SPECIES RICHNESS ESTIMATION

Abstract. Various models and estimation procedures for estimating the number of species in a community are reviewed under the following sampling schemes: sampling by continuous-type of efforts, sampling by individuals, and sampling by quadrats (or multiple occasions). Applications and relevant software are briefly reviewed.

Keywords and Phrases. Species richness, Abundance, Alpha diversity, Diversity indices.

AMS Subject Classification. Primary: 62P10; Secondary: 62F10, 62G05

Species richness (i.e., the number of species) is the simplest and the most intuitive concept for characterizing community diversity. We focus on the estimation of species richness based on a sample from a local community. This is also refered to as alpha diversity in ecological science. The topic is important for comparing communities in conservation and management of biodiversity, for assessing the effects of human disturbance on biodiversity, and for making environmental policy decisions. See references [21, 31, 34, 38, 44] for reviews on general ecological diversity as well as references [5, 16] for reviews specifically on species richness estimation. See also a recent book [49] for various sampling aspects and relevant methodologies.

In biological and ecological sciences, the compilation of complete species census and inventories often requires extraordinary efforts and is an almost unattainable goal in practical applications. There are undiscovered species in almost every taxonomic survey or species inventory. Traditional non-sampling-based approaches to estimating species richness include the following: (1) Extrapolating a species-accumulation or species-area curve to predict its asymptote, which is used as an estimate of species

richness. This approach has a long history and various curves have been presented in [23]; a summary is provided in "NON-SAMPLING-BASED EXTRAPOLATION," below. (2) Fitting a truncated distribution or functional form to the observed species abundances to obtain an estimate of species richness. The earliest approach was proposed by Preston [40], who fitted a truncated log-normal curve to the (properly grouped) frequencies and used the integrated value of the fitted curve over the real line as an estimate of the total number of species. Several major drawbacks have been noted regarding the non-sampling-based approaches; see [5, 16].

The work by Fisher, Corbet and Williams [22] provided the mathematical foundation on statistical sampling approaches to estimate species richness. Since then, a large body of literature discussing models and estimation under various sampling plans has been published. In addition to estimating the species richness for communities of plants or animals, the topic has a wide range of applications in various disciplines, as will be outlined in "APPLICATIONS," below.

There are two types of samplings: continuous-type (in which sampling efforts are continuous such as time, area or water volume) and discrete-type (sampling unit is an individual, quadrat or a trapping occasion). Most estimation procedures under both sampling and non-sampling frameworks require the use of a computer to obtain various estimates and their variances. Thus, user-friendly software has become an essential need in practical applications.

NOTATION

S total number of species in a community.

X_i number of times (frequency) the ith species is observed in the sample, $i = 1, 2, \cdots, S$; (Only those species with $X_i > 0$ are observable in the sample).

$I[A]$ the usual indicator function, i.e., $I[A] = 1$ if the event A occurs, 0 otherwise.

f_k number of species that are represented exactly k times in the sample, $k = 0, 1, \cdots, n$, $f_k = \sum_{i=1}^{S} I[X_i = k]$. ($f_0$ denotes the number of unobserved species).

n sample size, $n = \sum_{i=1}^{S} X_i = \sum_{k\geq 1} k f_k$.

D number of distinct species discovered in the sample, ($D = \sum_{i=1}^{S} I[X_i > 0] = \sum_{k\geq 1} f_k$).

t number of samples/quadrats or occasions.

Q_k number of species that are observed in exactly k samples, $k = 0, 1, \cdots, t$, based on presence/absence data.

SAMPLING BY CONTINUOUS-TYPE OF EFFORTS

Assume that the community is sampled by a continuous-type of effort and that the amount of efforts is increased from 0 to T. A common approach is based on the Poisson and mixed Poisson models. This approach can be traced back to Fisher, Corbet and Williams [22]. Assume that the S species are labeled from 1 to S. Individuals of the ith species arrive in the sample according to a Poisson process with a discovery rate λ_i. Here the rate is a combination of species abundance and individual detectability. If the detectability of individuals can be assumed to be equal across all species, then the rates can be interpreted as species abundances. In some applications, the exact arrival times for each individual are available, but in most biological samplings, only the frequencies of discovered species are recorded and would be sufficient for estimating species richness [35].

When multiple sets of frequency data are available, they can be pooled by species identities and analyzed under a mixed Poisson model. This is a payback for expending efforts on counting individuals per species in the sample.

In this sampling scheme, the sample size n (the number of individuals observed in the experiment) is a random variable. It is well-known that the conditional frequencies $(X_1, X_2, \cdots, X_S | \sum_{i=1}^S X_i = n)$ follow a multinomial distribution with cell total n and cell probabilities $p_k = \lambda_k / \sum_{i=1}^S \lambda_i$, $k = 1, 2, \cdots, S$. This is also the reason that many estimators are shared in both the continuous-type models and discrete-type (multinomial) models.

