Unveiling the species-rank abundance distribution by generalizing the Good-Turing sample coverage theory

ANNE CHAO,1,6 T. C. HSIEH,1 ROBIN L. CHAZDON,2,3 ROBERT K. COLWELL,2,4 AND NICHOLAS J. GOTELLI5

1Institute of Statistics, National Tsing Hua University, Hsin-Chu 30043 Taiwan
2Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut 06269 USA
3Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, Colorado 80309 USA
4University of Colorado Museum of Natural History, Boulder, Colorado 80309 USA
5Department of Biology, University of Vermont, Burlington, Vermont 05405 USA

Abstract. Based on a sample of individuals, we focus on inferring the vector of species relative abundance of an entire assemblage and propose a novel estimator of the complete species-rank abundance distribution (RAD). Nearly all previous estimators of the RAD use the conventional “plug-in” estimator \( \hat{p}_i \) (sample relative abundance) of the true relative abundance \( p_i \) of species \( i \). Because most biodiversity samples are incomplete, the plug-in estimators are applied only to the subset of species that are detected in the sample. Using the concept of sample coverage and its generalization, we propose a new statistical framework to estimate the complete RAD by separately adjusting the sample relative abundances for the set of species detected in the sample and estimating the relative abundances for the set of species undetected in the sample but inferred to be present in the assemblage. We first show that \( \hat{p}_i \) is a positively biased estimator of \( p_i \) for species detected in the sample, and that the degree of bias increases with increasing relative rarity of each species. We next derive a method to adjust the sample relative abundance to reduce the positive bias inherent in \( \hat{p}_i \). The adjustment method provides a nonparametric resolution to the longstanding challenge of characterizing the relationship between the true relative abundance in the entire assemblage and the observed relative abundance in a sample. Finally, we propose a method to estimate the true relative abundances of the undetected species based on a lower bound of the number of undetected species. We then combine the adjusted RAD for the detected species and the estimated RAD for the undetected species to obtain the complete RAD estimator. Simulation results show that the proposed RAD curve can unveil the true RAD and is more accurate than the empirical RAD. We also extend our method to incidence data. Our formulas and estimators are illustrated using empirical data sets from surveys of forest spiders (for abundance data) and soil ciliates (for incidence data). The proposed RAD estimator is also applicable to estimating various diversity measures and should be widely useful to analyses of biodiversity and community structure.

Key words: Good-Turing theory; relative abundance; sample coverage; species abundance distribution (SAD); species-rank abundance distribution (RAD).

INTRODUCTION

Most plant and animal assemblages are characterized by a few common species and many uncommon or rare species. A major research aim of ecology is to understand the mechanisms and processes that generate and shape the differences among species abundances (Whittaker 1965, 1970, 1972; see McGill et al. 2007 for a review). A broad array of conceptual and methodological frameworks has been proposed to model and interpret species abundance patterns among assemblages. These previous approaches encompass a wide range of biological and statistical models, from classic analyses of the log series (Fisher et al. 1943), log-normal distribution (Preston 1948), and broken-stick distribution (MacArthur 1957, 1960) to more recent treatments of mechanistic neutral (Caswell 1976, Hubbell 2001) and niche-partitioning (Sugihara 1980, Tokeshi 1990) models; see Magurran (2004) and Magurran and McGill (2011) for overviews.

In this study, we mainly focus on inferring the relative abundance or frequency of every species in an entire focal assemblage, including species undetected by sampling. Based on a sample of \( n \) individuals, ecologists often use the conventional “plug-in” estimator \( \hat{p}_i = X_i/n \) (sample relative abundance/frequency) to estimate the true relative abundance \( p_i \) or probability of species \( i \), where \( X_i \) is the number of individuals observed of species \( i \) in the sample. These sample relative abundances have routinely been used to compute species diversity and evenness measures.
We provide a method to reduce this inherent positive bias. The empirical RAD curve depicts a so-called Whittaker (1965) plot: the sample relative abundance on the y-axis (often with a $\log_{10}$-transformation to accommodate several orders of magnitude), based on a sample of species abundances from an assemblage, we propose a new statistical framework for inferring the SAD/RAD of the entire assemblage. We focus on the RAD estimation because the RAD conveys the same information as the SAD, and the RAD can be used visually to demonstrate the advantages of our approach and to reveal the novelty of our method.

