal to 400 in the above). When S is large enough, say larger than $S = 250$, the estimate of N is quite stable, as expected.

Table 5.1. Dataset **eurodip**: Rounded posterior expectation of the dipper population size, N, under a uniform prior $\mathcal{U}(\{1, \ldots, 400\})$

m_2	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$\mathbb{E}^\pi[N\|m_2]$	355	349	340	329	316	299	277	252	224	197	172	152	135	122	110	101

Table 5.2. Dataset **eurodip**: Rounded posterior expectation of the dipper population size, N, 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

Leaving the Darroch model and getting back to the two-stage capture model with probability p of capture, the posterior distribution of (N, p) associated with the noninformative prior $\pi(N, p) = 1/N$ is proportional to

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

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

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

the latter being a negative binomial distribution. Indeed, as a function of N,

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

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:

Algorithm 5.8 Two-stage Capture–Recapture Gibbs Sampler

Initialization: Generate $p^{(0)} \sim \mathscr{U}([0,1])$.
Iteration i $(i \geq 1)$:
1. Generate $N^{(i)} - n^+ \sim \mathscr{N}eg(n^+, 1-(1-p^{(i-1)})^2)$.
2. Generate $p^{(i)} \sim \mathscr{B}e(n^c+1, 2N^{(i)}-n^c+1)$.

5.2.3 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. In that case, 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 (with the convention that $m_1 = 0$), under the same assumptions as in the two-stage model, then $n_1 \sim \mathscr{B}(N,p)$ and, conditionally on the $j-1$ previous captures and recaptures $(2 \leq j \leq T)$,

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

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

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

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 a noninformative prior such as $\pi(N,p) = 1/N$, the joint posterior satisfies

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

Therefore, the conditional posterior distribution of p is

$$p|N,n^+,n^c \sim \mathscr{B}e(n^c+1, TN-n^c+1)$$

and the marginal posterior distribution of N

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

is computable. Note that the normalization coefficient can also be approximated by summation with an arbitrary precision unless N and n^+ are very large.

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

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

For the whole set of observations in **eurodip**, we have $T = 7$, $n^+ = 294$, and $n^c = 519$. Under the uniform prior with $S = 400$, the posterior expectation of N is given by

```
> sum((1:400)*pcapture(7,294,519))
[1] 372.7384
```

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 $S = 2500$

```
> S=2500;T=7;nplus=294;nc=519
> lprob=lchoose(max(nplus,1):S,nplus)+
+   lgamma(T*max(nplus,1):S-nc+1)-lgamma(T*max(nplus,1):S+2)
> prob=c(rep(0,max(nplus,1)-1),exp(lprob-max(lprob)))
> sum((1:S)*prob)/sum(prob)
[1] 373.9939
```

leads to 373.99, which shows the limited impact of this hyperparameter S.

Using even a slightly more advanced sampling model may lead to genuine computational difficulties. For instance, consider a heterogeneous capture–recapture model where the individuals are captured at time $1 \le t \le T$ with probability p_t and where both the size N of the population and the probabilities p_t are unknown. The corresponding likelihood is

$$\ell(N, p_1, \ldots, p_T | n_1, n_2, m_2 \ldots, n_T, m_T) \propto \frac{N!}{(N-n^+)!} \prod_{t=1}^{T} p_t^{n_t} (1-p_t)^{N-n_t}.$$

If the associated prior on $(N, p_1, \ldots, p_T)$ is such that

$$N \sim \mathscr{P}(\lambda)$$

and $(1 \le t \le T)$,

$$\alpha_t = \log\left(\frac{p_t}{1-p_t}\right) \sim \mathscr{N}(\mu_t, \sigma^2),$$

where both σ^2 and the μ_t's are known,[5] the posterior distribution satisfies

$$\begin{aligned} \pi(\alpha_1, \ldots, \alpha_T, N|, n_1, \ldots, n_T) \;\propto\; & \frac{N!}{(N-n^+)!} \frac{\lambda^N}{N!} \prod_{t=1}^{T} (1+e^{\alpha_t})^{-N} \\ & \times \prod_{t=1}^{T} \exp\left\{\alpha_t n_t - \frac{1}{2\sigma^2}(\alpha_t - \mu_t)^2\right\}. \end{aligned} \tag{5.3}$$

It is thus much less manageable from a computational point of view, especially when there are many capture episodes. A corresponding Gibbs sampler could simulate easily from the conditional posterior distribution on N since

$$N - n^+ | \alpha, n^+ \sim \mathscr{P}\left(\lambda \prod_{t=1}^{T} (1+e^{\alpha_t})\right),$$

but the conditionals on the α_t's $(1 \le t \le T)$ are less conventional,

$$\alpha_t | N, \mathbf{n} \sim \pi_t(\alpha_t | N, \mathbf{n}) \propto (1+e^{\alpha_t})^{-N} e^{\alpha_t n_t - (\alpha_t - \mu_t)^2/2\sigma^2},$$

and they require either an accept–reject algorithm (Sect. 5.4) or a Metropolis–Hastings algorithm in order to be simulated.

For the prior

$$\pi(N, p) \propto \frac{\lambda^N}{N!} \mathbb{I}_{\mathbb{N}}(N) \mathbb{I}_{[0,1]}(p),$$

the conditional posteriors are then

[5] This assumption can be justified on the basis that each capture probability is only observed once on the tth round (and so cannot reasonably be associated with a noninformative prior).

$$p|N, n^c \sim \mathscr{B}e(n^c + 1, TN - n^c + 1) \quad \text{and} \quad N - n^+|p, n^+ \sim \mathscr{P}(\lambda(1-p)^T)$$

and a Gibbs sampler similar to the one developed in Algorithm 5.8 can easily be implemented, for instance via the code

```
> lambda=200
> nsimu=10^4
> p=rep(1,nsimu); N=p
> N[1]=2*nplus
> p[1]=rbeta(1,nc+1,T*lambda-nc+1)
> for (i in 2:nsimu){
+  N[i]=nplus+rpois(1,lambda*(1-p[i-1])^T)
+  p[i]=rbeta(1,nc+1,T*N[i]-nc+1)
+  }
```

For **eurodip**, we used this Gibbs sampler and obtained the results illustrated by Fig. 5.1. When the chain is initialized at the (unlikely) value $N^{(0)} = \lambda = 200$ (which is the prior expectation of N), the stabilization of the chain is quite clear: It only takes a few iterations to converge toward the proper region that supports the posterior distribution. We can thus visually confirm the convergence of the algorithm and approximate the Bayes estimators of N and p by the Monte Carlo averages

```
> mean(N)
[1] 326.9831
> mean(p)
[1] 0.2271828
```

The precision of these estimates can be assessed as in a regular Monte Carlo experiment, but the variance estimate is biased because of the correlation between the simulations. A simple way to assess this effect is to call R function `acf()` for each component θ_i of the parameter, as

$$\nu = 1 + 2\sum_{t=1}^{\infty} \text{cor}(\theta_i^{(1)}, \theta_i^{(t+1)})$$

evaluates the loss of efficiency due to the correlation. The corresponding *effective sample size*, given by $T_{\text{ess}} = T/\nu$, provides the equivalent size of an iid sample. For instance,

```
> 1/(1+2*sum(acf(N)$acf[-1]))
[1] 0.599199
> 1/(1+2*sum(acf(p)$acf[-1]))
[1] 0.6063236
```

shows that the current Gibbs sampler offers an efficiency of 60% compared with an iid sample from the posterior distribution.