Based on different assumptions regarding the species discovery rates $(\lambda_1, \lambda_2, \cdots, \lambda_S)$, we classify all models into the following three categories:

(1) Homogeneous Models

In practical applications, the assumption of equal-rate $\lambda_1 = \lambda_2 = \cdots = \lambda_S \equiv \lambda$, is unlikely to be valid but this homogeneous model forms a basis for extension to more general models. Under the model, there are only two parameters S and λ. The likelihood over the effort $[0, T]$ can be expressed as $L(S, \lambda) \propto [S!/(S-D)!]\lambda^n \exp(-S\lambda T)$ (see [17]) and traditional inference procedures can be applied. The statistics D and n are complete and sufficient for S and λ. However, no unbiased estimators based on the sufficient statistics exist (see [35]). The profile likelihood for S is $L(S, \hat{\lambda}) \propto [S!/(S-D)!]S^{-n}$, where $\hat{\lambda} = n/(ST)$ denotes the maximum likelihood estimator (MLE) of λ in terms of S. It follows from ([17], p. 172) that the MLE of S is the solution of the equation $\sum_{j=1}^D (S-j+1)^{-1} = n/S$ when S is treated as a real number and the condition for differentiation is satisfied.

There are two approximations to the MLE in the literature, and they are, respectively, the solution of the two equations $S = D/(1-e^{-n/S})$ and $S = D/[1-(1-1/S)^n]$ [17, 18]. It can be shown that they correspond to, respectively, the conditional (on D) MLE [42] under the full likelihood and the profile likelihood. See subsequent material for a conditional MLE under parametric models. Both unconditional and conditional MLE's have identical asymptotic variance obtained by inverting the expected Fisher

information matrix from the corresponding likelihood [42].

Another useful estimator was suggested by Darroch and Ratcliff [19]. They provided a simple and explicit estimator with an asymptotic variance. The estimator is given by $\hat{S} = D/(1 - f_1/n)$. This estimator is highly efficient with respect to the MLE and was recommended in a comparative study [50]. It can also be regarded as a coverage-based estimator for a homogeneous case [13].

(2) Parametric and Bayes Models

In this approach, the species rates $(\lambda_1, \lambda_2, \cdots, \lambda_S)$ are modeled as a random sample from a mixing distribution with density $f(\lambda; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a low-dimensional vector. Many researchers have adopted a gamma density $f(\lambda; \alpha, \beta) = \beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}/\Gamma(\alpha)$ [22]. In the special case of $\alpha = 1$ (i.e., exponential distribution), the model is equivalent to a broken-stick model ([38] p. 285). Other parametric models include the log-normal [4], inverse-Gaussian [37], and generalized inverse-Gaussian [46].

An advantage of the parametric models is that the estimation reduces to an inference with only a few parameters and traditional estimation procedures can be applied. The likelihhood can be formulated as follows. For any mixture density $f(\lambda; \boldsymbol{\theta})$, define $P_{\boldsymbol{\theta}}(k)$, $k = 0, 1, \cdots$ as the probability that any species is observed k times in the sample, then

$$P_{\boldsymbol{\theta}}(k) = \int_0^\infty [(T\lambda)^k e^{-T\lambda}/k!] f(\lambda; \theta) d\lambda, \tag{1}$$

and $E(f_k) = S P_{\boldsymbol{\theta}}(k)$. The likelihood function for S and $\boldsymbol{\theta}$ can be written as

$$L(S, \boldsymbol{\theta}) = \frac{S!}{(S-D)! \prod_{k\geq 1}(f_k!)} [P_{\boldsymbol{\theta}}(0)]^{S-D} \prod_{k\geq 1} [P_{\boldsymbol{\theta}}(k)]^{f_k}. \tag{2}$$

The (unconditional) MLE and its asymptotic variance are obtained based on the above likelihood. Note that likelihood can be factored as $L(S, \boldsymbol{\theta}) = L_b(S, \boldsymbol{\theta}) L_c(\boldsymbol{\theta})$, where $L_b(S, \boldsymbol{\theta})$ is a likelihood with respect to D, a binomial$(S, 1 - P_{\boldsymbol{\theta}}(0))$, and $L_c(\boldsymbol{\theta})$

is a multinomial likelihood with respect to $\{f_k; k \geq 1\}$ with cell total D and zero-truncated cell probabilities $P_{\boldsymbol{\theta}}(k)/[1 - P_{\boldsymbol{\theta}}(0)]$, $k \geq 1$. The first likelihood $L_b(S, \boldsymbol{\theta})$ results in the conditional MLE $\hat{S} = D/[1 - P_{\hat{\boldsymbol{\theta}}}(0)]$, where $\widehat{\boldsymbol{\theta}}$ maximizes the second likelihood $L_c(\boldsymbol{\theta})$ [42]. These two types of MLE's can also be regarded as empirical Bayes estimators if we think of the mixing distribution as a prior having unknown parameters that must be estimated.

