PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2005) **272**, 285–287
doi:10.1098/rspb.2004.2955
Published online 04 February 2005

On estimating the number of species from the discovery record

Andrew R. Solow[1*] and Woollcott K. Smith[2]

[1] *Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA* (*asolow@whoi.edu*)
[2] *Temple University, Philadelphia, PA 19122, USA* (*wksmith@temple.edu*)

A common approach to estimating the number of species in a taxonomic or other group is to extrapolate the temporal pattern of historical species discoveries or descriptions. A formal statistical approach to this problem is described. This approach involves fitting an explicit model of the discovery record by maximum likelihood and using the fitted model to estimate the number of undiscovered species. The approach is applied to a description record of large marine animals covering the period 1828–1996. The estimated number of undiscovered species in this group is around 10 with an upper 0.95 confidence bound of around 16.

Keywords: large marine animals; maximum-likelihood estimation; non-stationary Poisson process; taxonomy

1. INTRODUCTION

A perennial problem in biology is estimating the number of species in a taxonomic or other group. May (1988, 1990) reviewed a variety of methods. One common approach is to estimate the number of species by extrapolating to infinite time the temporal pattern of historical discoveries or descriptions (Simon 1983; Diamond 1985). Briefly, in applying this approach, the standard practice is to fit a parametric model with an asymptote to the cumulative discovery record and to estimate the total number of species from the asymptote of the fitted model. For example, in estimating the number of large (i.e. in excess of 2 m in length) marine animals, Paxton (1998) fitted a rectangular hyperbola to the cumulative record of descriptions. The form of this model was chosen for convenience. The purpose of this paper is to place this practice on a firmer statistical footing by proposing an explicit statistical model for the discovery process and fitting this model by the method of maximum likelihood (ML).

The remainder of this paper is organized in the following way. The basic model of the discovery record is outlined in § 2. Estimation under this model is covered in § 3. In § 4, the method is applied to the description record of large marine animals compiled by Paxton (1998), and § 5 contains some concluding remarks.

2. A MODEL OF THE DISCOVERY RECORD

We will assume that the sightings of species j, follow a non-stationary Poisson process with rate function:

$$\lambda_j(t) = \lambda_j\, g(t), \tag{2.1}$$

where λ_j is an unknown positive constant that, roughly speaking, measures the visibility of species j, and the unknown positive function, $g(t)$, is intended to capture the trend over time in sighting skill and effort. The properties of the non-stationary Poisson process relevant here are given in Cox & Lewis (1978). Under this model, the sighting probability of species j in the interval $(t,\, t + \Delta t)$ is approximately $\lambda_j\, g(t)\, \Delta t$ for small Δt. It is useful to adopt the restriction $g(0) = 1$, so that λ_j has the interpretation of the mean sighting rate of species j at $t = 0$. A specific model for $g(t)$ and λ_j is considered below.

Let the random variable, T_j, be the discovery time—i.e. the time of the *first* sighting—of species j. Under the model outlined above, the cumulative distribution function (CDF) of T_j conditional on λ_j is:

$$\text{prob}(T_j \leqslant t_j | \lambda_j) = 1 - \exp(-\lambda_j\, G(t_j)), \tag{2.2}$$

where

$$G(t) = \int_t^0 g(u)\, \mathrm{d}u \tag{2.3}$$

is the cumulative effort through time t.

We will assume that λ_j represents the realization of an exponentially distributed random variable Λ with probability density function (PDF):

$$f(\lambda) = \theta\, \exp(-\theta\, \lambda), \tag{2.4}$$

with unknown mean θ^{-1} . It follows that the unconditional CDF of T_j is

$$\text{prob}(T_j \leqslant t_j) = 1 - \frac{\theta}{\theta + G(t_j)}, \tag{2.5}$$

