Statistica Sinica **17**(2007), 1591-1600

ESTIMATING THE NUMBER OF SPECIES WITH MULTIPLE INCIDENCE-BASED SUBSAMPLES

Chang Xuan Mao

University of California, Riverside

Abstract: Estimating the number of species from multiple incidence-based subsamples is of great importance in ecological and environmental sciences. The problem is investigated in a mixture model of multivariate binomial densities. A sequence of algebraic lower bounds to the odds that a species is unseen in the survey is proposed. Estimating a lower bound leads to an estimator for the number of species. The nonparametric bootstrap method can be used to compute lower confidence limits. The asymptotic standard error for the first order estimator is provided. An example is investigated and a simulation experiment is carried out to assess the proposed estimators.

Key words and phrases: Biodiversity, truncated moment problem.

1. Introduction

Conservation or management of biodiversity has been a central problem in the ecological and environmental sciences. Among a variety of measures of biodiversity (Colwell and Coddington (1994), Magurran (2004) and Mao (2007)), the number of species is the most important one. Estimating the number of species can be based on incomplete surveys (Bunge and Fitzpatrick (1993) and Colwell and Coddington (1994)). Survey data are either abundance-based or incidence-based. Abundance-based data concern the number of individuals from each species found in the survey, and incidence-based data concern the occurrence of species over selected representative sites in a species assemblage (Colwell and Coddington (1994)).

A single sampling method can access only a certain number of species that constitute a latent subuniverse. Scientists often use a variety of methods to access species in a species assemblage (e.g., Longino, Coddington and Colwell (2002)). One method yields a single subsample from its subuniverse. These subuniverses are overlapping in the sense that some species are accessible to two or more methods. Analysis of a survey with multiple subsamples has been a long-time challenging task.

For an abundance-based sample with a random size, the number of individuals from a species is usually treated as a Poisson random variable. The numbers

of individuals from the same species in multiple subsamples are Poisson random variables with possibly different means, and their sum is also a Poisson random variable. This means that multiple abundance-based subsamples can be pooled to form a single aggregated sample to which existing estimators can be applied. When an abundance-based sample has a fixed size, a multinomial model arises. For example, Chao, Hwang, Chen and Kuo (2000) developed an estimator for the number of species shared by two subsamples in a multinomial model, which can be used to derive an estimator for the total number of species. Note that a Poisson model can still be useful for a multinomial sample, similar to the practice of using log-linear models to analyze multinomial data.

We focus on multiple incidence-based subsamples. Consider that, in a species assemblage, there are s species labeled by $a = 1, \ldots, s$. Let a survey contain K subsamples labeled by $b = 1, \ldots, K$. In the bth subsample, there are m_b representative sites. Let X_{ab} denote the number of representative sites where the ath species is detected. Assume that a species has the same probability of being detected in each site of the same subuniverse, so that X_{ab} is a binomial random variable with detection probability π_{ab}. Note that, for each a, $\sum_{b=1}^{K} X_{ab}$ is not a binomial random variable unless the π_{ab} are identical over b. This means that it is usually inappropriate to pool multiple incidence-based subsamples.

The detection pattern of the ath species is $\boldsymbol{X}_a = (X_{a1}, \ldots, X_{aK})'$. If $\boldsymbol{X}_a = \mathbf{0}$, then the ath species is unseen in the survey, where $\mathbf{0}$ is the origin of $\mathcal{R}^K$. The $\boldsymbol{X}_a$ arise as a random sample from a mixture when the $\boldsymbol{\Lambda}_a$ are assumed to follow a mixing distribution P, where $\boldsymbol{\Lambda}_a = (\Lambda_{a1}, \ldots, \Lambda_{aK})'$ and $\Lambda_{ab} = \pi_{ab}/(1-\pi_{ab})$. The special case with $m_b = 1$ for all b has been extensively studied in the capture-recapture literature (e.g., Chao (2001)). The case with $m_b \geqslant 2$ for all b will be investigated here. Let $n_+ = \sum_{i=1}^{s} I(\boldsymbol{X}_i \neq \mathbf{0})$ be the number of observed species, where $I(\cdot)$ is the indicator function. Conditioning on n_+, the problem of estimating the number of species s can be reduced to estimating the odds that a species is unseen in the survey. On the one hand, the odds is nonparametrically nonidentifiable, and an arbitrarily small perturbation to the mixture can be associated with an arbitrarily large increment in the odds. On the other hand, it is appropriate to develop lower bounds to the odds, and to construct lower confidence limits. To this end, a truncated moment problem will be formulated, in which the odds becomes the total mass of a measure over the positive half line, and finitely many higher order moments of this measure are functionals of a mixture. Based on such a truncated moment sequence, we consider an algebraic approach to construction of lower bounds to the odds. Estimating an algebraic lower bound leads to a pseudo maximum likelihood estimator for the number of species. The bootstrap method can be used to obtain lower confidence limits. For the first order lower bound estimator, the asymptotic standard error is also given.

