SORT 41 (1) January-June 2017, 3-54

Thirty years of progeny from Chao's inequality: Estimating and comparing richness with incidence data and incomplete sampling

Anne Chao[1,*] and Robert K. Colwell[2,3,4]

Abstract

In the context of capture-recapture studies, Chao (1987) derived an inequality among capture frequency counts to obtain a lower bound for the size of a population based on individuals' capture/non-capture records for multiple capture occasions. The inequality has been applied to obtain a non-parametric lower bound of species richness of an assemblage based on species incidence (detection/non-detection) data in multiple sampling units. The inequality implies that the number of undetected species can be inferred from the species incidence frequency counts of the uniques (species detected in only one sampling unit) and duplicates (species detected in exactly two sampling units). In their pioneering paper, Colwell and Coddington (1994) gave the name "Chao2" to the estimator for the resulting species richness. (The "Chao1" estimator refers to a similar type of estimator based on species abundance data). Since then, the Chao2 estimator has been applied to many research fields and led to fruitful generalizations. Here, we first review Chao's inequality under various models and discuss some related statistical inference questions: (1) Under what conditions is the Chao2 estimator an unbiased point estimator? (2) How many additional sampling units are needed to detect any arbitrary proportion (including 100%) of the Chao2 estimate of asymptotic species richness? (3) Can other incidence frequency counts be used to obtain similar lower bounds? We then show how the Chao2 estimator can be also used to guide a non-asymptotic analysis in which species richness estimators can be compared for equally-large or equally-complete samples via sample-size-based and coverage-based rarefaction and extrapolation. We also review the generalization of Chao's inequality to estimate species richness under other sampling-without-replacement schemes (e.g. a set of quadrats, each surveyed only once), to obtain a lower bound of undetected species shared between two or multiple assemblages, and to allow inferences about undetected phylogenetic richness (the total length of undetected branches of a phylogenetic tree connecting all species), with associated rarefaction and extrapolation. A small empirical dataset for Australian birds is used for illustration, using online software SpadeR, iNEXT, and PhD.

MSC: 62D07, 62P07.

Keywords: Cauchy-Schwarz inequality, Chao2 estimator, extrapolation, Good-Turing frequency formula, incidence data, phylogenetic diversity, rarefaction, sampling effort, shared species richness, species richness.

[*] Corresponding author. E-mail: chao@stat.nthu.edu.tw

[1] Institute of Statistics, National Tsing Hua University, Hsin-Chu 30043, Taiwan.

[2] Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, USA.

[3] University of Colorado Museum of Natural History, Boulder, CO 80309, USA.

[4] Departmento de Ecologia, Universidade Federal de Goiás, CP 131, 74.001-970, Goiânia, GO, Brasil.

Received: December 2016.

1. Introduction

Thirty years ago, Chao (1987) developed an inequality among capture frequency counts to obtain a lower bound of population size based on individuals' capture/non-capture records in multiple-stage, closed capture-recapture studies. An earlier version of Chao's inequality and the corresponding lower bound (Chao, 1984) estimated the number of classes under a classic occupancy problem. Those inequalities and lower bounds were derived for their pure mathematical interest, as the models are simple and elegant, and also for their statistical interest, because these inequalities can be used to make inference about the richness of the undetected portion of a biological assemblage based on incomplete data.

In the first decade after their publication, these Chao-type lower bounds were rarely applied in other disciplines. In 1994, Colwell and Coddington published a seminal paper on estimating terrestrial biodiversity through extrapolation. They applied both of Chao's formulas (1984, 1987) to estimate species richness, because there is a simple analogy between the incidence data in species richness estimation for a multiple-species assemblage and the capture-recapture data in population size estimation for a single species. Chao (1984) had suggested that her occupancy-based estimator might be applied to estimating species richness, and offered examples of its application to capture-recapture data, the focus of Chao (1987). Colwell and Coddington distinguished two types of data: individual-based abundance data (counts of the number of individuals of each species within a single sampling unit) and multiple sampling-unit-based incidence data (counts of occurrences of each species among sampling units). They gave the name "Chao1" to the estimator of species richness specifically for abundance data, based on the Chao (1984) formula, and the name "Chao2" for incidence data based on the Chao (1987) formula. Colwell also featured these two estimators along with others in the widely used software EstimateS (Colwell, 2013; Colwell and Elsensohn, 2014). Since then, both the Chao1 and Chao2 estimators have been increasingly applied to many research fields, not only in ecology and conservation biology, but also in other disciplines; see Chazdon et al. (1998), Magurran (2004), Chao (2005), Gotelli and Colwell (2011), Magurran and McGill (2011), Gotelli and Chao (2013) and Chao and Chiu (2016) for various applications. Chao's inequalities also led to numerous generalizations under different models or frameworks; some closely related generalizations were accomplished by Mao (2006, 2008), Mao and Lindsay (2007), Rivest and Baillargeon (2007), Pan, Chao and Foissner (2009), Böhning and van der Heijden (2009), Lanumteangm and Böhning (2011), Böhning et al. (2013), Mao et al. (2013), Chiu et al. (2014), and Puig and Kokonendji (2017). In addition to EstimateS, these two estimators have now been included in other software and several R packages in CRAN (e.g. packages Species, Specpool, entropart, fossil, SpadeR, iNEXT, among others).

During the past 30 years, Chao and her students and collaborators have developed a number of population size and species richness estimators based on several other statistical models, including Chao and Lee's (1992) abundance- or incidence-based coverage estimators (ACE and ICE, two names bestowed by Chazdon et al., 1998), martingale estimators, estimating-function estimators, maximum quasi-likelihood estimators, and Horvitz-Thompson-type estimators; see Chao (2001) and Chao and Chiu (2016) for a review. These developments are more complicated and mathematically sophisticated than the estimators derived from Chao's inequalities. Surprisingly, it turns out that the earliest and simplest estimators are the most useful ones for biological applications.

In this paper, we mainly focus on Chao's (1987) inequality and its subsequent developments for multiple incidence data. For both practical and biological reasons, recording species detection/non-detection in multiple sampling units is often preferable to enumerating individuals in a single sampling unit (abundance data). For microbes, clonal plants, and sessile invertebrates, individuals are difficult or impossible to define. For mobile organisms, replicated incidence data are less likely to double-count individuals. For social animals, counting the individuals in a flock, herd, or school may be difficult or impractical. Also, replicated incidence data support statistical approaches to richness estimation that are just as powerful as corresponding abundance-based approaches (Chao et al., 2014b). Moreover, a further advantage is that replicated incidence records account for spatial (or temporal) heterogeneity in the data (Colwell et al., 2004, 2012).

In Sections 2.1 and 2.2, we first review the general model formulation for incidence data and the Chao (1987) inequality. Three related statistical inference problems are discussed:

1. In Section 2.3, we ask under what conditions the Chao2 estimator is an unbiased point estimator. Chao et al. (2017) recently provided an intuitive answer to this question for abundance data, from a Good-Turing perspective. Here we use a generalization of the Good-Turing frequency formula to answer the same question for incidence data.
2. In Section 2.4, we ask how many additional sampling units are needed to detect any arbitrary proportion (including 100%) of the Chao2 estimate. The Chao2 species richness estimator does not indicate how much sampling effort (additional sampling units) would be necessary to answer the question. Here we review the solution proposed by Chao et al. (2009).
3. In Section 2.5, we review approaches that use other incidence frequency counts to obtain similar-type lower bounds. In Chao's (1987) formula, the estimator for the number of undetected species is based only on the frequency counts of the uniques (species detected in only one sampling unit) and duplicates (species detected in exactly two sampling units). Lanumteangm and Böhning (2011), Chiu et al. (2014), Puig and Kokonendji (2017) made advances by extending Chao's inequality to use higher-order incidence frequency counts. Here we mainly review Puig and Kokonendji's (2017) extension, which leads to a series of lower bounds for species richness. Their framework was based mainly on abundance data, but it can be readily applied to multiple incidence data.

In Section 3, we show that, no matter whether the Chao2 formula is unbiased or biased low, it can always be used to guide a non-asymptotic analysis in which a species richness estimator can be compared for equally-large samples (based on a common number of sampling units) or equally-complete samples (based on a common value of sample completeness, as measured by coverage; see later text). Sample-size-based and coverage-based rarefaction and extrapolation provide a unified sampling approach to fairly comparing species richness across assemblages.

In the subsequent three sections we review three generalizations of Chao's inequality to estimate species richness under other sampling schemes (Section 4), to estimate shared species richness between two or multiple assemblages (Section 5), and also to make inferences about phylogenetic diversity, which incorporates species evolutionary history (Section 6). The next three paragraphs introduce these generalizations.

Chao's original inequality was developed under the assumption that sampling units are assessed with replacement. When sampling is done without replacement, e.g. quadrats or time periods are not repeatedly selected/surveyed, or mobile species are collected by lethal sampling methods, suitable modification is needed. In Section 4, we review the modifications developed by Chao and Lin (2012).

Compared with estimating species richness in a single assemblage, the estimation of shared species richness, taking undetected species into account, has received relatively little attention; see Chao and Chiu (2012) for a review. For two assemblages, shared species richness plays an important role in assessing assemblage overlap and forms a basis for constructing various types of beta diversity and (dis)similarity measures, such as the classic Sørensen and Jaccard indices (Colwell and Coddington, 1994; Magurran, 2004; Chao et al., 2005, 2006; Jost, Chao and Chazdon, 2011; Gotelli and Chao, 2013). In Section 5, we review the work by Pan et al. (2009), who extended Chao's inequality to the case of multiple assemblages to obtain a lower bound of undetected species shared between two or multiple assemblages.

A rapidly growing literature discusses phylogenetic diversity, which incorporates evolutionary histories among species into diversity analysis (see Faith, 1992; Warwick and Clarke, 1995; Crozier, 1997; Webb and Nonoghue, 2005; Petchey and Gaston, 2002; Cadotte et al., 2009; Cavender-Bares, Ackerly and Kozak, 2012). The most widely used phylogenetic metric is Faith's (1992) *PD* (phylogenetic diversity), which is defined as the sum of the branch lengths of a phylogenetic tree connecting all species in the target assemblage. As shown by Chao et al. (2010, 2015), *PD* can be regarded as a measure of phylogenetic richness, i.e. a phylogenetic generalization of species richness. Throughout this paper, *PD* refers to Faith's (1992) *PD*. When some species are present, but undetected by a sample, the lineages/branches associated with these undetected species are also missing from the phylogenetic tree spanned by the observed species. The undetected *PD* in an incomplete sample was not discussed until recent years (Cardoso et al., 2014; Chao et al., 2015). In Section 6, we review the phylogenetic version of Chao's inequality, developed recently by Chao et al. (2015), and the associated phylogenetic version of the rarefaction/extrapolation approach.

In Section 7, a small empirical dataset for Australian birds is used for illustration using online software, including Chao's SpadeR, iNEXT, and PhD. Section 8 provides discussion and conclusions. The diversity measures discussed in this review (species richness, shared species richness, and *PD*) do not take species abundances into account. We briefly discuss the extension of these measures to incorporate species abundances, and refer readers to relevant papers. Major notation used in each section is shown in Table 1.

2. Species richness estimation

2.1. A general framework: Sampling-unit-based incidence data and model

As indicated in the Introduction, Chao's (1987) original inequality was formulated based on a capture-recapture model to estimate the size of a population, but here we consider a framework based on species incidence (detection/non-detection) data to estimate species richness. These two statistical inference problems are equivalent. Assume that there are S species indexed $1,2,\ldots,S$ in the focal assemblage, where S is the estimating target in species richness estimation. Here we mainly consider the model developed by Colwell et al. (2012) for multiple incidence data. Assume that there are Tsampling units, and that they are indexed $1,2,\ldots,T$. The sampling unit is usually a trap, net, quadrat, plot, or timed survey, and it is these sampling units, not the individual organisms, that are sampled randomly and independently. The observed data consist of species detection/non-detection in each sampling unit. In a typical spatial study, these sampling units are deployed randomly in space within the area encompassing the assemblage. However, in a temporal study of diversity, the T sampling units would be deployed in one place at different independent points in time (such as an annual breeding bird census at a single site).

For any sampling unit, the model assumes that the ith species has its own unique incidence or detection probability π_i that is constant among all randomly selected sampling units. The incidence probability π_i is the probability that species i is detected in a sampling unit. Here $\sum_{i=1}^{S} \pi_i$ will generally not be equal to unity.

The incidence records consist of a species-by-sampling-unit incidence matrix $\{W_{ij};\ i = 1,2,\ldots,S,\ j = 1,2,\ldots,T\}$ with S rows and T columns; here $W_{ij} = 1$ if species i is detected in sampling unit j, and $W_{ij} = 0$ otherwise. Let Y_i be the number of sampling units in which species i is detected, $Y_i = \sum_{j=1}^{T} W_{ij}$; here Y_i is referred to as the sample *species incidence frequency*. Species present in the assemblage but not detected in any sampling unit yield $Y = 0$. See Section 6.1 for a hypothetical example and Appendices A and B for real data. Details about these data are provided in subsequent sections.

Table 1: *Major notation used in each section.*

Common notation and/or one-assemblage species richness estimation (Section 2)	
S	Number of species in an assemblage.
π_i	Detection or incidence probability of species i, $i = 1, 2, \ldots, S$ in a sampling unit.
T	Number of sampling units taken from an assemblage.
U	Total number of incidences in T sampling units.
ϕ_r	Mean detection probability of species that appeared in r sampling units, $r = 0, 1, \ldots, T$.
W_{ij}	Species detection/non-detection: $W_{ij} = 1$ if species i is detected in sampling unit j, and $W_{ij} = 0$ otherwise, $i = 1, 2, \ldots, S$, $j = 1, 2, \ldots, T$.
S_{obs}	Number of observed species in T sampling units.
Y_i	Species incidence frequency (number of sampling units in which species i is detected).
Q_k	Number of species detected in exactly k sampling units in the data, $k = 0, 1, \ldots, T$.
$\hat{}$	"Hat" above a parameter: an estimator of the parameter, e.g. $\hat{S}$, $\hat{\pi}_i$ and $\hat{\phi}_r$ denote, respectively, estimators of S, π_i and ϕ_r.
Rarefaction and extrapolation of one-assemblage species richness (Section 3)	
$C(T)$	Coverage for a reference sample of size T.
$C(t)$	Coverage in a hypothetical rarefied sample of t sampling units if $t < T$.
$C(T+t^*)$	Coverage in a hypothetical augmented sample of $T + t^*$ sampling units.
$S(t)$	Expected number of species in a hypothetical rarefied sample of t sampling units if $t < T$.
$S(T+t^*)$	Expected number of species in a hypothetical augmented sample of $T + t^*$ sampling units.
One-assemblage species richness under sampling without replacement (Section 4)	
T^*	Total number of sampling units in the entire assemblage (e.g. total number of disjoint, equal-area quadrats in a region).
U_i	Number of sampling units (or quadrats) that species i can be detected.
q	Known sampling fraction, $q = T/T^*$.
Two-assemblage shared species richness estimation (Section 5)	
S_{12}	Number of shared species between Assemblages I and II.
π_{i1}, π_{i2}	Detection or incidence probability of species i, $i = 1, 2, \ldots, S$, in any sampling unit taken, respectively, from Assemblages I and II.
T_1, T_2	Number of sampling units in Samples I and II taken, respectively, from Assemblages I and II.
Y_{i1}, Y_{i2}	Species incidence frequency (i.e. number of sampling units in which species i is detected), respectively, in Samples I and II.
Q_{rv}	Number of shared species that are detected in r sampling units in Sample I and v sampling units in Sample II, r, $v = 0, 1, 2, \ldots$.
Q_{r+}	Number of shared species that are detected in r sampling units in Sample I and that are detected in at least one sampling unit in Sample II, $r = 0, 1, 2, \ldots, T_1$.
Q_{+v}	Number of shared species that are detected in v sampling units in Sample II and that are detected in at least one sampling unit in Sample I, $v = 0, 1, 2, \ldots, T_2$.
Q_{++}	Total number of observed species shared between Samples I and II.

Table 1 (cont.): Major notation used in each section.

One-assemblage phylogenetic diversity (*PD*) Estimation (Section 6)	
B	Number of branches/nodes in the phylogenetic tree spanned by all species of an assemblage.
L_i	Length of the ith branch/node.
PD	Sum of branch lengths in a phylogenetic tree.
λ_i	Detection or incidence probability of branch/node i, i.e. the probability of detecting at least one species descended from branch/node i in a sampling unit.
W_{ij}^*	Node detection/non-detection: $W_{ij}^* = 1$ if at least one species descended from branch i is detected in jth sampling unit, and $W_{ij}^* = 0$ otherwise, $i = 1,2,\ldots,B$, $j = 1,2,\ldots,T$.
PD_{obs}	PD in the observed tree.
Y_i^*	branch/node incidence frequency for branch/node i, $i = 1,2,\ldots,B$.
R_k	Sum of branch lengths for the branches with node incidence frequency $= k$, $k = 0,1,\ldots,T$.
Q_1^*, Q_2^*	Number of nodes/branches with incidence frequency $= 1$ and $= 2$, respectively, in the observed tree.
Rarefaction and extrapolation of one-assemblage *PD* (Section 6 and Table 2)	
$PD(t)$	Expected PD in a hypothetical rarefied sample of t sampling units if $t < T$.
$PD(T+t^*)$	Expected PD in a hypothetical augmented sample of $T+t^*$ sampling units.

Following Colwell et al. (2012), we assume, given the set of detection probabilities $(\pi_1, \pi_2, \ldots, \pi_S)$, that each element W_{ij} in the incidence matrix is a Bernoulli random variable with probability π_i. The probability distribution for the incidence matrix can be expressed as

$$P(W_{ij} = w_{ij};\ i = 1,2,\ldots,S,\ j = 1,2,\ldots,T) = \prod_{j=1}^{T}\prod_{i=1}^{S} \pi_i^{w_{ij}}(1-\pi_i)^{1-w_{ij}} = \prod_{i=1}^{S} \pi_i^{y_i}(1-\pi_i)^{T-y_i}. \tag{1a}$$

The marginal distribution for the incidence-based frequency Y_i for the i-th species follows a binomial distribution characterized by T and the detection probability π_i:

$$P(Y_i = y_i) = \binom{T}{y_i}\pi_i^{y_i}(1-\pi_i)^{T-y_i}, \quad i = 1,2,\ldots,S. \tag{1b}$$

Denote the *incidence frequency counts* by $(Q_1, Q_2, \ldots, Q_T)$, where Q_k is the number of species detected in exactly k sampling units in the data, $k = 0,1,\ldots,T$. Here, Q_1 represents the number of “unique” species (those that are detected in only one sampling unit), and Q_2 represents the number of “duplicate” species (those that are detected in exactly two sampling units). The unobservable zero frequency count Q_0 denotes the number of species among the S species present in the assemblage that are not detected in any of the T sampling units. Then the number of observed species in the sample is $S_{obs} = \sum_{i>0} Q_i$ and $S_{obs} + Q_0 = S$.