In the special case of a gamma-mixed Poisson model, $P_{\boldsymbol{\theta}}(k)$, or equivalently $E(f_k)$, $k = 0, 1, 2, \cdots$ correspond to individual terms of a negative-binomial distribution. When $\alpha = 1$, they correspond to the terms of a geometric distribution. As α tends to 0, $P_{\boldsymbol{\theta}}(k)$, $k = 0, 1, \cdots$ tends to the well-known logarithmic series, but this model does not yield an estimate of species richness ([38], p. 274).

By assigning various priors for parameters (S, α, β) in a gamma-Poisson model, a fully Bayesian hierarchical approach was proposed in [41]. Complicated calculations are handled by computer-intensive algorithms through the use of Gibbs sampling, a Markov Chain Monte Carlo method. The reader is referred to the above reference for previous work in the Bayesian direction.

A difficulty in the parametric or Bayesian approach lies in the selection of a mixing or a prior distribution. Two models with different mixing distributions may fit the data equally well, but they yield widely different estimates. Also, a model which gives a good fit to the data does not necessarily result in a satisfactory species richness estimate.

(3) Non-parametric Approaches

The above concerns have led to the non-parametric approaches, which avoid making assumptions about species discovery rates. In the following, we review six methods:

- *Jackknife Estimator (Burnham and Overton [7])*

Jackknife techniques were developed as a general method to reduce the bias of a biased estimator. Here the biased estimator is the number of species observed. The basic idea with the jth-order jackknife method is to consider sub-data by successively deleting j individuals from the original data. The first-order jackknife turns out to be $\hat{S}_{j1} = D + (n-1)f_1/n$. That is, only the number of singletons is used to estimate the number of unseen species. The second-order jackknife estimator for which the estimated number of unseen species is in terms of singletons and doubletons has the form $\hat{S}_{j2} = D + (2n-3)f_1/n - (n-2)^2 f_2/[n(n-1)]$. Higher orders of the jackknife estimators were given in Burnham and Overton [7]. A sequential testing procedure was also presented to select the best order. They recommended an interpolated jackknife estimator. All estimators can be expressed as linear combinations of frequencies and thus variances can be obtained.

• *Estimator by Chao [9]*

Based on the concept that rare species carry the most information about the number of missing ones, Chao [9] used only the singletons and doubletons to estimate the number of missing species. The estimator has a simple form $\hat{S} = D + f_1^2/(2f_2)$, and a variance formula is provided [10]. This estimator was originally proposed to be a lower bound. This bound is quite sharp and its use as a point estimate has been recently justified under practical assumptions; see [45]. However, this estimator breaks down when $f_2 = 0$. A modified bias-corrected version is $\hat{S} = D + f_1(f_1-1)/[2(f_2+1)]$, which is always obtainable.

• *Bootstrap Method (Smith and van Belle [47])*

A bootstrap estimator and its variance were developed [47] originally for quadrat samplings (see below), but the procedure can be applied directly to others. Given the n individuals who were already observed in the experiment, draw a random sample of size n from these individuals with replacement. Assume the proportion of the

individuals for the ith species in the generated sample is $\hat{p}_i$. Then a bootstrap estimate of species richness is calculated by the formula $\hat{S} = D + \sum_{i=1}^{S}(1-\hat{p}_i)^n$. After a sufficient number of bootstrap estimates are computed, their average is taken as the final estimate.

• *Abundance-based Coverage Estimator(ACE) (Chao and Lee [13], Chao et al., [12])*

The concept of sample coverage was originally proposed by Turing and Good [24]. In a mixed Poisson model, the sample coverage is defined as $C = \sum_{i=1}^{S} \lambda_i I[X_i > 0]/\sum_{i=1}^{S} \lambda_i$, which represents the sum of the rates associated with the discovered species. This approach aims to estimate S via the sample coverage estimation; see below. It is also assumed in this approach that the species discovery rates are fully characterized by their mean $\bar{\lambda} = \sum_{i=1}^{S} \lambda_i/S$ and CV (coefficient of variation). The squared CV, γ^2 , is defined as $\gamma^2 = \sum_{i=1}^{S}(\lambda_i - \bar{\lambda})^2/(S\bar{\lambda}^2)$. The larger the CV, the greater the degree of heterogeneity among species rates.

The approach seperates the observed frequencies into two groups: abundant and rare. Abundant species are those having more than κ individuals in the sample, and the observed rare species are those represented by only one, two, $\cdots$, and up to κ individuals in the sample. A value of the cut-off point, $\kappa = 10$, is suggested based on empirical evidence [15]. For abundant species, only the presence/absence information is needed because they would be discovered anyway. Hence, it is not necessary to record the exact frequencies for those species that have already reached a sufficient number (say, 10) of representatives in the sample. The exact frequencies for the rare species are required because the estimation of the number of missing species is based entirely on these frequencies. For long-tailed data, separation is essential; and no separation usually results in positively biased estimates [16].