This article is organized as follows. The mixture model is formulated in Section 2. The estimation method is presented in Section 3. In Section 4, an example is studied, and a simulation experiment is reported.

2. A Mixture Model

Let $\boldsymbol{m} = (m_1, \ldots, m_K)'$. A multivariate binomial density is given, dependence on $\boldsymbol{m}$ being suppressed for notational convenience, by

$$f_{\boldsymbol{x}}(\boldsymbol{\lambda}) = \prod_{b=1}^{K} \binom{m_b}{x_b} \frac{\lambda_b^{x_b}}{(1+\lambda_b)^{m_b}}, \quad \boldsymbol{x} \in \mathscr{F},$$

where $\mathscr{F}$ is the set of all detection patterns, as a finite subset of the mesh $\mathscr{I}$ of non-negative integer-coordinated vectors in the first closed orthant of $\mathcal{R}^K$. Let $n_{\boldsymbol{x}} = \sum_{a=1}^{s} I(\boldsymbol{X}_a = \boldsymbol{x})$ for $\boldsymbol{x} \in \mathscr{F}$. Because the $\boldsymbol{X}_a$ arise from a mixture $f_{\boldsymbol{x}}(P) = \int f_{\boldsymbol{x}}(\boldsymbol{\lambda})\, dP(\boldsymbol{\lambda})$, with $n_+ = s - n_{\mathbf{0}}$, the full likelihood of the number of species s and the mixing distribution P is, with $\mathscr{G} = \mathscr{F} \backslash \{\mathbf{0}\}$,

$$L(s, P) = \frac{s!}{(s-n_+)! \prod_{\boldsymbol{x} \in \mathscr{G}} n_{\boldsymbol{x}}!} \{f_{\mathbf{0}}(P)\}^{s-n_+} \prod_{\boldsymbol{x} \in \mathscr{G}} \{f_{\boldsymbol{x}}(P)\}^{n_{\boldsymbol{x}}}. \tag{1}$$

To reformulate the problem, we reparameterize P by Q, where

$$dQ(\boldsymbol{\lambda}) = \{1 - f_{\mathbf{0}}(P)\}^{-1} \{1 - f_{\mathbf{0}}(\boldsymbol{\lambda})\}\, dP(\boldsymbol{\lambda}). \tag{2}$$

This yields a mixture $g = g(Q)$ of truncated multivariate densities $g(\boldsymbol{\lambda})$, where

$$g_{\boldsymbol{x}}(Q) = \int g_{\boldsymbol{x}}(\boldsymbol{\lambda})\, dQ(\boldsymbol{\lambda}),\ g_{\boldsymbol{x}}(\boldsymbol{\lambda}) = f_{\boldsymbol{x}}(\boldsymbol{\lambda}) / \{1 - f_{\mathbf{0}}(\boldsymbol{\lambda})\}, \boldsymbol{x} \in \mathscr{G}.$$

Note that $g_{\boldsymbol{x}}(Q) = f_{\boldsymbol{x}}(P)/\{1 - f_{\mathbf{0}}(P)\}$ for $\boldsymbol{x} \in \mathscr{G}$. Let $\theta = f_{\mathbf{0}}(P)/\{1 - f_{\mathbf{0}}(P)\}$ be the odds that a species is unseen in the survey, which can be written as

$$\theta = \theta(Q) = \int f_{\mathbf{0}}(\boldsymbol{\lambda}) / \{1 - f_{\mathbf{0}}(\boldsymbol{\lambda})\}\, dQ(\boldsymbol{\lambda}). \tag{3}$$

Because the number of observed species n_+ is a binomial random variable with size s and probability $1 - f_{\mathbf{0}}(P) = (1+\theta)^{-1}$, the maximum likelihood estimator (MLE) for the number of species s given Q is $\lfloor n_+ + n_+\theta \rfloor$ (Lindsay and Roeder (1987)), where $\lfloor x \rfloor$ stands for the greatest integer no larger than $x \in \mathcal{R}$. An estimator for θ yields a pseudo MLE for s (Gong and Samaniego (1981)).