2.2. Chao's inequality

Treating the incidence probabilities $(\pi_1, \pi_2, \ldots, \pi_S)$ as fixed, unknown parameters, we first present Chao's (1987) inequality under the model (1a) or (1b). Note the following expected value for the incidence frequency count Q_k:

$$E(Q_k) = E\left[\sum_{i=1}^{S} I(Y_i = k)\right] = \sum_{i=1}^{S} \binom{T}{k} \pi_i^k (1-\pi_i)^{T-k}, \quad k = 0, 1, 2, \ldots, T, \tag{1c}$$

where $I(A)$ is the indicator function, i.e. $I(A) = 1$ if the event A occurs, and is 0 otherwise. In particular, the expected number of undetected species, uniques and duplicates are respectively:

$$E(Q_0) = \sum_{i=1}^{S} (1-\pi_i)^T,$$

$$E(Q_1) = \sum_{i=1}^{S} T\pi_i (1-\pi_i)^{T-1},$$

$$E(Q_2) = \sum_{i=1}^{S} \binom{T}{2} \pi_i^2 (1-\pi_i)^{T-2}.$$

Chao (1987) proposed a lower bound of $E(Q_0)$ based on the following Cauchy-Schwarz inequality:

$$\left[\sum_{i=1}^{S} (1-\pi_i)^T\right] \left[\sum_{i=1}^{S} \pi_i^2 (1-\pi_i)^{T-2}\right] \geq \left[\sum_{i=1}^{S} \pi_i (1-\pi_i)^{T-1}\right]^2, \tag{2a}$$

equivalently,

$$E(Q_0) \times \frac{E(Q_2)}{\binom{T}{2}} \geq \left(\frac{E(Q_1)}{T}\right)^2.$$

Thus, a theoretical lower bound for $E(Q_0)$ is derived as

$$E(Q_0) \geq \frac{(T-1)}{T} \frac{[E(Q_1)]^2}{2E(Q_2)},$$

implying a theoretical lower bound for species richness:

$$S = E(S_{obs}) + E(Q_0) \geq E(S_{obs}) + \frac{(T-1)}{T} \frac{[E(Q_1)]^2}{2E(Q_2)}.$$

Replacing the expected values in the above with the observed data, we then obtain an estimated lower bound of species richness, with a slight modification when $Q_2 = 0$ (Colwell and Coddington, 1994, gave the name *Chao*2 to this estimator):

$$\hat{S}_{Chao2} = \begin{cases} S_{obs} + \dfrac{(T-1)}{T}\dfrac{Q_1^2}{2Q_2}, & \text{if } Q_2 > 0, \\ S_{obs} + \dfrac{(T-1)}{T}\dfrac{Q_1(Q_1-1)}{2}, & \text{if } Q_2 = 0. \end{cases} \tag{2b}$$

The estimated number of undetected species is based exclusively on the information on the least frequent species (the number of uniques and duplicates). This is based on a basic concept that the frequent/abundant species (those that occur in many sampling units) carry negligible information about the undetected species; only rare/infrequent species carry such information.

When does the Chao2 formula provide a nearly unbiased estimator? The Cauchy-Schwarz inequality in Eq. (2a) becomes an equality if and only if the species detection probabilities are homogeneous, that is, $\pi_1 = \pi_2 = \cdots = \pi_S$. Homogeneity of detection probabilities would be a very restrictive condition, one that is almost never satisfied in most practical applications, such as species abundance or incidence distributions in nature. However, as we will show in Section 2.3, this condition can be considerably relaxed from a different derivation/perspective. Note that in Chao's inequality (2a), only three expected frequency counts are involved: $E(Q_0)$, $E(Q_1)$ and $E(Q_2)$. The frequent species (species with relatively large detection probabilities) would tend to occur in many sampling units and thus generally do not contribute to any of these three terms. On the other hand, only rare/infrequent species (species with relatively low detection probabilities) would either be undetected or detected in only one or two sampling units and thus are those species that contribute to the three terms. Therefore, a relaxed condition for an unbiased Chao2 estimator is that *very rare/infrequent* species have approximately the same detection probabilities, and frequent species are allowed to be highly heterogeneous without affecting the estimates. A more rigorous justification is given in Section 2.3.

Applying a standard asymptotic approach (Chao, 1987), the following estimated variance estimators can be obtained if Q_1, $Q_2 > 0$:

$$\widehat{\text{var}}(\hat{S}_{Chao2}) = Q_2\left[\frac{1}{4}\left(\frac{T-1}{T}\right)^2\left(\frac{Q_1}{Q_2}\right)^4 + \left(\frac{T-1}{T}\right)^2\left(\frac{Q_1}{Q_2}\right)^3 + \frac{1}{2}\left(\frac{T-1}{T}\right)\left(\frac{Q_1}{Q_2}\right)^2\right], \tag{3a}$$

If $Q_1 > 0$, $Q_2 = 0$, the variance becomes

$$\widehat{\text{var}}(\hat{S}_{Chao2}) = \frac{1}{4}\left(\frac{T-1}{T}\right)^2 Q_1(2Q_1-1)^2 + \frac{1}{2}\left(\frac{T-1}{T}\right)Q_1(Q_1-1) - \frac{1}{4}\left(\frac{T-1}{T}\right)^2\frac{Q_1^4}{\hat{S}_{Chao2}}. \tag{3b}$$

In the special case that $Q_1 = 0$, we have $\hat{S}_{Chao2} = S_{obs}$, implying that sampling is complete and there are no undetected species in the data; an approximate variance of S_{obs} can be obtained using an analytic method (Colwell, 2013) or a bootstrap method (see Section 3.3). When $Q_1 > 0$ so that $\hat{S}_{Chao2} > S_{obs}$, the distribution of $\hat{S}_{Chao2} - S_{obs}$ is generally skewed to the right. Using a log-transformation by treating $\log(\hat{S}_{Chao2} - S_{obs})$ as an approximately normal random variable, we obtain a 95% confidence interval for S: (Chao, 1987)

$$[S_{obs} + (\hat{S}_{Chao2} - S_{obs})/R,\ \ S_{obs} + (\hat{S}_{Chao2} - S_{obs})R], \tag{3c}$$

where $R = \exp\{1.96[\log(1 + \widehat{\text{var}}(\hat{S}_{Chao2})/(\hat{S}_{Chao2} - S_{obs})^2)]^{1/2}\}$. In this case, the resulting lower confidence limit is always greater than or equal to the observed species richness, a sensible result.

The Chao2 estimator is also valid in a binomial-mixture model in which incidence probabilities $(\pi_1, \pi_2, \ldots, \pi_S)$ are assumed to be a random sample from an unknown distribution with density $h(\pi)$. Under this model, we have

$$E(Q_k) = S \int_0^1 \binom{T}{k} \pi^k (1-\pi)^{T-k} h(\pi)\, d\pi, \quad k = 0, 1, 2, \ldots T. \tag{4a}$$

The summation terms in the Cauchy-Schwarz inequality (2a) are replaced by integral terms:

$$\left[\int_0^1 (1-\pi)^T h(\pi) d\pi\right] \left[\int_0^1 \pi^2 (1-\pi)^{T-2} h(\pi) d\pi\right] \geq \left[\int_0^1 \pi (1-\pi)^{T-1} h(\pi) d\pi\right]^2. \tag{4b}$$

The above two formulas also lead to the same Chao2 formula given in Eq. (2b). In the special case that $h(\pi)$ is a beta distribution with parameters α and β, the resulting expected incidence-frequency count $\{E(Q_k),\, k = 0, 1, 2, \ldots, n\}$ correspond to the probabilities of a beta-binomial distribution. Under the two conditions (i) T is large and π is small, such that $T\pi$ tends to a positive constant, and (ii) β/T tends to a positive constant c, Skellam (1948) proved that $E(Q_k)$ tends to $(\alpha + k - 1)![(\alpha - 1)!k!]^{-1}[1/(1 + c)]^k[c/(1+c)]^\alpha$, which is the probability of a negative binomial variable taking the value k. This result theoretically justifies the inference that Chao's inequality is also valid for beta-binomial and negative binomial distributions. It is well known that beta-binomial and negative binomial can be used to describe spatially clustered (if sampling units are quadrats in an area) or temporally aggregated (if sampling units are different times) pattern of species; see Hughes and Madden (1993) and Shiyomi, Takahashi and Yoshimura (2000). Therefore, even though there is spatial/temporal heterogeneity pattern for species incidences, the lower bound and the associated estimation are still valid.

2.3. When is the Chao2 estimator nearly unbiased?

Alan Turing and I. J. Good, in their famous cryptanalysis to crack German ciphers during World War II, developed novel statistical methods to estimate the true frequencies of rare code elements (including still-undetected code elements), based on the observed frequencies in "samples" of intercepted Nazi code. After the War, Turing gave permission to Good to publish their statistical work. An influential paper by Good (1953) and one by Good and Toulmin (1956) presented Turing's wartime statistical work on the frequency formula and related topics; see Good (1983, 2000) for more details. The frequency formula is now referred to as the Good-Turing frequency formula, which has a wide range of applications in biological sciences, statistics, computer sciences, information sciences, and linguistics, among others (McGrayne, 2011, p. 100).

In an ecological context, Turing's statistical problem can be formulated as an estimation of the true frequencies of rare species when a random sample of individuals is drawn from an assemblage. In Turing's case, there were almost infinitely many rare species so that all samples have undetected species. The Good-Turing formula answers the following question: given a species that appears r times ($r = 0, 1, 2, \dots$) in a sample of n individuals that fails to detect all species present, what is its true relative frequency in the entire assemblage? Turing and Good focussed on the case of small r, i.e. rare species. Turing gave a surprisingly simple and remarkably effective answer that is contrary to most people's intuition; see Chao et al. (2017) for a review.

The Good-Turing original frequency formula was based on abundance data. We here extend their formula to incidence data to answer the following question: Given species incidence data of T sampling units, for those species that appeared in r ($r = 0, 1, 2, \dots$) out of T sampling units, what is the mean detection probability of species that appeared in r sampling units, ϕ_r? Such a mean detection probability can be mathematically expressed as

$$\phi_r = \sum_{i=1}^{S} \pi_i I(Y_i = r)/Q_r, \quad r = 0, 1, 2, \dots \tag{5a}$$

The numerator in Eq. (5a) represents the total incidence probabilities of those species that appeared in r sampling units. Dividing the total by Q_r, we obtain the mean detection probability per species, among those that each appeared in r sampling units. Note that, for the special case of $r = 0$, Eq. (5a) implies

$$\phi_0 Q_0 = \sum_{i=1}^{S} \pi_i I(Y_i = 0), \tag{5b}$$

which is the total detection probabilities of the undetected species. If one additional sampling unit can be added, then we can interpret it as the expected number of species in the additional sampling unit that are undetected in the original sample.

Here we derive the corresponding Good-Turing incidence frequency formula for multiple incidence data by treating $(\pi_1, \pi_2, \dots, \pi_S)$ as fixed, unknown parameters, although a similar derivation is also valid for binomial-mixture models. Under the model (Eq. 1b), in which the incidence frequencies Y_i, $i = 1,2,\ldots,S$, follow a binomial distribution characterized by T and detection probability π_i, we can express the sum of the odds of π_i for those species that each appeared in r sampling units as follows:

$$\begin{aligned}
E\left[\sum_{i=1}^{S}\frac{\pi_i}{1-\pi_i}I(Y_i=r)\right] &= \sum_{i=1}^{S}\frac{\pi_i}{1-\pi_i}\binom{T}{r}\pi_i^r(1-\pi_i)^{T-r} \\
&= \sum_{i=1}^{S}\binom{T}{r}\pi_i^{r+1}(1-\pi_i)^{T-(r+1)} \\
&= \frac{\binom{T}{r}}{\binom{T}{r+1}}\left[\sum_{i=1}^{S}\binom{T}{r+1}\pi_i^{r+1}(1-\pi_i)^{T-(r+1)}\right] \\
&= \frac{(r+1)}{(T-r)}E(Q_{r+1}). \qquad (5c)
\end{aligned}$$

Assume that all species that appeared in r sampling units have approximately the same incidence probabilities. Then we have the following approximation formula:

$$E\left[\sum_{i=1}^{S}\frac{\pi_i}{1-\pi_i}I(Y_i=r)\right] \approx Q_r\frac{\phi_r}{1-\phi_r}.$$

Thus, ϕ_r can be obtained by solving the equation: $Q_r\phi_r/(1-\phi_r) \approx (r+1)Q_{r+1}/(T-r)$, based on Eq. (5c). We then obtain the corresponding Good-Turing formula for incidence data:

$$\hat{\phi}_r = \frac{(r+1)Q_{r+1}}{(T-r)Q_r+(r+1)Q_{r+1}} \approx \frac{(r+1)Q_{r+1}}{(T-r)Q_r}. \qquad (5d)$$

The original Good-Turing frequency formula for abundance data has a similar form as the above approximation, but with incidence frequency counts being replaced by abundance frequency counts.

Good (1983, p. 28) provided an intuitive justification for the abundance-based Good-Turing frequency formula. Here we follow Good's approach to give a similar justification for incidence data. Given an original sample, consisting of T sampling units, suppose one additional sampling unit can be added. We ask how many species that had appeared r times in the original sample would occur in the additional sampling unit. Based on Eq. (5a), the answer is simply $\sum_{i=1}^{S}\pi_i I(Y_i=r) = \phi_r Q_r$, which can be estimated by $(r+1)Q_{r+1}/(T-r)$ using the following simple reasoning. Notice that any species that appeared r times in the original sample and also occurs in the additional sampling unit must occur in $r+1$ sampling units in the enlarged sample consisting of $T+1$ sampling units. Then the total number of incidences of such species is $(r+1)Q_{r+1}$. Because the order in which sampling units were taken is assumed to be irrelevant, the average number of such species occurring in a single sampling unit is thus $(r+1)Q_{r+1}/(T+1)$, which is approximately equal to $(r+1)Q_{r+1}/(T-r)$ if r is small. Dividing this ratio by the number of such species, Q_r, we obtain the incidence-data-based Good-Turing frequency formula for ϕ_r as given in Eq. (5d).

For the special cases of $r=0$ and $r=1$, Eqs. (5b) and (5d) lead to

$$\widehat{\phi_0 Q_0} = \frac{Q_1}{T}, \hat{\phi}_1 = \frac{2Q_2}{(T-1)Q_1},$$

where $\widehat{\phi_0 Q_0}$ denotes the estimate of the product of ϕ_0 and Q_0. Intuitively, we expect that the mean incidence probability of all undetected species should not be more than that of all uniques in the sample, i.e. $\phi_0 \le \phi_1$, and this ordering is preserved by the corresponding estimates. Then we obtain the Chao2 lower bound for the number of undetected species by the following inequality:

$$\hat{Q}_0 = \frac{\widehat{\phi_0 Q_0}}{\hat{\phi}_0} \ge \frac{\widehat{\phi_0 Q_0}}{\hat{\phi}_1} = \frac{\frac{Q_1}{T}}{\frac{2Q_2}{(T-1)Q_1}} = \frac{(T-1)}{T}\frac{Q_1^2}{2Q_2}. \tag{5e}$$

Notice that, in the above derivation, if $\hat{\phi}_0 \approx \hat{\phi}_1$, then the inequality sign in Eq. (5e) becomes an equality sign. Therefore, from the Good-Turing perspective, the Chao2 lower bound is a nearly unbiased point estimator if all undetected and unique species in samples have the same mean detection probabilities. Such a conclusion is valid if very rare/infrequent species have approximately homogenous detection probabilities in any sampling unit (because this implies $\hat{\phi}_0 \approx \hat{\phi}_1$); in this case, frequent species could be highly heterogeneous without affecting the estimator.

2.4. How many sampling units are needed to reach the Chao2 estimate?

As discussed earlier, the Chao2 formula (in Eq. 2b) implies that sampling is complete when all species have been found in at least two sampling units, i.e. $Q_1 = 0$; in such a case, the estimated undetected species richness is 0 and the estimated species richness reduces simply to the observed number of species. This result also reveals that, whenever at least one species is found in only one sample ($Q_1 > 0$), sampling is not complete and some species remain undetected. However, the Chao2 species richness estimator does not indicate how much sampling effort (how many additional sampling units) would be necessary to reach the Chao2 estimate (i.e. the first point at which there are no longer any singletons).

For incidence data, “sample size” means the number of sampling units. Chao et al. (2009) developed a non-parametric method for estimating the minimum sample size required to detect any arbitrary proportion (including 100%) of the estimated Chao2 species richness based on the Good-Turing formula discussed in Section 2.3. When the target is the Chao2 estimate, Chao et al. (2009) approach is to predict the minimum sample size t to achieve the following stopping rule: there are no uniques in the enlarged sample of size $T+t$, or equivalently, the expected number of uniques in the enlarged sample of size $T+t$ is less than 0.5, because the theoretical expected value may not be an integer.

