Aust. N. Z. J. Stat. **48**(2), 2006, 117–128 doi: 10.1111/j.1467-842X.2006.00430.x

APPLICATION OF LAPLACE'S BOUNDARY-MODE APPROXIMATIONS TO ESTIMATE SPECIES AND SHARED SPECIES RICHNESS

ANNE CHAO[1]*, TSUNG-JEN SHEN[2] AND WEN-HAN HWANG[3]

National Tsing Hua University, National Chung Hsing University and Feng-Chia University

Summary

The Laplace method for approximating integrals is a useful technique in a number of research fields. This paper shows that it also has interesting applications in biological and ecological statistical inferences. When sample abundance or replicated incidence (i.e., presence or absence) records of each species are available, the expected low-order frequency counts in heterogeneous communities can be approximated by the Laplace method when the species discovery or detection probabilities are bounded from below by a constant. The approximation formulae as applied to one community can then be used to derive estimators of species richness and to examine their performance. The approach is also extended to obtain simple and new estimators for the number of shared species in two communities. The replicated species incidence data recorded by competing teams of the Hong Kong Big Bird Race for the years 1999 and 2000 are used to estimate the number of resident birds in Hong Kong and to illustrate the method of estimation.

Key words: diversity indices; heterogeneity; species abundance; species incidence.

1. Introduction

The Laplace method and the closely related saddle-point technique are useful tools for approximating integrals arising in various fields of research. Goutis & Casella (1999) provided an excellent introduction to the topic along with its intuitive motivation and several illustrative examples. These methods have been extensively used in Bayesian statistics, generalized linear models, conditional inference and engineering applications. However, as far as the authors are aware, the technique has not yet been applied to the biological or ecological sciences. In this paper, we focus on the Laplace approximation and show that it also has interesting applications to the estimation of species richness in one community and shared species richness in two communities.

Consider the following k-dimensional integral in a form of the Laplace transform:

$$I_k = \int e^{nh(u)} g(u) du, \tag{1}$$

Received May 2004; revised August 2005; accepted December 2005.

*Author to whom correspondence should be addressed.

[1] Institute of Statistics, National Tsing Hua University, Hsin-Chu 30043, Taiwan.
e-mail: chao@stat.nthu.edu.tw

[2] Department of Applied Mathematics, National Chung Hsing University, Tai-Chung 402, Taiwan.

[3] Department of Statistics, Feng-Chia University, Tai-Chung 40724, Taiwan.

Acknowledgments. The authors sincerely thank Mr Michael Chalmers and Ms Maxine Chu of the World Wide Fund for Nature, Hong Kong for providing the Hong Kong Big Bird Race data with the assistance of Dr Paul S. F. Yip, University of Hong Kong. The comments from the Managing Editor, Dr Chris Lloyd, and editorial suggestions from the Technical Editor, Dr Ken Russell, were helpful in the revision of the original paper. The research was supported by the National Science Council of Taiwan.

where the integration domain is a subset of the k-dimensional Euclidean space. The Laplace approximation provides an asymptotic formula for the integral when $n \to \infty$ and $h(u)$ is a smooth function with a unique maximum in the domain of integration. The basic concept is as follows: as $n \to \infty$, the function $e^{nh(u)}$ exponentially decays near its maximum point and the integrand becomes increasingly concentrated near the maximum point, so the main contribution for the integral occurs near the maximum.

Typical Laplace approximation formulae deal with cases in which the maximum occurs in the interior of the integration domain. In our applications, as we show later, the maximum occurs at the boundary of the integrals. There has been relatively little published research on the boundary mode of Laplace formulae, with the exception of Erkanli (1994, 1997). The boundary-mode formulae for one and two dimensions are presented in Section 2.

The assessment of community biodiversity is essential for conservation biology, wildlife management and environmental policy decisions. Species richness in one community (alpha diversity) and the number of shared species in two communities (beta diversity) are the simplest and the most intuitive concepts for characterizing community diversity. A significant body of research has been conducted to estimate species richness in a single community (see Bunge & Fitzpatrick, 1993; Colwell & Coddington, 1994; and Chao, 2005 for suitable reviews). However, when compared with species richness, the estimation of shared species richness has received little attention. Chao *et al.* (2000) proposed an estimator of the number of shared species using abundance data, but there are no other shared species estimators reported in the literature.

It is intuitively understood that, if there are infinitely many undetectable or 'invisible' species in a hyper-diverse community, then it is impossible to obtain an accurate estimator of species richness. As suggested by Seber (1982 p. 72), species richness should only refer to the number of detectable species. If it is not possible to observe or catch a portion of the species, then those unobservable or undetectable species cannot be included in the estimating target. Therefore, it is reasonable to set a lower bound on species discovery or detection probabilities. This bounded-below assumption is critical for dealing with heterogeneous communities. For example, Lloyd (1992) models species heterogeneity with a family of beta B(α, β) distributions, but restricts the parameter α to be greater than 1 (such that cases with nonzero density at 0 are excluded). Similarly, restrictive assumptions are employed by Alho (1990 Proposition 2.1) and Norris & Pollock (1996 p. 640).