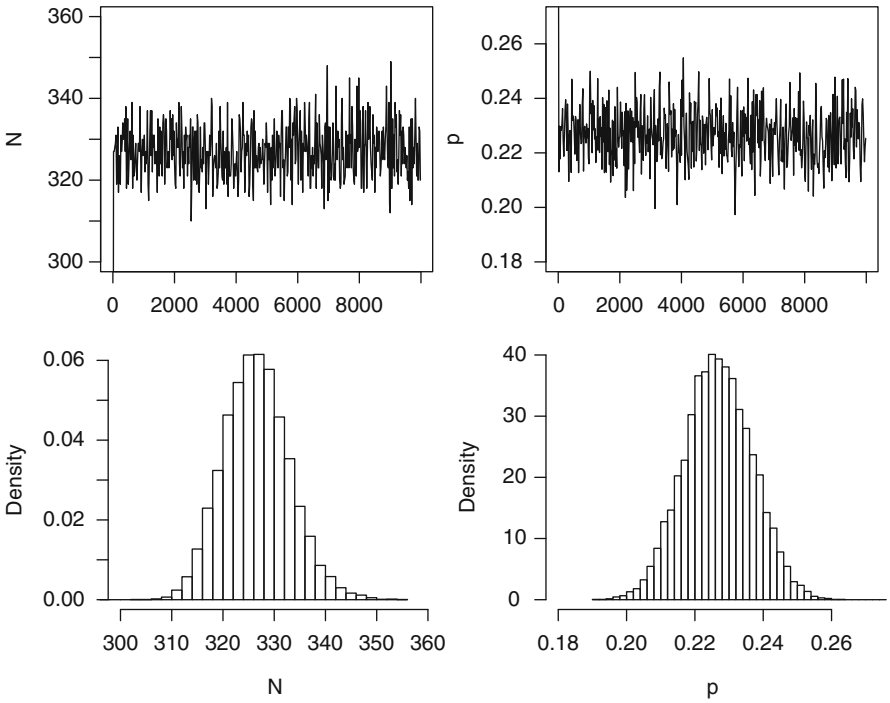


Fig. 5.1. Dataset **eurodip**: Representation of the Gibbs sampling output for the parameters p (*first column*) and N (*second column*)

5.3 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 (or, more accurately, between any two capture episodes). Given that the associated likelihood involves unobserved indicators (namely, indicators of survival; see Exercise 5.14), we study here a simpler model where only the individuals captured during the first capture experiment are marked and subsequent recaptures are registered. For three successive capture experiments, we thus have

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

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

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

for the number of recaptured individuals during the second and the third experiments, respectively. 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 available and must therefore be simulated, as well as the parameters N, p, and q.[6] The likelihood $\ell(N,p,q,r_1,r_2|n_1,c_2,c_3)$ is given by

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

and, if we use the prior $\pi(N,p,q) \propto N^{-1}\mathbb{I}_{[0,1]}(p)\mathbb{I}_{[0,1]}(q)$, the associated conditionals are

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

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

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

(u stands for *unobserved*, even though these variables can be computed conditional on the remaining unknowns). Therefore, the full conditionals are

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

which are very easily simulated, while r_1 has a less conventional distribution. However, this difficulty is minor since, in our case, n_1 is not extremely

[6] From a theoretical point of view, r_1 and r_2 are *missing variables* rather than true parameters. This obviously does not change anything either for simulation purposes or for Bayesian inference.

large. It is thus possible to compute the probability that r_1 is equal to each of the values in $\{0, 1, \ldots, \min(n_1 - r_2 - c_3, n_1 - c_2)\}$. This means that the corresponding Gibbs sampler can be implemented as well.

```
gibbscap1=function(nsimu,n1,c2,c3){

N=p=q=r1=r2=rep(0,nsimu)
N[1]=round(n1/runif(1))
r1[1]=max(c2,c3)+round((n1-c2)*runif(1))
r2[1]=round((n1-r1[1]-c3)*runif(1))
nplus=n1+c2+c3
for (i in 2:nsimu){
  uplus=N[i-1]-r1[i-1]-c2+n1-r1[i-1]-r2[i-1]-c3
  p[i]=rbeta(1,nplus+1,uplus+1)
  q[i]=rbeta(1,r1[i-1]+r2[i-1]+1,2*n1-2*r1[i-1]-r2[i-1]+1)
  N[i]=n1+rnbinom(1,n1,p[i])
  rbar=min(n1-r2[i-1]-c3,n1-c2)
  pq=q[i]/((1-q[i])*(1-p[i]))^2
  pr=lchoose(n1-c2,0:rbar)+(0:rbar)*log(pq)+
     lchoose(n1-(0:rbar),r2[i-1]+c3)
  r1[i]=sample(0:rbar,1,prob=exp(pr-max(pr)))
  r2[i]=rbinom(1,n1-r1[i]-c3,q[i]/(q[i]+(1-q[i])*(1-p[i])))
  }
  list(N=N,p=p,q=q,r1=r1,r2=r2)
}
```

We stress that R is quite helpful in simulating from unusual distributions and in particular from those with finite support. For instance, the conditional distribution of r_1 above can be simulated using the following representation of $\mathbb{P}(r_1 = k|p, q, n_1, c_2, c_3, r_2)$ $(0 \le k \le \overline{r} = \min(n_1 - r_2 - c_3, n_1 - c_2))$,

$$\binom{n_1 - c_2}{k} \left\{ \frac{q}{(1-q)^2(1-p)^2} \right\}^k \binom{n_1 - k}{r_2 + c_3}, \tag{5.4}$$

up to a normalization constant, since the binomial coefficients and the power in k can be computed for all values of k at once, thanks to the matrix capabilities of R, through the command `lchoose`. The above quantity corresponding to

```
pr=lchoose(n=n1 - c2,k=0:r̄) + (0:r̄)*log(q1)
          + lchoose(n=n1-(0:r̄),k=r2 + c3)
```

is the whole vector of the log-probabilities, with $q_1 = q/(1-q)^2(1-p)^2$.

⚡ In most computations, it is safer to use logarithmic transforms to reduce the risk of running into overflow or underflow error messages. For instance, in the example above, the probability vector can be recovered by

```
pr=exp(pr-max(pr))/sum(exp(pr-max(pr)))
```

while a direct computation of `exp(pr)` may well produce an `Inf` value that invalidates the remaining computations.[7]

Once the probabilities are transformed as in the previous R code, a call to the R command

```
> sample(0:mm,n,prob=exp(pr-max(pr)))
```

is sufficient to provide n simulations of r_1. The production of a large Gibbs sample is immediate:

```
> system.time(gibbscap1(10^5,22,11,6))
   user  system elapsed
 12.816   0.000  12.830
```

Even a large value such as $n_1 = 1612$ used below does not lead to computing difficulties since we can run 10,000 iterations of the corresponding Gibbs sampler in a few seconds on a laptop:

```
> system.time(gibbscap1(10^4,1612,811,236))
   user  system elapsed
 10.245   0.028  10.294
```

For **eurodip**, we have $n_1 = 22$, $c_2 = 11$, and $c_3 = 6$. We obtain the Gibbs output

```
> gg=gibbscap1(10^5,22,11,6)
```

summarized in Fig. 5.2. The sequences for all components are rather stable and their mixing behavior (i.e., the speed of exploration of the support of the target) is satisfactory, even though we can still detect a trend in the first three rows. Since r_1 and r_2 are integers with only a few possible values, the last two rows show apparently higher jumps than the three other parameters. The MCMC approximations to the posterior expectations of N and p are equal

```
> mean(gg$N)
[1] 57.52955
> mean(gg$p)
[1] 0.3962891
```

respectively.