To complete the model, we will assume that

$$g(t) = \exp(\beta\, t) \tag{2.6}$$

for unknown β. Under this complete model, the CDF of T_j is

$$\text{prob}(T_j \leqslant t_j) = 1 - \frac{\theta}{\theta + \frac{1}{\beta}(\exp(\beta\, t_j) - 1)} = P(t_j) \tag{2.7}$$

and the corresponding PDF of T_j is

$$p(t_j) = \frac{\theta\, \exp(\beta\, t_j)}{z\left(\theta + \frac{1}{\beta}\, (\exp(\beta\, t_j) - 1)\right)^2}. \tag{2.8}$$

* Author for correspondence (asolow@whoi.edu)

Received 8 July 2004
Accepted 28 September 2004

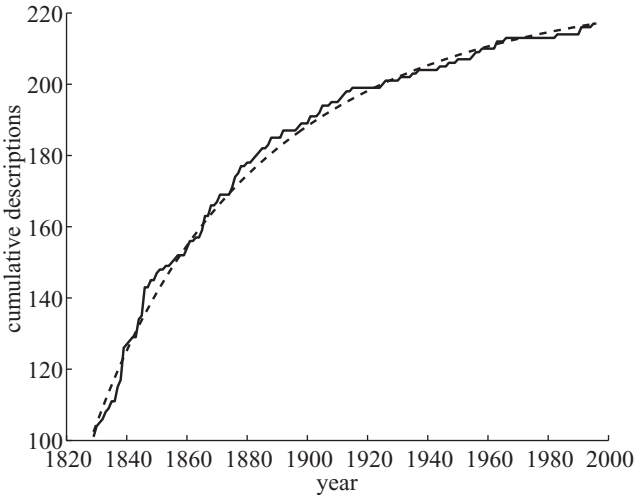

Figure 1. Cumulative description record of large marine animals, 1828–1996 (solid line) and fitted model (dashed line). Both curves include 100 species described prior to 1828.

A convenient property of the exponential form, equation (2.6) for $g(t)$ is that the time origin $t = 0$ need not coincide with the true beginning of the discovery period, but can be taken as any time after the beginning of this period. In that case, the relevant group consists of the species that were undiscovered as of $t = 0$.

The model outlined above can be thought of as an extension of the Jelinski–Moranda model used in software reliability (Jelinski & Moranda 1972). Under this model, a software program contains an unknown number of errors and the times at which these errors are discovered (and repaired) are independent and identically distributed exponential random variables. In the model outlined here, the assumption that the discovery times have the same distribution is relaxed through the so-called mixing distribution, f, and allowance is made for increasing discovery effort through the function g.

3. ESTIMATION

Suppose that over the period $(0, t_0)$ a total of n species are discovered. Let $t_1, t_2, \ldots, t_n$ be the discovery times of these species. These discovery times are assumed to have been generated independently from the model outlined in § 2 with the important *proviso* that they are all no later than t_0. The likelihood is defined as the joint PDF of these discovery times regarded as a function of the unknown parameters θ and β:

$$L(\theta, \beta) = \prod_{j=1}^{n} \frac{p(t_j)}{P(t_0)}, \tag{3.1}$$

where $p(t)$ and $P(t)$ are given in equations (2.8) and (2.7), respectively. The ML estimates of θ and β are found by maximizing $L(\theta, \beta)$ or its logarithm.

Let $\hat{\theta}$ and $\hat{\beta}$ be the ML estimates of θ and β. There is usually little direct interest in these parameters themselves. However, by treating their estimates as correct, it is possible to construct a rough estimate of the number of undiscovered species in the group. Let m be the unknown number of undiscovered species, so that the total number of species is $n + m$. The number of these species that are discovered has a binomial distribution with $n + m$ trials and success probability $P(t_0)$. A simple estimate of m can be found by equating an estimate of the expected number of discoveries to its observed value and solving for m. This point estimate is given by

$$\hat{m} = n\,\frac{1 - \hat{P}(t_0)}{\hat{P}(t_0)}, \tag{3.2}$$

where $\hat{P}(t_0)$ is the estimate of $P(t_0)$ found by replacing θ and β by their ML estimates.