Under the bounded-below assumption, approximation formulae for the expected low-order frequency counts in heterogeneous communities can be obtained using the Laplace method. The resulting formulae are subsequently used to quantify the biases of two previous lower bounds of species richness (Chao 1984, 1987) and to justify their use as estimators under some conditions. The approach can also be extended to obtain new and simple estimators of the number of shared species in two communities based on abundance or replicated incidence data.

In Section 2, the Laplace boundary-mode approximations for one and two dimensions are described briefly. The applications to species richness estimation for one community are presented for abundance data in Section 3.1, and for replicated incidence data in Section 3.2. In Section 4, we develop the estimation of the number of species in common to two communities separately with two types of data. In Section 5, the species incidence data recorded by competing teams of the Hong Kong Big Bird Race (BBR) for the years 1999 and 2000

are used to illustrate the estimation procedure. Some concluding remarks are provided in Section 6.

2. Boundary-Mode Laplace Method

Consider a one-dimensional case for the integral (1) with an integration domain of the interval $[a, 1]$. Assume that the function h has a maximum attained at the lower boundary a and $h'(a) < 0$. Extending the formula in Erkanli (1997 Theorem 1) to include one more term, we have the following expansion under some regularity conditions:

$$I_1 = \int_{u \geq a} e^{nh(u)} g(u) du \tag{2a}$$

$$= g(a) \frac{e^{nh(a)}}{(-h'(a))} \frac{1}{n} \left(1 + \frac{A}{n} + \frac{B}{n^2} + O(n^{-3}) \right), \tag{2b}$$

where (all functions are evaluated at the point a)

$$A = A(g, h') = \frac{g'}{g} \frac{1}{(-h')} + \frac{h''}{(h')^2}, \tag{3}$$

and

$$B = B(g, h') = \frac{g''}{g} \frac{1}{(h')^2} + 3 \frac{g'}{g} \frac{h''}{(-h')^3} + 3 \frac{(h'')^2}{(h')^4} + \frac{h^{(3)}}{(-h')^3}. \tag{4}$$

For a two-dimensional integral of the form $\iint e^{nh(u,v)} g(u, v) du dv$, we assume that the integral domain is $\{(u, v): a \leq u \leq 1, b \leq v \leq 1\}$ and $h(u, v)$ has a maximum at the boundary point of (a, b). Erkanli (1997) provided an asymptotic approximation for this type of higher-dimensional integral, but a critical condition is that the boundary of the integral domain must be smooth. However, in our case, the boundary of the integral domain does not satisfy the condition. Accordingly, Erkanli's formula cannot be applied and another expansion is needed. Define $h_1 = \partial h(u, v)/\partial u$, $h_2 = \partial h(u, v)/\partial v$, $h_{11} = \partial^2 h(u, v)/\partial u^2$, $h_{12} = \partial^2 h(u, v)/\partial u \partial v$ and $h_{22} = \partial^2 h(u, v)/\partial v^2$ for the function $h(u, v)$ and similarly define g_1, g_2, g_{11}, g_{12} and g_{22} for the function $g(u, v)$. In the special case that $h_{12} = 0$, $h_1(a, b) < 0$ and $h_2(a, b) < 0$, we obtain the following approximation formula (details of the derivation are omitted here):

$$I_2 = \int_{u \geq a} \int_{v \geq b} e^{nh(u,v)} g(u, v) du dv \tag{5a}$$

$$= e^{nh(a,b)} \frac{g(a, b)}{h_1(a, b) h_2(a, b)} \frac{1}{n^2} \left(1 + \frac{C}{n} + \frac{D}{n^2} + O(n^{-3}) \right), \tag{5b}$$

where $C = C(g, h) = A(g, h_1) + A(g, h_2)$ and

$$D = D(g, h) = B(g, h_1) + B(g, h_2) + \frac{g_{12}}{g} \frac{1}{h_1 h_2} + \frac{g_1}{g} \frac{h_{22}}{(-h_1)(h_2)^2} + \frac{g_2}{g} \frac{h_{11}}{(-h_2)(h_1)^2} + \frac{h_{11} h_{22}}{(h_1)^2 (h_2)^2},$$

and A and B are defined as in (3) and (4). All of the above functions are evaluated at the point (a, b).

3. Species Richness Estimation in One Community

3.1. Abundance Data

Assume that there are S species in a community labelled from 1 to S. Denote the probabilities of species discovery by $(\theta_1, \theta_2, \ldots, \theta_S)$, where $\sum_{i=1}^{S} \theta_i = 1$. Here, the species discovery probability of any given species is generally a combination of species abundance and individual detectability. If all individuals in the community have the same probabilities of being detected, then the discovery probabilities represent the true relative abundances. Suppose a fixed number of n individuals are independently observed. Let X_i (species frequency) be the number of times, or individuals, that the ith species is discovered in the sample, $i = 1, 2, \ldots, S$. Only those species with $X_i > 0$ are observable in the sample. The species frequencies $(X_1, X_2, \ldots, X_S)$ are assumed to follow a multinomial distribution with cell total n and probabilities $(\theta_1, \theta_2, \ldots, \theta_S)$.