Given the large difference between n_1 and c_2 and the proximity between c_2 and c_3, high values of q are rejected, and the difference can be attributed

[7] This recommendation also applies to the computation of likelihoods that tend to take absolute values that exceed the range of the computer representation of real numbers, while only the relative values are relevant for Bayesian computations. Using a transform such as `exp(loglike-max(loglike))` thus helps in reducing the risk of overflows.

with high likelihood to a poor capture rate. One should take into account the fact that there are only three observations for a model that involves three true parameters plus two missing variables. Figure 5.3 gives another insight into the posterior distribution by representing the joint distribution of the sample of (r_1, r_2)'s

```
> plot(jitter(gg$r1,factor=1),jitter(g2$r2,factor=1),cex=0.5,
+ xlab=expression(r[1]),ylab=expression(r[2]))
```

using for representation purposes the R function jitter(), which moves each point by a tiny random amount. There is a clear positive correlation between r_1 and r_2, despite the fact that r_2 is simulated on an $(n_1 - c_3 - r_1)$ scale. The mode of the posterior is $(r_1, r_2) = (0, 0)$, which means that it is likely that no dipper died or left the observation area over the 3-year period.

5.4 Accept–Reject Algorithms

In Chap. 2, we mentioned standard random number generators used for the most common distributions and presented importance sampling (Algorithm 2.2) as a possible alternative when such generators are not available. While MCMC algorithms always offer a solution when facing nonstandard distributions, there often exists a possibility that is in fact used in most of the standard random generators and which we now present. It also relates to the independent Metropolis–Hastings algorithm of Sect. 4.2.2.

Given a density g that is defined on an arbitrary space (of any dimension), a fundamental identity is that simulating X distributed from $g(x)$ is completely equivalent to simulating (X, U) uniformly distributed on the set

$$\mathscr{S} = \{(x, u) : 0 < u < g(x)\}$$

(this is called the *Fundamental Theorem of Simulation* in Robert and Casella, 2004, Chap. 3). The reason for this equivalence is simply that

$$\int_0^\infty \mathbb{I}_{0<u<g(x)}\, \mathrm{d}u = g(x)\,.$$

Since $\mathscr{S}$ usually has complex features, direct simulation from the uniform distribution on $\mathscr{S}$ is most often impossible (Exercise 5.16). The idea behind the accept–reject method is to find a simpler set $\mathscr{G}$ that contains $\mathscr{S}$, $\mathscr{S} \subset \mathscr{G}$, and then to simulate uniformly on this set $\mathscr{G}$ until the value belongs to $\mathscr{S}$. In practice, this means that one needs to find an upper bound on g; that is, another density f and a constant M such that

$$g(x) \leq M f(x) \tag{5.5}$$

on the support of the density g. (Note that $M > 1$ necessarily.) Implementing the following algorithm then leads to a simulation from g.

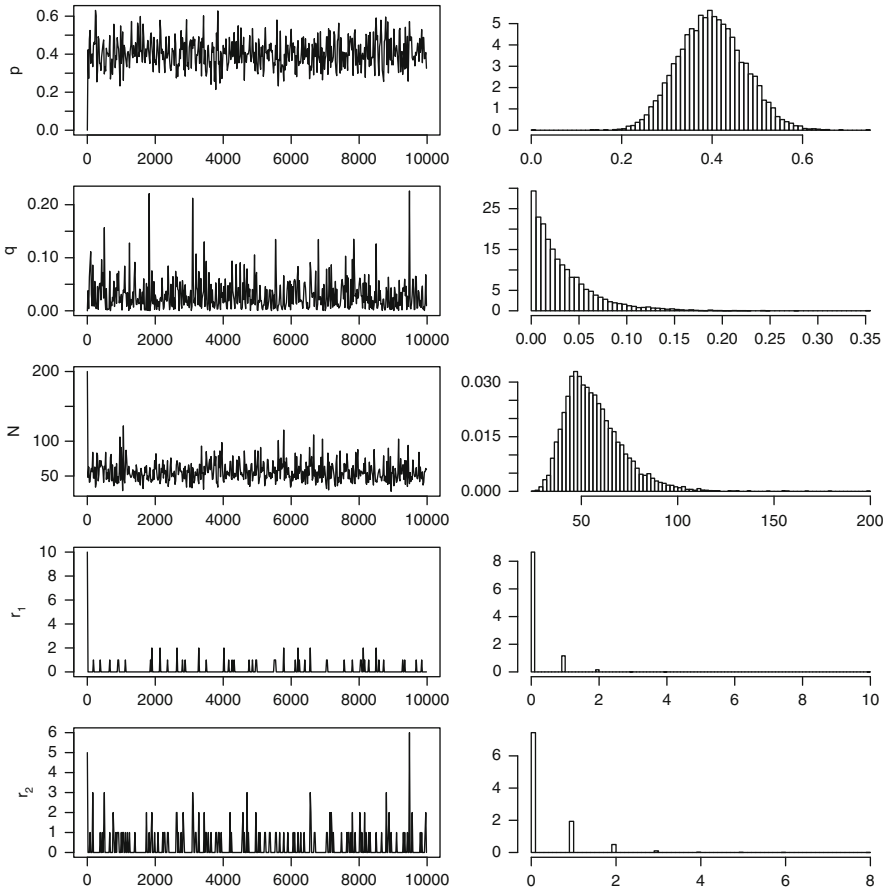


Fig. 5.2. Dataset **eurodip**: Representation of the Gibbs sampling output for the five parameters of the open population model, based on 10,000 iterations, with raw plots *(first column)* and histograms *(second column)*

Algorithm 5.9 Accept–Reject Sampler

1. Generate $X \sim f$, $U \sim \mathcal{U}_{[0,1]}$.
2. Accept $Y = x$ if $u \leq g(x)/(Mf(x))$.
3. Return to 1 otherwise.

This method provides a random generator for densities g that are known up to a multiplicative factor, which is a feature that occurs particularly often in Bayesian calculations since the posterior distribution is usually specified up to a normalizing constant.