Beginning with seminal work by R. A. Fisher and F. W. Preston in the 1940s, ecologists have fit various statistical models to species or species-rank abundance data; see Magurran (2004) for a review. These distribution-fitting approaches to estimating the complete RAD are entirely dependent on the use of the plug-in estimator for detected species. This approach seems natural and intuitive, because the sample relative abundance is considered to be an unbiased estimator of the true species relative abundance under popular sampling models (Lehmann and Casella 1998). As we explain by simple examples and statistical theory, “unbiasedness” can be achieved only by averaging out all possible species occurrences, including both nonzero occurrences (which are detected in the sample) and zero occurrences (which are not). In nearly all practical applications, however, data consist of the detected species only. The undetected species cannot be included in the data because we do not know whether or not the focal assemblage includes any unobserved species.

This study first addresses the following questions: given the detection of a species in a sample, is its sample relative abundance an unbiased estimator of that species’ true relative abundance? If not, can the bias be reduced or eliminated? These questions are related to a longstanding challenge in community ecology of characterizing the relationship between the SAD in the entire assemblage and the observed SAD in a sample. Most previous approaches (e.g., Dewdney 2000, Green and Plotkin 2007) are based on a parametric assumption about the SAD of the entire assemblage. In this study, we provide a simple and transparent nonparametric relationship. For any species detected in the sample, we demonstrate that the plug-in estimator is a positively biased estimator of the true relative abundance of the species when the sample is not complete. We provide a method to reduce this inherent positive bias.

The next question this study addresses is, without assuming a particular statistical distribution for the underlying SAD/RAD, is it feasible to estimate the relative abundances of the undetected species? In Preston’s (1948) pioneering work, a log-normal model was used to estimate the portion of the assemblage behind a lower limit of observed abundance that he called the “veil line.” The fitted log-normal distribution is used to push back the veil line to estimate the number and relative proportions of the undetected species. But Preston’s analysis depends on the restrictive assumption of a known log-normal model. In different contexts, Gotelli et al. (2010) and Chazdon et al. (2011) addressed this problem in a nonparametric way, but it has not previously been applied to the estimation of the complete RAD.

Here, we describe a general method for estimating the RAD for both detected and undetected species to address these questions. Our method is based on the Good-Turing sample coverage theory and a generalization of that theory that is derived for the first time in this study. The basic theory was originally developed by A. Turing and I. J. Good for their famous cryptographic analyses during World War II. Turing never published this theory, but gave permission to Good to publish it (Good 1953, 2000). Good and Turing discovered that the total probabilities (total true relative abundances) for those species detected in a sample (sample coverage) can be very accurately estimated based only on the sample data themselves. This result implies that the complement of sample coverage (the total probabilities for those species undetected in the sample; coverage deficit sensu Chao and Jost [2012]) can also be very accurately estimated. However, as we will show, this information, although essential, is not in itself sufficient to properly adjust for the biases caused by using the plug-in estimator $\hat{p}_i$ of species relative abundance, nor is it sufficient to accurately estimate the relative abundances for undetected species. We generalize the Good-Turing sample coverage theory to show that there are other aspects of undetected species that we can estimate accurately, and that these measures of information are required to construct a complete RAD.

We separately estimate the RAD for species detected and undetected in a sample. Based on the Good-Turing sample coverage theory and its generalization, we show how to adjust the sample relative abundance of each detected species to better estimate its true relative abundance. Using an estimate of the number of undetected species in the sample (the Chao1 estimator; Chao 1984), we assume that the functional form of the relative abundances of undetected species follows a simple geometric series model (although any other models or distributions could be used instead) and derive an estimated RAD for undetected species. We then combine the adjusted relative abundances for detected species and the estimated part for undetected species to obtain an estimator of the complete RAD (or SAD). Using simulations, we compare the empirical RAD based on $\hat{p}_i$ and the proposed, estimated RAD.
May 2015 1191

SPECIES-RANK ABUNDANCE DISTRIBUTION

[66x710]replacement. Let $X_i$ be the random sample of $n$ individuals.

