Bayesian Estimation of the Number of Unknown Species: Incorporating a Model for the Discovery Process

Simon P. Wilson¹, Brett Houlding¹, and Mark Costello²

¹ Discipline of Statistics, Trinity College Dublin, Ireland. Correspondence to houldinb@tcd.ie.
² Department of Marine Sciences, University of Auckland, Leigh Marine Laboratory, New Zealand.

Abstract

The problem of estimating the number of existing species has been reviewed many times within the statistics literature. The typical case is to estimate the unknown number of species \( S \) from a sample of \( m \) individuals in which \( R \leq S \) distinct species are observed with abundances \( n_1, \ldots, n_R \). Unfortunately, records of animal and plant species discovery usually contain only the number of sampled species \( R \) and their dates of discovery \( t_1, \ldots, t_R \). As such, here we propose a model for estimating the number of species \( S \) using only the species discovery dates \( t_1, \ldots, t_R \) and the yearly number of distinct author names \( a_1, \ldots, a_T \), the latter being considered a proxy for a latent 'effort process' in species sampling.

1. Species Estimation

A community contains an unknown number of species \( S \) with unknown abundances \( N_1, \ldots, N_S \). The total population of life within the community is thus \( N = \sum_{i=1}^{S} N_i \). Typically, species abundance models define probability distributions for the \( N_i \) in terms of a set of parameters \( \theta \). Here we assume species abundances follow a Log-normal distribution, which, even though being a continuous distribution whilst populations are integer-valued, is a common modelling assumption. Hence:

\[
\log(N_i) \sim N(\theta_i, \theta_i^2), \quad i = 1 : S
\]  

(1)

A sample of individuals is assumed to have been taken sequentially over a period of \( T \) years, with \( n(t) \) individuals being sampled in year \( t \). This has resulted in a total of \( B \) distinct species being identified. Furthermore, it is assumed that each individual had an equal chance of being sampled, and that sampling was with replacement, so that the probability of a randomly sampled individual being of species \( i \) is \( p_i = N_i/N \).

Denoting by \( M_t \) the sample number in which the \( i \)th species is first observed, (hence 1 = \( M_1 < M_2 < \cdots < M_B \leq N \)), we define \( m_i = M_i - M_{i-1} \) to be the number of individuals sampled between the \( (i-1) \)-th and \( i \)-th species discoveries.

The distribution of \( m_1, \ldots, m_B \) given \( N_1, \ldots, N_S \) can then be determined as follows: There are \( m_i \) individuals sampled between the discovery of the \( (i-1) \)-th and \( i \)-th species. This happens if and only if \( m_i - 1 \) individuals are sampled from the first \( i - 1 \) species (meaning no new species are observed) and then the \( m_i \)-th sample is a first observation of a new species \( i \). Under the assumption of sampling with replacement, this occurs with probability \((\sum_{j=1}^{i-1} p_j)^{m_i-1} p_i \). Hence, assuming that \( m_i \geq 1 \) for all \( i \) and that \( S \geq R \), we have for all discovered species:

\[
P(m_1, \ldots, m_B | N_1, \ldots, N_S, R) = \prod_{i=1}^{B} \left( \sum_{j=1}^{i-1} p_j \right)^{m_i-1} p_i
\]  

(2)

Figure 1: DAG for Species Estimation using inter-discovery sample sizes.

As such, if the values of \( m_1, \ldots, m_B \) were available, then such data, taken in combination with a prior distribution over \( \theta \), could be implemented in Bayes’ Theorem to determine a posterior distribution for \( N_1, \ldots, N_S \). Moreover, techniques such as Reversible Jump Markov Chain Monte Carlo (RJCMC) can then be implemented to permit inference over the total number of species \( S \). Figure 1 above provides a graphical representation for this part of the species estimation model.

Unfortunately, however, records of animal and plant species discovery usually only last the year of discovery, and not the number of samples collected between discoveries. e.g., the most comprehensive database of known species: the Catalogue of Life. As such, to estimate the total number of unknown species using such limited information, a model is required for the discovery process itself.