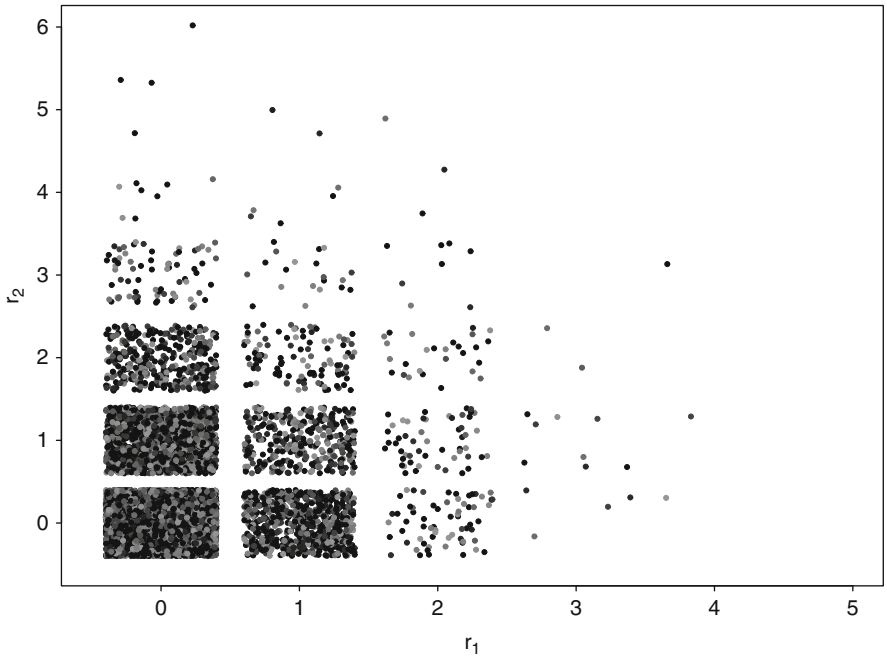


Fig. 5.3. Dataset **eurodip**: Representation of the Gibbs sampling output of the (r_1, r_2)'s by a jitterplot: to translate the density of the possible values of (r_1, r_2) on the $\mathbb{N}^2$ grid, each simulation has been randomly moved using the R jitter procedure and colored at random using *grey levels* to help distinguish the various simulations

For the open population model, we found the full conditional distribution of r_1 to be rather non-standard, as shown by (5.4). Rather than using an exhaustive enumeration of all probabilities $\mathbb{P}(m_1 = k) = g(k)$ and then sampling from this distribution, we can instead try to use a proposal based on a binomial upper bound. Take for instance f that corresponds to the binomial distribution $\mathscr{B}(\bar{r}, q_2)$ with

$$q_2 = q/\{q + (1-q)^2(1-p)^2\}\,.$$

The ratio $g(k)/f(k)$ is proportional to

$$\frac{\binom{n_1-c_2}{k}\binom{n_1-k}{r_2+c_3}}{\binom{\bar{r}}{k}} \propto \frac{(n_1-k)!}{(\max(n_1-c_2, n_1-r_2-c_3)-k)!}\,,$$

which is decreasing in k. The ratio is therefore bounded by

$$\frac{\binom{n_1-c_2}{0}\binom{n_1-0}{r_2+c_3}}{\binom{\bar{r}}{0}} = \frac{(n_1-c_2)!}{(r_2+c_3)!(n_1-r_2-c_3)!}$$

(up to the same normalizing constant). Note that this is *not* the constant M introduced in Algorithm 5.9 because we use unnormalized densities (the bound M may therefore also depend on q_2). Therefore we cannot derive the average acceptance rate from this ratio and we have to use a Monte Carlo experiment to check whether or not the method is really efficient (see Exercise 5.20).

If we use the values from **eurodip**—that is, $n_1 = 22$, $c_2 = 11$ and $c_3 = 6$, with $r_2 = 1$ and $q_1 = 0.1$—, we can use R functions like

```
thresh=function(k,n1,c2,c3,r2,barr){
  choose(n1-c2,k)*choose(n1-k,c3+r2)/choose(barr,k)
  }

ardipper=function(nsimu=1,n1,c2,c3,r2,q2){

  barr=min(n1-c2,n1-r2-c3)
  boundM=thresh(0,n1,c2,c3,r2,barr)
  echan=1:nsimu
  for (i in 1:nsimu){
    test=TRUE
    while (test){
      y=rbinom(1,size=barr,prob=q2)
      test=(runif(1)>thresh(y,n1,c2,c3,r2,barr))
      }
    echan[i]=y
    }
  echan
}
```

the average of the acceptance ratios $g(k)/Mf(k)$ is equal to 0.12. This is a relatively small value since it corresponds to a rejection rate of about 9/10. The simulation process could thus be a little slow, although

```
> system.time(ardipper(10^5,n1=22,c2=11,c3=6,r2=1,q1=.1))
   user  system elapsed
  8.148   0.024   8.1959
```

shows this is not the case. (Note that the code `ardipper` provided here does not produce the rejection rate. It has to be modified for this purpose.) An histogram of accepted values is shown in Fig. 5.4.

Obviously, this method is not hassle-free. For complex densities g, it may prove impossible to find a density f such that $g(x) \leq Mf(x)$ and M is small enough. However, there exists a large class of univariate distributions for which a generic choice of f is possible (see Robert and Casella, 2004, Chap. 2).

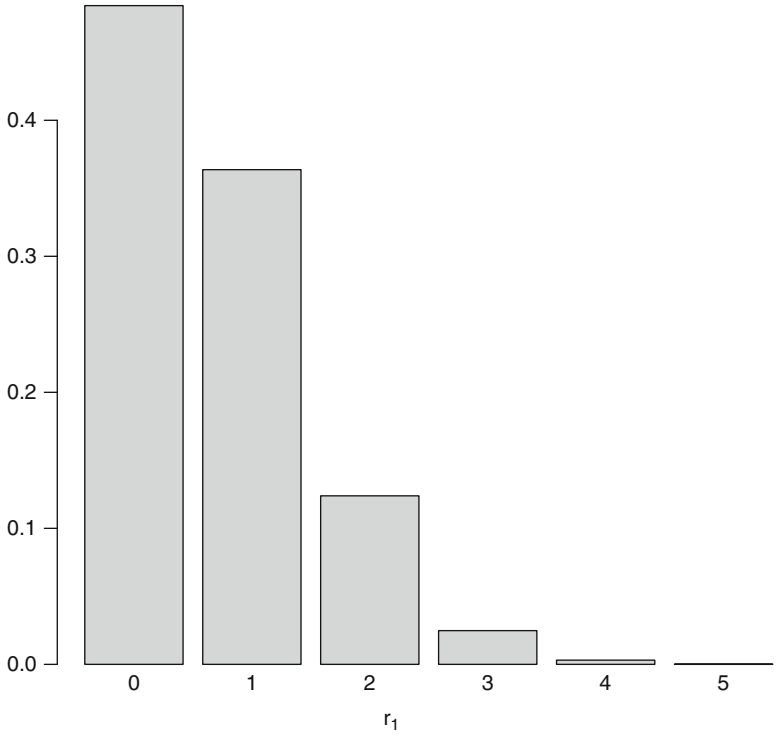


Fig. 5.4. Dataset **eurodip**: Sample from the distribution (5.4) obtained by accept–reject and based on the simulation of 10,000 values from a $\mathscr{B}(n_1, q_1)$ distribution for $n_1 = 22$, $c_2 = 11$, $c_3 = 6$, $r_2 = 1$, and $q_1 = 0.1$

5.5 The Arnason–Schwarz Capture–Recapture Model

We consider in this final section a more advanced capture–recapture model based on the realistic assumption that, in most capture–recapture experiments, we can tag individuals one by one; that is, we can distinguish each individual at the time of its first capture and thus follow its capture history. For instance, when tagging mammals and birds, differentiated tags can be used, so that there is only *one* individual with tag, say, 23131932.[8]

The *Arnason–Schwarz model* thus considers a capture–recapture experiment as a collection of individual histories. For each individual that has been

[8] In a capture–recapture experiment used in Dupuis (1995), a population of lizards was observed in the south of France (Lozère). When it was found that plastic tags caused necrosis on those lizards, the biologists in charge of the experiment decided to cut a phalange of one of the fingers of the captured lizards to identify them later. While the number of possibilities, 2^{20}, is limited, it is still much larger than the number of captured lizards in this study. Whether or not the lizards appreciated this ability to classify them is not known.

captured at least once during the experiment, individual characteristics of interest are registered at each capture. For instance, this may include location, weight, sexual status, pregnancy occurrence, social status, and so on. The probabilistic modeling includes this categorical decomposition by adding what we will call *movement* probabilities to the survival probabilities already used in the Darroch open population model of Sect. 5.2.2. From a theoretical point of view, this is a first example of a (partially) hidden Markov model, a structure studied in detail in Chap. 7. In addition, the model includes the possibility that individuals vanish from the population between two capture experiments. (This is thus another example of an open population model.)