In the nonparametric mixture model, the number of support points, the support points and the mixing weights of Q are treated as unknown parameters. The odds θ is nonidentifiable in the sense that there exist two mixing distributions

Q and O such that $g_{\boldsymbol{x}}(Q) = g_{\boldsymbol{x}}(O)$ for all $\boldsymbol{x} \in \mathscr{G}$, but $\theta(Q) \neq \theta(O)$. The odds θ can also change dramatically. For example, consider $Q_\varepsilon = \sqrt{\varepsilon}\Delta(\varepsilon\mathbf{1}) + (1-\sqrt{\varepsilon})Q$, $\varepsilon \in (0,1)$, where $\mathbf{1} = (1,\ldots,1) \in \mathcal{R}^K$ and $\Delta(\boldsymbol{\lambda})$ is a distribution degenerate at $\boldsymbol{\lambda}$. By choosing a sufficiently small ε, the two mixture densities $g(Q)$ and $g(Q_\varepsilon)$ can be arbitrarily close while $\theta(Q_\varepsilon)$ can be arbitrarily large, because

$$\sum_{\boldsymbol{x}\in\mathscr{G}} |g_{\boldsymbol{x}}(Q_\varepsilon) - g_{\boldsymbol{x}}(Q)| = \sqrt{\varepsilon} \sum_{\boldsymbol{x}\in\mathscr{G}} |g_{\boldsymbol{x}}(\Delta(\varepsilon\mathbf{1})) - g_{\boldsymbol{x}}(Q)| \leqslant 2\sqrt{\varepsilon},$$

$$\theta(Q_\varepsilon) = \sqrt{\varepsilon}\theta(\Delta(\varepsilon\mathbf{1})) + (1-\sqrt{\varepsilon})\theta(Q) \geqslant \Big[\sqrt{\varepsilon}\Big\{\|\boldsymbol{m}\| + \sum_{i=2}^{\|\boldsymbol{m}\|} \binom{\|\boldsymbol{m}\|}{i} \varepsilon^{i-1}\Big\}\Big]^{-1},$$

where $\|\boldsymbol{x}\| = \sum_{b=1}^K |x_b|$ is the ℓ_1 norm of $\boldsymbol{x} \in \mathcal{R}^K$. One consequence is the nonexistence of genuine two-sided nonparametric confidence intervals (Mao and Lindsay (2004)). These observations invite us to develop lower bounds to θ.

3. Algebraic Lower Bounds

Given Q, the sharpest lower bound to the odds θ is

$$\varphi = \varphi(g) = \inf\{\theta(O) : g_{\boldsymbol{x}}(Q) = g_{\boldsymbol{x}}(O), \forall \boldsymbol{x} \in \mathscr{G}, \forall O\}. \tag{4}$$

Because it is difficult to calculate φ, it is of interest to develop lower bounds to θ that can be easily calculated, although maybe smaller than φ. To achieve this goal, we consider generalized moments of functions of $\boldsymbol{\lambda}$. In particular, we focus on $\|\boldsymbol{\lambda}\|$ with a weight function $\theta(\Delta(\boldsymbol{\lambda})) = f_{\mathbf{0}}(\boldsymbol{\lambda})/\{1 - f_{\mathbf{0}}(\boldsymbol{\lambda})\}$.

Write a discrete mixing distribution as $Q = \sum_{i=1}^{\iota} \pi_i \Delta(\boldsymbol{\lambda}_i)$, where Q has ι distinct support points $\boldsymbol{\lambda}_i$ with corresponding mixing weights π_i, $i = 1,\ldots,\iota$. There are κ distinct values among the $\|\boldsymbol{\lambda}_i\|$, denoted by ω_k, $k = 1,\ldots,\kappa$. The mixing distribution Q induces a finite univariate measure ν, where

$$\nu = \nu_Q = \sum_{k=1}^{\kappa} \Big\{ \sum_{i=1}^{\iota} \theta(\Delta(\boldsymbol{\lambda}_i))\pi_i I(\|\boldsymbol{\lambda}_i\| = \omega_k)\Big\}\Delta(\omega_k). \tag{5}$$

The moments $\mu(h)$ of ν are generalized moments of Q, where

$$\mu(h) = \mu(h,\nu) = \int \omega^h \, d\nu(\omega) = \int \|\boldsymbol{\lambda}\|^h \theta(\Delta(\boldsymbol{\lambda})) \, dQ(\boldsymbol{\lambda}), \quad h = 0,1,\ldots. \tag{6}$$