2. The Discovery Process

If we assume that the available data consists not of the inter-discovery sample sizes, but instead the years \( t_1, \ldots, t_R \) of species discovery, then if the number of individuals sampled in any given year were also known (i.e., the values \( n(1), \ldots, n(T) \) were available), the following constraints would bound the values of \( M_1, \ldots, M_B \):

\[
\max \left\{ M_j - \sum_{i=1}^{j-1} n(i) \right\} < M_j \leq \sum_{i=1}^{j} n(i)
\]  

(3)

That is to say, for the \( M_j \)-th individual to have been sampled in year \( t_j \), the cumulative number of individuals sampled up to and inclusive of year \( t_j - 1 \) must be less than \( M_j \). In addition, the cumulative number of individuals sampled up to and inclusive of the year in which the \( M_j \)-th individual was sampled, must be greater than or equal to \( M_j \).

As such, and observing that \( M_1 = 1 + \sum_{j=1}^{T-1} m_j \), the inter-discovery sample sizes can be considered as latent variables whose probability mass function follows that defined in Equation (2), but truncated so as to satisfy the constraints of Equation (3). The information leading to inference on the species abundances, and hence the number of unknown species, would then be provided through the species discovery years and the yearly sample sizes. Figure 2 below represents this aspect of the inference problem graphically.

Figure 2: DAG displaying the inference model for the latent inter-discovery sample sizes.

Again, however, it is unlikely that ecologists or databases will maintain a record of the number of samples taken annually, and so these too must be assumed unavailable. Moreover, models for how many individuals have been sampled, as a function of time, do not appear to have been considered within the statistical or biological literatures.

To account for this problem, we propose the use of a latent 'effort process' that is used to represent an underlying rate of sampling effort \( E_t \). The advantage of defining the 'effort process' is that we can hopefully measure some proxy of it, and relate that to the effort rate \( E_t \). One such possibility is the yearly number of distinct author names in papers concerning species discovery. Defining \( a_t \) to be the number of distinct author names in year \( t \), a simple model is to assume that \( a_t \) is Poisson distributed with mean \( \exp(\alpha - \alpha) \), for some parameter \( \alpha \).

\[
a_t | a_t \sim \text{Poisson} \left( \exp(\alpha - \alpha) \right), \quad t = 1 : T
\]  

(5)

The parameter \( \alpha \) can be interpreted in the following way: Since we expect \( \exp(\alpha) \) individuals to be sampled in year \( t \), then we expect \( \exp(\alpha) \) individuals to be sampled per author. In other words, \( \alpha \) is the natural logarithm of the number of individuals we expect an author to sample in a given year. Finally, we are able to generate the \( \alpha_t \) by examining databases of species discovery. Figure 3 below represents the proposed inference procedure concerning yearly sample sizes.

Figure 3: DAG displaying the inference model for the yearly sample sizes.

Taken collectively, the model for estimating the number of unknown species by incorporating a model for the discovery process which utilizes a proxy for a latent effort process, is presented graphically in Figure 4 below.

Figure 4: Full DAG for species estimation problem.

3. A Proxy for the Effort Process

The number of individuals sampled in a given year is then assumed to be a random process whose probability distribution is determined by \( e_t \). A common technique that is suitable for the discrete effort process suggested here is to assume the number of individuals sampled in a given year follows a Poisson distribution with mean \( E_t = \exp(\varepsilon) \):

\[
n(t) | e_t \sim \text{Poisson} \left( \exp(e_t) \right), \quad t = 1 : T
\]  

(6)

The parameter \( \varepsilon \) can be interpreted in the following way: Since we expect \( \exp(\varepsilon) \) individuals to be sampled in year \( t \), then we expect \( \exp(\varepsilon) \) individuals to be sampled per effort unit. In other words, \( \varepsilon \) is the natural logarithm of the number of effort units we expect an effort unit to sample in a given year. Finally, we are able to generate the \( \varepsilon_t \) by examining databases of species discovery. Figure 5 below represents the proposed inference procedure concerning yearly sample sizes.

Figure 5: DAG displaying the inference model for the yearly sample sizes.

Acknowledgement

This work is funded by the STATICA project, a Principal Investigator program of Science Foundation Ireland, Grant number 08/IN.1/11879.