Let the total number of abundant and rare species in the sample be $S_{abun} = \sum_{i=\kappa+1}^{n} f_i = \sum_{i=1}^{S} I[X_i > \kappa]$ and $S_{rare} = \sum_{i=1}^{\kappa} f_i = \sum_{i=1}^{S} I[0 < X_i \leq \kappa]$. Then

the estimator of species richness based on the estimated sample coverage $\hat{C} = 1 - f_1/\sum_{i=1}^{\kappa} i f_i$ is given by $\hat{S} = S_{abun} + (S_{rare} + f_1\hat{\gamma^2})/\hat{C}$, where $\hat{\gamma^2} = \max\{S_{rare}\sum_{i=1}^{\kappa} i(i-1)f_i/[\hat{C}(\sum_{i=1}^{\kappa} i f_i)^2] - 1, 0\}$ denotes the estimated squared CV ([12], Section 2). For highly heterogeneous communities, a bias-corrected CV estimator is provided in [13].

- *Non-parametric MLE (Norris and Pollock [36])*

A mixed Poisson model with a non-parametric mixing distribution F is considered in this approach. By substituting $P_{\boldsymbol{\theta}}(k) = \int (e^{-T\lambda}T^k\lambda^k/k!)dF(\lambda)$ for $k = 0, 1, \cdots$ into Equation (2), the likelihood can be expressed as a function of S and the entire distribution F. Based on an EM algorithm, the non-parametric MLE of F turns out to be a discrete distribution with a finite number of support points. This is equivalent to dividing the species rates into several classes, with the rates in each class being identical. A bootstrap method was proposed in [36] to obtain variance estimators.

- *Coverage-based Horvitz-Thompson Estimator (Ashbridge and Goudie [1])*

In sampling theory, the Horvitz-Thompson estimator has been used to adjust the effect of unobserved sampling units in an unequal sampling scheme. When it is applied to species richness estimation, the estimator takes the form $\hat{S} = \sum_{k\geq 1} f_k/[1 - \exp(-k\hat{C})]$, where $\hat{C} = 1 - f_1/n$ denotes the estimated sample coverage. The concept of sample coverage is used here for adjustment of the sample fraction of unseen species. A bootstrap procedure is used to obtain a variance estimator and confidence interval.

SAMPLING BY INDIVIDUALS

In many biological studies (e.g., bird, insect, mammal and plant), it is often the case that one individual is observed or encountered at a time and classified as to species identity. Suppose a fixed number of n individuals are independently observed from the study site. The commonly used models are the multinomial model (in which an individual may be observed repeatedly) and the multivariate hypergeometric model (any individual can only be observed or counted once). In the

former case, the frequencies $(X_1, X_2, \cdots, X_S)$ are assumed to have a multinomial distribution with cell total n and probabilities $(p_1, p_2, \cdots, p_S)$, where p_k denotes the species discovery probability of the kth species, $k = 1, 2, \cdots, S$ and $\sum_{i=1}^{S} p_i = 1$. In the latter case, the frequencies $(X_1, X_2, \cdots, X_S)$ are assumed to have a multivariate hypergeometric with a likelihood $\binom{N}{n}^{-1} \prod_{i=1}^{S} \binom{N_i}{X_i}$, where N_k denotes the total number of individuals for the kth species in the community and $N = \sum_{i=1}^{S} N_i$. Most researchers have assumed that N is known, but this information is rarely available in biological sampling. When only a small portion of individuals is selected for each species, the multinomial provides a good approximation with $p_i = N_i/N$. Thus, we focus on the multinomial model. Parallel to the mixed Poisson model, there are three classes of models here too:

(1) Homogeneous Model

This model assumes that $p_1 = p_2 = \cdots = p_S = 1/S$. There is only one parameter S and the likelihood is $L(S) \propto [S!/(S-D)!]S^{-n}$. Note that this likelihood is identical to the profile likelihood of S in an equal-rate Poisson model; thus the MLE and its properties are the same as those discussed there. In contrast to a homogeneous continuous-effort model, the minimum variance unbiased estimator of S does exist in a multinomial model if $n \geq S$ [18].

(2) Parametric and Bayes Models

Ecologists usually present species frequencies graphically in two different ways. One way is to rank the frequencies $(X_1, X_2, \cdots, X_S)$ from the most abundant to least abundant and plot the frequency of each species with respect to its rank (1 means the most abundant species). To characterize the theoretical patterns, a functional form is selected to model $(p_1, p_2, \cdots, p_S)$. The most popular functional forms include the geometric $p_i \propto \alpha(1-\alpha)^{i-1}$ and the Zipf-Mandelbrot law $p_i \propto (i+\alpha)^{-\theta}$, where α and θ are parameters. Although these types of models can produce species richness

estimates [5], they are mainly useful for describing the features of abundant species especially for applications in linguistics. Moreover, simulation studies have shown [6] that these estimates generally do not perform satisfactorily. A random-effect model assuming that $(p_1, p_2, \cdots, p_S)$ follows a Dirichlet distribution leads to $E(p_i) = S^{-1}\sum_{k=i}^{S}(1/k)$, which is equivalent to a broken stick model.