There is a multinomial expansion of $\|\boldsymbol{\lambda}\|^h$, i.e.,

$$\|\boldsymbol{\lambda}\|^h = \Big\{\sum_{b=1}^K \lambda_b\Big\}^h = \sum_{\{\boldsymbol{x}\in\mathscr{I}:\, \|\boldsymbol{x}\|=h\}} u_{\boldsymbol{x}} \prod_{b=1}^K \lambda_b^{x_b}, \quad h = 0,1,\ldots, \tag{7}$$

where $u_{\boldsymbol{x}} = \|\boldsymbol{x}\|! \prod_{b=1}^{K} (x_b!)^{-1}$ is a multinomial coefficient. For $h = 1, \ldots, m$, because $\{\boldsymbol{x} \in \mathscr{G} : \|\boldsymbol{x}\| = h\} = \{\boldsymbol{x} \in \mathscr{I} : \|\boldsymbol{x}\| = h\}$, write

$$\sum_{\{\boldsymbol{x} \in \mathscr{G} : \|\boldsymbol{x}\| = h\}} w_{\boldsymbol{x}} g_{\boldsymbol{x}}(Q) = \int \Big\{ \sum_{\{\boldsymbol{x} \in \mathscr{I} : \|\boldsymbol{x}\| = h\}} u_{\boldsymbol{x}} \prod_{b=1}^{K} \lambda_b^{x_b} \Big\} \theta(\Delta(\boldsymbol{\lambda})) dQ(\boldsymbol{\lambda}), \tag{8}$$

where $m = \min(m_1, \ldots, m_K) \geqslant 2$ and

$$w_{\boldsymbol{x}} = u_{\boldsymbol{x}} \prod_{b=1}^{K} \binom{m_b}{x_b}^{-1} = \Big(\sum_{b=1}^{K} x_b \Big)! \prod_{b=1}^{K} \frac{(m_b - x_b)!}{m_b!}. \tag{9}$$

It is clear that (3), (6), (7) and (8) yield the following.

Proposition 1. *The moments $\mu(0), \ldots, \mu(m)$ satisfy*

$$\mu(0) = \theta, \tag{10}$$

$$\mu(h) = \sum_{\{\boldsymbol{x} \in \mathscr{G} : \|\boldsymbol{x}\| = h\}} w_{\boldsymbol{x}} g_{\boldsymbol{x}}(Q), h = 1, \ldots, m. \tag{11}$$

There are numerous families of moments of univariate measures induced by real-valued functions of $\boldsymbol{\lambda}$. For example, given $\boldsymbol{d} \in \mathcal{R}^K$, consider

$$(\boldsymbol{d}'\boldsymbol{\lambda})^h = \Big\{ \sum_{b=1}^{K} d_b \lambda_b \Big\}^h = \sum_{\{\boldsymbol{x} \in \mathscr{I} : \|\boldsymbol{x}\| = h\}} \Big\{ u_{\boldsymbol{x}} \prod_{b=1}^{K} d_b^{x_b} \Big\} \prod_{b=1}^{K} \lambda_b^{x_b}.$$

While $\mu(h)$, for $h = 1, \ldots, m$, is nonparametrically identifiable, $\mu(h)$ for $h > m$ is not identifiable, since there exists $\boldsymbol{x} \in \mathscr{I}$ with $\|\boldsymbol{x}\| = h$ but $\boldsymbol{x} \notin \mathscr{G}$, or

$$\mu(h) = \sum_{\{\boldsymbol{x} \in \mathscr{I} : \|\boldsymbol{x}\| = h\}} u_{\boldsymbol{x}} \int \prod_{b=1}^{K} \lambda_b^{x_b} \, \theta(\Delta(\boldsymbol{\lambda})) dQ(\boldsymbol{\lambda}) > \sum_{\{\boldsymbol{x} \in \mathscr{G} : \|\boldsymbol{x}\| = h\}} w_{\boldsymbol{x}} g_{\boldsymbol{x}}(Q).$$

From Proposition 1, we consider determining lower bounds to the total mass of a finite measure from its truncated higher order moment sequence, an issue in the scope of the truncated Stieltjes moment problem (Curto and Fialkow (1991)).

By the Cauchy-Schwartz inequality, $\mu(0)\mu(2) \geqslant \mu^2(1)$, which implies that $\mu(0)$ has a lower bound $\mu^2(1)/\mu(2)$. This lower bound is identical to θ if ν in (5) is degenerate, which means that, while the support points of Q has the same ℓ_1 norm, Q is not necessarily degenerate.