Let f_k, $k = 0, 1, \ldots, n$, (frequency counts) be the number of species represented by k times, or individuals, in the sample. That is, $f_k = \sum_{i=1}^{S} I(X_i = k)$, where $I(A)$ is the usual indicator function, i.e., $I(A) = 1$ if the event A occurs, and 0 otherwise. Here, f_0 denotes the number of undiscovered species in the sample. Thus, we have $n = \sum_{i=1}^{S} X_i = \sum_{k\geq 1} k f_k$. Let D denote the number of distinct species discovered in the sample, that is, $D = \sum_{i=1}^{S} I(X_i > 0) = \sum_{k\geq 1} f_k$.

Since the homogeneous assumption $(\theta_1 = \theta_2 = \cdots = \theta_S = 1/S)$ is unlikely to be valid for most communities, we assume that $(\theta_1, \theta_2, \ldots, \theta_S)$ are a random sample from a joint symmetric distribution with a common marginal density $g(\theta)$. This general model includes the well-known form proposed by Fisher, Corbet & Williams (1943) and the broken-stick model put forward by MacArthur (1957) as special cases. Note that, in the Fisher *et al.* (1943) model, species abundances are modelled as a random sample from a gamma distribution, such that the relative abundances (i.e., the sum of abundances is normalized to one) are distributed as a Dirichlet distribution and $g(\theta)$ is then a beta density. In the broken-stick model, $g(\theta)$ is a uniform density.

As discussed, we assume that the species discovery probabilities are bounded from below by a fixed constant. That is, there exists a constant $a > 0$ such that $\theta_i \geq a$, $i = 1, 2, \ldots, S$. In other words, every species has a discovery probability of at least a. Under these assumptions, the expected number of missing species can be expressed as $\mathrm{E}(f_0) = \sum_{i=1}^{S} \mathrm{EE}(I(X_i = 0)|\theta_i) = \sum_{i=1}^{S} \mathrm{E}(1 - \theta_i)^n$. This implies the following expression:

$$\mathrm{E}(f_0) = S \int_a^1 (1-\theta)^n g(\theta) d\theta = S \int_a^1 e^{n \log(1-\theta)} g(\theta) d\theta,$$

which is a form of the integral in (2a). Based on (2b), we obtain

$$\mathrm{E}(f_0) = S g(a)(1-a)^{n+1} \frac{1}{n} \left(1 + \frac{A_0}{n} + \frac{B_0}{n^2} + O(n^{-3})\right), \tag{6}$$

where $A_0 = A(g, h')$ and $B_0 = B(g, h')$ with $h(\theta) = \log(1 - \theta)$. For the expected number of singletons in the sample, we have

$$\mathrm{E}(f_1) = Sn \int_a^1 \theta(1-\theta)^{n-1} g(\theta)\, d\theta = Sn \int_a^1 e^{n\log(1-\theta)} (\theta(1-\theta)^{-1} g(\theta)) d\theta.$$

This is also a form of (2a) with $h(\theta) = \log(1 - \theta)$ and $g(\theta)$ replaced by $g^*(\theta) = \theta(1 - \theta)^{-1}\, g(\theta)$. Thus, formula (2b) leads to

$$\mathrm{E}(f_1) = Sag(a)(1-a)^n \left(1 + \frac{A_1}{n} + \frac{B_1}{n^2} + O(n^{-3})\right), \tag{7}$$

where $A_1 = A(g^*, h')$ and $B_1 = B(g^*, h')$. A similar derivation gives

$$\mathrm{E}(f_2) = \frac{1}{2} Sa^2 g(a)(1-a)^{n-1}(n-1)\left(1 + \frac{A_2}{n} + \frac{B_2}{n^2} + O(n^{-3})\right), \tag{8}$$

where $A_2 = A(g^{**}, h')$ and $B_2 = B(g^{**}, h')$ with $g^{**}(\theta) = \theta^2(1-\theta)^{-2}\, g(\theta)$. From expansions (6), (7) and (8), we have

$$\frac{\mathrm{E}(f_0)}{\mathrm{E}(f_1)} = \frac{1}{n}\frac{(1-a)}{a}\left(1 + \frac{A_0 - A_1}{n} + \frac{B_0 - A_0A_1 - B_1 + A_1^2}{n^2} + O(n^{-3})\right),$$

and

$$\frac{n-1}{n}\frac{\mathrm{E}(f_1)}{2\mathrm{E}(f_2)} = \frac{1}{n}\frac{(1-a)}{a}\left(1 + \frac{A_1 - A_2}{n} + \frac{B_1 - A_1A_2 - B_2 + A_2^2}{n^2} + O(n^{-3})\right).$$

Comparing the two expansion formulae above, we can verify the following relationships for the coefficients: $A_0 - A_1 = A_1 - A_2$ and $(B_0 - A_0A_1 - B_1 + A_1^2) - (B_1 - A_1A_2 - B_2 + A_2^2) = 1/a^2$. Then

$$\frac{\mathrm{E}(f_0)}{\mathrm{E}(f_1)} = \frac{n-1}{n}\frac{\mathrm{E}(f_1)}{2\mathrm{E}(f_2)} + \frac{1}{n^3}\frac{(1-a)}{a^3} + O(n^{-4}). \tag{9}$$

Consequently, we obtain

$$\mathrm{E}(f_0) - \frac{n-1}{n}\frac{(\mathrm{E}(f_1))^2}{2\mathrm{E}(f_2)} = Sg(a)(1-a)^{n+1}\frac{1}{a^2}\left(\frac{1}{n^3} + O(n^{-4})\right). \tag{10}$$

Replacing the expected values by the observed data in (10), we see that $(n-1)f_1^2/(2nf_2)$ provides an approximate lower bound for the number of missing species, with the bound being achieved under a homogeneous community. This lower bound was proposed in Chao (1984) using an alternative derivation. For abundance data, the sample size n is often large so that the term $(n-1)/n$ in the bound can be dropped. If $f_2 > 0$, the corresponding lower bound for species richness is $\hat{S}_{Chao1} = D + f_1^2/(2f_2)$, which has been referred to as the Chao1 estimator in the biological and ecological literature (e.g., Colwell & Coddington, 1994; Hughes *et al.*, 2001) and related software (Colwell, 2005).