The probability that any species is detected is $p$. The expected relative abundance of any species

is given by $E[X_i] = np$. The distribution of $X_i$ is a multinomial distribution

with parameters $(p_1, p_2, \ldots, p_S)$, where $\sum_{i=1}^{S} p_i = 1$. Here, $p_i$ can also be interpreted as the probability that any individual is classified to the $i$th species. Assume a random sample of $n$ individuals is selected with replacement. Let $X_i$ denote the sample abundance of the $i$th species in the sample, $i = 1, 2, \ldots, S$. Then $(X_1, X_2, \ldots, X_n)$ is a multinomial distribution with parameters $(p_1, p_2, \ldots, p_S)$. A total of 10 000 samples of size 100 were generated from the assemblage. Among the

10 000 samples, the species sample relative abundances for the first three samples and the last sample are shown. A blank cell means that a species was not detected in that sample. For each particular species, the averages of sample relative abundances over 10 000 samples are shown as the unconditional average, i.e., those samples in which that species was not detected (as shown by a blank, for which the species’ sample relative abundance is thus simply 0) are also counted in the divisor. For each species, only the divisor differs between the two averages.

Most biological survey data can be classified as abundance data (in which individuals are randomly selected) or incidence data (in which sampling units are randomly selected). For the latter, the sampling unit is often a trap, net, quadrat, plot, or timed survey. For incidence data, the abundance of each species is not recorded; only its detection or non-detection in each sampling unit. Although our study deals primarily with abundance data, we briefly discuss parallel derivations that extend our approach to incidence data.

We illustrate the application of our estimators to an empirical data set of pitfall trap catches of temperate forest spiders for abundance data (Sackett et al. 2011), and a data set of soil ciliates for incidence data based on soil samples (Foissner et al. 2002). The formulas for our estimated RAD are relatively simple to calculate and should improve estimation for a variety of ecological questions in which an estimator of the true RAD is desired. We discuss the potential application of our method to the estimation of various diversity measures derived from the RAD and the assessment of sampling errors of complicated estimators.

**Problems with Sample Relative Abundances for Detected Species**

Assume that there are $S$ species in the assemblage and that the true species relative abundances or probabilities are $(p_1, p_2, \ldots, p_S)$, $\sum_{i=1}^{S} p_i = 1$. Here, $p_i$ can also be interpreted as the probability that any individual is classified to the $i$th species. Assume a random sample of $n$ individuals is selected with replacement. Let $X_i$ denote the sample abundance of the $i$th species in the sample, $i = 1, 2, \ldots, S$. Then $(X_1, X_2, \ldots, X_S)$ is a multinomial distribution with parameters $(p_1, p_2, \ldots, p_S)$, where $\sum_{i=1}^{S} X_i = n$. Only those species with abundance $X \geq 1$ are detected in sample; those species with abundance $X = 0$ are undetected in sample and are therefore not included in the data.

We use a simple example to explain the problem with the familiar plug-in estimator of relative abundances. Assume that an assemblage consists of 10 species labeled A, B, . . ., I, J, as in Table 1, with $p = 0.3, 0.1, 0.4, 0.03, 0.05, 0.065, 0.025, 0.015, 0.010,$ and 0.005, respectively (Table 1). Some species are common and some are rare. Assume we take a random sample of 100 individuals, with replacement, from this assemblage. The expected abundances for the 10 species would be 30, 10, 40, 3, 5, 6.5, 2.5, 1.5, 1.0, and 0.5, respectively. However, some of the species with small expected abundances will likely be undetected in any particular sample. We generate 10 000 samples, each with sample size 100. Of the 10 000 samples, we illustrate in Table 1 the sample relative abundances for the first three samples and the last sample. Note that in each sample, some species are not detected. For example, in the first sample, species G and I are not detected (and are thus indicated as blank in Table 1).