To present more lower bounds to $\mu(0)$, for any natural number $k = 1, \ldots, \lfloor m/2 \rfloor$, consider the moment matrices

$$H_k(g) = (\mu(i+j))_{i,j=0}^{k}, a_k(g) = (\mu(j))_{j=1}^{k}, B_k(g) = (\mu(i+j))_{i,j=1}^{k}.$$

The moment matrix $H_k(g)$ is non-negative definite (Curto and Fialkow (1991)). When $H_k(g)$ is positive definite, we define

$$\gamma_k = \gamma_k(g) = a_k'(g)B_k^{-1}(g)a_k(g). \tag{12}$$

Note that $\gamma_1 = \mu(1)^2/\mu(2)$. For each k, $\gamma_k \leqslant \mu(0)$, because it is the unique zero of $|H_k(g)|$ as a function of $\mu(0)$, and $|H_k(g)| \geqslant 0$ for $\mu(0) \geqslant \gamma_k$. Let

$$\chi = \max\Big\{k \in \{1, \ldots, \lfloor \frac{m}{2} \rfloor\} : |B_k(g)| > 0\Big\}.$$

Theorem 2. *For $k = 1, \ldots, \chi$, γ_k exists and*

$$\gamma_k = \gamma_k(g(Q)) = \inf\{\theta(O) : \mu(h, \nu_O) = \mu(h, \nu_Q), \quad h = 0, \ldots, 2k, \forall O\}. \tag{13}$$

As an application of Mao and Lindsay (2007), Theorem 2 provides the condition under which γ_k can be defined. For each k, γ_k is also the infimum of the odds over all mixing distributions whose induced univariate measures have $2k$ specified moments identical to those of ν_Q induced from Q. From (13), it is clear that

$$\gamma_1 \leqslant \gamma_2 \leqslant \cdots \leqslant \gamma_\chi \leqslant \varphi. \tag{14}$$

The γ_k will be called *algebraic lower bounds*. The γ_k lead to a sequence of lower bounds s_k to the number of species s, where

$$s_k = \mathscr{E}(n_+) \cdot (1 + \gamma_k) = \frac{c(1 + \gamma_k)}{1 + \theta} \leqslant c, \quad k = 1, \ldots, \chi. \tag{15}$$

Because $n_{\boldsymbol{x}}/n_+$ estimates $g_{\boldsymbol{x}}(Q)$, the empirical moment $\hat{\mu}(h)$ estimates $\mu(h)$,

$$\hat{\mu}(h) = n_+^{-1} \sum_{\{\boldsymbol{x} \in \mathscr{G} : \|\boldsymbol{x}\| = h\}} w_{\boldsymbol{x}} n_{\boldsymbol{x}}, \quad h = 1, \ldots, m. \tag{16}$$

When $\mu(h)$ in the algebraic lower bound γ_k is replaced with $\hat{\mu}(h)$, we obtain $\hat{\gamma}_k$, which leads to a pseudo MLE $\hat{s}_k$ for the number of species s,

$$\hat{s}_k = n_+ + n_+\hat{\gamma}_k, \quad k = 1, \ldots, \lfloor \frac{m}{2} \rfloor. \tag{17}$$

Because the $\hat{\mu}(h)$ do not necessarily arise as moments from a finite measure over $(0, \infty)$, the $\hat{\gamma}_k$ may not share the same properties as the γ_k. For each $k = 1, \ldots, \lfloor m/2 \rfloor$, $\hat{s}_k$ can be calculated when the estimate for the matrix $B_k(g)$ is non-singular, so that the $\hat{s}_k$ may be not non-decreasing and they can even be smaller than n_+. A monotonized sequence of estimators can be defined by

$$\hat{s}_k = \begin{cases} n_+ + n_+ \max(0, \hat{\gamma}_1) & (k = 1), \\ \max(\hat{s}_{k-1}, n_+ + n_+\hat{\gamma}_k) & (k \geqslant 2). \end{cases} \tag{18}$$

The nonparametric bootstrap method can be used to construct lower confidence limits for s_k, which can be treated as conservative lower confidence limits for s.