Note that the number of uniques in the enlarged sample of size $T+t$ includes two groups of species: (1) any species observed in only one sampling unit in the original sample (i.e. those species with $Y_i = 1$) for which no additional incidences are detected in the additional t samples with probability $(1-\pi_i)^t$, and (2) any species not detected in the original sample (i.e. those species with $Y_i = 0$) for which detection in exactly one sampling unit is observed in the additional t sampling units with probability $t\pi_i(1-\pi_i)^{t-1}$. That is, the expected number of uniques in the enlarged $T+t$ sampling units is:

$$\sum\nolimits_{i=1}^{S}(1-\pi_i)^t I(Y_i=1)+\sum\nolimits_{i=1}^{S} t\pi_i(1-\pi_i)^{t-1} I(Y_i=0).$$

As discussed in Section 2.3, we assume that all uniques in the original sample have mean detection probability ϕ_1, and all previously undetected species have mean detection probability ϕ_0. Then the number of uniques in the enlarged $T+t$ sampling units will decline to < 0.5 when t satisfies

$$Q_1(1-\phi_1)^t + Q_0\, t\phi_0(1-\phi_0)^{t-1} < 0.5.$$

When we apply the Good-Turing incidence frequency formula to this equation, and substitute ϕ_1, ϕ_0 and Q_0 by $\hat{\phi}_1 = 2Q_2/[2Q_2+(T-1)Q_1]$, $\hat{\phi}_0 = Q_1/[Q_1+T\hat{Q}_0]$ and $\hat{Q}_0 = (1-1/T)Q_1^2/(2Q_2)$, then the required t must satisfy the following equation:

$$Q_1\left(1+\frac{t}{T}\right)\left[1-\frac{2Q_2}{(T-1)Q_1+2Q_2}\right]^t < 0.5.$$

The additional number of sampling units needed to reach the Chao2 estimate is approximately equal to $t = Tx^*$, where x^* is the solution of the following equation:

$$2Q_1(1+x) = \exp\left[x\frac{2Q_2}{(1-1/T)Q_1+2Q_2/T}\right]. \tag{6a}$$

If g is the fraction of $\hat{S}_{Chao2}$ that is desired $(0 < g < 1)$, then the objective is to find the number of additional m_g sampling units such that the number of species reaches the target value $g\hat{S}_{Chao2}$, i.e. the expected number of previously undetected species that will be discovered in the additional m_g sampling units is $g\hat{S}_{Chao2} - S_{obs}$. This expected number, given the observed data, is

$$\sum_{i=1}^{S}[1-(1-\pi_i)^{m_g}]I(Y_i=0)\approx Q_0[1-(1-\phi_0)^{m_g}]. \tag{6b}$$

Applying the Good-Turing incidence frequency formula and substituting ϕ_0 and Q_0, we obtain that the required number of additional sampling units to reach a fraction g of $\hat{S}_{Chao2}$ (if $g\hat{S}_{Chao2} > S_{obs}$) is the number m_g such that $\hat{Q}_0[1-(1-\hat{\phi}_0)^{m_g}] = g\hat{S}_{Chao2} - S_{obs}$, i.e.

$$m_g \approx \frac{\log\left[1-\dfrac{T}{(T-1)}\dfrac{2Q_2}{Q_1^2}(g\hat{S}_{Chao2}-S_{obs})\right]}{\log\left[1-\dfrac{2Q_2}{(T-1)Q_1+2Q_2}\right]}. \tag{6c}$$

Chao et al. (2009) also provided an Excel spreadsheet for calculating necessary sampling effort for either abundance data or replicated incidence data.

2.5. A class of lower bounds

In the Chao2 approach (Eq. 2b), the estimator for undetected species richness is only in terms of the species incidence frequency counts of the uniques and duplicates in data. Several authors extended this approach to higher-order incidence frequency counts. Lanumteang and Böhning (2011) proposed using an additional incidence frequency count, i.e. the number of species that are detected in exactly three sampling units. They applied the above estimator to a variety of real data sets and concluded that the new estimator is especially useful for large populations and heterogeneous detection probabilities.

When the Chao2 estimator only provides a lower bound, its bias can be evaluated and assessed by using the Good-Turing frequency formula. In this case, an improved reduced-bias lower bound, which makes use of the additional information of Q_3 and Q_4, was derived by Chiu et al. (2014). The corresponding lower bound of species richness is referred to as *iChao2 estimator* (here the sub-index*i* stands for “improved”):

$$\hat{S}_{iChao2} = \hat{S}_{Chao2} + \frac{(T-3)}{4T}\frac{Q_3}{Q_4}\times \max\left(Q_1 - \frac{(T-3)}{2(T-1)}\frac{Q_2Q_3}{Q_4}, 0\right). \tag{6d}$$

They also provided an analytic variance estimator to construct the associated confidence intervals.

Puig and Kokonendji (2017) extended Chao’s inequality to a broader class of distributions that have log-convex probability generating functions. They obtained a series of lower bounds for the undetected species richness. This class of distribution includes compound Poisson distribution and Poisson-mixture distributions. Their framework is mainly based on abundance data, but it can be readily applied to multiple incidence data, as shown below.

Following the proof of Puig and Kokonendji (2017), we assume that the incidence probabilities $(\pi_1, \pi_2, \ldots, \pi_S)$ are a random sample from an unknown distribution with density $h(\pi)$, and we have $E(Q_k)$ given in Eq. (4a). Consider a probability density function:

$$H(\pi) = \frac{(1-\pi)^T h(\pi) d\pi}{\int_0^1 (1-u)^T h(u) du}, \quad 0 < \pi < 1.$$

Puig and Kokonendji (2017) showed the following moment inequality for $r,\, v = 0, 1, 2, \ldots$

$$\int_0^1 \left(\frac{\pi}{1-\pi}\right)^{r+v} H(\pi) d\pi \geq \int_0^1 \left(\frac{\pi}{1-\pi}\right)^{r} H(\pi) d\pi \times \int_0^1 \left(\frac{\pi}{1-\pi}\right)^{v} H(\pi) d\pi,$$

equivalently,

$$\left[\int_0^1 (1-\pi)^T h(\pi) d\pi\right] \left[\int_0^1 \pi^{r+v} (1-\pi)^{T-(r+v)} h(\pi) d\pi\right]$$
$$\geq \left[\int_0^1 \pi^{r} (1-\pi)^{T-r} h(\pi) d\pi\right] \left[\int_0^1 \pi^{v} (1-\pi)^{T-v} h(\pi) d\pi\right].$$

Then we have

$$E(Q_0) \geq \frac{\binom{T}{r+v} E(Q_r) \times E(Q_v)}{\binom{T}{r}\binom{T}{v} E(Q_{r+v})}, \quad r,\, v = 0, 1, 2, \ldots \tag{6e}$$

A series of lower bounds of S can then be obtained if $Q_{r+v} > 0$:

$$S_{obs} + \frac{\binom{T}{r+v} Q_r \times Q_v}{\binom{T}{r}\binom{T}{v} Q_{r+v}}, \quad r,\, v = 1, 2, \ldots$$

In the special case of $r = v = 1$, the above lower bound reduces to the Chao2 estimator. Puig and Kokonendji (2017) proved that, under a Poisson-mixture model, the greatest lower bound attains at the special case $r = v = 1$. This also provides a justification for the use of the Chao2 lower bound.

3. Species richness estimation for standardized samples: non-asymptotic analysis

Species richness estimation represents an "asymptotic" analysis; here "asymptotic" means that, as sample size tends to infinity, sample completeness approaches unity. When the Chao2 estimates are nearly unbiased under the conditions given in Section 2.3, they can be compared across multiple assemblages. However, when rare/infrequent species are highly heterogeneous and sample size is not sufficiently large, the Chao2 formula can provide only a lower bound, which cannot be compared accurately across assemblages, because the data provide insufficient information to accurately estimate species richness due to high heterogeneity of infrequent species. No matter whether or not Chao2 is unbiased, in any particular case, we can always use it to perform "non-asymptotic" analysis, in which samples are standardized based on a common finite sample size or on sample completeness via rarefaction and extrapolation. Again for incidence data, sample size refers to the number of sampling units.

The objective of a non-asymptotic approach is to control the dependence of the empirical species counts on sampling effort and sample completeness. The earliest development of standardization of sample size for abundance data by rarefaction was proposed by Sanders (1968), but see Chiarucci et al. (2008) for a historical review. Subsequent developments include studies by Hurlbert (1971), Simberloff (1972), Heck, van Belle and Simberloff (1975) and Coleman et al. (1982); see Gotelli and Colwell (2001, 2011) for details. Ecologists typically use rarefaction to down-sample the larger samples until they are the same size as the smallest sample. Ecologists then compare richness of these equally-large samples, but this approach implies that some data in larger samples are thrown away. To avoid discarding data, Colwell et al. (2012) proposed using a unified sample-size-based rarefaction (interpolation) and extrapolation (prediction) sampling curve for species richness, that can be rarefied to smaller sample sizes or extrapolated to larger sample sizes.

Chao and Jost (2012) indicated that a sample of a given size may be sufficient to fully characterize a low-diversity assemblage, but insufficient to characterize a rich-assemblage. Thus, when the species counts of two equally-large samples are compared, one might be comparing a nearly complete sample to a very incomplete one. In this case, any difference in diversity between the sites will generally be underestimated. They proposed rarefaction and extrapolation to a comparable degree of sample completeness (as measured by sample coverage; see below) and developed a coverage-based rarefaction and extrapolation methodology. The sample-size-based and coverage-based integration of rarefaction and extrapolation of species richness represent a unified sampling framework for quantifying and comparing species richness across multiple assemblages.

Here we review the sample-size-based and coverage-based rarefaction and extrapolation of species richness; all formulas are tabulated in the first and the third columns of Table 2.

Table 2: *The theoretical formulas and analytic estimators for rarefaction and extrapolation of species richness (left column), Faith's PD (middle column), and sample coverage (right column) based on incidence data, given a reference sample with observed species richness* $= S_{obs}$*, observed PD* $= PD_{obs}$*, and estimated coverage* $\hat{C}(T)$ *for incidence data. Here the sample size means the number of sampling units. See Colwell et al. (2012) and Chao and Jost (2012) for derivation details.*

Species richness	Faith's PD	Coverage
(a) Theoretical formula for any hypothetical sample size of t		
$S(t) = \sum_{i=1}^{S}[1-(1-\pi_i)^t]$	$PD(t) = \sum_{i=1}^{B} L_i[1-(1-\lambda_i)^t]$	$C(t) = 1 - \frac{\sum_{i=1}^{S} \pi_i (1-\pi_i)^t}{\sum_{i=1}^{S} \pi_i}$
(b) Rarefaction estimator for $t < T$		
$\hat{S}(t) = S_{obs} - \sum_{1 \le Y_i \le T-t} \frac{\binom{T-Y_i}{t}}{\binom{T}{t}}$	$\widehat{PD}(t) = PD_{obs} - \sum_{1 \le Y_i \le T-t} L_i \frac{\binom{T-Y_i}{t}}{\binom{T}{t}}$	$\hat{C}(t) = 1 - \sum_{1 \le Y_i \le T-t} \frac{Y_i}{U} \frac{\binom{T-Y_i}{t}}{\binom{T-1}{t}}$
(c) Reference sample of size T		
$\hat{S}(T) = S_{obs}$	$\widehat{PD}(T) = PD_{obs}$	$\hat{C}(T) = 1 - \frac{Q_1}{U}\left[\frac{(T-1)Q_1}{(T-1)Q_1+2Q_2}\right]$
(d) Extrapolation estimator for sample size $T+t^*$		
$\hat{S}(T+t^*) = S_{obs} + \hat{Q}_0\left[1-\left(1-\frac{Q_1}{T\hat{Q}_0+Q_1}\right)^{t^*}\right]$	$\widehat{PD}(T+t^*) = PD_{obs} + \hat{R}_0\left[1-\left(1-\frac{R_1}{T\hat{R}_0+R_1}\right)^{t^*}\right]$	$\hat{C}(T+t^*) = 1 - \frac{Q_1}{U}\left[\frac{(T-1)Q_1}{(T-1)Q_1+2Q_2}\right]^{t^*+1}$

Notes: $U = \sum_{Y_i>0} Y_i = \sum_{j=1}^{T} jQ_j$ denotes the total number of incidences in T sampling units; $\hat{Q}_0$ and $\hat{R}_0$ denote the estimated number of undetected species richness in Eq. (2b) and undetected PD in Eq. (11c).

3.1. Sample-size-based rarefaction and extrapolation

Following Colwell et al. (2012), we refer to the observed sample of T sampling units as a *reference sample*. Let $S(t)$ be the expected number of species in a hypothetical sample of t sampling units, randomly selected from the sampling units that represent the assemblage. If we knew the true species detection probabilities $(\pi_1, \pi_2, \ldots, \pi_S)$ of the S species in each sampling unit, we could compute the following expected value:

$$S(t) = S - \sum_{i=1}^{S} (1-\pi_i)^t, \quad t = 1, 2, \ldots \tag{7a}$$

The plot of $S(t)$ with respect to the number of sampling units t is the sampling-unit-based species accumulation curve. Note that the true species richness represents the "asymptote" of the curve, i.e. $S = S(\infty)$. The rarefaction (interpolation) part estimates the expected species richness for a smaller number of sampling units $t < T$. On the basis of a reference sample of T sampling units, an unbiased estimator $\hat{S}(t)$ for $S(t)$, $t < T$, is

$$\hat{S}(t) = S_{obs} - \sum_{1 \le Y_i \le T-t} \binom{T-Y_i}{t} \Big/ \binom{T}{t}, \quad t < T. \tag{7b}$$

This analytic formula was first derived by Shinozaki (1963) and rediscovered multiple times (Chiarucci et al., 2008).

The extrapolation is to estimate the expected number of species $S(T+t^*)$ in a hypothetical sample of $T+t^*$ sampling units ($t^* > 0$) from the assemblage. Rewrite

$$\begin{aligned} S(T+t^*) &= \sum_{i=1}^{S} [1-(1-\pi_i)^{T+t^*}] \\ &= \sum_{i=1}^{S} [1-(1-\pi_i)^{T}] + \sum_{i=1}^{S} [1-(1-\pi_i)^{t^*}](1-\pi_i)^T \\ &= E\ (S_{obs}) + E\left[\sum_{i=1}^{S} [1-(1-\pi_i)^{t^*}] I(Y_i = 0)\right]. \end{aligned}$$

The first term in the above formula represents the observed species richness. For the second term, we can apply the Good-Turing incidence frequency formula (Section 2.3) by assuming that all previously undetected species have mean detection probability ϕ_0. Then for the second term, we have

$$\sum_{i=1}^{S} [1-(1-\pi_i)^{t^*}] I(Y_i = 0) \approx Q_0[1-(1-\phi_0)^{t^*}].$$

Based on Eq. (5d), we have the extrapolated species richness for a sample of size $T+t^*$:

$$\hat{S}(T+t^*) = S_{obs} + \hat{Q}_0 \left[1 - \left(1 - \frac{Q_1}{T\hat{Q}_0 + Q_1}\right)^{t^*}\right], \quad t^* \geq 0. \tag{7c}$$

Colwell et al. (2012) linked rarefaction and extrapolation to form an integrated smooth curve. The integrated sample-size-based sampling curve includes a rarefaction part (which plots $\hat{S}(t)$ as a function of $t < T$), and an extrapolation part (which plots $\hat{S}(T+t^*)$ as a function of $T+t^*$), joining smoothly at the reference point (T, S_{obs}). The confidence intervals based on the bootstrap method (Section 3.3) also join smoothly.

For a short-range prediction (e.g. t^* is much less than T), the extrapolation formula is independent of the choice of $\hat{Q}_0$ as indicated by the approximation formula $\hat{S}(T+t^*) \approx S_{obs} + (Q_1/T)t^*$. This implies that the extrapolation formula in Eq. (7c) is very robust and reliable even though the species richness estimator is subject to bias. Previous experiences by Colwell et al. (2012) suggested that the prediction size can be extrapolated at most to double the observed sample size.

3.2. Coverage-based rarefaction and extrapolation

Turing and Good developed the very important concept of "sample coverage" to characterize the sample completeness of an observed set of individual-based abundance data. Their concept was extended by Chao et al. (1992) to capture-recapture data. For multiple incidence data, the *sample coverage* of a reference sample of T sampling units is defined as

$$C \equiv C(T) = \frac{\sum_{i=1}^{S} \pi_i I(Y_i > 0)}{\sum_{i=1}^{S} \pi_i} = 1 - \frac{\sum_{i=1}^{S} \pi_i I(Y_i = 0)}{\sum_{i=1}^{S} \pi_i},$$

which represents the fraction of the total incidence probabilities in the assemblage (including undetected species) that is represented by species detected in the reference sample. Note that under the binomial model (Eq. 1b), an unbiased estimator for the denominator in $C(T)$ is U/T, where $U = \sum_{k=1}^{T} kQ_k = \sum_{i=1}^{S} Y_i$ denotes the total number of incidences in the reference sample. For the numerator, we can apply the Good-Turing incidence frequency formula (Section 2.3) by assuming that all uniques in the sample have approximately the same detection probabilities, ϕ_1. Then we can write

$$\begin{aligned} E\left[\sum_{i=1}^{S} \pi_i I(Y_i = 0)\right] &= \sum_{i=1}^{S} \pi_i (1-\pi_i)^T \\ &= \frac{1}{T} E\left[\sum_{i=1}^{S} (1-\pi_i) I(Y_i = 1)\right] \approx \frac{E(Q_1)}{T}(1-\phi_1). \end{aligned}$$

Applying the Good-Turing formula $\hat{\phi}_1 = 2Q_2/[2Q_2 + (T-1)Q_1]$ (Eq. 5d), we obtain a very accurate estimator of the sample coverage for the reference sample size, if $Q_2 > 0$:

$$\hat{C}(T) = 1 - \frac{Q_1}{U}\left[\frac{(T-1)Q_1}{(T-1)Q_1 + 2Q_2}\right]. \tag{7d}$$

If $Q_2 = 0$, a modified formula based on Chao et al. (2014b, Appendix G) is:

$$\hat{C}(T) = 1 - \frac{Q_1}{U}\left[\frac{(T-1)(Q_1-1)}{(T-1)(Q_1-1)+2}\right]. \tag{7e}$$

In addition to the reference sample, we also need to consider the estimation of the expected sample coverage, $E[C(t)]$, for any hypothetical sample of t sampling units, $t = 1, 2, \ldots$. This expected sample coverage is a function of t as given below:

$$E[C(t)] = 1 - \frac{\sum_{i=1}^{S} \pi_i (1-\pi_i)^t}{\sum_{i=1}^{S} \pi_i}, \quad t \geq 1. \tag{7f}$$

For a rarefied sample ($t < T$), an unbiased estimator exists for the denominator and numerator in Eq. (7f), respectively, but their ratio $\hat{C}(t)$, given below, is only a nearly unbiased estimator of $E[C(t)]$:

$$\hat{C}(t) = 1 - \sum_{1 \leq Y_i \leq T-t} \frac{Y_i}{U} \frac{\binom{T-Y_i}{t}}{\binom{T-1}{t}}, \quad t < T.$$

An estimator for the expected coverage of an extrapolated sample with $T + t^*$ sampling units if $Q_2 > 0$ is

$$\hat{C}(T+t^*) = 1 - \frac{Q_1}{U}\left[\frac{(T-1)Q_1}{(T-1)Q_1 + 2Q_2}\right]^{t^*+1}. \tag{7g}$$

The above estimator is based on the following approximation formula:

$$E[C(T+t^*)] = 1 - \frac{\sum_{i=1}^{S} \pi_i (1-\pi_i)^{T+t^*}}{\sum_{i=1}^{S} \pi_i} \approx 1 - \frac{E[\sum_{i=1}^{S} (1-\pi_i)^{t^*+1} I(Y_i = 1)]}{T \sum_{i=1}^{S} \pi_i},$$

$$\approx 1 - \frac{[E(Q_1)](1-\phi_1)^{t^*+1}}{T \sum_{i=1}^{S} \pi_i}.$$

Replacing $\sum_{i=1}^{S} \pi_i$ and ϕ_1 with their respective estimators, U/T and $\hat{\phi}_1 = 2Q_2/[2Q_2 + (T-1)Q_1]$, we obtain Eq. (7g). If $Q_2 = 0$, a similar modification as in Eq. (7e) can be applied. Note that when $t^* = 0$, Eq. (7g) reduces to the sample coverage estimator for the reference sample. The coverage-based sampling curve includes a rarefaction part (which plots $\hat{S}(t)$ as a function of $\hat{C}(t)$), and an extrapolation part (which plots $\hat{S}(T+t^*)$ as a function of $\hat{C}(T+t^*)$), joining smoothly at the reference sample point

$(\hat{C}(T), S_{obs})$. The confidence intervals based on the bootstrap method (Section 3.3) also join smoothly. To equalize coverage among multiple, independent reference samples, their coverage-based curves can be extended to the coverage of the maximum size used in the corresponding sample-size-based sampling curve.