The new approach here provides more insights about the role of the lower bound as an estimator, mainly because its theoretical performance can be seen more readily. It follows from (10) that the magnitude of the relative bias (relative to the true species richness) of the Chao1 estimator is approximately equal to $g(a)e^{-an}/(a^2n^3)$. The bias depends on the

rarity a of the rarest species, density $g(a)$ of the rarest species, and the sample size n. The magnitude of the bias could become substantial if there is a non-negligible number of very rare species with rarities tending to zero. In this case, the Chao1 method only provides a lower bound of species richness. On the other hand, if the species probabilities are bounded from below by a constant and the density at this bound is finite, then an implication from (10) is that the lower bound is very sharp if n is large enough. Hence, our approach also justifies the use of the Chao1 estimator in such situations as a valid estimator for large n. An estimated variance formula in Chao (1987) for the Chao1 estimator is $\hat{\text{var}}(\hat{S}_{Chao1}) = f_2(0.25(f_1/f_2)^4 + (f_1/f_2)^3 + 0.5(f_1/f_2)^2)$. When $f_2 = 0$, a bias-corrected form is given by $\hat{S}^*_{Chao1} = D + f_1(f_1 - 1)/2$. In this instance, the variance formula is modified to $\hat{\text{var}}(\hat{S}^*_{Chao1}) = 0.25 f_1(2f_1 - 1)^2 + 0.5 f_1(f_1 - 1) - 0.25 f_1^4/\hat{S}^*_{Chao1}$.

One advantage of the Chao1 estimator is that the estimated number of unseen species depends only on the first two frequency counts, i.e., the numbers of singletons and doubletons. This implies that ecologists do not need to obtain the exact frequency of any species that has at least three individuals in the sample. The estimator is especially useful if counting the exact number of individuals for each species appearing in the sample requires substantial effort, such as the case for estimating diversity for microbial communities (e.g., Bohannan & Hughes, 2003; Hughes *et al.*, 2001; Stach *et al.*, 2003; Walther & Morand, 1998). The modified Chao1 estimator for $f_2 = 0$ can be directly applied to data with only singletons, including DNA frequencies from hyper-diverse communities. For example, Borneman & Triplett (1997) examined a sample of Amazonian soil and found 100 unique clones. Our formula for the special case of $f_2 = 0$ then provides a plausible lower bound of 5050 for soil diversity in the Amazon.

3.2. Replicated Incidence Data

In many biological surveys, only species presence/absence data can be collected because it is impossible (e.g., in plant communities) to count individuals. In such cases, sampling is often conducted by several teams, or on multiple occasions. For example, in the Hong Kong Big Bird Race, only presence/absence was required for each species. There are a number of teams participating in the Race, and so replicated presence/absence data for each species were recorded each year. Another example is quadrat sampling in which the study area is divided into a number of quadrats, and species presence/absence data for a sample of quadrats are recorded. We use the general term 'sample' to refer to a team, a quadrat, an occasion, a site, a transect line, a fixed period of time or an investigator.

Assume that there are t samples and they are indexed $1, 2, \ldots, t$. The presence or absence of any species for these t samples is recorded to form a species-by-sample incidence matrix. In most applications, sufficient statistics from the species-by-sample incidence matrix are the incidence-based frequency counts $(Q_1, Q_2, \ldots, Q_t)$, where Q_k denotes the number of species that are detected in exactly k samples, $k = 1, 2, \ldots, t$. Hence, Q_1 represents the number of 'unique' species (those that are detected in only one sample) and Q_2 represents the number of 'duplicate' species (those that are detected in only two samples).

Assume that the species detection probabilities, defined as the chance of encountering at least one individual of a given species, are $(\theta_1, \theta_2, \ldots, \theta_S)$ and these probabilities are kept constant across samples. Under the assumption that $(\theta_1, \theta_2, \ldots, \theta_S)$ are a random sample from a distribution with density $g(\theta)$, and all the detection probabilities

are bounded from below by a constant, then parallel derivations to those in Section 3.1 can be made with n being replaced by t, and the counts $(f_1, f_2, \ldots, f_n)$ replaced by $(Q_1, Q_2, \ldots, Q_t)$. Therefore, an estimator based on presence/absence records for multiple samples has the form $\hat{S}_{Chao2} = D + ((t-1)/t)Q_1^2/(2Q_2)$, which is referred to in the literature as the Chao2 estimator. Unlike abundance data, the number of samples t for incidence data may not be large, so we suggest retaining the term $(t-1)/t$ in the estimator. This estimator was originally derived by Chao (1987) for capture-recapture data as a lower bound. With arguments analogous to those in the abundance case, its role as an estimator can also be justified under some conditions and its associated properties can be similarly derived.