Finally, we will further investigate the first order estimator $\hat{s}_1$. Write

$$\hat{s}_1 = n_+ + \Big\{\sum_{b=1}^{K} \frac{n_{\boldsymbol{e}(b)}}{m_b}\Big\}^2 \Big\{\sum_{b=1}^{K}\sum_{\ell=b}^{K} \frac{n_{\boldsymbol{e}(b)+\boldsymbol{e}(\ell)}}{m_b(m_\ell - \delta_{b\ell})2^{-1}}\Big\}^{-1}, \tag{19}$$

where $\{\boldsymbol{e}(b)\}_{b=1}^{K}$ is the set of standard bases of $\mathcal{R}^K$ and $\delta_{b\ell} = I(b = \ell)$. The following theorem can be easily obtained by the delta method.

Theorem 3. *The pseudo MLE $\hat{s}_1$ is asymptotically normally distributed with mean s_1 and variance σ^2, where*

$$\sigma^2 = sf_{\mathbf{0}}(P) + 4\frac{\alpha_{12}\alpha_{11}^2}{\alpha_{21}^2} + \frac{\alpha_{22}\alpha_{11}^4}{\alpha_{21}^4} - s^{-1}\left\{\frac{\alpha_{11}^2}{\alpha_{21}} - sf_{\mathbf{0}}(P)\right\}^2,$$

$$\alpha_{1t} = \sum_{b=1}^{K} \frac{sf_{\boldsymbol{e}(b)}(P)}{m_b^t}, \quad \alpha_{2t} = \sum_{b=1}^{K}\sum_{\ell=b}^{K} \frac{sf_{\boldsymbol{e}(b)+\boldsymbol{e}(\ell)}(P)}{\{m_b(m_\ell - \delta_{b\ell})2^{-1}\}^t}, \quad t = 1, 2.$$

When s is estimated by $\hat{s}_1$, the asymptotic standard error $\mathrm{se}\,(\hat{s}_1)$ is given by

$$\mathrm{se}\,(\hat{s}_1) = \left\{\frac{\hat{\alpha}_{11}^2}{\hat{\alpha}_{21}} + \frac{4\hat{\alpha}_{12}\hat{\alpha}_{11}^2}{\hat{\alpha}_{21}^2} + \frac{\hat{\alpha}_{22}\hat{\alpha}_{11}^4}{\hat{\alpha}_{21}^4}\right\}^{\frac{1}{2}}, \tag{20}$$

where, for $t = 1$ and 2, and for $i = 1$ and 2, $\hat{\alpha}_{it}$ estimates α_{it}, with

$$\hat{\alpha}_{1t} = \sum_{b=1}^{K} \frac{n_{\boldsymbol{e}(b)}}{m_b^t}, \quad \hat{\alpha}_{2t} = \sum_{b=1}^{K}\sum_{\ell=b}^{K} \frac{n_{\boldsymbol{e}(b)+\boldsymbol{e}(\ell)}}{\{m_b(m_\ell - \delta_{b\ell})2^{-1}\}^t}, \ t = 1, 2.$$

The asymptotic normality of $\hat{s}_1$ or that of $\log \hat{s}_1$ can be used to construct confidence intervals. For example, the $1 - \alpha$ lower confidence limit is given by

$$\hat{s}_1 \exp\{-z_{1-\alpha}\hat{s}_1^{-1}\mathrm{se}\,(\hat{s}_1)\}, \tag{21}$$

where $z_{1-\alpha}$ is the $1 - \alpha$ quantile of the standard normal distribution.

The special case $K = 1$ corresponds to a single incidence-based sample. When $K = 1$, with $m = m_1$ and $\boldsymbol{e}(1) = 1$, (19) and (20) are written as

$$\hat{s}_1 = n_+ + \frac{n_1^2}{2n_2}\cdot\frac{m-1}{m}, \ \mathrm{se}\,(\hat{s}_1) = \left\{\frac{n_1^2}{2n_2}\cdot\frac{m-1}{m} + \left(\frac{n_1^3}{n_2^2} + \frac{n_1^4}{4n_2^3}\right)\frac{(m-1)^2}{m^2}\right\}^{\frac{1}{2}},$$

which can be compared with the estimator and standard error in Chao (1989):

$$\hat{s}_1^{\mathrm{Chao}} = n_+ + \frac{n_1^2}{2n_2}, \ \mathrm{se}\,(\hat{s}_1^{\mathrm{Chao}}) = \left\{\frac{n_1^2}{2n_2} + \frac{n_1^3}{n_2^2} + \frac{n_1^4}{4n_2^3}\right\}^{\frac{1}{2}}.$$

The results in Chao (1989) are obtained via $(m-1)/m \approx 1$ for a large m.