As in **eurodip**, the interest that drives the capture–recapture experiment may be to study the movements of individuals within a zone $\mathfrak{K}$ divided into $k = 3$ strata denoted by $1, 2, 3$. (This structure is generic: Zones are not necessarily geographic and can correspond to anything from social status, to HIV stage, to university degree.) For instance, four consecutive rows of possible **eurodip** (individual) capture–recapture histories look as follows:

45	0 3 0 0 0 0 0
46	0 2 2 2 2 1 1
47	0 2 0 0 0 0 0
48	2 1 2 1 0 0 0

where 0 denotes a failure to capture. This means that, for dipper number 46, the first location was not observed but this dipper was captured for all the other experiments. For dippers number 45 and 47, there was no capture after the second time and thus one or both of them could be dead (or outside the range of the capture area) at the time of the last capture experiment. We also stress that the Arnason–Schwarz model often assumes that individuals that were not part of the population on the first capture experiments can be identified as such.[9] We thus have *cohorts* of individuals that entered the study in the first year, the second year, and so on.

5.5.1 Modeling

A description of the basic Arnason–Schwarz model involves two types of variables for each individual i ($i = 1, \ldots, n$) in the population: first, a variable that describes the location of this individual,

$$\mathbf{z}_i = (z_{(i,t)}, t = 1, .., \tau)\,,$$

where τ is the number of capture periods; and, second, a binary variable that describes the capture history of this individual,

$$\mathbf{x}_i = (x_{(i,t)}, t = 1, .., \tau)\,.$$

[9] This is the case, for instance, with newborns or new mothers in animal capture experiments.

We[10] assume that $z_{(i,t)} = r$ means the animal i is alive in stratum r at time t and that $z_{(i,t)} = \dagger$ denotes the case when the animal i is dead at time t. The variable $\mathbf{z}_i$ is sometimes called the *migration* process of individual i by analogy with the special case where one is considering animals moving between geographical zones, like some northern birds in spring and fall. Note that $\mathbf{x}_i$ is entirely observed, while $\mathbf{z}_i$ is not. For instance, we may have

$$\mathbf{x}_i = 1\,1\,0\,1\,1\,1\,0\,0\,0$$

and

$$\mathbf{z}_i = 1\,2\,\cdot\,3\,1\,1\,\cdot\cdot\cdot,$$

for which a possible completed $\mathbf{z}_i$ is

$$\mathbf{z}_i = 1\,2\,1\,3\,1\,1\,2\,\dagger\,\dagger,$$

meaning that the animal died between the seventh and the eighth capture events. In particular, the Arnason–Schwarz model assumes that dead animals are never observed (although this type of assumption can easily be modified when processing the model, in what are called *tag-recovery experiments*). Therefore $z_{(i,t)} = \dagger$ always corresponds to $x_{(i,t)} = 0$.

Moreover, we assume that the $(\mathbf{x}_i, \mathbf{z}_i)$'s $(i = 1, \ldots, n)$ are independent and that each random vector $\mathbf{z}_i$ is a Markov chain taking values in $\mathfrak{K} \cup \{\dagger\}$ with uniform initial probability on $\mathfrak{K}$ (unless there is prior information to the contrary). The parameters of the Arnason–Schwarz model are thus of two kinds: the capture probabilities

$$p_t(r) = \mathbb{P}\left(x_{(i,t)} = 1 | z_{(i,t)} = r\right)$$

on the one hand and the transition probabilities

$$q_t(r,s) = \mathbb{P}\left(z_{(i,t+1)} = s | z_{(i,t)} = r\right) \quad r \in \mathfrak{K}, s \in \mathfrak{K} \cup \{\dagger\}, \quad q_t(\dagger,\dagger) = 1$$

on the other hand. We derive two further sets of parameters, $\varphi_t(r) = 1 - q_t(r,\dagger)$ the *survival* probabilities and $\psi_t(r,s)$ the interstrata *movement* probabilities, defined as

$$q_t(r,s) = \varphi_t(r) \times \psi_t(r,s) \quad r \in \mathfrak{K}, s \in \mathfrak{K}\,.$$

The likelihood corresponding to the complete observation of the $(\mathbf{x}_i, \mathbf{z}_i)$'s, $\ell(p_1, \ldots, p_\tau, q_1, \ldots, q_\tau | (\mathbf{x}_1, \mathbf{z}_1), \ldots, (\mathbf{x}_n, \mathbf{z}_n))$, is then given by

$$\prod_{i=1}^{n} \left[\prod_{t=1}^{\tau} p_t(z_{(i,t)})^{x_{(i,t)}} \{1 - p_t(z_{(i,t)})\}^{1 - x_{(i,t)}} \times \prod_{t=1}^{\tau-1} q_t(z_{(i,t)}, z_{(i,t+1)}) \right], \quad (5.6)$$

[10] Covariates registered once or at each time will not be used here, although they could be introduced via a generalized linear model as in Chap. 4, so we abstain from adding further notations in an already dense section.

up to a constant. The complexity of the likelihood corresponding to the data actually observed is due to the fact that the $\mathbf{z}_i$'s are not fully observed, hence that (5.6) would have to be summed over all possible values of the missing components of the $\mathbf{z}_i$'s. This complexity can be bypassed by a simulation alternative described below in Sect. 5.5.2.

The prior modeling corresponding to these parameters will depend on the information that is available about the population covered by the capture–recapture experiment. For illustration's sake, consider the use of conjugate priors

$$p_t(r) \sim \mathscr{B}e(a_t(r), b_t(r))\,, \qquad \varphi_t(r) \sim \mathscr{B}e(\alpha_t(r), \beta_t(r))\,,$$

where the hyperparameters, $a_t(r)$, $b_t(r)$ and so on, depend on both time t and location r, and

$$\psi_t(r) \sim \mathscr{D}ir(\gamma_t(r))\,,$$

a Dirichlet distribution, where $\psi_t(r) = (\psi_t(r,s); s \in \mathfrak{K})$ with

$$\sum_{s\in\mathfrak{K}} \psi_t(r,s) = 1\,,$$

and $\gamma_t(r) = (\gamma_t(r,s); s \in \mathfrak{K})$. The determination of these (numerous) hyperparameters is also case-dependent and varies from a noninformative modeling, where all hyperparameters are taken to be equal to 1 or 1/2, to a very informative setting where exact values of these hyperparameters can be chosen from the prior information. The following example is an illustration of the latter.