It is possible to go beyond point estimation to construct an approximate confidence interval for m. Specifically, the upper bound m_u of an approximate $1 - \alpha$ confidence interval of the form $(0, m_u)$ is given by the smallest value of m for which:

$$\sum_{i=0}^{n} \binom{m+n}{i} \hat{P}(t_0)^i\,(1 - \hat{P}(t_0))^{m+n-i} \geqslant \alpha. \tag{3.3}$$

By using $\hat{P}(t_0)$ in place of $P(t_0)$, this confidence interval ignores the variability in the estimates of θ and β and, therefore, has coverage less than $1 - \alpha$.

4. APPLICATION

Paxton (1998) compiled a record of the discovery times of 117 large marine animals. The first discovery in this record was made in 1829 and the last in 1995. We will take the beginning of the period of observation to be 1828 and the end to be 1996, so that $t_0 = 169$. The cumulative discovery record is shown in figure 1. This record includes 100 species discovered prior to the beginning of the observation period. We fit the model described above to this record. The ML estimates are $\hat{\theta} = 52.6$ and $\hat{\beta} = 0.013$. Thus, the estimated mean sighting rate at the beginning of the observation period was $\hat{\theta}^{-1} = 0.02$ and grows thereafter at an estimated annual rate of 1.3% to reach 0.17 at the end. The estimated probability that a previously undiscovered species is discovered during the observation period is $\hat{P}(t_0) = 0.92$. The point estimate of the number, m, of undiscovered species given in equation (3.2) is $\hat{m} = 10.2$. Finally, the upper bound of the *ca.* 0.95 confidence interval for m based on equation (3.3) is 16.

The goodness of the fitted model can be graphically assessed in the following way. Under the model described in this paper, the expected cumulative number of species discovered by time t is $(m + n)\,F(t)$. As in figure 1, goodness of fit can be assessed by plotting $(\hat{m} + n)\,\hat{F}(t)$ along with the cumulative discovery record. In this case, the fitted model appears to capture the behaviour of the data well.

5. DISCUSSION

The purpose of this paper has been to describe and illustrate a formal statistical approach to estimating the number of species in a group by extrapolating the temporal pattern of historical species discoveries or descriptions. The general idea of estimating species number in this way is not entirely satisfactory, being based as it is on human activity and not on any biological considerations. Despite this, it is a common approach and therefore worth putting on a firm statistical footing. This paper appears to represent the first step in this direction.

The method described in this paper uses only the discovery record. A variety of methods for estimating species

number based on the abundance of each species in a random sample of individuals have been proposed. These methods, which were reviewed by Bunge & Fitzpatrick (1993), are especially appropriate in situations in which the sampling is controlled. By contrast, the method described here is best suited for analysing historical taxonomic records that presumably were collected for quite different reasons.

Finally, the method described here could be extended in a number of ways. For example, external information about collection effort could be used either to construct the function g directly or to guide the specification of its parametric form. On the technical side, it would be possible—at the expense of additional computation—to include the effect of variability in the estimated parameters in constructing a confidence interval for m.

The authors are very grateful to Charles Paxton for sharing his data and to two anonymous referees for helpful comments.

REFERENCES

Bunge, J. & Fitzpatrick, M. 1993 Estimating the number of species: a review. *J. Am. Statist. Assoc.* **88**, 364–373.

Cox, D. R. & Lewis, P. A. W. 1978 *The statistical analysis of series of events*. London: Chapman & Hall.

Diamond, J. M. 1985 How many unknown species are yet to be discovered? *Nature* **315**, 538–539.

Jelinski, Z. & Moranda, P. B. 1972 Software reliability research. In *Statistical computer performance evaluation* (ed. W. Freiberger), pp. 465–484. London: Academic.

May, R. M. 1988 How many species are there on Earth? *Science* **247**, 1441–1449.

May, R. M. 1990 How many species? *Phil. Trans. R. Soc. B* **330**, 293–304.

Paxton, C. G. M. 1998 A cumulative species description curve for large open water marine animals. *J. Mar. Biol. Assoc.* **78**, 1389–1391.

Simon, H. R. 1983 Research and publication trends in systematic zoology. PhD thesis, The City University, London.