4. Numerical Studies

4.1. An example

An example is taken from a rain forest ant study at the La Selva Biological Station in Costa Rica (Longino et al. (2002)). There were several methods used to sample ant species, among which the Barger and Malaise subsamples are considered here. The Barger subsample was obtained by N. Barger who used a mixture of honey and solid vegetable oil as the bait. The Malaise subsample was obtained using Malaise traps. The size of the Barger subsample is $m_1 = 40$ and that of the Malaise subsample is $m_2 = 62$. The number of observed species is $n_+ = 157$, among which 54 were found only in the Barger subsample, 92 only in the Malaise subsample, and 11 in both subsamples. The counts are presented in Table 1.

Table 1. The ant data. There are 38 observed detection patterns $\boldsymbol{x}$ with $n_{\boldsymbol{x}} \neq 0$.

x_1	x_2	$n_{\boldsymbol{x}}$	x_1	x_2	$n_{\boldsymbol{x}}$	x_1	x_2	$n_{\boldsymbol{x}}$	x_1	x_2	$n_{\boldsymbol{x}}$
1	0	24	15	0	2	0	4	7	9	14	1
2	0	6	25	0	1	0	5	4	0	15	1
3	0	7	0	1	37	0	6	2	4	15	1
4	0	2	3	1	1	0	7	1	2	16	1
5	0	2	14	1	1	0	9	3	0	18	1
6	0	3	0	2	17	2	9	1	0	19	2
7	0	2	5	2	1	0	10	2	0	20	1
8	0	3	0	3	10	0	12	1	5	29	1
9	0	1	1	3	2	0	13	1	–	–	–
11	0	1	23	3	1	0	14	2	–	–	–

The first order lower bound estimate is $\hat{s}_1 = 242 = \lfloor 242.9 \rfloor$, with a standard error $\mathrm{se}\,(\hat{s}_1) = 31.3$. The estimate $\hat{s}_2$ calculated from (17) is negative (-26.1). The 95% asymptotic lower confidence limit for the number of species s from (21) is 196. Although the standard error of $\hat{s}_1$ from 1,000 resamples is 35.5, a little larger than the asymptotic counterpart 31.3, the 95% bootstrap lower confidence limit for s from these resamples is 200, also a little larger than the asymptotic counterpart 196. This happens because the distribution of $\log \hat{s}_1$ can be skewed.

4.2. Simulation

The simulation experiment consisted of eight trials from a 2^3 design of three factors for a fixed $K = 3$. The number of species s was 200 or 1,000. The vector of binomial sizes $\boldsymbol{m}$ was $\boldsymbol{m}_1 = (10, 10, 10)'$ or $\boldsymbol{m}_2 = (20, 20, 20)'$. The mixing

distribution P was P_1 or P_2, where P_1 and P_2 are uniform distributions with six support points. The support points of P_1 were $0.05\boldsymbol{e}_1$, $0.05\boldsymbol{e}_2$, $.05\boldsymbol{e}_3$, $0.1\boldsymbol{e}_1$, $0.1\boldsymbol{e}_2$, and $0.1\boldsymbol{e}_3$, and those of P_2 were $0.1\boldsymbol{e}_1$, $0.1\boldsymbol{e}_2$, $0.1\boldsymbol{e}_3$, $0.1(\boldsymbol{e}_1+\boldsymbol{e}_2)$, $0.1(\boldsymbol{e}_1+\boldsymbol{e}_2)$, and $0.1(\boldsymbol{e}_2+\boldsymbol{e}_3)$. The measure ν in (5) determined by either P_1 or P_2 had two support points.

The lower bounds γ_k in (12) do not depend on the number of species. The first two lower bounds γ_1 and γ_2 were given, respectively, by 0.786 and 0.900 $(\boldsymbol{m}_1, P_1)$, 0.277 and 0.316 $(\boldsymbol{m}_2, P_1)$, 0.274 and 0.307 $(\boldsymbol{m}_1, P_2)$, and 0.068 and 0.073 $(\boldsymbol{m}_2, P_2)$. The odds θ was identical to γ_2 for all pairs of $(\boldsymbol{m}, P)$, which implies that γ_2 was also identical to the sharpest lower bound φ in (4). There was no approximation bias when γ_2 was used, while there is a negative approximation bias when γ_1 was used.