The other way to present frequency data is to plot f_k with respect to k, $k = 1, 2, \cdots$. The theoretical patterns can be examined by fitting a discrete zero-truncated distribution or a functional form to the histogram of frequencies. The three widely used distributions are the zero-truncated negative-binomial, geometric and logarithmic series models; these models have been discussed in the mixed Poisson models.

Bayesian models under a Dirichlet prior for $(p_1, p_2, \cdots, p_S)$ and a negative binomial for S were considered in [31]. See reference [48] for other types of priors and relevant Bayesian estimators.

(3) Non-parametric Approaches

All the non-parametric approaches described for the mixed Poisson models are valid here except that the Horvitz-Thompson estimator is modified to $\hat{S} = \sum_{k\geq 1} f_k/[1-(1-k\hat{C}/n)^n]$. The exact variances for any estimator under the mixed Poisson and multinomial models are different because the sample size n in the latter case is fixed. However, the asymptotic variances are very close.

MULTIPLE SAMPLES OR MULTIPLE OCCASIONS

Counting the exact number of individuals for each species appearing in the sample requires substantial effort or may become impossible (e.g., in plant communities). In such cases, incidence (presence/absence) data are commonly collected over repeated samples in time and space. Quadrat sampling provides an example in which the study area is divided into a number of quadrats, and a sample of quadrats are randomly selected for observation. There are other examples: similar sampling is conducted by

several investigators, or trapping records are collected over multiple occasions. We use the general term "sample" in what follows to refer to a quadrat, occasion, site, transect line, a period of fixed time, a fixed number of traps, or an investigator, etc.

Assume that there are t samples and they are indexed by $1, 2, \cdots, t$. The presence and absence of any species in any sample are recorded to form a species-by-sample incidence matrix. This $S \times t$ matrix is similar to a capture-recapture matrix in estimating the size of an animal population. For most applications, the sufficient statistics from the species-by-sample incidence matrix are the incidence counts $(Q_1, Q_2, \cdots, Q_t)$, where Q_k denotes the number of species that are detected in k samples, $k = 1, 2, \cdots, t$. There is a simple analogy between species richness estimation for multiple-species communities and population size estimation for single species. The capture probability in a capture-recapture study corresponds to species detection probability, which is defined as the chance of encountering at least one individual of a given species. Therefore, the estimation techniques in the capture-recapture technique can be directly applied to estimate species richness. There has been an explosion of methodological research on capture-recapture in the past two decades. A recent comprehensive review of methodology and applications is provided by Schwarz and Seber [43].

A sequence of useful models was proposed by Pollock [39] for analyzing capture-recapture data and has been used in [2, 7, 49] to estimate species richness. Three sources of variations in species detection probability are considered: (i) model $\mathcal{M}_t$, which allows probabilities to vary by time or sample; (ii) model $\mathcal{M}_b$, which allows behavioral responses to previous records; and (iii) model $\mathcal{M}_h$, which allows heterogeneous detection probabilities. Various combinations of the above three variations (i.e., models $\mathcal{M}_{tb}$, $\mathcal{M}_{th}$, $\mathcal{M}_{bh}$ and $\mathcal{M}_{tbh}$) are also considered. A wide range of statistical estimation methods have been proposed in the literature. These estimators

rely on many different approaches: the maximum likelihood, the jackknife method, the bootstrap method, log-linear or generalized log-linear models, Bayesian methods, mixture models, sample coverage procedures, and martingale estimating functions [11, 43, 44].

Models with behavioral response (i.e., models $\mathcal{M}_b$, $\mathcal{M}_{tb}$, $\mathcal{M}_{tb}$ and $\mathcal{M}_{tbh}$) allow the detection probability of any species to depend on whether the observer has already recorded it at "previous" samples. Thus ordering is implicitly involved in these four models. Meanwhile, almost all estimation procedures derived under these models depend on the ordering of the samples. These models are useful only for temporally replicated samples, especially when the sampling is conducted by a single investigator or when only data on the accumulation of previously undiscovered species are used (see below). Therefore, models $\mathcal{M}_t$, $\mathcal{M}_h$ and $\mathcal{M}_{th}$ are more potentially useful for species estimation. Since heterogeneity is expected in natural communities, this leaves models $\mathcal{M}_h$ and $\mathcal{M}_{th}$.