Table 5.3. Prior information about the capture and survival parameters of the Arnason–Schwarz model, represented by prior expectation and prior confidence interval, for a capture–recapture experiment on the migrations of lizards (*source*: Dupuis, 1995)

Episode		2	3	4	5	6
p_t	Mean	0.3	0.4	0.5	0.2	0.2
	95% cred. int.	[0.1, 0.5]	[0.2, 0.6]	[0.3, 0.7]	[0.05, 0.4]	[0.05, 0.4]

Site		A		B,C	
Episode		$t = 1, 3, 5$	$t = 2, 4$	$t = 1, 3, 5$	$t = 2, 4$
$\varphi_t(r)$	Mean	0.7	0.65	0.7	0.7
	95% cred. int.	[0.4, 0.95]	[0.35, 0.9]	[0.4, 0.95]	[0.4, 0.95]

Example 5.1. For the capture–recapture experiment described in Footnote 8 on the migrations of lizards between three adjacent zones, there are six capture episodes. The prior information provided by the biologists on the capture and survival probabilities, p_t (which are assumed to be zone independent) and $\varphi_t(r)$, is given by Table 5.3. While this may seem very artificial, this

construction of the prior distribution actually happened that way because the biologists in charge were able to quantify their beliefs and intuitions in terms of prior expectation and prior confidence interval. (The differences in the prior values on p_t are due to differences in capture efforts, while the differences between the group of episodes 1, 3 and 5, and the group of episodes 2 and 4 are due to the fact that the odd indices correspond to spring and the even indices to fall and mortality is higher over the winter.) Moreover, this prior information can be perfectly translated in a collection of beta priors by the R divide-and-conquer function

```
probet=function(a,b,c,alpha){

  coc=(1-c)/c
  pbeta(b,alpha,alpha*coc)-pbeta(a,alpha,alpha*coc)
  }

solbeta=function(a,b,c,prec=10^(-3)){

  coc=(1-c)/c
  detail=alpha=1
  while (probet(a,b,c,alpha)<.95) alpha=alpha+detail
  while (abs(probet(a,b,c,alpha)-.95)>prec){

    alpha=max(alpha-detail,detail/10)
    detail=detail/10
    while (probet(a,b,c,alpha)<.95) alpha=alpha+detail
    }
  list(alpha=alpha,beta=alpha*coc)
  }
```

(see Exercise 5.23 for details). Repeated calls to `solbeta` as in

```
> solbeta(.1,.5,.3,10^(-4))
[1]  5.45300 12.72367
```

then leads to the hyperparameters given in Table 5.4. ◀

Table 5.4. Hyperparameters of the beta priors corresponding to the information contained in Table 5.3 (*source*: Dupuis, 1995)

Episode	2	3	4	5	6
Dist.	$\mathcal{B}e(5,13)$	$\mathcal{B}e(8,12)$	$\mathcal{B}e(12,12)$	$\mathcal{B}e(3.5,14)$	$\mathcal{B}e(3.5,14)$

Site	A		B	
Episode	$t=1,3,5$	$t=2,4$	$t=1,3,5$	$t=2,4$
Dist.	$\mathcal{B}e(6.0,2.5)$	$\mathcal{B}e(6.5,3.5)$	$\mathcal{B}e(6.0,2.5)$	$\mathcal{B}e(6.0,2.5)$

5.5.2 Gibbs Sampler

Given the presence of missing data in the Arnason–Schwarz model, a Gibbs sampler is a natural solution to handle the complexity of the likelihood. It needs to include simulation of the missing components in the vectors $\mathbf{z}_i$ in order to simulate the parameters from the full conditional distribution

$$\pi(\theta|\mathbf{x},\mathbf{z}) \propto \ell(\theta|\mathbf{x},\mathbf{z}) \times \pi(\theta)\,,$$

Algorithm 5.10 Arnason–Schwarz Gibbs Sampler
Iteration l $(l \geq 1)$:

1. **Parameter simulation**
   Simulate $\theta^{(l)} \sim \pi(\theta|\mathbf{z}^{(l-1)},\mathbf{x})$ as $(t = 1,\ldots,\tau)$,

$$p_t^{(l)}(r)|\mathbf{x},\mathbf{z}^{(l-1)} \sim \mathscr{B}e\left(a_t(r)+u_t(r), b_t(r)+v_t^{(l)}(r)\right)\,,$$

$$\varphi_t^{(l)}(r)|\mathbf{x},\mathbf{z}^{(l-1)} \sim \mathscr{B}e\left(\alpha_t(r)+\sum_{j\in\mathfrak{K}} w_t^{(l)}(r,j), \beta_t(r)+w_t^{(l)}(r,\dagger)\right)\,,$$

$$\psi_t^{(l)}(r)|\mathbf{x},\mathbf{z}^{(l-1)} \sim \mathscr{D}ir\left(\gamma_t(r,s)+w_t^{(l)}(r,s); s\in\mathfrak{K}\right)\,,$$

   where

$$w_t^{(l)}(r,s) = \sum_{i=1}^{n} \mathbb{I}_{(z_{(i,t)}^{(l-1)}=r, z_{(i,t+1)}^{(l-1)}=s)}\,,$$

$$u_t^{(l)}(r) = \sum_{i=1}^{n} \mathbb{I}_{(x_{(i,t)}=1, z_{(i,t)}^{(l-1)}=r)}\,,$$

$$v_t^{(l)}(r) = \sum_{i=1}^{n} \mathbb{I}_{(x_{(i,t)}=0, z_{(i,t)}^{(l-1)}=r)}\,.$$

2. **Missing location simulation**
   Generate the unobserved $z_{(i,t)}^{(l)}$'s from the full conditional distributions

$$\mathbb{P}(z_{(i,1)}^{(l)} = s|x_{(i,1)}, z_{(i,2)}^{(l-1)}, \theta^{(l)}) \propto q_1^{(l)}(s, z_{(i,2)}^{(l-1)})(1-p_1^{(l)}(s))\,,$$

$$\mathbb{P}(z_{(i,t)}^{(l)} = s|x_{(i,t)}, z_{(i,t-1)}^{(l)}, z_{(i,t+1)}^{(l-1)}, \theta^{(l)}) \propto q_{t-1}^{(l)}(z_{(i,t-1)}^{(l)}, s) \times q_t(s, z_{(i,t+1)}^{(l-1)})(1-p_t^{(l)}(s))\,,$$

$$\mathbb{P}(z_{(i,\tau)}^{(l)} = s|x_{(i,\tau)}, z_{(i,\tau-1)}^{(l)}, \theta^{(l)}) \propto q_{\tau-1}^{(l)}(z_{(i,\tau-1)}^{(l)}, s)(1-p_\tau(s)^{(l)})\,.$$

where $\mathbf{x}$ and $\mathbf{z}$ denote the collections of the vectors of capture indicators and locations, respectively. This is thus a particular case of *data augmentation*, where the missing data $\mathbf{z}$ are simulated at each step t in order to reconstitute a complete sample $(\mathbf{x}, \mathbf{z}^{(t)})$ for which conjugacy applies. In the setting of the Arnason–Schwarz model, we can simulate the full conditional distributions both of the parameters and of the missing components. The Gibbs sampler is as follows:

Note that simulating the missing locations in the $\mathbf{z}_i$'s conditionally on the other locations and on the parameters is not a very complex task because of the good conditioning properties of these vectors (which stem from their Markovian nature). As shown in Step 2 of Algorithm 5.10, the full conditional distribution of $z_{(i,t)}$ only depends on the previous and next locations $z_{(i,t-1)}$ and $z_{(i,t+1)}$ (and obviously on the fact that it is not observed; that is, that $x_{(i,t)} = 0$). The corresponding part of the R code is based on a `latent` matrix containing the current values of both the observed and missing locations:

```
for (i in 1:n){

  if (z[i,1]==0) latent[i,1]=sample(1:(m+1),1,
     prob=q[,latent[i,2]]*(1-c(p[s,],0)))
  for (t in ((2:(T-1))[z[i,-c(1.T)]==0]))
    latent[i,t]=sample(1:(m+1),1,
      prob=q[latent[i,t-1],]*q[,latent[i,t+1]]*(1-c(p[s,],0)))
  if (z[i,T]==0) latent[i,T]=sample(1:(m+1),1,
     prob=q[latent[i,T-1],]*(1-c(p[s,],0)))
  }
```

(The convoluted range for the inner loop replaces an `if (z[i,t]==0)`.) When the number of states $s \in \mathfrak{K}$ is moderate, it is straightforward to simulate from such a distribution.

Take $\mathfrak{K} = \{1, 2\}$, $n = 4$, $m = 8$ and assume that, for $\mathbf{x}$, we have the following histories:

1	1	1	·	·	1	·	·	·
2	1	·	1	·	1	·	2	1
3	2	1	·	1	2	·	·	1
4	1	·	·	1	2	1	1	2

Assume also that all (prior) hyperparameters are taken equal to 1. Then one possible instance of a simulated $\mathbf{z}$ is

1	1	1	2	1	1	2	†
1	1	1	2	1	1	1	2
2	1	2	1	2	1	1	1
1	2	1	1	2	1	1	2

and it leads to the following simulation of the parameters:

$$p_4^{(l)}(1)|\mathbf{x}, \mathbf{z}^{(l-1)} \sim \mathscr{B}e(1+2, 1+0)\,,$$
$$\varphi_7^{(l)}(2)|\mathbf{x}, \mathbf{z}^{(l-1)} \sim \mathscr{B}e(1+0, 1+1)\,,$$
$$\psi_2^{(l)}(1,2)|\mathbf{x}, \mathbf{z}^{(l-1)} \sim \mathscr{B}e(1+1, 1+2)\,,$$

in the Gibbs sampler, where the hyperparameters are therefore derived from the (partly) simulated history above. Note that because there are only two possible states, the Dirichlet distribution simplifies into a beta distribution.

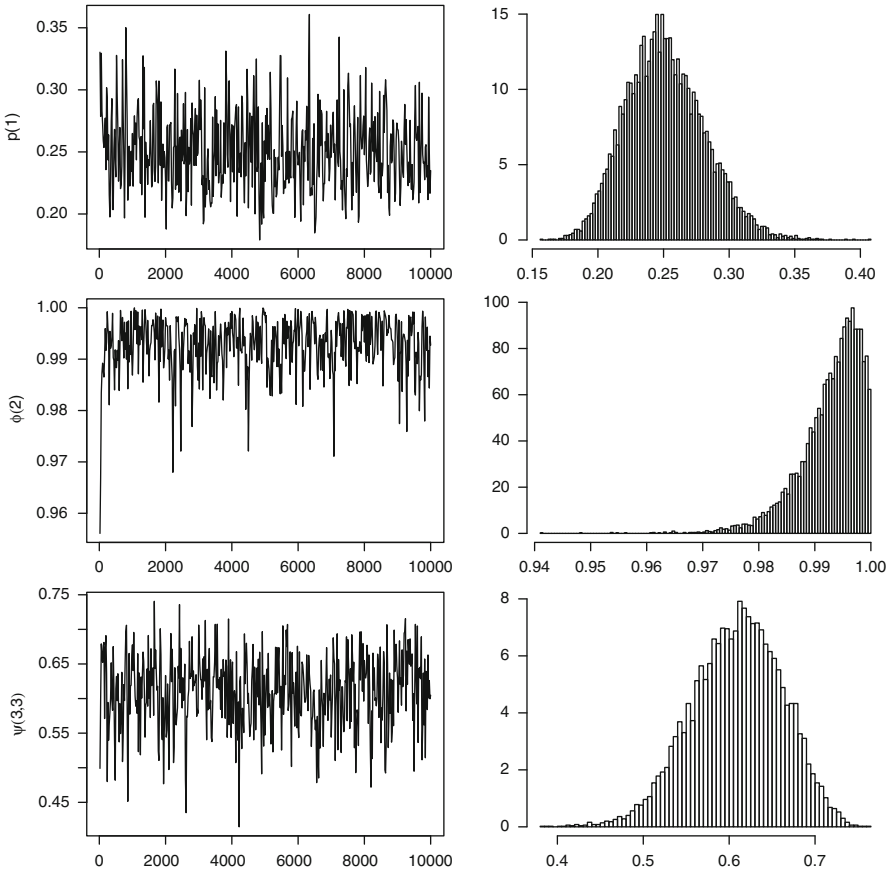


Fig. 5.5. Dataset **eurodip**: Representation of the Gibbs sampling output for some parameters of the Arnason–Schwarz model, based on 10,000 iterations, with raw plots (*first column*) and histograms (*second column*)

For **eurodip**, Lebreton et al. (1992) argue that the capture and survival rates should be constant over time. If we assume that the movement probabilities are also time independent, we are left with $3+3+3\times 2=12$ parameters. Figure 5.5 gives the Gibbs output for the parameters $p(1)$, $\varphi(2)$, and $\psi(3,3)$ using noninformative priors with $a(r)=b(r)=\alpha(r)=\beta(r)=\gamma(r,s)=1$. The simulation of the parameters is obtained by the following piece of R code, where `s` is the current index of the Gibbs iteration in the R code below:

```
for (r1 in 1:m){

  for (r2 in 1:(m+1))
    omega[r2]=sum(latent[,1:(T-1)]==r1 & latent[,2:T]==r2)
  u=sum(z!=0 & latent==r1)


  v=sum(z==0 & latent==r1)
  p[s,r1]=rbeta(1,1+u,1+v)
  phi[s,r1]=rbeta(1,1+sum(omega[1:m]),1+omega[m+1])
  psi[r1,,s]=rdirichlet(1,rep(1,m)+omega[1:m])
  }
```

The transition probabilities $q_t(r,s)$ are then reconstructed from the survival and movement probabilities, with the special case of the `m+1` column corresponding to the absorbing † state:

```
tt=matrix(rep(phi[s,],m),m,byrow=T)
q=rbind(tt*psi[,,s],rep(0,m))
q=cbind(q,1-apply(q,1,sum))
```

The convergence of the Gibbs sampler to the region of interest occurs very quickly, even though we can spot an approximate periodicity in the raw plots on the left-hand side. The MCMC approximations of the estimates of $p(1)$, $\varphi(2)$, and $\psi(3,3)$, the empirical mean over the last 8,000 simulations, are equal to 0.25, 0.99, and 0.61, respectively.