For each trial, the results from 1,000 samples that concerned the number of observed species n_+, and the monotonized estimates $\hat{s}_1$ and $\hat{s}_2$ in (18) are presented in Table 2. Both $\hat{s}_1$ and $\hat{s}_2$ have a positive estimation bias. The estimation bias and the standard error of $\hat{s}_2$ are larger than those of $\hat{s}_1$. While $\hat{s}_1$ has a non-positive total bias, $\hat{s}_2$ has a positive total bias. The 5% quantiles of $\hat{s}_1$ and $\hat{s}_2$ are smaller than those of s_1 and $s_2 = s$, respectively. The difference between the 5% quantile of $\hat{s}_1$ and that of $\hat{s}_2$ is small.

Table 2. Simulation results. Note that $s_2 = s$, and "ave", "se" and "lq" stand for the sample mean, the standard error, and the lower 5% quantile, respectively.

	$s = 200$				$s = 1,000$			
	P_1		P_2		P_1		P_2	
	$\boldsymbol{m}_1$	$\boldsymbol{m}_2$	$\boldsymbol{m}_1$	$\boldsymbol{m}_2$	$\boldsymbol{m}_1$	$\boldsymbol{m}_2$	$\boldsymbol{m}_1$	$\boldsymbol{m}_2$
s_1	188	194	195	199	940	970	975	995
ave (n_+)	105	152	153	186	526	760	765	932
se (n_+)	7	6	6	3	16	13	14	8
ave $(\hat{s}_1)$	194	194	196	200	945	972	976	996
se $(\hat{s}_1)$	30	14	14	6	59	32	31	14
lq $(\hat{s}_1)$	151	172	174	190	849	925	926	974
ave $(\hat{s}_2)$	212	207	211	205	1,106	1,044	1,039	1,038
se $(\hat{s}_2)$	92	78	65	26	1,957	464	417	478
lq $(\hat{s}_2)$	157	177	178	191	868	930	932	976

5. Discussion

A multivariate binomial mixture model is developed for an incomplete survey with multiple incidence-based subsamples. A sequence of lower bound estimators for the number of species is proposed. The approximation bias of a lower bound

estimator has a known direction. These lower bounds have analytic expressions that can be easily computed. Although higher order lower bounds have smaller approximation bias, it may be difficult to estimate them accurately, the first order lower bound estimator $\hat{s}_1$ is recommended.

The proposed estimators $\hat{s}_k$ are developed based on the method of moments. There are other possibilities, such as the nonparametric maximum likelihood method, that may be computationally challenging, but deserve further investigation. There may be alternative approaches by which lower bounds to the number of species can be defined (e.g., Mao (2006)).

References

Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: a review. *J. Amer. Statist. Assoc.* **88**, 364-373.

Chao, A. (1989). Estimating population size for sparse data in capture-recapture experiments. *Biometrics* **45**, 427-438.

Chao, A. (2001). An overview of closed capture-recapture models. *J. Agric. Biol. Environ. Stat.* **6**, 138-155.

Chao, A., Hwang, W. H., Chen, Y. C. and Kuo, C. Y. (2000). Estimating the number of shared species in two communities. *Statist. Sinica* **10**, 227-246.

Colwell, R. K. and Coddington, J. A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philos. Trans. R. Soc. Biol. Sci.* **345**, 101-118.

Curto, R. E. and Fialkow, L. A. (1991). Recursiveness, positivity, and truncated moment problems. *Houston J. Math.* **17**, 603-635.

Gong, G. and Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: theory and applications. *Ann. Statist.* **9**, 861-869.

Lindsay, B. G. and Roeder, K. (1987). A unified treatment of integer parameter models. *J. Amer. Statist. Assoc.* **82**, 758-764.

Longino, J. T., Coddington, J. A., and Colwell, R. K. (2002). The ant fauna of a tropical estimating species richness three different ways. *Ecology* **83**, 689-702.

Magurran, A. E. (2004). *Measuring Biological Diversity.* Blackwell.

Mao, C. X. (2006). Inference on the number of species via geometric lower bounds. *J. Amer. Statist. Assoc.* **101**, 1663-1670.

Mao, C. X. (2007). Estimating species accumulation curves and diversity indices. *Statist. Sinica* **17**, 761-774.

Mao, C. X. and Lindsay, B. G. (2004). Estimating the number of classes in multiple populations: a geometric analysis. *Canad. J. Statist.* **32**, 303-314.

Mao, C. X. and Lindsay, B. G. (2007). Estimating the number of classes. *Ann. Statist.* **35**, 917-930.

Department of Statistics, University of California, Riverside, CA, 92521, U.S.A.

E-mail: cmao@stat.ucr.edu

(Received July 2005; accepted February 2006)