The sample-size-based approach plots the estimated species richness as a function of sample size, whereas the corresponding coverage-based approach plots the same richness estimate with respect to sample coverage. Therefore, the two types of sampling curves can be bridged by a sample completeness curve, which shows how the sample coverage estimate varies with sample size and also provides an estimate of the sample size needed to achieve a fixed degree of completeness. The two types of sampling curves along with the associated sample completeness curve are illustrated in Section 7 through an example. There, we also illustrate the use of the online software iNEXT (iNterpolation/EXTrapolation) to compute and plot the integrated sampling curves for incidence data. These methods allow researchers to efficiently use all available data to make more robust and more detailed inferences about species richness of the sampled assemblages, and also to make objective comparisons of species richness across assemblages.

3.3. Bootstrap method to obtain variance estimator and confidence intervals

The interpolated and extrapolated estimators are complicated functions of incidence data. Thus, it is not possible to derive analytic variance estimators. A bootstrap procedure can be applied to approximate the variance of any estimator based on incidence data. The estimated variance estimator can be subsequently used to construct a confidence interval of the expected species richness. Here we use the rarefied estimator $\hat{S}(t)$ given in Eq. (7b) as an example. Parallel steps can be formulated for any extrapolated estimator, coverage estimators, and for Chao2-type estimators.

First, we construct the *bootstrap assemblage*, which aims to mimic the true entire assemblage. Given a reference sample of size T and species sample incidence frequencies $(Y_1, Y_2, \ldots, Y_S)$, let $\hat{Q}_0$ be the Chao2-type estimator of the number of undetected species. Since the number of species in the bootstrap assemblage must be an integer, we define $\hat{Q}_0^*$ as the smallest integer that is greater than or equal to $\hat{Q}_0$. Thus, there are $S_{obs} + \hat{Q}_0^*$ species in the bootstrap assemblage.

Next we determine the detection probabilities in any sampling unit for the species in the bootstrap assemblage. Given that the ith species is detected in $Y_i > 0$ sampling units (there are S_{obs} of such species), the sample detection probability Y_i/T of an observed species ($Y_i > 0$), on average, overestimates the true detection probability π_i. This overestimation is due to the following conditional expectation:

$$E\left(\frac{Y_i}{T}\middle| Y_i > 0\right) = \frac{\pi_i}{1-(1-\pi_i)^T} > \pi_i.$$

The above conditional expectation leads to

$$\pi_i = E\left(\frac{Y_i}{T}\middle| Y_i > 0\right)[1-(1-\pi_i)^T].$$

If we replace the expected value in the above equation by the observed data, then we have the following approximation:

$$\pi_i \approx \frac{Y_i}{T}[1-(1-\pi_i)^T]. \tag{7h}$$

For any given $Y_i > 1$, one can numerically solve the above equation for π_i; but for $Y_i = 1$ (singletons, the most important count in our analysis), the only solution is $\pi_i = 0$, which is not reasonable. Therefore, Chao et al. (2014b, Appendix G) recommended the following analytic approach. Note that Eq. (7h) reveals that the approximate adjustment factor for the sample detection probability Y_i/T would be $[1-(1-\pi_i)^T]$. However, the adjustment factor $[1-(1-\pi_i)^T]$ cannot be estimated simply by substituting the sample detection probability for π_i, because the sample detection probability does not estimate π_i well for rare species. Chao et al. (2014b) suggested a more flexible adjustment factor, $[1-\tau(1-Y_i/T)^T]$. Applying this factor, we obtain that the species incidence probabilities for the S_{obs} observed species in the bootstrap assemblage can be estimated by

$$\hat{\pi}_i = \frac{Y_i}{T}\left[1-\hat{\tau}\left(1-\frac{Y_i}{T}\right)^T\right], \quad Y_i > 0, \tag{8a}$$

where $\hat{\tau}$ can be obtained from the sample coverage estimate:

$$\hat{C}(T)\times\frac{U}{T} = \sum_i \hat{\pi}_i I(Y_i > 0) = \sum_{Y_i>0}\frac{Y_i}{T}\left[1-\hat{\tau}\left(1-\frac{Y_i}{T}\right)^T\right],$$

Then we can solve for $\hat{\tau}$:

$$\hat{\tau} = \frac{\frac{U}{T}[1-\hat{C}(T)]}{\sum\limits_{Y_i\geq 1}\frac{Y_i}{T}\left(1-\frac{Y_i}{T}\right)^T} = \frac{[1-\hat{C}(T)]}{\sum\limits_{Y_i\geq 1}\frac{Y_i}{U}\left(1-\frac{Y_i}{T}\right)^T}. \tag{8b}$$

We assume that each of the remaining $\hat{Q}_0^*$ species in the bootstrap assemblage (i. e. those species that were not detected in any sampling unit but exist in the bootstrap assemblage) has a common detection probability of $(U/T)[1-\hat{C}(T)]/\hat{Q}_0^*$. This assumption may seem restrictive, but the effect on the resulting variance estimator is limited, based on our extensive simulations.

After the bootstrap assemblage is determined, a random sample of T sampling units is generated from the assemblage, and a bootstrap estimate $\hat{S}(t)$ is calculated for the generated sample. The procedure is repeated B times to obtain B bootstrap estimates ($B = 200$ is suggested). The bootstrap variance estimator $\hat{S}(t)$ is the sample variance of these B estimates. The resulting bootstrap *s.e.* of $\hat{S}(t)$ is then used to construct a 95% confidence interval $\hat{S}(t) \pm 1.96$ *s.e.* $[\hat{S}(t)]$ for the expected species richness in a sample of size t. Similar procedures can be used to derive variance estimators for any other estimator and its associated confidence intervals.

4. Species richness estimation under sampling without replacement

Chao's original inequality was developed under the binomial (Eq. 1b) model, which assumes that sampling units are taken with replacement. When sampling is done without replacement, e.g. quadrats or time periods that are not repeatedly selected/surveyed, or mobile species are collected by lethal sampling methods, Chao's inequality and the Chao2 estimator require modification, unless the sampling fraction is small. For simplicity, we assume quadrat sampling in the following derivation, but the term "quadrat," here, may refer to any sampling unit that is not sampled with replacement, such as a trap, net, team, observer, occasion, transect line, or fixed period of time in other sampling protocols. Suppose that the region under investigation consists of T^* disjoint, equal-area quadrats, and a sample of T quadrats is randomly selected. Then each quadrat is surveyed, and species detection/non-detection data are recorded for each of these T quadrats.

The model assumes that species i can be detected in only U_i quadrats (U_i is unknown). We restrict our analysis to the case $U_i > 1$. (For any species with $U_i = 0$, there is no chance to detect this species in any sample, so it should be excluded from the estimating target.) In the other $T^* - U_i$ quadrats, species i is either absent or it is present but cannot be detected. Because U_i may vary independently among species, our model holds even if species are spatially aggregated, associated, or dissociated in the study area.

Assume that detection/non-detection of all species for each of the T quadrats is recorded to form a species-by-quadrat incidence matrix. Using the same notation as in Section 2, we let Y_i (sample incidence frequency) be the number of quadrats in which the ith species is observed in the sample, $i = 1, 2, \ldots, S$. Under sampling without replacement, the sample frequencies $(Y_1, Y_2, \ldots, Y_S)$ given $U_i = u_i$, follow a product-hypergeometric distribution:

$$P(Y_i = y_i, i = 1, 2, \ldots, S) = \prod_{i=1}^{S} \left\{ \binom{u_i}{y_i} \binom{T^* - u_i}{T - y_i} \Big/ \binom{T^*}{T} \right\}, \quad 1 \le u_i \le T^*. \qquad (9a)$$

That is, $(Y_1, Y_2, \ldots, Y_S)$ are independent but non-identically distributed random variables, each of which follows a hypergeometric distribution. If the sampling fraction is relatively small (i.e. $T^* \gg T$), then equation (9a) approaches the product binomial distribution:

$$P(Y_i = y_i, i = 1, 2, \ldots, S) \to \prod_{i=1}^{S} \left\{ \binom{T}{y_i} \left(\frac{u_i}{T^*}\right)^{y_i} \left(1 - \frac{u_i}{T^*}\right)^{T-y_i} \right\}.$$

This is a model for sampling with replacement with incidence probabilities $\pi_i = u_i/T^*$. The above approximation shows that, if there are many quadrats, and only a small number of the quadrats are sampled, then the inferences for the two types of sampling schemes differ little. Based on the general model (9a), the marginal distribution for each species' frequency is a hypergeometric distribution. The expected value of the frequency counts is

$$E(Q_k) = \sum_{i=1}^{S} P(Y_i = k) = \sum_{i=1}^{S} \frac{\binom{u_i}{k}\binom{T^* - u_i}{T - k}}{\binom{T^*}{T}}. \tag{9b}$$

In particular, we have

$$E(Q_0) = \sum_{i=1}^{S} \frac{\binom{T^* - u_i}{T}}{\binom{T^*}{T}},$$

$$E(Q_1) = \sum_{i=1}^{S} \frac{\binom{u_i}{1}\binom{T^* - u_i}{T - 1}}{\binom{T^*}{T}} = \sum_{i=1}^{S} \frac{T u_i}{T^* - u_i - T + 1} \frac{\binom{T^* - u_i}{T}}{\binom{T^*}{T}}$$

$$E(Q_2) = \sum_{i=1}^{S} \frac{\binom{u_i}{2}\binom{T^* - u_i}{T - 2}}{\binom{T^*}{T}} = \sum_{i=1}^{S} \frac{T(T-1)u_i(u_i - 1)}{2(T^* - u_i - T + 1)(T^* - u_i - T + 2)} \frac{\binom{T^* - u_i}{T}}{\binom{T^*}{T}}$$

The Cauchy-Schwarz inequality leads to

$$\left\{ \sum_{i=1}^{S} \frac{\binom{T^* - u_i}{T}}{\binom{T^*}{T}} \right\} \left\{ \sum_{i=1}^{S} \left(\frac{T u_i}{T^* - u_i - T + 1} \right)^2 \frac{\binom{T^* - u_i}{T}}{\binom{T^*}{T}} \right\} \geq \left\{ \sum_{i=1}^{S} \frac{T u_i}{T^* - u_i - T + 1} \frac{\binom{T^* - u_i}{T}}{\binom{T^*}{T}} \right\}^2,$$

The right side in the above inequality is $\{E(Q_1)\}^2$, and the first sum on the left side is $E(Q_0)$. For the second sum, we rewrite

$$\left(\frac{Tu_i}{T^*-u_i-T+1}\right)^2=\frac{T}{T-1}\left(\frac{T(T-1)u_i(u_i-1)}{(T^*-u_i-T+1)^2}\right)+\frac{T^2u_i}{(T^*-u_i-T+1)^2}.$$

Thus the second sum becomes

$$\left\{\sum_{i=1}^{S}\left(\frac{Tu_i}{T^*-u_i-T+1}\right)^2\frac{\binom{T^*-u_i}{T}}{\binom{T^*}{T}}\right\}$$

$$\approx\frac{2T}{T-1}E(Q_2)+\sum_{i=1}^{S}\left[\frac{T}{T^*-u_i-T+1}\right]\frac{Tu_i}{T^*-u_i-T+1}\frac{\binom{T^*-u_i}{T}}{\binom{T^*}{T}}.$$

The contribution of species with large u_i (frequent species) to any term involved in the above Cauchy-Schwarz inequality is almost negligible. For infrequent species (with u_i much less than T^*), we have

$$\frac{T}{T^*-u_i-T+1}=\frac{T/T^*}{(T^*-u_i-T+1)/T^*}\approx\frac{T/T^*}{1-(T/T^*)}=\frac{q}{1-q},$$

where $q=T/T^*$ denotes the sampling fraction. We then obtain the following approximate inequality

$$\{E(Q_0)\}\left(\frac{T}{T-1}2E(Q_2)+\frac{q}{1-q}E(Q_1)\right)\geq\{E(Q_1)\}^2,$$

which is equivalent to

$$E(Q_0)\geq\frac{\{E(Q_1)\}^2}{\frac{T}{T-1}2E(Q_2)+\frac{q}{1-q}E(Q_1)}.$$

Replacing the expected value by the observed frequencies, we thus obtain the following lower bound for the true species richness.

$$\hat{S}_{wor2}=S_{obs}+\frac{Q_1^2}{2wQ_2+rQ_1},\tag{9c}$$

where $w=T/(T-1)$ and $r=q/(1-q)$, and the subscript "*wor*" refers to "without replacement". When the sample fraction q approaches zero, then r approaches zero, and our lower bound approaches the Chao2 estimator. On the other hand, when q approaches 1, $r=q/(1-q)$ approaches infinity and our lower bound reduces to the number of observed species, which is the true parameter for complete sampling.

An approximate variance formula for $\hat{S}_{wor2}$ can be obtained by using an asymptotic approach based on the hypergeometric distribution. The resulting variance estimator is:

$$\widehat{\text{var}}(\hat{S}_{wor2}) = \hat{Q}_0 + \frac{(2wQ_2\hat{Q}_0^2 + Q_1^2\hat{Q}_0)^2}{Q_1^5} + 4w^2Q_2\left(\frac{\hat{Q}_0}{Q_1}\right)^4,$$

where $\hat{Q}_0 = \hat{S}_{wor2} - S_{obs}$ denotes the estimator of the undetected species in the sample. When $\hat{S}_{wor2}$ is used as an estimator of species richness, a confidence interval of S can be constructed by a log-transformation (Eq. 3c), so that the lower bound is always greater than the number of observed species.

5. Shared species richness estimation

We now extend the one-assemblage model formulation and data framework to two assemblages (I and II), which can differ not only in their species richness, but also in their species composition. Suppose that there are S species in the *pooled* assemblage. Assume that T_1 sampling units (Sample I) are randomly taken from Assemblage I, and T_2 sampling units (Sample II) are taken from Assemblage II. In each sampling unit, only species detection/non-detection data are recorded. The two sets of probabilities $(\pi_{11}, \pi_{21}, \ldots, \pi_{S1})$ and $(\pi_{12}, \pi_{22}, \ldots, \pi_{S2})$ in the incidence case represent species detection probabilities in any sampling unit from Assemblages I and II, respectively, $\pi_{i1}, \pi_{i2} \geq 0$, $i = 1, 2, \ldots, S$. Let the true number of shared species between the two assemblages be S_{12}. Without loss of generality, we assume that the first S_{12}species in the pooled assemblage are these shared species.

Let Y_{i1} and Y_{i2} denote the number of sampling units in which the *i*th species is detected in Samples I and II, respectively. For any two non-negative integers r and v, define

$$Q_{rv} = \sum\nolimits_{i=1}^{S_{12}} I(Y_{i1} = r, Y_{i2} = v), \quad r,\ v = 0, 1, 2, \ldots$$

That is, Q_{rv} denotes the number of *shared* species that are detected in r sampling units in Sample I and v sampling units in Sample II. In particular, Q_{11} denotes the number of shared species that are uniques in both samples, and Q_{00} denotes the number of shared species that are present in both samples, but detected in neither. Also, let Q_{r+} denote the number of shared species that are detected in r sampling units in Sample I and that are detected in at least one sampling unit (using a "+" sign to replace the index v) in Sample II, with a similar symmetric definition for Q_{+v}. Thus, Q_{++} becomes the total number of observed species shared between the two samples. Mathematically, we have the following expressions:

$$Q_{r+} = \sum\nolimits_{i=1}^{S_{12}} I(Y_{i1} = r,\ Y_{i2} \geq 1) = \sum\nolimits_{v>0} Q_{rv}, \quad r = 0, 1, 2, \ldots$$

$$Q_{+v} = \sum_{i=1}^{S_{12}} I(Y_{i1} \geq 1,\, Y_{i2} = v) = \sum_{r>0} Q_{rv}, \quad v = 0, 1, 2, \ldots$$

Here, Q_{+0} denotes the number of *shared* species that are detected in Sample I but not detected in Sample II, and a similar interpretation for Q_{0+}.

Since $S_{12} = Q_{++} + Q_{+0} + Q_{0+} + Q_{00}$ but only Q_{++} is observable, our approach is to find a lower bound for each of the expected values of the other three terms, i.e. $E(Q_{+0})$, $E(Q_{0+})$ and $E(Q_{00})$. Assuming the binomial models (Eq. 1b) for species incidence frequencies for each of the two independent sets of frequencies, we have

$$E(Q_{00}) = \sum_{i=1}^{S_{12}} (1-\pi_{i1})^{T_1}(1-\pi_{i2})^{T_2},$$

$$E(Q_{+0}) = \sum_{i=1}^{S_{12}} [1-(1-\pi_{i1})^{T_1}](1-\pi_{i2})^{T_2},$$

$$E(Q_{0+}) = \sum_{i=1}^{S_{12}} (1-\pi_{i1})^{T_1}[1-(1-\pi_{i2})^{T_2}].$$

We now derive a lower bound for each term as follows.