5.6 Exercises

5.1 Show that the posterior distribution $\pi(N|n^+)$ given by (5.1), while associated with an improper prior, is defined for all values of n^+. Show that the normalization factor of (5.1) is $n^+ \vee 1$, and deduce that the posterior median is equal to $2(n^+ \vee 1)-1$. Discuss the relevance of this estimator and show that it corresponds to a Bayes estimate of p equal to $1/2$.

5.2 Under the same prior as in Sect. 5.2.1, derive the marginal posterior density of N in the case where $n_1^+ \sim \mathscr{B}(N,p)$ and

$$n_2^+, \ldots, n_k^+ \overset{\text{iid}}{\sim} \mathscr{B}(n_1^+, p)$$

are observed (the later are in fact recaptures). Apply to the sample

$$(n_1^+, n_2^+, \ldots, n_{11}^+) = (32, 20, 8, 5, 1, 2, 0, 2, 1, 1, 0)\,,$$

which describes a series of tag recoveries over 11 years.

5.3 Show that the conditional distribution of m_2 conditional on both sample sizes n_1 and n_2 is given by (5.2) and does not depend on p. Deduce the expectation $\mathbb{E}^\pi[m_2|n_1,n_2,N]$.

5.4 In order to determine the number N of buses in a town, a capture–recapture strategy goes as follows. We observe $n_1 = 20$ buses during the first day and keep track of their identifying numbers. Then we repeat the experiment the following day by recording the number of buses that have already been spotted on the previous day, say $m_2 = 5$, out of the $n_2 = 30$ buses observed the second day. For the Darroch model, give the posterior expectation of N under the prior $\pi(N) = 1/N$.

5.5 Show that the maximum likelihood estimator of N for the Darroch model is $\hat{N} = n_1/(m_2/n_2)$, and deduce that it is not defined when $m_2 = 0$.

5.6 Give the likelihood of the extension of Darroch's model when the capture–recapture experiments are repeated K times with capture sizes and recapture observations n_k $(1 \le k \le K)$ and m_k $(2 \le k \le K)$, respectively. (*Hint*: Exhibit first the two-dimensional sufficient statistic associated with this model.)

5.7 Give both conditional posterior distributions involved in Algorithm 5.8 in the case $n^+ = 0$.

5.8 Show that, for the two-stage capture model with probability p of capture, when the prior on N is a $\mathscr{P}(\lambda)$ distribution, the conditional posterior on $N - n^+$ is $\mathscr{P}(\lambda(1-p)^2)$.

5.9 Reproduce the analysis of **eurodip** summarized by Fig. 5.1 when switching the prior from $\pi(N,p) \propto \lambda^N/N!$ to $\pi(N,p) \propto N^{-1}$.

5.10 An extension of the T-stage capture–recapture model of Sect. 5.2.3 is to consider that the capture of an individual modifies its probability of being captured from p to q for future recaptures. Give the likelihood $\ell(N,p,q|n_1,n_2,m_2\ldots,n_T,m_T)$.

5.11 Another extension of the two-stage capture–recapture model is to allow for mark loss.[11] If we introduce q as the probability of losing the mark, r as the probability of recovering a lost mark and k as the number of recovered lost marks, give the associated likelihood $\ell(N,p,q,r|n_1,n_2,m_2,k)$.

[11] Tags can be lost by marked animals, but the animals themselves could also be lost to recapture either by changing habitat or dying. Our current model assumes that the population is *closed*; that is, that there is no immigration, emigration, birth, or death within the population during the length of the study. These other kinds of extension are dealt with in Sects. 5.3 and 5.5.

5.12 Show that the conditional distribution of r_1 in the open population model of Sect. 5.3 is proportional to the product (5.4).

5.13 Show that the distribution of r_2 in the open population model of Sect. 5.3 can be integrated out from the joint distribution and that this leads to the following distribution on r_1:

$$\pi(r_1|p,q,n_1,c_2,c_3) \propto \frac{(n_1-r_1)!(n_1-r_1-c_3)!}{r_1!(n_1-r_1-c_2)!} \times \left(\frac{q}{(1-p)(1-q)[q+(1-p)(1-q)]}\right)^{r_1}.$$

Compare the computational cost of a Gibbs sampler based on this approach with a Gibbs sampler using the full conditionals.

5.14 Show that the likelihood associated with an open population as in Sect. 5.3 can be written as

$$\ell(N,p|\mathscr{D}^*) = \sum_{(\epsilon_{it},\delta_{it})_{it}} \prod_{t=1}^{T}\prod_{i=1}^{N} q_{\epsilon_{i(t-1)}}^{\epsilon_{it}}(1-q_{\epsilon_{i(t-1)}})^{1-\epsilon_{it}} \times p^{(1-\epsilon_{it})\delta_{it}}(1-p)^{(1-\epsilon_{it})(1-\delta_{it})},$$

where $q_0 = q$, $q_1 = 1$, and δ_{it} and ϵ_{it} are the capture and exit indicators, respectively. Derive the order of complexity of this likelihood; that is, the number of elementary operations necessary to compute it.[12]

5.15 In connection with the presentation of the accept–reject algorithm in Sect. 5.4, show that, for $M > 0$, if g is replaced with Mg in $\mathscr{S}$ and if (X,U) is uniformly distributed on $\mathscr{S}$, the marginal distribution of X is still g. Deduce that the density g only needs to be known up to a normalizing constant.

5.16 For the function $g(x) = (1+\sin^2(x))(2+\cos^4(4x))\exp[-x^4\{1+\sin^6(x)\}]$ on $[0,2\pi]$, examine the feasibility of running a uniform sampler on the set $\mathscr{S}$ associated with the accept–reject algorithm in Sect. 5.4.

5.17 Show that the probability of acceptance in Step 2 of Algorithm 5.9 is $1/M$ and that the number of trials until a variable is accepted has a geometric distribution with parameter $1/M$. Conclude that the expected number of trials per simulation is M.

5.18 For the conditional distribution of α_t derived from (5.3), construct an accept–reject algorithm based on a normal bounding density f and study its performances for $N = 532$, $n_t = 118$, $\mu_t = -0.5$, and $\sigma^2 = 3$.

5.19 When uniform simulation on the accept–reject set $\mathscr{S}$ of Sect. 5.4 is impossible, construct a Gibbs sampler based on the conditional distributions of u and x. (*Hint*: Show that both conditionals are uniform distributions.) This special case of the Gibbs sampler is called the *slice sampler* (see Robert and Casella, 2004, Chap. 8). Apply to the distribution of Exercise 5.16.

[12] We will see in Chap. 7 a derivation of this likelihood that enjoys an $O(T)$ complexity.

5.20 Show that the normalizing constant M of a target density f can be deduced from the acceptance rate in the accept–reject algorithm (Algorithm 5.9) under the assumption that g is properly normalized.

5.21 Reproduce the analysis of Exercise 5.20 for the marginal distribution of r_1 computed in Exercise 5.13.

5.22 Modify the function `ardipper` used in Sect. 5.4 to return the acceptance rate as well as a sample from the target distribution.

5.23 Show that, given a mean and a 95% confidence interval in $[0, 1]$, there exists at most one beta distribution $\mathscr{B}e(a, b)$ with such a mean and confidence interval.

5.24 Show that, for the Arnason–Schwarz model, groups of consecutive unknown locations are independent of one another, conditional on the observations. Devise a way to simulate these groups by blocks rather than one at a time; that is, using the joint posterior distributions of the groups rather than the full conditional distributions of the states.