A multiplicative form of model $\mathcal{M}_{th}$ assumes that the detection probability P_{ij}, the probability of detecting the ith species in the jth sample, has the form $P_{ij} = \pi_i e_j$, $0 < \pi_i e_j < 1$; here the parameters $\{e_1, e_2, \cdots, e_t\}$, $\{\pi_1, \pi_2, \cdots, \pi_S\}$ are used, respectively, to denote the unknown sample effects and heterogeneity pattern. The latter is mostly determined by species abundance structure whereas the former is closely related to sampling efforts, quadrat area, sampling method, landscape and other environmental variables associated with each sample. When the sample effects can be assumed to be identical (e.g, equal-size quadrats, equal-effort sampling with similar protocols), this model reduces to model $\mathcal{M}_h$, i.e., $P_{ij} = \pi_i$. In this model, the number of incidences (occurrences) for any species is a binomial random variable. A common parametric approach is the beta-binomial model, where the heterogeneity effects are assumed to be a random sample from a beta distribution. The likelihood

is similar to that in Equation (2) with $P_{\boldsymbol{\theta}}(k)$ being replaced by a beta-binomial form. Therefore, the maximum likelihood or empirical Bayes estimation procedures can be similarly obtained.

A major advantage of the non-parametric methods is that they can be applied to various types of samplings with only slight modifications. All the non-parametric approaches presented for the two previous sampling schemes can be adapted for use in model $\mathcal{M}_h$ with n being replaced by t, and the capture frequencies $\{f_1, f_2, \cdots, f_n\}$ there replaced by the incidence counts $(Q_1, Q_2, \cdots, Q_t)$. Actually most of the non-parametric estimators were originally derived for closed capture-recapture experiments. The coverage-based method can be directly extended [14, 32] to yield estimators for model $\mathcal{M}_{th}$ when a sufficient number of samples (say, 5) are available. The resulting estimators are referred to as ICE (Incidence-based Coverage Estimator) in the program EstimateS (see below). There is relatively little literature for model $\mathcal{M}_{th}$. See [11] for recent advances. Kendall (in [30]) provided valuable discussion on the robustness of some methods to violation of the closure assumption.

We remark that a logistic model $\mathcal{M}_{th}$ was proposed by Huggins [28] and can be expressed as $P_{ij} = \pi_i e_j/(1 + \pi_i e_j)$, which is also known as the Rasch model in educational statistics. There are several approaches to this model including the log-linear approach, mixture models and latent class models [11]. The relevant covariates or auxiliary variables can be easily incorporated to explain heterogeneity effects in analysis.

NON-SAMPLING-BASED EXTRAPOLATION

The earliest attempts to study communities started with finding the relationship between species richness and the area that the survey covered. A species-area or species-accumulation curve (or collector's curve, species-cover curve) is a plot of the accumulated number of species found with respect to the number of units of effort

expended. The effort may correspond to either a continuous type (area, trap-time, volumes) or a discrete-type (individuals, sampling occasions, quadrats, number of nets). This curve as a function of effort monotonically increases and typically approaches an asymptote, which is the total number of species. The species-accumulation curve has been used by biologists or ecologists to assess inventory completeness, to estimate the minimum effort needed to reach a certain level of completeness, to standardize the comparison of various inventories, and to use the estimated asymptote as a species richness estimate.

There is extensive literature on the various functional forms used to fit the curves [23]. Let D_t denote the cumulative number of species for t units of effort. Two early models proposed in the literature are $D_t = \alpha t^\beta$ and $D_t = \alpha + \beta \log t$, where α, β are parameters to be estimated from data. These two non-asymptotic models are useful for species richness estimation when the study area is known or a finite number of efforts would result in a complete census.

For the models with an asymptotic value S, we group them into the following three categories: (In each category, α, β and μ are additional parameters.)

(1) Negative exponential model and its generalizations: These include the exponential model $D_t = S[1 - \exp(-\alpha t)]$, and two generalized forms $D_t = S[1 - \exp(-\alpha t)]^\beta$ and $D_t = S\{1 - \exp[-\alpha(t - \beta)^\mu]\}$ (Weibull model).

(2) Hyperbolic curve and its generalization: These include the Michaelis-Menten equation $D_t = St/(\beta + t)$, and two generalized forms $D_t = (\alpha + St)/(\beta + t)$ and $D_t = St^\alpha/(\beta + t^\alpha)$ (logistic model).

(3) Other models include $D_t = S(1 - \alpha\beta^t)$ and $D_t = S\{1 - [1 + (t/\alpha)^\beta]^{-\mu}\}$.