1. A lower bound for $E(Q_{+0})$: Since

$$E(Q_{+1}) = \sum_{i=1}^{S_{12}} [1-(1-\pi_{i1})^{T_1}]\; T_2\, \pi_{i2}(1-\pi_{i2})^{T_2-1},$$

$$E(Q_{+2}) = \sum_{i=1}^{S_{12}} [1-(1-\pi_{i1})^{T_1}]\; [T_2(T_2-1)/2]\pi_{i2}^2(1-\pi_{i2})^{T_2-2}.$$

The following Cauchy-Schwarz inequality

$$\left[\sum_{i=1}^{S_{12}} [1-(1-\pi_{i1})^{T_1}](1-\pi_{i2})^{T_2}\right] \left[\sum_{i=1}^{S_{12}} [1-(1-\pi_{i1})^{T_1}]\; \pi_{i2}^2(1-\pi_{i2})^{T_2-2}\right]$$
$$\geq \left[\sum_{i=1}^{S_{12}} [1-(1-\pi_{i1})^{T_1}]\; \pi_{i2}(1-\pi_{i2})^{T_2-1}\right]^2$$

leads to a lower bound

$$E(Q_{+0}) \geq \frac{(T_2-1)}{T_2}\frac{[E(Q_{+1})]^2}{2E(Q_{+2})}. \tag{10a}$$

2. Similarly, a lower bound for $E(Q_{0+})$ is

$$E(Q_{0+}) \geq \frac{(T_1-1)}{T_1}\frac{[E(Q_{1+})]^2}{2E(Q_{2+})}. \tag{10b}$$

3. A lower bound for $E(Q_{00})$ is obtained by noting

$$E(Q_{11}) = \sum_{i=1}^{S_{12}} T_1\; \pi_{i1}(1-\pi_{i1})^{T_1-1}\; T_2\; \pi_{i2}(1-\pi_{i2})^{T_2-1},$$

$$E(Q_{22}) = \sum\nolimits_{i=1}^{S_{12}} [T_1(T_1-1)/2]\pi_{i1}^2(1-\pi_{i1})^{T_1-2}[T_2(T_2-1)/2]\pi_{i2}^2(1-\pi_{i2})^{T_2-2}.$$

Again, a similar Cauchy-Schwarz inequality

$$\left[\sum\nolimits_{i=1}^{S_{12}} (1-\pi_{i1})^{T_1}(1-\pi_{i2})^{T_2}\right] \left[\sum\nolimits_{i=1}^{S_{12}} \pi_{i1}^2(1-\pi_{i1})^{T_1-2}\pi_{i2}^2(1-\pi_{i2})^{T_2-2}\right]$$
$$\geq \left[\sum\nolimits_{i=1}^{S_{12}} \pi_{i1}(1-\pi_{i1})^{T_1-1}\pi_{i2}(1-\pi_{i2})^{T_2-1}\right]^2$$

gives

$$E(Q_{00}) \geq \frac{(T_1-1)}{T_1}\frac{(T_2-1)}{T_2}\frac{[E(Q_{11})]^2}{4E(Q_{22})}. \tag{10c}$$

Combining the above three lower bounds and letting $K_i = (T_i-1)/T_i$, we thus have a lower bound for the shared species richness:

$$\hat{S}_{12} = Q_{++} + K_2\frac{Q_{+1}^2}{2Q_{+2}} + K_1\frac{Q_{1+}^2}{2Q_{2+}} + K_1K_2\frac{Q_{11}^2}{4Q_{22}}. \tag{10d}$$

The above estimator is referred to as the *Chao*2-*shared* estimator because it can be regarded as an extension of the single-assemblage Chao2 estimator (Eq. 2b) to the case of two assemblages. A bias-corrected estimator to avoid zero divisor is

$$\tilde{S}_{12} = Q_{++} + K_2\frac{Q_{+1}(Q_{+1}-1)}{2(Q_{+2}+1)} + K_1\frac{Q_{1+}(Q_{1+}-1)}{2(Q_{2+}+1)} + K_1K_2\frac{Q_{11}(Q_{11}-1)}{4(Q_{22}+1)}. \tag{10e}$$

Note that only observed, shared species are involved in the formulas (10a) to (10e), thus observed non-shared species play no role in our estimation, although any species observed in one Sample but not in the other could actually be a shared species. Because the proposed estimator can be regarded as a function of the statistics $(Q_{++}, Q_{11}, Q_{22}, Q_{1+}, Q_{2+}, Q_{+1}, Q_{+2})$, we obtain a variance estimator by using a standard asymptotic approach under a multinomial distribution. Then the estimated variance can be used to construct a confidence interval for the true parameter using a log-transformation (Chao, 1987).

The above approach has an obvious extension to the case of more than two assemblages. For example, in the case of three assemblages, a "shared" species is defined as that the species belongs to all three assemblages. Assume that there are S_{123} species shared by all three assemblages (I, II and III), and a random sample of sampling units is taken from each of the three assemblages. The three samples are called Samples I, II and III with sizes T_1, T_2 and T_3 respectively. Then

$$S_{123} = Q_{+++} + Q_{++0} + Q_{+0+} + Q_{0++} + Q_{00+} + Q_{0+0} + Q_{+00} + Q_{000},$$

where Q_{+++} denotes the observed shared species richness in the three samples, Q_{++0} denotes the number of shared species that are observed in Samples I, II but not observed in Sample III, Q_{000} denotes the number of shared species that are not detected in any of the three samples, and a similar interpretation for other terms in the above formula. Parallel derivations (with self-explanatory notation) lead to a lower bound for S_{123} as follows:

$$\begin{aligned}\hat{S}_{123} = Q_{+++} + K_3\frac{Q_{++1}^2}{2Q_{++2}} + K_2\frac{Q_{+1+}^2}{2Q_{+2+}} + K_1\frac{Q_{1++}^2}{2Q_{2++}} \\ + K_1K_2\frac{Q_{11+}^2}{4Q_{22+}} + K_1K_3\frac{Q_{1+1}^2}{4Q_{2+2}} + K_2K_3\frac{Q_{+11}^2}{4Q_{+22}} + K_1K_2K_3\frac{Q_{111}^2}{8Q_{222}}.\end{aligned}$$

We can formulate a bias-corrected version to avoid zero divisor in the same manner as that given in Eq. (10e). An estimated variance can be obtained by an asymptotic method.

6. Phylogenetic richness estimation

6.1. Framework

In traditional measures of species diversity, all species (or taxa at some other rank) are considered to be equally distinct from one another. However, in an evolutionary context, species differences can be based directly on their evolutionary relationships, either in the form of taxonomic classification or well-supported phylogenetic trees. Species that are closely related are generally less distinct in important ecological characteristics than are distantly-related species. A wide range of phylogenetic diversity metrics and related (dis)similarity measures have been proposed in the literature. The most widely used phylogenetic metric is Faith's (1992) *PD* (phylogenetic diversity), which is defined as the sum of the branch lengths of a phylogenetic tree connecting all species in the focal assemblage.

Chao et al. (2010, 2015) proposed a class of abundance-sensitive phylogenetic measures and showed that Faith's *PD* is a phylogenetic generalization of species richness. In other words, Faith's *PD* is a phylogenetic diversity of order zero in which species abundances are not considered. From this perspective, Faith's *PD* is a measure of *phylogenetic richness*. Throughout this paper, *PD* refers to Faith's (1992) *PD*. When some species that are present in an assemblage are not detected in a sample, the lineages/branches associated with these undetected species are also missing from the phylogenetic tree of the observed species. The undetected *PD* in an incomplete sample was not discussed until recent years (Cardoso et al., 2014; Chao et al., 2015).

Model formulation and *PD* estimation based on abundance data were developed in Chao et al. (2015). The corresponding framework for incidence data, introduced in their Appendix S7 and presented here, is a generalization of the framework for species richness. As discussed in Section 2.1, suppose, in the focal assemblage, that there are S species indexed by $1,2,\ldots,S$, and T sampling units are surveyed from the assemblage. In each sampling unit, we assume that only incidence (detection or non-detection) of each species is recorded. For any sampling unit, assume that the ith species has its own unique incidence (or detection) probability π_i that is constant for any randomly selected sampling unit. We also assume that a rooted ultrametric or non-ultrametric phylogenetic tree of the S species (as tip nodes) can be constructed. Here we assume that all phylogenetic measures are computed from a fixed, basal reference point in the tree that is ancestral to all taxa considered in the study.

Assume that there are B branch segments in the corresponding tree, $B \geq S$, descendant to the given basal reference point. Let L_i denote the length of the ith branch. We expand the set of detection probabilities $(\pi_1, \pi_2, \ldots, \pi_S)$ of the S species (as tip nodes) to a larger set of branch/node detection probabilities $\{\lambda_i,\, i=1,2,\ldots,B\}$ with $(\pi_1, \pi_2, \ldots, \pi_S)$ as the first S elements. Here we define λ_i as the probability of detecting at least one species descended from branch i in a sampling unit, $i=1,2,\ldots,B$, and refer to λ_i as the incidence (or detection) probability of branch/node i. The true PD for the fixed reference point is expressed as $PD = \sum_{i=1}^{B} L_i$.

The species-by-sampling-unit incidence matrix $\{W_{ij};\, i=1,2,\ldots,S,\, j=1,2,\ldots,T\}$ and the species incidence frequencies $Y_i = \sum_{j=1}^{T} W_{ij}$ are defined exactly the same as those in Section 2.1. Here we expand the $S \times T$ incidence matrix $\{W_{ij};\, i=1,2,\ldots,S,\, j=1,2,\ldots,T\}$ to a larger $B \times T$ matrix $\{W_{ij}^*,\, i=1,2,\ldots,B,\, j=1,2,\ldots,T\}$ by specifying that $W_{ij}^* = 1$ if at least one species descended from branch i is detected in jth sampling unit, and $W_{ij}^* = 0$ otherwise. This specification also expands the set of the observed species incidence frequencies $\{Y_1, Y_2, \ldots, Y_S\}$ to a larger set $\{Y_i^*,\, i=1,2,\ldots,B\}$, which consists of the row sums of the expanded incidence matrix $[W_{ij}^*]$. We refer to Y_i^* as the sample *branch/node incidence frequency* for branch/node i, $i=1,2,\ldots,B$. See Table 3 for a simple, hypothetical dataset for nine species in six sampling units, illustrating the expansion of the matrix $[W_{ij}]$ to $[W_{ij}^*]$.

Suppose that the incidence data for all the sampling units are independent. Then Y_i^*, $i=1,2,\ldots,B$, follows a binomial distribution:

$$P(Y_i^* = y_i) = \binom{T}{y_i} \lambda_i^{y_i} (1-\lambda_i)^{T-y_i}, \quad y_i = 0,1,2,\ldots,T.$$

Define R_k as the sum of branch lengths for those branches with branch/node incidence frequency k, i.e.

$$R_k = \sum\nolimits_{i=1}^{B} L_i\, I(Y_i^* = k), \quad k = 0,1,\ldots,T. \tag{11a}$$

Thus, R_0 represents the total length of branches that are not detected in the observed tree (i.e. not detected by the tree spanned by the observed species in the reference sample), and R_1 denotes the total branch length of the uniques in the branch incidence frequency set of the observed tree. A similar interpretation is valid for R_2. Let PD_{obs} denote the

Table 3: *Species detection/non-detection data for the hypothetical tree in Figure 1. Species 4, 7, 8, and 9 (grey shaded area) are not observed in the sample; Node 14 (grey shaded area) is not observed in the tree spanned by the observed species.*

Species/node/branch	Detection/non-detection in six sampling units (1 means detection; blank means non-detection)						Species/node/branch incidence frequency
	1	2	3	4	5	6	
1	1	1	1	1	1	1	$Y_1 = 6$
2		1					$Y_2 = 1$
3		1					$Y_3 = 1$
4							$Y_4 = 0$
5		1	1				$Y_5 = 2$
6				1			$Y_6 = 1$
7							$Y_7 = 0$
8							$Y_8 = 0$
9							$Y_9 = 0$
10		1					$Y_{10}^* = 1$
11	1	1	1	1	1	1	$Y_{11}^* = 6$
12		1	1				$Y_{12}^* = 2$
13				1			$Y_{13}^* = 1$
14							$Y_{14}^* = 0$
15				1			$Y_{15}^* = 1$

observed PD. Then we have $PD_{obs} = \sum_{i>0} R_i$ and $PD = PD_{obs} + R_0$. See Figure 1 for a hypothetical tree spanned by 9 species for an example.

6.2. Chao's inequality for PD

The undetected PD in the reference sample is R_0, which is unknown. However, $\{R_1, R_2, \dots\}$ can be computed from the reference sample and the tree spanned by the observed species. Following the same approach that Chao et al. (2015) used for abundance data, we have the expected value of R_k:

$$E(R_k) = E\left[\sum_{i=1}^{B} L_i I(Y_i^* = k)\right] = \binom{T}{k} \sum_{i=1}^{B} L_i \lambda_i^k (1-\lambda_i)^{T-k}, \quad k = 0, 1, \dots, T. \tag{11b}$$

In particular, we have

$$E(R_0) = \sum_{i=1}^{B} L_i (1-\lambda_i)^T,$$

$$E(R_1) = T \sum_{i=1}^{B} L_i \lambda_i (1-\lambda_i)^{T-1},$$

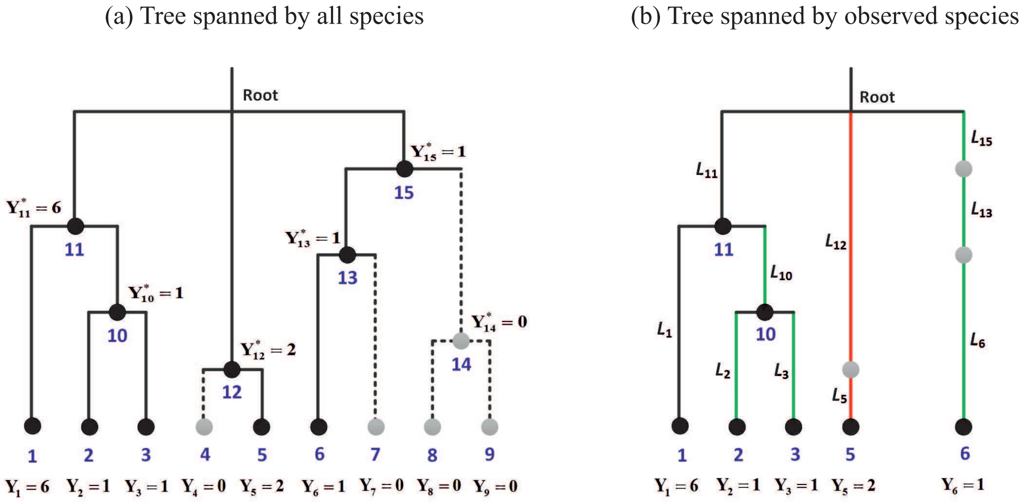

(a) Tree spanned by all species
(b) Tree spanned by observed species
Root
Y*11 = 6
11
Y*10 = 1
10
Y*12 = 2
12
Y*13 = 1
13
Y*15 = 1
15
Y*14 = 0
14
1 2 3 4 5 6 7 8 9
Y1 = 6 Y2 = 1 Y3 = 1 Y4 = 0 Y5 = 2 Y6 = 1 Y7 = 0 Y8 = 0 Y9 = 0
Root
L11
L15
L13
L12
L10
L1
L6
L2
L3
L5
11
10
1 2 3 5 6
Y1 = 6 Y2 = 1 Y3 = 1 Y5 = 2 Y6 = 1

Figure 1: *(a) A hypothetical tree spanned by 9 species (tip nodes) indexed by 1, 2, ..., 9 in an assemblage. The ancestor of the entire assemblage is the "root" at the top, with time progressing towards the branch tips at the bottom. Here the root of the entire assemblage is selected as the reference point for illustration. Species detection/non-detection records in six sampling units are given in Table 3. A black dot means a node with species incidence frequency* > 0*; a grey dot means a node with frequency* $= 0$*. (b) A sub-tree spanned by the observed 5 species (1, 2, 3, 5 and 6). Species 4, 7, 8 and 9 are not detected in any of the six sampling units, so only a portion of the tree (solid branches in the left panel) is observed as shown in Panel (b). Black dots in Panel (b) are nodes in the observed tree; grey dots are not observed in the tree. The sample incidence frequency vector in 6 sampling units for 9 species is* $(Y_1, Y_2, \ldots, Y_9)$ *= (6, 1, 1, 0, 2, 1, 0, 0, 0); only non-zero frequencies represent observed species. The branch set B in the assemblage includes 15 branches (indexed from 1 to 15) with branch lengths* $(L_1, L_2, \ldots, L_{15})$ *and the corresponding 15 nodes. The corresponding node/branch incidence frequencies are* $(Y_1^*, Y_2^*, \ldots, Y_9^*, Y_{10}^*, Y_{11}^*, \ldots, Y_{15}^*)$ *= (6, 1, 1, 0, 2, 1, 0, 0, 0, 1, 6, 2, 1, 0, 1) with* $(Y_1, Y_2, \ldots, Y_9)$ *as the first 9 elements (see Table 3). The dotted branches in Panel (a) are not detected in the sample, and the total length of the undetected branches is* $R_0 = L_4 + L_7 + L_8 + L_9 + L_{14}$*. In Panel (b), the total length of those branches with* $Y_i^* = 1$ *(there are four uniques in the node/branch incidence frequency set of the observed tree) is* $R_1 = L_2 + L_3 + L_6 + L_{10} + L_{13} + L_{15}$ *(as shown by green lines in the observed tree in Panel (b)); the total length with* $Y_i^* = 2$ *(only one duplicate in the node/branch incidence frequency set of the observed tree) is* $R_2 = L_5 + L_{12}$ *(as shown by red lines in the observed tree in Panel (b)).*

$$E(R_2) = \frac{T(T-1)}{2} \sum_{i=1}^{B} L_i \lambda_i^2 (1-\lambda_i)^{T-2}.$$

The Cauchy-Schwarz inequality

$$\left[\sum_{i=1}^{B} L_i (1-\lambda_i)^T\right] \left[\sum_{i=1}^{B} L_i \lambda_i^2 (1-\lambda_i)^{T-1}\right] \geq \left[\sum_{i=1}^{B} L_i \lambda_i (1-\lambda_i)^{T-1}\right]^2$$

leads to the following inequality:

$$E(R_0) \geq \frac{(T-1)}{T} \frac{[E(R_1)]^2}{2E(R_2)}.$$

Thus, a direct estimator of the undetected *PD* would be $\frac{(T-1)}{T}\frac{R_1^2}{2R_2}$. However, when R_2 is relatively small, including the case of $R_2 = 0$, this estimator may yield an extremely large value and thus exhibit a large variance. To cope with such cases, Chao et al. (2015) and Hsieh and Chao (2017) proposed the following Chao2-*PD* estimator:

$$\widehat{PD}_{Chao2} = PD_{obs} + \hat{R}_0 = \begin{cases} PD_{obs} + \dfrac{(T-1)}{T}\dfrac{R_1^2}{2R_2}, & \text{if } R_2 > \dfrac{R_1 Q_2^*}{2Q_1^*}; \\[2ex] PD_{obs} + \dfrac{(T-1)}{T}\dfrac{R_1(Q_1^*-1)}{2(Q_2^*+1)}, & \text{if } R_2 \leq \dfrac{R_1 Q_2^*}{2Q_1^*}. \end{cases} \tag{11c}$$

where Q_1^* and Q_2^* denote, respectively, the number of nodes/branches with incidence frequency $= 1$ and frequency $= 2$ in the observed tree; see Figure 1 for an example.