4. Two Communities

4.1. Abundance Data

This section extends our approach to the estimation of shared species between two communities. Assume that there are S_1 species in community I and there are S_2 species in community II. The probabilities of species discovery in communities I and II are denoted $(\theta_1, \theta_2, \ldots, \theta_{S_1})$ and $(\lambda_1, \lambda_2, \ldots, \lambda_{S_2})$, respectively, where $\theta_i \geq a$, $\lambda_i \geq b$, $\sum_{i=1}^{S_1} \theta_i = 1$ and $\sum_{i=1}^{S_2} \lambda_i = 1$. Let the number of shared species be S_{12}. Without loss of generality, we assume that the first S_{12} species are the shared species. We further assume that (θ_i, λ_i), $i = 1, 2, \ldots, S_{12}$ are identically distributed with a density function $g(\theta, \lambda)$ on $\{(\theta, \lambda) : a \leq \theta \leq 1, b \leq \lambda \leq 1\}$.

Two random samples (samples I and sample II of sizes n and m) are taken from communities I and II, respectively. Assume that D_{12} shared species are observed. Denote the observed frequencies in the two communities by $(X_1, X_2, \ldots, X_{S_1})$ and $(Y_1, Y_2, \ldots, Y_{S_2})$. Let $f_{jk} = \sum_{i=1}^{S_{12}} I(X_i = j, Y_i = k)$ be the number of shared species that are observed j times in sample I and k times in sample II. Thus, $f_{11} = \sum_{i=1}^{S_{12}} I(X_i = 1, Y_i = 1)$ denotes the number of shared species that are singletons in both samples. Define $f_{j+} = \sum_{i=1}^{S_{12}} I(X_i = j, Y_i \geq 1)$ as the number of shared species that are observed in sample II and appear exactly j times in sample I, and define f_{+k} in a similar manner. Since $S_{12} = D_{12} + f_{+0} + f_{0+} + f_{00}$ and only D_{12} is observable, our approach is to find estimators for the expectations of the other three terms. We first estimate $\mathrm{E}(f_{00})$. Letting $\alpha = m/n$, we write

$$\begin{aligned}\mathrm{E}(f_{00}) &= S_{12} \int_b^1 \int_a^1 (1-\theta)^n (1-\lambda)^m g(\theta, \lambda)\, d\theta d\lambda \\ &= S_{12} \int_b^1 \int_a^1 e^{n(\log(1-\theta) + \alpha \log(1-\lambda))} g(\theta, \lambda)\, d\theta d\lambda,\end{aligned}$$

which is a form of the integral in (5a) with $h(\theta, \lambda) = \log(1-\theta) + \alpha \log(1-\lambda)$. Thus from (5b), we have

$$\mathrm{E}(f_{00}) = S_{12} \frac{g(a,b)}{h_1(a,b)h_2(a,b)} (1-a)^n (1-b)^m \frac{1}{n^2} \left(1 + \frac{C_0}{n} + \frac{D_0}{n^2} + O(n^{-3})\right),$$

where $C_0 = C(g, h)$, $D_0 = D(g, h)$, h_1, h_2, C and D are defined in Section 2. We can also write

$$\mathrm{E}(f_{11}) = S_{12}mn \int_b^1 \int_a^1 e^{n(\log(1-\theta)+\alpha\log(1-\lambda))}(\theta\lambda(1-\theta)^{-1}(1-\lambda)^{-1}g(\theta,\lambda))d\theta d\lambda,$$

Define $g^{***}(\theta, \lambda) = \theta\lambda(1-\theta)^{-1}(1-\lambda)^{-1} g(\theta, \lambda)$. It follows from (5b) that

$$\mathrm{E}(f_{11}) = S_{12}mn\frac{g^{***}(a,b)}{h_1(a,b)h_2(a,b)}(1-a)^n(1-b)^m\frac{1}{n^2}\left(1+\frac{C_1}{n}+\frac{D_1}{n^2}+O(n^{-3})\right),$$

where $C_1 = C(g^{***}, h)$ and $D_1 = D(g^{***}, h)$. Using these two expansions, we obtain the following generalized form of (9) (details of calculation are omitted):

$$\frac{\mathrm{E}(f_{00})}{\mathrm{E}(f_{11})} = \frac{(n-1)}{n}\frac{\mathrm{E}(f_{1+})}{2\mathrm{E}(f_{2+})}\frac{(m-1)}{m}\frac{\mathrm{E}(f_{+1})}{2\mathrm{E}(f_{+2})} + O(n^{-4}).$$

In many cases, the sample sizes m and n are large for abundance data; thus, an estimator for E (f_{00}) based on the above formula is $f_{11}\ f_{1+}\ f_{+1}/(4f_{2+}\ f_{+2})$. Following the same procedures as in the one-community case, we show that $\mathrm{E}(f_{0+})$ and $\mathrm{E}(f_{+0})$ can be estimated by $f_{1+}^2/(2f_{2+})$ and $f_{+1}^2/(2f_{+2})$, respectively. The proposed estimator for the number of shared species becomes

$$\hat{S}_{12} = D_{12} + f_{11}\frac{f_{1+}f_{+1}}{4f_{2+}f_{+2}} + \frac{f_{1+}^2}{2f_{2+}} + \frac{f_{+1}^2}{2f_{+2}}. \tag{11}$$