In addition to the uses mentioned above, there are other reasons for researchers adopting an extrapolation method: (1) Only presence/absence data are required and

thus efforts to count individuals of each species in the sample can be avoided. (2) No specification about species abundance structure is needed. (3) It can be applied to all sampling schemes. However, there are some concerns regarding this approach: (1) A sufficient amount of data is needed to construct the accumulation shape, so it can not be used on sparsely sampled communities. (2) Various forms may fit the data well, but the asymptotic values are drastically different. (3) A good fit does not imply the extrapolated asymptote is a good estimate because the prediction is out of the range for which data are available. (4) The shape of the curve depends on the sequential order in which efforts are accumulated. When different orders are used, the curve may be totally different. As a result, the estimates may vary. (5) The variance of the resulting extrapolated value cannot be theoretically justified without further assumptions, and theoretical difficulties arise for model selection.

Sampling-based approaches (i.e., removal model) have recently been introduced for dealing with species accumulated data [8]. The removal model is statistically equivalent to model $\mathcal{M}_b$ or $\mathcal{M}_{bh}$ discussed earlier. This new direction thus links the traditional extrapolation with the capture-recapture models.

APPLICATIONS

In the following, we list some application areas along with specific goals in each:

- Population biology: estimation of the size (i.e., the total number of individuals) of biological populations [49].

- Genetics: estimation of the number of genes or alleles based on sample frequency counts [27].

- Medical science and epidemiology: estimation of the number of different cases for a specific disease by merging several incomplete lists of individuals [11, 26].

- Environmental science: estimation of the number of organic pollutants that were discharged to a water environment [29].

- Software reliability: estimation of the number of undiscovered bugs in a piece of software when data in debugging processes are available [3].

- Numismatics and archaeology: estimation of the number of die types for ancient coins found in a hoard [25].

- Linguistics: estimation of the size of vocabulary for an author based on his/her known writings [20].

SOFTWARE

A program EstimateS which calculates various estimators of species richness is available from Robert Colwell's website at http://viceroy.eeb.uconn.edu/estimates. Another program SPADE (Species Prediction And Diversity Estimation) developed by the author and colleagues is downloadable from the author's website at http://chao.stat.nthu.edu.tw.

A widely used program, CAPTURE, for capture-recapture analysis can be applied to estimate species richness for incidence data collected on multiple sampling occasions; the program is provided at Gary White's website at http://www.cnr.colostate.edu/~gwhite/software.html. An additional program CARE (for CApture-REcapture) which accommodates some recently developed estimators is available from the author's website given above.

Acknowledgements. This work was supported by the National Science Council of Taiwan.

Anne Chao

Institute of Statistics
National Tsing Hua University
Hsin-Chu, TAIWAN 30043
chao@stat.nthu.edu.tw

References

[1] Ashbridge, J. and Goudie, I. B. J. (2000). Coverage-adjusted estimators for mark-recapture in heterogeneous populations. *Commun. Statist.-Simul. Comput.*, **29**, 1215-1237.

[2] Boulinier, T., Nichols, J. D., Sauer, J. R., Hines, J. E., and Pollock, K. H. (1998). Estimating species richness: the importance of heterogeneity in species detectability. *Ecology*, **79**, 1018-1028.

[3] Briand, L. C., El Emam, K., Freimut, B. G., and Laitenberger, O. (2000). A comprehensive evaluation of capture-recapture models for estimating software defect content. *IEEE Trans. Software Engrg.*, **26**, 518-540.

[4] Bulmer, M. G. (1974). On fitting the Poisson lognormal distribution to species abundance data. *Biometrics*, **30**, 101-110.

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

[6] Bunge, J., Fitzpatrick. M., and Handley, J. (1995). Comparison of three estimators of the number of species. *J. Appl. Stat.*, **22**, 45-59.

[7] Burnham, K. P. and Overton, W. S. (1979). Robust estimation of population size when capture probabilities vary among animals. *Ecology*, **60**, 927-936.

[8] Cam, E., Nichols, J. D., Sauer, J. R., and Hines, J. E. (2002). On the estimation of species richness based on the accumulation of previously unrecorded species. *Ecography*, **25**, 102-108.

[9] Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scand. J. Statist.*, **11**, 265-270.

[10] Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, **43**, 783-791.

[11] Chao, A. (2001). An overview of closed capture-recapture models. *J. Agric. Bio. Environ. Stat.*, **6**, 158-175.

[12] 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.

[13] Chao, A. and Lee, S.-M. (1992). Estimating the number of classes via sample coverage. *J. Amer. Statist. Ass.*, **87**, 210-217.

[14] Chao, A., Lee, S.-M., and Jeng, S.-L. (1992). Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. *Biometrics*, **48**, 201-216.

[15] Chao, A., Ma, M.-C., and Yang, M. C. K. (1993). Stopping rules and estimation for recapture debugging with unequal failure rates. *Biometrika*, **80**, 193-201.

[16] Colwell, R. K. and Coddington, J. A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philos. Trans. Royal Soc., London, Series B*, **345**, 101-118.

[17] Craig, C. C. (1953). On the utilization of marked specimens in estimating population of flying insects. *Biometrika*, **40**, 170-176.