As with the Chao2 estimator, this lower bound is a nearly unbiased point estimator if unique and undetected branches/nodes have approximately identical mean detection probabilities. A sufficient condition is that rare/infrequent node/branch detection probabilities are approximately homogeneous, while other nodes/branches can be highly heterogeneous. When the detection probabilities for rare nodes/branches are heterogeneous and the sample is not sufficiently large, negative bias exists. The variance of the Chao1-*PD* estimator can be obtained using Eqs. (3a) and (3b) with $\{Q_1, Q_2\}$ being replaced by $\{R_1, R_2\}$. The construction of the confidence interval for Faith *PD* based on the Chao1-*PD* estimator can be similarly obtained as that given in Eq. (3c).

Comparing the derivations for the above phylogenetic version of Chao's inequality with those in Section 2.3 for species richness, we see that all estimation steps are parallel and the analogy between the two estimation frameworks is transparent. The analogy was first proposed by Faith (1992). From Faith's perspective, each unit-length branch is regarded as a "feature" in phylogenetic diversity (like a "species" in species diversity). Chao et al. (2014a) subsequently referred to each unit-length branch segment as a *phylogenetic entity*. All entities are phylogenetically equally distinct, just as all species are assumed taxonomically equally distinct in computing simple species richness. Instead of species, for *PD* we are measuring the total number of phylogenetic entities, or equivalently, the total branch length (because each entity has length of unity). Based on this perspective, for incidence data the measures of branch lengths $\{R_k, k = 0, 1, \ldots,\}$ used to estimate *PD* play the same role as the frequency counts $\{Q_k, k = 0, 1, \ldots\}$ in estimating species richness. This analogy to counting up species means that most ecological indices defined at the species level can be converted to *PD* equivalents (by counting phylogenetic entities rather than species).

6.3. Rarefaction/extrapolation guided by the Chao2-PD estimator

Because of the analogy between counting up species richness and counting up branch lengths, all the species richness estimation tools for standardized samples in Section 3 can be directly extended to their phylogenetic equivalents, and similar sample-size-based and coverage-based rarefaction and extrapolation sampling curves can be constructed. Table 2 gives all the corresponding formulas; thus we omit all details except for the following two notes.

The theoretical formula for $PD(t)$, the expected PD when a set of t sampling units is taken from the assemblage, is a generalization of Eq. (7a):

$$PD(t) = \sum_{i=1}^{B} L_i[1-(1-\lambda_i)^t], \quad t = 1,2,\ldots$$

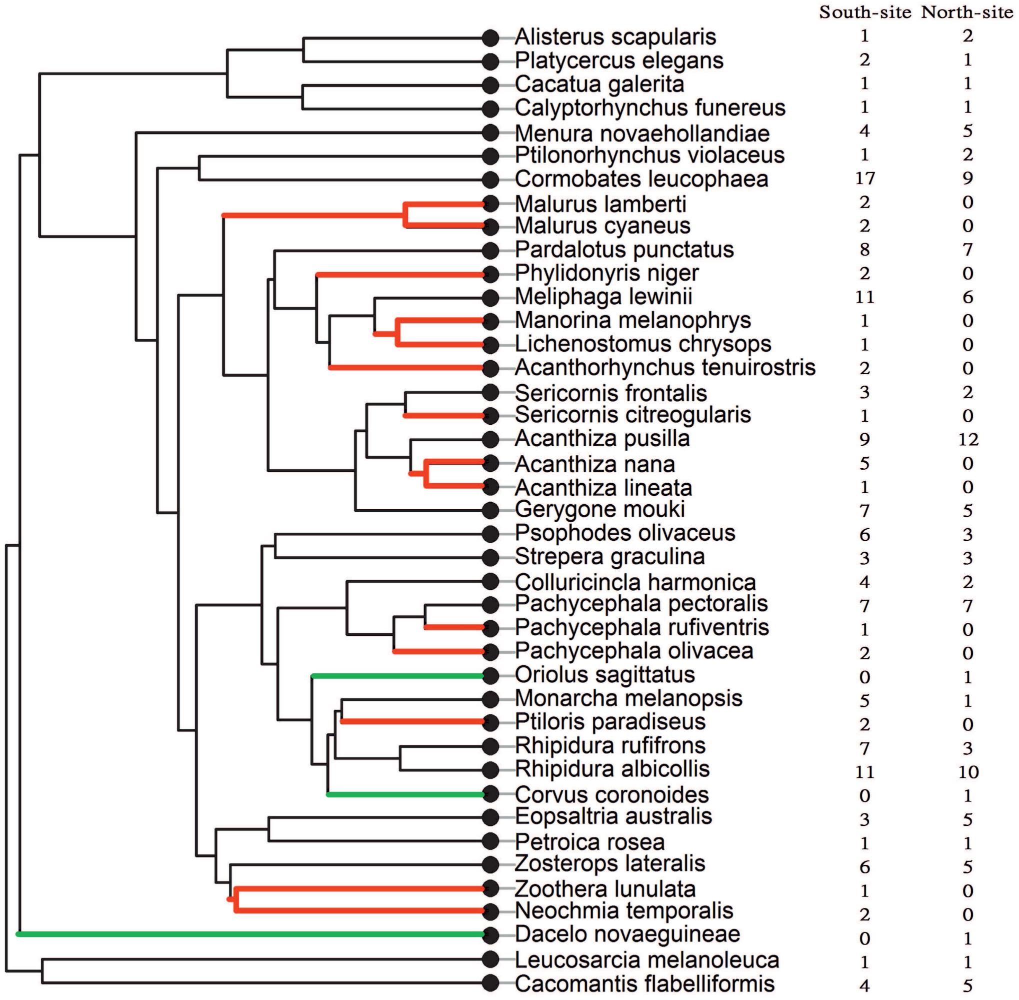
The plot of $PD(t)$ as a function of t is a non-decreasing function and is referred to as the *sampling-unit-based PD accumulation curve*. As sample size t tends to infinity, $PD(t)$ approaches the true PD. Thus the true PD represents the "asymptote" of the PD accumulation curve, i.e. the true $PD = PD(\infty)$. When there are no internal nodes, and all S lineages are equally distinct with branch lengths of unity (i.e. branch lengths are normalized to unity), the sampling-unit-based PD accumulation curve reduces to the species accumulation curve.

The bootstrap method to assess the variance and confidence interval associated with the PD estimator for rarefied and extrapolated samples is similar to that in Section 3.3, except that a "bootstrap tree" should be constructed in the resampling procedure. Recall that, in the bootstrap assemblage discussed in Section 3.3 for species richness, there are $S_{obs} + \hat{Q}_0^*$ species, where $\hat{Q}_0^*$ is the smallest integer that is greater than or equal to the estimated undetected species richness $\hat{Q}_0$ based on the Chao2 estimator in Eq. (2b). The PD bootstrap tree includes two portions: the known tree spanned by the observed species, and the undetected tree spanned by the remaining $\hat{Q}_0^*$ species in the bootstrap assemblage. The latter portion of tree is estimated by assuming that the undetected species in the bootstrap tree all diverged directly from the root of the observed tree with a constant branch length $\hat{\bar{L}}_{(0)}$, where $\hat{\bar{L}}_{(0)} = \hat{R}_0/\hat{Q}_0^*$, and $\hat{R}_0$ is the estimated undetected PD based on Eq. (11c). This augmented portion of tree may seem to be restrictive, but the effect on the resulting variance is limited; see Chao et al. (2015) for details.

7. Example

7.1. Data description (Figure 2, Appendices A and B)

A small empirical data set for birds observed in November 2012 in Australian Barrington Tops National Park is used for illustration. The original data were described in Chao et al. (2015). At each data sampling point, the abundance of each bird species observed



	South-site	North-site
Alisterus scapularis	1	2
Platycercus elegans	2	1
Cacatua galerita	1	1
Calyptorhynchus funereus	1	1
Menura novaehollandiae	4	5
Ptilonorhynchus violaceus	1	2
Cormobates leucophaea	17	9
Malurus lamberti	2	0
Malurus cyaneus	2	0
Pardalotus punctatus	8	7
Phylidonyris niger	2	0
Meliphaga lewinii	11	6
Manorina melanophrys	1	0
Lichenostomus chrysops	1	0
Acanthorhynchus tenuirostris	2	0
Sericornis frontalis	3	2
Sericornis citreogularis	1	0
Acanthiza pusilla	9	12
Acanthiza nana	5	0
Acanthiza lineata	1	0
Gerygone mouki	7	5
Psophodes olivaceus	6	3
Strepera graculina	3	3
Colluricincla harmonica	4	2
Pachycephala pectoralis	7	7
Pachycephala rufiventris	1	0
Pachycephala olivacea	2	0
Oriolus sagittatus	0	1
Monarcha melanopsis	5	1
Ptiloris paradiseus	2	0
Rhipidura rufifrons	7	3
Rhipidura albicollis	11	10
Corvus coronoides	0	1
Eopsaltria australis	3	5
Petroica rosea	1	1
Zosterops lateralis	6	5
Zoothera lunulata	1	0
Neochmia temporalis	2	0
Dacelo novaeguineae	0	1
Leucosarcia melanoleuca	1	1
Cacomantis flabelliformis	4	5

Figure 2: *The phylogenetic tree of 41 bird species and the sample species incidence frequencies for two sites (the North Site with 12 point-counts and the South Site with 17 point-counts) in Australian Barrington Tops National Park (Chao et al., 2015). The phylogenetic tree is a Maximum Clade Credibility tree from the Bayesian analysis of Jetz et al. (2012). Branch lengths are scaled to millions of years since divergence. The phylogenetic tree for the species observed in the North Site includes black branches and green branches. The phylogenetic tree for the species observed in the South Site includes black branches and red branches. (Black branches are shared by both sites; red and green branches are non-shared.) A zero-frequency in a site means that the species was not observed in that site. The age of the root (i.e. tree depth) is 82.9 millions of years.*

over a 30-minute period in a 50 m radius was recorded – called a *point-count* in ornithology. We treat each point-count as a sampling unit. There were 12 point counts conducted along the Barrington Tops Forest Road in the northern part of the national park. The corresponding records, shown in Appendix A, form the reference sample for the North Site. There were 17 point counts conducted along the Gloucester Tops

Road in the southern part of the Barrington Tops National Park; the raw detection/non-detection records (ignoring abundances) for the 17 point counts are listed in Appendix B. Those records form the reference sample for the South Site. Vegetation at both sites ranged from wet eucalypt forest to rainforest, with an average canopy cover of 80% for the North Site and 60% for the South Site. The sampling points comprising the North Site had an average elevation of 1078 m, while those of the South Site had an average elevation of 928 m. A total of 41 species were observed, for both sites combined, and all species incidence frequencies are shown in Figure 2 and in the last column of Appendices A and B. A phylogenetic tree of these species (Figure 2) was constructed from a Maximum Clade Credibility tree of the Bayesian analysis of Jetz et al. (2012). The age of the root for the phylogenetic tree spanned by the observed species is 82.9 million years (Myr). Chao et al. (2015) analyzed these data based on species abundance data. Here we focus on species incidence frequency data which can account for spatial heterogeneity in the data, whereas abundance-based approach often cannot.

7.2. Species richness and shared species richness estimation (Table 4)

In the North Site ($T = 12$ sampling units), the reference sample includes 102 incidences ($U = 102$) representing 27 observed species; in the South Site ($T = 17$ sampling units), the reference sample includes 148 incidences ($U = 148$) representing 38 observed species. The species incidence frequency counts (Q_1 to Q_T) for the two sites are summarized in Table 4. Based on Eq. (7d), the estimated sample coverage values for the North Site and the South Site are nearly identical at a level of 92% (specifically, 91.8% for the North Site and 92.5% for the South Site) in spite of the difference in the number of sampling units. Thus, the raw data imply that the South Site is more diverse than the North site for a standardized fraction of approximately 92% of the individuals in each assemblage.

In each site, some species were each observed in only one point-count. The existence of such "uniques" signifies that some species were undetected in each site. In the North Site, 9 species were observed in only one point-count ($Q_1 = 9$) and 4 species were observed in two point-counts ($Q_2 = 4$). These 13 rare species contain most of the available information about the number of undetected species. The Chao2 formula in Eq. (2b) implies a species richness estimate for the North Site of 36.3, with a 95% confidence interval of (29.1, 68.8). In the South Site, 12 species were observed in only one point ($Q_1 = 12$), and 8 species were observed in two points ($Q_2 = 8$). The Chao2 formula in Eq. (2b) yields a species richness estimate of 46.5 for the South Site, with a 95% confidence interval of (40.3, 69.8). Richness estimates based on the improved iChao2 estimator (38.6 for North and 48.2 for South), derived by Chiu et al. (2014) in Eq. (6d), differ little from the corresponding Chao2 estimates, so our interpretation is mainly based on the Chao2 estimates. All estimates were computed from the SpadeR Online (Species-richness Prediction And Diversity Estimation Online) software, which is available from Anne Chao's website at `http://chao.stat.nthu.edu.tw/wordpress/software_download/`.

Table 4: *A summary of raw data and species richness estimation for bird species in two sites (the South Site and the North Site in Australian Barrington Tops National Park); see Chao et al. (2015).*

(a) Species incidence frequency counts in the North Site ($S_{obs} = 27$, $T = 12$, total number of incidences $U = 102$, sample coverage estimate = 91.8%); Q_i: the number of species detected in exactly i sampling units (point counts).

i	1	2	3	5	6	7	9	10	11	12
Q_i	9	4	3	5	1	2	1	1	0	1

(b) Species incidence frequency counts in the South Site ($S_{obs} = 38$, $T = 17$, total number of incidences $U = 148$, sample coverage estimate = 92.5%).

i	1	2	3	4	5	6	7	8	9	11	17
Q_i	12	8	3	3	2	2	3	1	1	2	1

(c) Undetected species richness and Chao2 point and interval estimates for each site; see Eq. (2b)

Site	T	Q_1	Q_2	Observed richness	Undetected richness	Chao2 richness	s.e. of Chao2	95% conf. interval
North	12	9	4	27	9.28	36.28	8.31	(29.06, 68.77)
South	17	12	8	38	8.47	46.47	6.43	(40.25, 69.78)

(d) Undetected shared species richness between the two sites and the corresponding Chao2-shared point and interval estimates for shared species richness; see Eq. (10e)

Observed shared richness	Q_{+1}	Q_{+2}	Q_{1+}	Q_{2+}	Q_{11}	Q_{22}
24	6	1	6	4	4	0

$\hat{Q}_{+0}$	$\hat{Q}_{0+}$	$\hat{Q}_{00}$	Undetected shared richness	Chao2-shared richness	s.e. of Chao2-shared	95% conf. interval
2.75	7.06	2.59	12.39	36.39	11.42	(26.67,81.64)

The above results reveal that a relatively high fraction of the species present in each site remain undetected. As discussed in Section 2.3, if we can assume for each site that all undetected and unique species have approximately the same probability to be detected in each point-count, then these asymptotic estimates represent nearly unbiased estimates and can be compared between the two sites. In this case, the data are not sufficient to detect statistically significant differences in richness between the two sites, as reflected by the overlapping confidence intervals associated with the two Chao2 estimates (Table 4). However, the data do support inference of a significance difference in species richness if only a fraction of the assemblages are compared, as shown by the disjoint confidence intervals in the coverage-based rarefaction and extrapolation in the next sub-section.

Table 4 also shows overlap information and shared species richness estimation between the two sites. Out of the 24 observed shared species, 4 were uniques in both sites ($Q_{11} = 4$), 12 shared species were uniques in one site or the other ($Q_{+1} = 6$, $Q_{1+} = 6$), one shared species was a duplicate in the South Site ($Q_{+2} = 1$), and 4 shared species were duplicates in the North Site ($Q_{2+} = 4$). The existence of such rare shared species signifies that there were undetected shared species. Based on the Chao2-shared formula (Eq. 10e), the minimum number of undetected shared species is estimated to be 12.4, and the minimum shared species richness is estimated to be 36.4, with a 95% confidence interval of (26.7, 81.6); see Table 4 for details. Our approach reveals the extent of underestimation and provides helpful information for understanding community/assemblage overlap.

7.3. Rarefaction and extrapolation of species richness (Figures 3, 4 and 5)

We use the data from these two sites to illustrate the construction of two types of rarefaction and extrapolation curves of species richness (sample-size-based and coverage-based), and the sample completeness curve; all formulas are given in Table 2. The constructed sampling curves are then used to compare species richness between the two sites. These sampling curves can be obtained using the online software iNEXT (iNterpolation and EXTrapolation, available from the website address is given in Section 7.2). iNEXT online returns the three sampling curves as shown in Figures 3, 4 and 5, along with some related statistics (omitted here). The omitted output includes basic data information and species richness estimates for some rarefied and extrapolated samples.