This estimator can be regarded as an extension of the Chao1 estimator to two communities. When $f_{2+} = 0$ or $f_{+2} = 0$, a bias-corrected form is

$$\hat{S}_{12}^{*} = D_{12} + f_{11}\frac{f_{1+}f_{+1}}{4(f_{2+}+1)(f_{+2}+1)} + \frac{f_{1+}(f_{1+}-1)}{2(f_{2+}+1)} + \frac{f_{+1}(f_{+1}-1)}{2(f_{+2}+1)}. \tag{12}$$

Because the proposed estimator can be regarded as a function of the statistics (D_{12}, f_{11}, f_{1+}, f_{2+}, f_{+1}, f_{+2}), we obtain a variance estimator by using a standard asymptotic approach under a multinomial distribution.

Chao *et al.* (2000) derived an estimator for shared species based on abundance data. The relative merits of their estimator and the one proposed are summarized as follows: (a) only the counts of singletons and doubletons are required for (11) or (12), whereas for the Chao *et al.* (2000) estimator more information (i.e., frequencies up to 10) is necessary; and (b) for highly heterogeneous communities, the Chao *et al.* (2000) estimator is generally less biased than the estimator (11) or (12), but it occasionally breaks down due to a zero estimate of sample coverage (so that a zero appears in a denominator), whereas the estimator (11) or (12) is always obtainable.

4.2. Replicated Incidence Data

The method developed for abundance data can be easily adapted to deal with the replicated incidence case. All notation and model formulations are similar to those in Section 4.1. The two sets of probabilities $(\theta_1, \theta_2, \ldots, \theta_{S_1})$ and $(\lambda_1, \lambda_2, \ldots, \lambda_{S_2})$ in the incidence case represent species detection probabilities in communities I and II, respectively. Assume that

there are t_1 samples taken from community I and there are t_2 samples from community II. In each sample, only presence/absence data are recorded. Let X_i and Y_i denote the number of samples in which the ith species are detected in communities I and II, respectively. Let $Q_{jk} = \sum_{i=1}^{S_{12}} I(X_i = j, Y_i = k)$ denote the number of shared species that are detected in j samples in community I and k samples in community II. Similarly, define Q_{j+} and Q_{+k} as in the abundance case. By applying a method analogous to that in Section 4.1, it can be shown that an estimator for the number of shared species based on incidence counts is

$$\hat{S}_{12} = D_{12} + Q_{11}\frac{(t_1-1)}{t_1}\frac{(t_2-1)}{t_2}\frac{Q_{1+}Q_{+1}}{4Q_{2+}Q_{+2}} + \frac{(t_1-1)}{t_1}\frac{Q_{1+}^2}{2Q_{2+}} + \frac{(t_2-1)}{t_2}\frac{Q_{+1}^2}{2Q_{+2}}. \quad (13)$$

The performance of this estimator was examined using simulations (see Section 6 for details). If $Q_{2+} = 0$ or $Q_{+2} = 0$, then a bias-corrected form is

$$\hat{S}_{12}^* = D_{12} + Q_{11}\frac{(t_1-1)}{t_1}\frac{(t_2-1)}{t_2}\frac{Q_{1+}Q_{+1}}{4(Q_{2+}+1)(Q_{+2}+1)} + \frac{(t_1-1)}{t_1}\frac{Q_{1+}(Q_{1+}-1)}{2(Q_{2+}+1)} + \frac{(t_2-1)}{t_2}\frac{Q_{+1}(Q_{+1}-1)}{2(Q_{+2}+1)}. \quad (14)$$

Variance estimators for both non-bias-corrected and bias-corrected estimators can be obtained from a standard asymptotic approach.

There have been no estimators of shared species richness based on replicated incidence data. When there are sufficiently many homogeneous samples available in each community, incidence frequencies may be good proxies of species abundance in each community, and the estimators in Section 4.1 can be applied. In this case, the total number of incidences is regarded as the sample size (i.e., n and m in Section 4.1) and thus the number of samples plays no role in the formulae. Our estimator in (13) or (14) shows the effect of the number of samples and is valid for any number of samples.

5. Numerical Example

The Hong Kong Big Bird Race (BBR) is an annual competition among teams of bird-watchers. The challenge is to record as many wild bird species in the Hong Kong territory as possible during a fixed interval of time. Birds in the Hong Kong territory include both resident and migratory birds (summer or winter visitors). Summer migratory visitors arrive in April and begin departing in September, whereas winter migratory visitors usually arrive at the end of October and return north in February/March. The Race was held annually in April up until 2000. The competition date was then moved to February because it was believed that there would be more species present. Since winter visitors in February belong to species not shared by summer visitors in April, the true number of shared species should be approximately equal to the species richness for the resident bird population in the Hong Kong territory.