[18] Darroch, J. N. (1958). The multiple-recapture census. I: estimation of a closed population. *Biometrika*, **45**, 343-359.

[19] Darroch, J. N. and Ratcliff, D. (1980). A note on capture-recapture estimation. *Biometrics*, **36**, 149-153.

[20] Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: how many words did Shakespeare know? *Biometrika*, **63**, 435-447.

[21] Engen, S. (1978). *Stochastic Abundance Models*. Chapman and Hall, London.

[22] Fisher, R. A., Corbet, A. S., and 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. Anim. Ecol.*, **12**, 42-58.

[23] Flather, C. H. (1996). Fitting species-accumulation functions and assessing regional land use impacts on avian diversity. *J. Biogeogr.*, **23**, 155-168.

[24] Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237-264.

[25] Holst, L. (1981). Some asymptotic results for incomplete multinomial or Poisson samples. *Scand. J. Statist.*, **8**, 243-246.

[26] Hook, E. B. and Regal, R. R. (1995). Capture-recapture methods in epidemiology: methods and limitations. *Epid. Reviews*, **17**, 243-264.

[27] Huang, S. P. and Weir, B. S. (2001). Estimating the total number of alleles using a sample coverage method. *Genetics*, **159**, 1365-1373.

[28] Huggins, R. M. (1991). Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics*, **47**, 725-732.

[29] Janardan, K. G. and Schaeffer, D. J. (1981). Methods for estimating the number of identifiable organic pollutants in the aquatic environment. *Water Resources Res.*, **17**, 243-249.

[30] Kendall, W. L. (1999). Robustness of closed capture-recapture methods to violations of the closure assumption. *Ecology*, **80**, 2517-2525.

[31] Krebs, C. J. (1999). *Ecological Methodology* (2nd Edition). Addison Wesley, Menlo Park, CA.

[32] Lee, S.-M. and Chao, A. (1994). Estimating population size via sample coverage for closed capture-recapture models. *Biometrics*, **50**, 88-97.

[33] Lewins, W. A. and Joanes, D. N. (1984). Bayesian estimation of the number of species. *Biometrics*, **40**, 323-328.

[34] Magurran, A. E. (1988). *Ecological Diversity and Its Measurement.* Princeton University Press, Princeton, New Jersey.

[35] Nayak, T. K. (1991). Estimating the number of component processes of a superimposed process. *Biometrika*, **78**, 75-81.

[36] Norris III, J. L. and Pollock, K. H. (1998). Non-parametric MLE for Poisson species abundance models allowing for heterogeneity between species. *Environ. Ecol. Statist.*, **5**, 391-402.

[37] Ord, J. K. and Whitmore, G. A. (1986). The Poisson-inverse Gaussian distribution as a model for species abundance. *Commun. Statist.-Theory Methods*, **15**, 853-871.

[38] Pielou, E. C. (1977). *Mathematical Ecology.* Wiley, New York.

[39] Pollock, K. H. (1991). Modeling capture, recapture, and removal statistics for estimation of demographic parameters for fish and wildlife populations: past, present, and future. *J. Amer. Statist. Ass.*, **86**, 225-238.

[40] Preston, F. W. (1948). The commonness and rarity of species. *Ecology*, **29**, 254-283.

[41] Rodrigues J., Milan L. A., and Leite, J. G. (2001). Hierarchical Bayesian estimation for the number of species. *Biometrical J.*, **43**, 737-746.

[42] Sanathanan, L. (1977). Estimating the size of a truncated sample. *J. Amer. Statist. Ass.*, **72**, 669-672.

[43] Schwarz, C. J. and Seber, G. A. F. (1999). A review of estimating animal abundance III. *Stat. Sci.*, **14**, 427-456.

[44] Seber, G. A. F. (1982). *The Estimation of Animal Abundance (2nd Edition)*, Griffin, London.

[45] Shen, T.-J., Chao, A., and Lin, J.-F. (2003). Predicting the number of new species in further taxonomic sampling. *Ecology*, **84**, 798-804.

[46] Sichel, H. S. (1997). Modelling species-abundance frequencies and species-individual functions with the generalized inverse Gaussian-Poisson distribution. *S. Afri. Statist. J.*, **31**, 13-37.

[47] Smith, E. P. and van Belle, G. (1984). Nonparametric estimation of species richness. *Biometrics*, **40**, 119-129.

[48] Solow, A. R. (1994). On the Bayesian estimation of the number of species in a community. *Ecology*, **75**, 2139-2142.

[49] Williams, B. K., Nichols, J. D., and Conroy, M. J. (2002). *Analysis and Management of Animal Populations.* Academic Press, San Diego, CA.

[50] Wilson, R. M. and Collins, M. F. (1992). Capture-recapture estimation with samples of size one using frequency data. *Biometrika*, **79**, 543-553.

Related Entries: Capture-Recapture Methods, Diversity Indices, Ecological Statistics