The sample-size-based sampling curve (Figure 3) includes a rarefaction part (which plots $\hat{S}(t)$ as a function of $t < T$), and an extrapolation part (which plots $\hat{S}(T + t^*)$ as a function of $T + t^*$), joining smoothly at the reference point (T, S_{obs}). The confidence intervals based on the bootstrap method also join smoothly. With this type of sampling curve, we can compare species richness for two equally-large samples along with 95% confidence intervals. For each site, the extrapolation is extended to 34 sampling units, double that of the reference sample size of the South Site. Extrapolation beyond the doubled reference sample size could theoretically be computed and used for ranking species richness, but the estimates may be subject to some prediction biases and should be used with caution in estimating species richness ratios or other measures. Figure 3 reveals that the curve for the South Site lies above that of the North Site. However, the confidence intervals of the two sites overlap, implying that comparing two equally-large samples

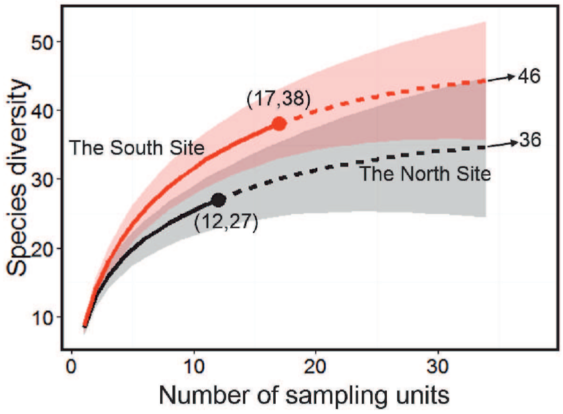


Figure 3: *Sample-size-based rarefaction (solid lines) and extrapolation (dashed lines) sampling curves with 95% confidence intervals (shaded areas, based on a bootstrap method with 200 replications) comparing species richness for Australian bird data in two sites (the South Site and the North Site in Barrington Tops National Park); see Chao et al. (2015). Observed (reference) samples are denoted by the solid dots. The extrapolation extends up to a maximum sample size of 34; here the sample size means the number of sampling units. The numbers in parentheses are the number of sampling units and the observed species richness for each reference sample. The estimated asymptote for each curve is shown next to the arrow at the right-hand end of each curve.*

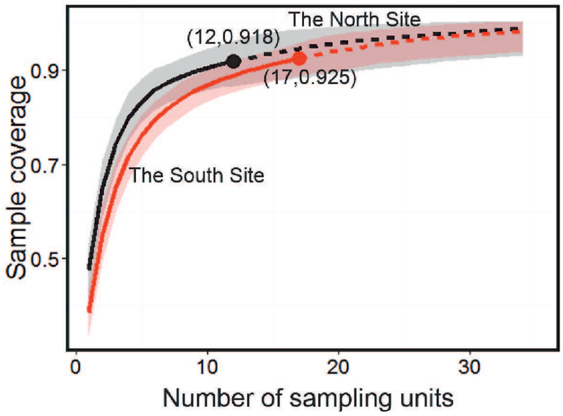


Figure 4: *Plot of sample coverage for rarefied samples (solid line) and extrapolated samples (dashed line) as a function of sample size for Australian bird data in two sites (the South Site and the North Site in Barrington Tops National Park); see Chao et al. (2015). Observed (reference) samples are denoted by solid dots. The 95% confidence intervals (shaded areas) are obtained by a bootstrap method based on 200 replications. Each of the two curves was extrapolated up to the base sample size of 34. The numbers in parentheses are the number of sampling units and the estimated sample coverage for each reference sample.*

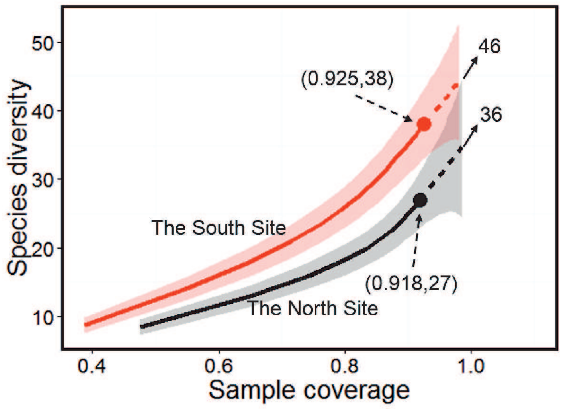


Figure 5: *Coverage-based rarefaction (solid lines) and extrapolation (dashed lines) sampling curves with 95% confidence intervals (shaded areas, based on a bootstrap method with 200 replications) for comparing species richness for Australian bird data in two sites (the South Site and the North Site in Barrington Tops National Park); see Chao et al. (2015). Observed (reference) samples are denoted by solid dots. The extrapolation extends up to the coverage value of the corresponding maximum number of sampling units of 34 in Figure 4 (98.5% in the North Site and 98.1% in the South Site). The numbers in parentheses are the estimated coverage and the observed species richness for each reference sample. The estimated asymptote for each curve is shown next to the arrow at the right-hand end of each curve.*

is inconclusive regarding the test of significant difference in species richness between the two sites. Generally, for any fixed sample size (or completeness) in the comparison range, if the 95% confidence intervals do not overlap, then significant differences at a level of 5% among the expected diversities (whether interpolated or extrapolated) are guaranteed. However, overlapping intervals do not guarantee non-significance (Colwell et al., 2012).

The sample completeness curve (Figure 4) shows how the sample coverage varies with the number of sampling units, along with 95% confidence intervals for each of the two sites, up to the sample size of 34. This curve includes a rarefaction part (which plots $\hat{C}(t)$ as a function of $t < T$), and an extrapolation part (which plots $\hat{C}(T+t^*)$ as a function of $T+t^*$), joining smoothly at the reference point $(T, \hat{C}(T))$. For any fixed number of sampling units, the curve of the North Site lies consistently above that of the South Site, but there is little difference between the two curves when the number of units exceeds 10. For the North Site, when the number of units is extended from 12 to 34, the sample coverage is extended from 91.8% to 98.5% (a number provided by the unreported iNEXT output). For the South Site, when the sample size is extended from 17 to 34 the coverage is extended from 92.5% to 98.1% (as shown in the unreported iNEXT output). The sample completeness curve provides a bridge between sample-size-based and coverage-based sampling curves.

The coverage-based sampling curve (Figure 5) includes a rarefaction part (which plots $\hat{S}(t)$ as a function of $\hat{C}(t)$ for $t < T$), and an extrapolation part (which plots

$\hat{S}(T+t^*)$ as a function of $\hat{C}(T+t^*)$), joining smoothly at the reference sample point $(\hat{C}(T),\ S_{obs})$. In this type of sampling curve, we compare species richness for two equally-complete samples along with 95% confidence intervals. The extrapolation is extended to 98.5% for the North Site and to 98.1% for the South Site, as explained in the preceding paragraph. One advantage of using coverage-based curves is that the South Site has significantly greater species richness than the North Site, as evidenced by the non-overlapping confidence intervals for any fixed coverage up to about 93% in Figure 5. This implies that, if we compare species richness for sample coverage up to 93%, the data do provide sufficient information to conclude that the South Site is significantly richer in species. Unlike the sample-sized-based standardization, in which sample size is determined by investigators, the coverage-based standardization compares equal population fractions of each assemblage. The population fraction is an assemblage-level characteristic that can be reliably estimated from data.

As demonstrated in the above-described example, the two R packages (SpadeR and iNEXT) supply useful information for both asymptotic and non-asymptotic analyses. These methods efficiently use all available data to make robust and meaningful comparisons of species richness between assemblages for a wide range of sample sizes/completeness.

7.4. Faith's PD estimation (Table 5)

Without loss of generality, we select the time depth at 82.9 Myr (the age of the root of the phylogenetic tree connecting the observed 41 species) as our temporal perspective for our phylogenetic diversity estimation in this sub-section and for rarefaction/extrapolation in the next sub-section. Although the root of the observed species varies with sampling data, we can easily transform all our estimates to those for a new reference point that is ancestral to all species; see Chao et al. (2015) for transformations.

In the North Site (27 species in 12 sampling units), the observed *PD* (PD_{obs}) is 1222.10 Myr. The total branch lengths for uniques in the sample branch/node incidence frequencies is calculated as $R_1 = 376.5$ Myr, and for duplicates is $R_2 = 153.5$ Myr. These two statistics and the two counts ($Q_1^* = 9$, $Q_2^* = 6$) in the observed tree produce (by Eq. 11c) an estimate of the undetected *PD* as $\hat{R}_0 = 423.3$ Myr, leading to a Chao2-*PD* estimate of the true *PD* of $\widehat{PD}_{Chao2} = PD_{obs} + \hat{R}_0 = 1645.4$, with an estimated s.e. of 465.81 and 95% confidence interval of (1296.0, 3647.9), based on a bootstrap method using 200 replications and a log-transformation.

In the South Site (38 species in 17 sampling units), the observed *PD* (PD_{obs}) is 1416.0 Myr. The corresponding statistics are $R_1 = 376.8$ Myr, $R_2 = 229.5$ Myr, $Q_1^* = 13$ and $Q_2^* = 10$. These yield an estimate of the undetected *PD* as $\hat{R}_0 = 291.2$ Myr, leading to a Chao2-*PD* estimate of the true *PD* of $\widehat{PD}_{Chao2} = PD_{obs} + \hat{R}_0 = 1707.2$, with an estimated s.e. of 206.45 and 95% confidence interval of (1499.4, 2433.1). Thus, signifi-

Table 5: *A summary of phylogenetic data and PD estimation based on the incidence frequency counts data (in Table 4) and the phylogenetic tree (in Figure 2) for bird species in two sites (the South Site and the North Site in Australian Barrington Tops National Park); see Chao et al. (2015). All calculations are based on a reference time point of 82.9 Myr, the age of the root of the phylogenetic tree connecting the observed 41 species.*

Site	Q_1^*	Q_2^*	R_1	R_2	Observed *PD*	Undetected *PD*	Chao2-*PD*	s.e. of Chao2-*PD*	95% conf. interval
North	9	6	376.5	153.5	1222.1	423.3	1645.4	465.81	(1296.0, 3647.9)
South	13	10	376.8	229.5	1416.0	291.2	1707.2	206.45	(1499.4, 2433.1)

cant difference in *PD* between the two sites cannot be guaranteed due to the overlapping confidence intervals.

7.5. Rarefaction and extrapolation of PD (Figures 6 and 7)

The two types of rarefaction and extrapolation curves, along with the sample completeness curves, can be obtained using the online software PhD (Phylogenetic Diversity), available from the website given in Section 7.2. The sample-size-based and coverage-based sampling curves are shown respectively in Figures 6 and 7. These two curves are plotted in the same manner as those for species richness in Section 7.3; the only difference lies in that species richness estimates were replaced by *PD* estimates (all *PD*

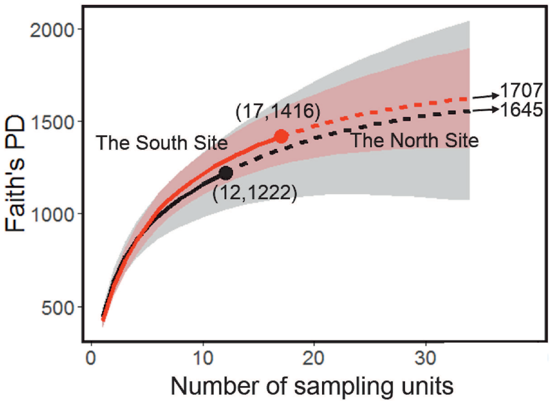


Figure 6: *Comparison of sample-size-based rarefaction (solid lines) and extrapolation (dotted curves) of Faith's PD, up to 34 sampling units for Australian bird data in two sites (the South Site and the North Site in Barrington Tops National Park); see Chao et al. (2015). The fixed time depth is 82.9 Myr (the age of the root of the observed tree.) Observed (reference) samples are denoted by solid dots. The 95% confidence intervals (shaded areas) are obtained by a bootstrap method based on 200 replications. The numbers in parentheses are the number of sampling units and the observed PD for each reference sample. The estimated asymptote of PD (Eq. 11c) for each curve is shown after an arrow sign.*

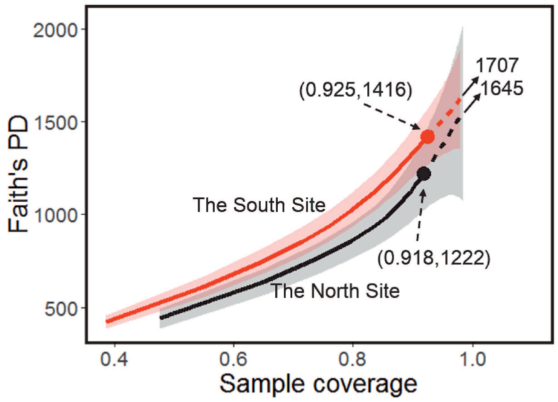


Figure 7: *(a) Comparison of the coverage-based rarefaction (solid lines) and extrapolation (dotted curves) of Faith's PD, up to the coverage 98.5% for the North Site and 98.1% for the South Site for Australian bird data in Barrington Tops National Park (Chao et al., 2015). The fixed time depth is 82.9 Myr (the age of the root of the observed tree.) Observed (reference) samples are denoted by solid dots. The 95% confidence intervals (shaded areas) are obtained by a bootstrap method based on 200 replications. The numbers in parentheses are the estimated sample coverage and the observed PD for each reference sample. The estimated asymptote of PD (Eq. 11c) for each curve is shown after an arrow sign.*

formulas for rarefied and extrapolated samples are provided in the second column of Table 2). The sample completeness curve is identical to that in Figure 4.

We first compare the integrated sample-size-based rarefaction and extrapolation curves for *PD* along with 95% confidence intervals (based on a bootstrap method of 200 replications) up to 34 sampling units. The estimated *PD* and confidence intervals then can be compared across sites for any sample size less than the size of 34. Across this range of sample size, Figure 6 reveals that the South Site has higher *PD* estimate than that of the North Site, but the two confidence intervals overlap and thus data do not provide evidence to support significant difference.

In Figure 7, we compare the corresponding coverage-based rarefaction and extrapolation curves for *PD* with 95% confidence intervals up to the coverage of 98.5% (for the North Site) and 98.1% (for the South Site). Although the estimated *PD* for the South Site still consistently lies above that for the North Site for any standardized sample coverage, the two confidence intervals overlap and thus significant difference cannot be concluded. Chao et al. (2015) analyzed the same data set but based on species abundance data. Although the two types of data yield generally consistent patterns for rarefaction and extrapolation curves, they found that species abundance data show that the PD in the South Site is significantly higher than that in the North-site for any standardized sample coverage less than 90%; see Chao et al. (2015) for analyses based on abundance data.

8. Conclusion and discussion

We have reviewed Chao's (1987) inequality and the associated Chao2 estimator (Eq. 2b) of species richness for multiple incidence data. Using an incidence-data-based generalization of the Good-Turing frequency formula, we have demonstrated that the Chao2 estimator is an unbiased point estimator as long as very rare/infrequent species (specifically, undetected species and unique species in the data) have approximately the same detection probabilities in any sampling unit; the other species (those detected in two or more sampling units) can be highly heterogeneous without affecting the estimator. On the other hand, if very rare/infrequent species are heterogeneous and the sample size is not sufficiently large, then the data do not contain sufficient information to accurately estimate species richness, and the Chao2 formula provides a universal nonparametric lower bound. We have also reviewed the work of Chao et al. (2009) on a related sampling issue, i.e. how many additional sampling units are needed to detect any arbitrary proportion (including 100%) of the Chao2 estimate. Higher-order incidence frequency counts can be also used to construct a series of Chao2-type lower bounds, as derived by Chiu et al. (2014) in Eq. (6d), and by Puig and Kokonendji (2017) in Eq. (6e).

We have also reviewed subsequent developments, including species richness estimators under sampling without replacement, specifically the Chao2-type species richness estimator under sampling without replacement is shown in Eq. (9c). When there are multiple assemblages, the Chao2-shared estimator (Eq. 10d) can be used to infer shared species richness. We also described the Chao2-*PD* estimator (Eq. 11c), which estimates the true *PD* for the phylogenetic tree spanned by all species in the focal assemblage. Similarly, for phylogenetic diversity, the Chao2-*PD* estimator is nearly unbiased if the detection probabilities of rare/infrequent nodes/branches are approximately homogeneous, even if other nodes/branches are heterogeneously detectable. These estimates can be computed from online software SpadeR, iNEXT, and PhD. We have illustrated the use of the software for a real data set in Section 7.

When rare/infrequent species or nodes are highly heterogeneous in their detection probabilities, such as in microbial assemblages or DNA sequencing data, all estimators derived in this paper underestimate the true diversities and can be regarded only as lower bounds. In such cases, a non-asymptotic approach via sample-size-based and coverage-based rarefaction and extrapolation on the basis of standardized sample size or sample completeness (as measured by sample coverage) is recommended. This non-asymptotic approach facilitates fair comparison of diversities (Sections 3 and Section 6.3) for equally-large or equally-complete samples across multiple assemblages. See the example data analysis for rarefaction/extrapolation curves (Figures 3–5 for species richness, and Figures 6 and 7 for *PD*).

None of the diversity measures discussed in this paper (species richness, shared species richness, and Faith's *PD*) considers species abundances. Hill (1973) integrated species richness and species relative abundances into a class of diversity measures later called *Hill numbers*, which include species richness for the diversity order zero. Hill numbers (or the effective number of species) have been increasingly used to quantify the species/taxonomic diversity of assemblages because they represent an intuitive and statistically rigorous alternative to other diversity indices. Hill numbers are parameterized by a diversity order q, which determines the measures' sensitivity to species relative abundances. Hill numbers include the three most widely used species diversity measures as special cases: species richness ($q = 0$), Shannon diversity ($q = 1$), and Simpson diversity ($q = 2$). Like species richness, a Hill number of any order q is dependent on sample size and sample completeness, and thus standardization is needed. The sample-size-based and coverage-based integration of rarefaction (interpolation) and extrapolation (prediction) of Hill numbers represent a unified standardization method for quantifying and comparing species diversity across multiple assemblages; see Chao et al. (2014b) for rarefaction and extrapolation methods based on Hill numbers.