Nineteen teams ($t_1 = 19$) competed in 1999 and 20 teams ($t_2 = 20$) in 2000. In 1999, a total of 217 species was observed by 19 teams and the winning team recorded 152 species. This means that the winning team missed 65 species that were observed by at least one of the other teams. In 2000, a total of 220 species was observed and the winning team recorded 154 species; thus, 66 species that were observed by the other teams were missed by the winning team. These records can be arranged as a 217×19 presence/absence data matrix in 1999

and as a 220×20 matrix in 2000. Merging the tables by species names, we found that there were $D_{12} = 115$ shared species. Our approach can then be applied to resolve the following questions. First, were there any species missed by all teams in each race? Second, did the data provide sufficient evidence that there are more species in February than in April? Finally, were there any unobserved shared species in the two sets of data?

We first present the species richness estimation for one community. The presence/absence data in 1999 shows that there were 20 unique species (that were detected by only one team) and 11 duplicated species (that were detected by only two teams). Based on the method for incidence data presented in Section 3.2, the estimated total number of species is $217 + (18/19)(400/22) = 234$ with an estimated standard error (s.e.) of 10.2 and a 95% confidence interval of (223, 267) based on a log-transformation (Chao, 1987). In 2000, there were 21 unique species and 16 duplicated species. A similar method yields an estimated species richness for 2000 of 233 (s.e. 7.5) with a 95% confidence interval of (225, 257). As a result, there appears insufficient evidence to support the conjecture that there are more bird species in February than in April in Hong Kong.

The replicated presence/absence data is similar to a capture-recapture matrix in estimating the size of an animal population. There is a simple analogy between species richness estimation for a multiple-species community and population size estimation for a single species. Consequently, the estimators in the capture-recapture models can be directly applied to the BBR data. The incidence-based coverage estimator (ICE) featured in SPADE (Chao, 2005; Chao & Shen, 2005) gives an estimate of 229 (s.e. 3.3) for 1999, and of 232 (s.e. 3.4) for 2000. For 1999, the first-order jackknife estimate (Burnham & Overton, 1979) is 236 (s.e. 6.1) and the second-order jackknife estimate is 245 (s.e. 10.2). For 2000, the corresponding first-order and second-order estimates are 240 (s.e. 6.2) and 245 (s.e. 10.4). The beta-binomial model (Lloyd & Yip, 1991; Lloyd, 1992) is also a useful model for capture-recapture studies. The maximum likelihood estimator (i.e., empirical Bayes estimator from the Bayesian viewpoint) under a beta-binomial model yields a species richness estimate of 284 (s.e. 39.3) for 1999 and of 368 (s.e. 177.9) for 2000. This empirical Bayes estimate for each year is significantly higher than the other estimates and its associated s.e. is also quite large.

We now proceed to estimate the shared species richness. From the merged (by species identity) data of the two years, we obtain frequency counts $Q_{1+} = 6$, $Q_{2+} = 4$, $Q_{+1} = 10$, $Q_{+2} = 7$ and $Q_{11} = 1$. Then formula (13) results in an estimate of 127 shared species (s.e. 7.6) with a 95% confidence interval of (119, 153). Thus, we can conclude that there were about 12 shared species not detected with a 95% confidence interval of (3, 38).

If the incidence counts are regarded as abundance frequency counts in each community, then based on the method presented in Section 4.1, we obtain the same estimate of 127 but a slightly higher s.e. of 8.0 for the shared species richness. The abundance-based Chao *et al*. (2000) method yields a lower estimate of 123 (s.e. 10.6). These two methods do not take the number of samples into account, although all the results do not differ much.

6. Concluding Remarks

Using the boundary-mode Laplace approximation formulae for the expected frequency counts based on abundance or replicated incidence data, we have developed useful simple estimators for species richness in one community and shared species richness in two communities when species discovery (or detection) probabilities are assumed to be bounded

from below. This assumption excludes cases in which there is a substantial proportion of undetectable or invisible species, for which no known method appears to offer accurate estimates.

For the case of one community, the relative merits of the various species richness estimators, including the two derived in this paper, has been extensively discussed (e.g., Colwell & Coddington, 1994; Walther & Morand, 1998). Since the proposed shared species estimators for two communities are new, we have conducted simulations to examine their performance specifically for replicated incidence data. Because of limited space, our simulation studies are not reported here. Generally, simulation results have shown that our estimators for the number of shared species perform well over a range of beta distributions for the detection probabilities when a sufficient amount of data is available. If there are not enough data, then our approach provides a reliable lower bound. Further work is needed to determine general guidelines about how many data are needed. Simulations also show that the estimated standard errors using an asymptotic method, although biased slightly downwards, are generally satisfactory when compared with the sample standard errors.

Our estimator of the number of unseen species is formulated in terms of the two lowest-order frequency counts. As a result, for abundance data, ecologists do not need to expend effort on recording the exact frequencies of those species that have at least three individuals in the sample. Similarly, our estimator for the shared species richness given in (11) and (12) implies that having the exact species frequency is not necessary for species that have at least three individuals in each of the two communities. Parallel conclusions are also valid for replicated incidence data.

The proposed estimators are non-parametric in the sense that they are not dependent on parametric assumptions for the heterogeneous species discovery (or detection) probabilities. Link (2003) demonstrated that two parametric models might result in the same conditional likelihood, defined as the likelihood conditional on the observed frequencies, but they could produce quite different population size estimates. This implies, from our point of view, one disadvantage of the parametric approach. However, Link (2003) also questioned the usefulness of non-parametric methods because 'they simply cannot be reliable in all cases' (Link, 2003 p. 1129). We believe that in many applications almost no estimators would work in all cases. For example, in estimating the centre of a symmetric density, one obtains different estimates under different parametric models (e.g., normal or double exponential). By way of contrast, a trimmed mean, as a non-parametric estimator, does not vary across models. However, it is expected that a trimmed mean performs well in some models, but less well in others. In a similar manner, we show that the proposed species richness estimators work satisfactorily under certain conditions, but otherwise our estimators serve only as lower bounds. In biological surveys, rigorously derived lower bounds are also informative in making inferences about community biodiversity.