Chao et al. (2010) extended Hill numbers to a class of phylogenetic diversity measures. This class of phylogenetic measures can be regarded as a generalization of Faith's *PD* to incorporate species abundances, because it includes Faith's *PD* as the diversity of order zero ($q = 0$). The corresponding sample-size-based and coverage-based integration of rarefaction and extrapolation of this class of phylogenetic diversity measures was recently developed by Hsieh and Chao (2017). In addition to abundances and evolutionary history, species are often described by a set of traits that affect organismal and/or ecosystem functioning. *Functional diversity* quantifies the diversity of species' traits among coexisting species in an assemblage and is regarded as key to understanding ecosystem processes and their response to environmental stress or disturbance (Tilman et al., 1997; Cadotte et al., 2009). The extension of rarefaction and extrapolation to functional diversity is still under development.

Acknowledgements

We thank the Editor (Pere Puig) for inviting us to write this review paper. We also thank the two reviewers for carefully reading an earlier version and providing very thoughtful and helpful comments and suggestions, which greatly improved the paper. AC was supported by the Taiwan Ministry of Science and Technology under Contracts 105-2628-M007-001.

Appendix A: *Species detection/non-detection records in 12 point-counts for the North Site at Barrington Tops National Park, Australia (Chao et al., 2015).*

Species name	Detection/non-detection record in 12 sampling units (point-counts)												Incidence frequency
Acanthiza_lineata	0	0	0	0	0	0	0	0	0	0	0	0	0
Acanthiza_nana	0	0	0	0	0	0	0	0	0	0	0	0	0
Acanthiza_pusilla	1	1	1	1	1	1	1	1	1	1	1	1	12
Acanthorhynchus_tenuirostris	0	0	0	0	0	0	0	0	0	0	0	0	0
Alisterus_scapularis	0	0	1	0	0	0	0	0	1	0	0	0	2
Cacatua_galerita	0	0	0	0	0	0	1	0	0	0	0	0	1
Cacomantis_flabelliformis	0	1	0	0	1	0	1	1	1	0	0	0	5
Calyptorhynchus_funereus	0	0	0	0	0	0	0	0	0	1	0	0	1
Colluricincla_harmonica	1	1	0	0	0	0	0	0	0	0	0	0	2
Cormobates_leucophaea	1	1	1	1	1	1	1	1	0	1	0	0	9
Corvus_coronoides	0	0	0	0	0	0	0	1	0	0	0	0	1
Dacelo_novaeguineae	0	0	0	0	0	0	0	0	0	0	1	0	1
Eopsaltria_australis	1	0	0	1	0	1	1	0	1	0	0	0	5
Gerygone_mouki	1	1	1	1	0	1	0	0	0	0	0	0	5
Leucosarcia_melanoleuca	0	1	0	0	0	0	0	0	0	0	0	0	1
Lichenostomus_chrysops	0	0	0	0	0	0	0	0	0	0	0	0	0
Malurus_cyaneus	0	0	0	0	0	0	0	0	0	0	0	0	0
Malurus_lamberti	0	0	0	0	0	0	0	0	0	0	0	0	0
Manorina_melanophrys	0	0	0	0	0	0	0	0	0	0	0	0	0
Meliphaga_lewinii	1	1	1	1	1	1	0	0	0	0	0	0	6
Menura_novaehollandiae	1	1	1	0	0	0	0	0	1	0	0	1	5
Monarcha_melanopsis	0	0	0	0	1	0	0	0	0	0	0	0	1
Neochmia_temporalis	0	0	0	0	0	0	0	0	0	0	0	0	0
Oriolus_sagittatus	0	0	1	0	0	0	0	0	0	0	0	0	1
Pachycephala_olivacea	0	0	0	0	0	0	0	0	0	0	0	0	0
Pachycephala_pectoralis	1	1	1	1	0	1	0	0	0	0	1	1	7
Pachycephala_rufiventris	0	0	0	0	0	0	0	0	0	0	0	0	0
Pardalotus_punctatus	1	0	1	1	1	0	1	1	0	0	1	0	7
Petroica_rosea	0	0	0	0	0	0	0	0	0	0	1	0	1
Phylidonyris_niger	0	0	0	0	0	0	0	0	0	0	0	0	0
Platycercus_elegans	0	0	1	0	0	0	0	0	0	0	0	0	1
Psophodes_olivaceus	0	0	1	1	0	0	0	1	0	0	0	0	3
Ptilonorhynchus_violaceus	0	0	0	0	1	0	0	0	0	0	0	1	2
Ptiloris_paradiseus	0	0	0	0	0	0	0	0	0	0	0	0	0
Rhipidura_albicollis	1	1	0	1	0	1	1	1	1	1	1	1	10
Rhipidura_rufifrons	0	0	1	0	1	0	0	0	0	1	0	0	3
Sericornis_citreogularis	0	0	0	0	0	0	0	0	0	0	0	0	0
Sericornis_frontalis	1	0	0	1	0	0	0	0	0	0	0	0	2
Strepera_graculina	0	0	1	0	0	0	0	0	0	1	0	1	3
Zoothera_lunulata	0	0	0	0	0	0	0	0	0	0	0	0	0
Zosterops_lateralis	1	1	0	0	1	1	1	0	0	0	0	0	5

Appendix B: *Species detection/non-detection records in 17 point-counts for the South Site at Barrington Tops National Park, Australia (Chao et al., 2015).*

Species name	Detection/non-detection record in 17 sampling units (point-counts)																	Incidence frequency
Acanthiza_lineata	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Acanthiza_nana	0	0	0	1	0	1	1	0	0	1	0	0	0	0	1	0	0	5
Acanthiza_pusilla	0	0	0	0	1	0	0	1	0	0	1	1	1	1	1	1	1	9
Acanthorhynchus_tenuirostris	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	2
Alisterus_scapularis	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Cacatua_galerita	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
Cacomantis_flabelliformis	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	4
Calyptorhynchus_funereus	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Colluricincla_harmonica	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	1	0	4
Cormobates_leucophaea	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	17
Corvus_coronoides	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Dacelo_novaeguineae	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Eopsaltria_australis	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	3
Gerygone_mouki	0	0	0	0	1	1	1	1	0	0	1	0	1	1	0	0	0	7
Leucosarcia_melanoleuca	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Lichenostomus_chrysops	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Malurus_cyaneus	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2
Malurus_lamberti	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	2
Manorina_melanophrys	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Meliphaga_lewinii	0	1	1	1	1	1	1	1	1	0	1	0	0	1	1	0	0	11
Menura_novaehollandiae	0	0	0	1	0	0	0	1	1	0	0	0	0	1	0	0	0	4
Monarcha_melanopsis	0	0	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	5
Neochmia_temporalis	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	2
Oriolus_sagittatus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pachycephala_olivacea	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	2
Pachycephala_pectoralis	0	0	0	0	1	0	1	0	0	1	0	0	1	1	1	1	0	7
Pachycephala_rufiventris	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Pardalotus_punctatus	0	0	1	1	0	1	1	1	1	0	0	0	0	0	1	0	1	8
Petroica_rosea	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
Phylidonyris_niger	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Platycercus_elegans	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2
Psophodes_olivaceus	0	1	0	0	0	0	1	1	0	0	1	1	0	0	0	1	0	6
Ptilonorhynchus_violaceus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
Ptiloris_paradiseus	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	2
Rhipidura_albicollis	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	1	1	11
Rhipidura_rufifrons	0	0	0	1	0	1	1	1	0	1	1	1	0	0	0	0	0	7
Sericornis_citreogularis	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
Sericornis_frontalis	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	3
Strepera_graculina	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	3
Zoothera_lunulata	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
Zosterops_lateralis	0	1	0	0	0	1	0	1	0	1	0	0	0	1	0	0	1	6

References

Böhning, D. and van der Heijden, P.G.M. (2009). A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations. *Annals of Applied Statistics*, 3, 595–610.

Böhning, D., Vidal-Diez, A., Lerdsuwansri, R., Viwatwongkasem, C. and Arnold, M. (2013). A Generalization of Chao's estimator for covariate information. *Biometrics*, 69, 1033–1042.

Cadotte, M.W., Cavender-Bares, J., Tilman, D. and Oakley, T.H. (2009). Using phylogenetic, functional and trait diversity to understand patterns of plant community productivity. *PLoS One*, 4, e5695.

Cardoso, P., Rigal, F., Borges, P.A. and Carvalho J.C. (2014). A new frontier in biodiversity inventory: a proposal for estimators of phylogenetic and functional diversity. *Methods in Ecology and Evolution*, 5, 452–461.

Cavender-Bares, J., Ackerly, D.D. and Kozak, K.H. (2012). Integrating ecology and phylogenetics: the footprint of history in modern-day communities. *Ecology*, 93, S1–S3.

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, 11, 265–270.

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

Chao, A. and Lee, S-M. (1992). Estimating the number of classes via sample coverage. *Journal of American Statistical Association*, 87, 210–217.

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

Chao, A. (2001). An overview of closed capture-recapture models. *Journal of Agricultural, Biological and Environmental Statistics*, 6, 158–175.

Chao, A. (2005). Species estimation and applications. In Balakrishnan, N., C. Read, B, and B. Vidakovic (eds) *Encyclopedia of Statistical Sciences*, 7907–7916. Wiley, New York.

Chao, A., Chazdon, R.L., Colwell, R.K. and Shen, T.-J. (2005). A new statistical approach for assessing compositional similarity based on incidence and abundance data. *Ecology Letters*, 8, 148–159.

Chao, A., Chazdon, R.L., Colwell, R.K. and Shen, T.-J. (2006). Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics*, 62, 361–371.

Chao, A., Colwell, R.K., Lin, C.-W. and Gotelli, N.J. (2009). Sufficient sampling for asymptotic minimum species richness estimators. *Ecology*, 90, 1125–1133.

Chao, A., Chiu, C.-H. and Jost, L. (2010). Phylogenetic diversity measures based on Hill numbers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365, 3599–3609.

Chao, A. and Chiu, C.-H. (2012). Estimation of species richness and shared species richness. In: Balakrishnan, N (ed) *Methods and Applications of Statistics in the Atmospheric and Earth Sciences*, 76–111. Wiley, New York.

Chao, A. and Jost, L. (2012). Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology*, 93, 2533–2547.

Chao, A. and Lin, C.-W. (2012). A nonparametric lower bound for species richness and shared species richness under sampling without replacement. *Biometrics*, 68, 912–921.

Chao, A., Chiu, C.-H. and Jost, L. (2014a). Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through Hill numbers. *Annual Reviews of Ecology, Evolution, and Systematics*, 45, 297–324.

Chao, A., Gotelli, N.J., Hsieh, T., Sander, E.L., Ma, K., Colwell, R.K. and Ellison, A.M. (2014b). Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs*, 84, 45–67.

Chao, A., Chiu, C.-H., Hsieh, T., Davis, T., Nipperess, D.A. and Faith, D.P. (2015). Rarefaction and extrapolation of phylogenetic diversity. *Methods in Ecology and Evolution*, 6, 380–388.

Chao, A. and Chiu, C.-H. (2016). Species richness: estimation and comparison. *Wiley StatsRef: Statistics Reference Online*. 1–26.

Chao, A., Chiu, C.-H., Colwell, R.K., Magnago, L.F.S., Chazdon, R.L. and Gotelli, N.J. (2017). Deciphering the enigma of undetected species, phylogenetic, and functional diversity based on Good-Turing theory. *Ecology*, under revision.

Chazdon, R.L., Colwell, R.K., Denslow, J.S. and Guariguata, M.R. (1998). Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of Northeastern Costa Rica. In: Dallmeier, F., Comiskey, J.A. (eds.). *Forest Biodiversity Research, Monitoring and Modeling: Conceptual Background and Old World Case Studies*. 285–309. Parthenon Publishing, Paris.

Chiarucci, A., Bacaro, G., Rocchini, D. and Fattorini, L. (2008). Discovering and rediscovering the sample-based rarefaction formula in the ecological literature. *Community Ecology*, 9, 121–123.

Chiu, C.H., Wang, Y.T., Walther, B.A. and Chao, A. (2014). An improved nonparametric lower bound of species richness via a modified Good–Turing frequency formula. *Biometrics*, 70, 671–682.

Coleman, B.D., Mares, M.A., Willig, M.R. and Hsieh, Y.H. (1982). Randomness, area, and species richness. *Ecology*, 63, 1121–1133.

Colwell, R.K. and Coddington, J.A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London B - Biological Sciences*, 345, 101–118.

Colwell, R.K., Mao, C.X. and Chang, J. (2004). Interpolating, extrapolating, and comparing incidence-based species accumulation curves. *Ecology*, 85, 2717–2727.

Colwell, R.K., Chao, A., Gotelli, N.J., Lin, S.-Y., Mao, C.X., Chazdon, R.L. and Longino, J.T. (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, 5, 3–21.

Colwell, R.K. (2013). EstimateS: Statistical estimation of species richness and shared species from samples. Version 9 and earlier. User's Guide and application. Published at: `http://purl.oclc.org/estimates`.

Colwell, R.K. and Elsensohn, J.E. (2014). EstimateS turns 20: statistical estimation of species richness and shared species from samples, with non-parametric extrapolation. *Ecography*, 37, 609–613.

Crozier, R.H. (1997). Preserving the information content of species: genetic diversity, phylogeny, and conservation worth. *Annual Review of Ecology and Systematics*, 28, 243–268.

Faith, D.P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*, 61, 1–10.

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

Good, I.J. (1983). *Good Thinking: The Foundations of Probability and Its Applications*. University of Minnesota Press.

Good, I.J. (2000). Turing's anticipation of empirical Bayes in connection with the cryptanalysis of the naval enigma. *Journal of Statistical Computation and Simulation*, 66, 101–111.

Good, I.J. and Toulmin G. (1956). The number of new species and the increase of population coverage when a sample is increased. *Biometrika*, 43, 45–63.

Gotelli, N.J. and Chao, A. (2013). Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. In Levin, S.A. (Ed). *Encyclopedia of Biodiversity*, 2$^{\text{nd}}$ Edition, Vol. 5, 195–211, Waltham, MA: Academic Press.

Gotelli, N.J. and Colwell, R.K. (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, 4, 379–391.

Gotelli, N.J. and Colwell, R.K. (2011). Estimating species richness. In: A. Magurran and B. McGill (eds). *Biological Diversity: Frontiers in Measurement and Assessment*, 39–54. Oxford University Press, Oxford.

Heck, K.L., Jr., van Belle G. and Simberloff, D. (1975). Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology*, 56, 1459–61.

Hill, M. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54, 427–432.

Hsieh, T.C. and Chao, A. (2017). Rarefaction and extrapolation: making fair comparison of abundance-sensitive phylogenetic diversity among multiple assemblages. *Systematic Biology*, 66, 100–111.

Hughes, G. and Madden, L.P. (1993). Using the beta-binomial distribution to discrete aggregated patterns of disease incidence. *Phytopathology*, 83, 759–763.

Hurlbert, S.H. (1971). The nonconcept of species diversity: a critique and alternative parameters. *Ecology*, 52, 577–586.

Jetz, W., Thomas, G.H., Joy, J.B., Hartmann, K. and Mooers, A.O. (2012). The global diversity of birds in space and time. *Nature*, 491, 444–448.

Jost, L., Chao, A. and Chazdon, R. (2011). Compositional similarity and beta diversity. In A. Magurran and B. McGill (eds). *Biological Diversity: Frontiers in Measurement and Assessment*, 66–84. Oxford University Press, Oxford.

Lanumteang, K. and Böhning, D. (2011). An extension of Chao's estimator of population size based on the first three capture frequency counts. *Computational Statistics & Data Analysis*, 7, 2302–2311.

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

Magurran, A.E. and McGill, B.J. (eds) (2011). *Biological Diversity: Frontiers in Measurement and Assessment*. Oxford University Press, Oxford.

Mao, C.X. (2006). Inference on the number of species through geometric lower bounds. *Journal of American Statistical Association*, 101, 1663–1670.

Mao, C.X. (2008). Lower bounds to the population size when capture probabilities vary over individuals. *Australian and New Zealand Journal of Statistics*, 50, 125–134.

Mao, C.X. and Lindsay, B.G. (2007). Estimating the number of classes. *Annals of Statistics*, 35, 917–930.

Mao, C.X., Yang, N. and Zhang, J. (2013). On population size estimators in the Poisson mixture model. *Biometrics*, 69, 758–765.

McGrayne, S.B. (2011). The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy. *Yale University Press*, New Haven, Connecticut.

Pan, H.Y., Chao, A. and Foissner, W. (2009). A non-parametric lower bound for the number of species shared by multiple communities. *Journal of Agricultural, Biological and Environmental Statistics*, 14, 452–468.

Petchey, O.L. and Gaston, K.J. (2002). Functional diversity (FD), species richness and community composition. *Ecology Letters*, 5,402–411.

Puig, P. and Kokonendji, C. (2017). Nonparametric estimation of the number of zeros in truncated count distributions. To appear in the *Scandinavian Journal of Statistics*.

Rivest, L.P. and Baillargeon, S. (2007). Applications and extensions of Chao's moment estimator for the size of a closed population. *Biometrics*, 63, 999–1006.

Sanders, H.L. (1968). Marine benthic diversity: a comparative study. *American Naturalist*, 102, 243–282.

Shinozaki, K. (1963). Notes on the species-area curve, *10th Annual Meeting of the Ecological Society of Japan* (Abstract), p. 5.

Shiyomi, M., Takahashi, S. and Yoshimura, J. (2000). A measure for spatial heterogeneity of a grassland vegetation based on the beta-binomial distribution. *Journal of Vegetation Science*, 11, 627–632.

Simberloff, D. (1972). Properties of the rarefaction diversity measurement. *American Naturalist*, 106, 414–418.

Skellam, J.G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society B*, 10, 257–261.

Tilman, D., Knops, J., Wedin, D., Reich, P., Ritchie, M. and Siemann, E. (1997). The influence of functional diversity and composition on ecosystem processes. *Science*, 277, 1300–1302.

Warwick, R.M. and Clarke, K.R. (1995). New 'biodiversity' measures reveal a decrease in taxonomic distinctness with increasing stress. *Marine Ecology Progress Series*, 129, 301–305.

Webb, C.O. and Donoghue, M.J. (2005). Phylomatic: tree assembly for applied phylogenetics. *Molecular Ecology Notes*, 5, 181–183.