The proposed methodology for replicated incidence data can be applied to estimate the population size of a single species. If each individual in a population can be uniquely marked or tagged, then each individual can be regarded as a species. In this case, the data consist of the capture frequency for each individual. Therefore, our approach can be applied in capture-recapture studies to estimate population sizes.

The estimators discussed in this paper for one community are featured in SPADE (Species Prediction And Diversity Estimation); see Chao & Shen (2005). The new shared species estimator for two communities will be included following publication of this paper.

Some comparison results will also be made available on the first-named author's website at http://chao.stat.nthu.edu.tw/.

References

ALHO, J.M. (1990). Logistic regression in capture-recapture models. *Biometrics* **46**, 623–635.

BOHANNAN, B.J.M. & HUGHES, J. (2003). New approaches to analyzing microbial biodiversity data. *Curr. Opin. Microbiol.* **6**, 182–187.

BORNEMAN, J. & TRIPLETT, E.W. (1997). Molecular microbial diversity in soils from eastern Amazonia: evidence from unusual microorganisms and microbial population shifts associated with deforestation. *Appl. Environ. Microbiol.* **63**, 2647–2653.

BUNGE, J. & FITZPATRICK, M. (1993). Estimating the number of species: A review. *J. Amer. Statist. Assoc.* **88**, 364–373.

BURNHAM, K.P. & OVERTON, W.S. (1979). Robust estimation of population size when capture probabilities vary among animals. *Ecology* **60**, 927–936.

CHAO, A. (1984). Nonparametric estimation of the number of classes in a population. *Scand. J. Statist.* **11**, 265–270.

CHAO, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43**, 783–791.

CHAO, A. (2005). Species richness estimation. In *Encyclopedia of Statistical Sciences*, Vol. 12, 2nd edn, eds N. Balakrishnan, C.B. Read & B. Vidakovic, pp. 7909–7916. New York: Wiley.

CHAO, A., HWANG, W.-H., CHEN, Y.-C. & KUO, C.-Y. (2000). Estimating the number of shared species in two communities. *Statist. Sinica* **10**, 227–246.

CHAO, A. & SHEN, T.-J. (2005). Program SPADE (Species Prediction And Diversity Estimation). Program and user's guide at http://chao.stat.nthu.edu.tw.

COLWELL, R.K. (2005). EstimateS: statistical estimation of species richness and shared species from samples. Version 7.5. User's guide and application published at http://viceroy.eeb.uconn.edu/estimates.

COLWELL, R.K. & CODDINGTON, J.A. (1994). Estimating terrestrial biodiversity through extrapolation. *Phil. Trans. R. Soc. Lond. B - Biol. Sci.* **345**, 101–118.

ERKANLI, A. (1994). Laplace approximations for posterior expectations when the mode occurs at the boundary of the parameter space. *J. Amer. Statist. Assoc.* **89**, 250–258.

ERKANLI, A. (1997). Boundary-mode approximations for posterior expectations. *J. Statist. Plann. Inference* **58**, 217–239.

FISHER, R.A., CORBET, A.S. & WILLIAMS, C.B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Animal Ecol.* **12**, 42–58.

GOUTIS, C. & CASELLA, G. (1999). Explaining the saddlepoint approximation. *Amer. Statist.* **53**, 216–224.

HUGHES, J.B., HELLMANN, J.J., RICKETTS, T.H. & BOHANNAN, B.J.M. (2001). Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl. Environ. Microbiol.* **67**, 4399–4406.

LINK, W.A. (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics* **59**, 1123–1130.

LLOYD, C.J. (1992). Modified martingale estimation for recapture experiments with heterogeneous capture probabilities. *Biometrika* **79**, 833–836.

LLOYD, C.J. & YIP, P.S.F. (1991). A unification of inference from capture-recapture studies through martingale estimating functions. In *Estimating Functions*, ed. V.P. Godambe, pp. 65–88. Oxford: Clarendon Press.

MACARTHUR, R.H. (1957). On the relative abundances of bird species. *Proc. Nat. Acad. Sci. U.S.A.* **43**, 193–295.

NORRIS, J.L. & POLLOCK, K.H. (1996). Nonparametric MLE under two closed capture-recapture models with heterogeneity. *Biometrics* **52**, 639–649.

SEBER, G.A.F. (1982). *The Estimation of Animal Abundance*, 2nd edn. London: Griffin.

STACH, J.E.M., MALDONADO, L.A., MASSON, D.G., WARD, A.C., GOODFELLOW, M. & BULL, A.T. (2003). Statistical approaches for estimating actinobacterial diversity in marine sediments. *Appl. Environ. Microbiol.* **69**, 6189–6200.

WALTHER, B.A. & MORAND, S. (1998). Comparative performance of species richness estimation methods. *Parasitology* **116**, 395–405.