For each species, we can calculate two types of averages or expectations for the sample relative abundance: the conditional (on detection) average and the unconditional average. The unconditional average is obtained by averaging over all 10 000 samples, including both detected and undetected species in the calculation: if a species occurs in a particular sample, the sample relative abundance is used in computing the average; if a species does not occur in a particular sample, its estimated relative abundance is 0. The divisor for this

<table>
<thead>
<tr>
<th>Species ID</th>
<th>Sample relative abundances with sample size $n = 100$</th>
<th>Average</th>
<th>True relative abundance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>A</td>
<td>0.3</td>
<td>0.41</td>
<td>0.29</td>
</tr>
<tr>
<td>B</td>
<td>0.08</td>
<td>0.11</td>
<td>0.13</td>
</tr>
<tr>
<td>C</td>
<td>0.42</td>
<td>0.34</td>
<td>0.38</td>
</tr>
<tr>
<td>D</td>
<td>0.02</td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>E</td>
<td>0.1</td>
<td>0.02</td>
<td>0.06</td>
</tr>
<tr>
<td>F</td>
<td>0.06</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>G</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>H</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>I</td>
<td>0.02</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>J</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes:** A total of 10 000 samples of size 100 were generated from the assemblage. Among the 10 000 samples, the species sample relative abundances for the first three samples and the last sample are shown. A blank cell means that a species was not detected in that sample. For each particular species, the averages of sample relative abundances over 10 000 samples are shown as the unconditional average, i.e., those samples in which that species was not detected (as shown by a blank, for which the species’ sample relative abundance is thus simply 0) are also counted in the divisor. For each species, only the divisor differs between the two averages.
unconditional average is always 10,000 for each species. However, in practice, in a single sample, we can only obtain sample relative abundances conditional on those species that are detected in that sample. This conditional average is obtained by averaging over only those samples in which that species is detected. Therefore, the divisor for the conditional average for rare species may be less than 10,000, because not all samples are included in the divisor.

Table 1 reveals that, for all species, the unconditional averages are very close to the true relative abundances. For abundant species, which are likely to be observed in nearly all samples (such as species A–F), the conditional and unconditional averages are almost identical. For rare species, however, which will be found in few or no samples, the conditional averages are consistently higher than the true relative abundances. For rare species G–J in Table 1, with relative abundances 0.025, 0.015, 0.010, and 0.005, the corresponding conditional averages are 0.0269, 0.0194, 0.0157, and 0.0127. These results imply that the sample relative abundance for any detected species overestimates its true value. The level of overestimation is not uniform, but scales inversely with abundance: estimates for rare detected species are more severely biased than estimates for common detected species. Sample relative abundances do not need to be adjusted for abundant species, but sample relative abundances for rare species have substantial positive relative biases and should be properly adjusted.

Statistical explanation

The level of overestimation of the sample relative abundance for any detected species can be seen by examining the following theoretical conditional average or statistical expectation (Chazdon et al. 2011: Appendix C):

\[
E(\hat{\pi}_i | X_i > 0) = E\left(\frac{X_i}{n} | X_i > 0\right) = \frac{\pi_i}{1 - (1 - \pi_i)^n}
\]

which is the expected proportion of individuals in a sample of size \(n\) that represent species \(i\), given that species \(i\) has been detected in sample. The denominator \(1 - (1 - \pi_i)^n\) in Eq. 1 is \(P(X_i > 0)\), the probability of detection of species \(i\) in the sample. Because this denominator is always less than 1, Eq. 1 proves that the sample relative abundance for any detected species consistently overestimates the true probability \(\pi_i\).

When \(\pi_i\) is relatively large, the denominator \(1 - (1 - \pi_i)^n\) tends to 1, because the species is sufficiently abundant that it would be observed in any sample. Therefore, for relatively common species, the sample relative abundance \(X_i/n\) works well as an estimate of \(\pi_i\) and almost no adjustment is required. In contrast, when \(\pi_i\) is very small, the denominator \(1 - (1 - \pi_i)^n\) is much less than 1, which generates a substantial bias. For example, with a sample size of 100, the probability of detecting the rarest species in Table 1 is \(1 - (1 - \pi_{10})^n = 1 - (1 - 0.005)^{100} = 0.394.\) The conditional average is \(0.005/0.394 = 0.0127,\) more than double the correct value (0.005). This theoretical value of 0.0127 is further confirmed by our simulation result (Table 1).

Now we can connect the foregoing discussion to the classic unbiasedness of sample relative abundance in the following sense. For any species \(i\), it will be detected in the sample with probability \(P(X_i > 0) = 1 - (1 - \pi_i)^n\) or it will be undetected with probability \(P(X_i = 0) = (1 - \pi_i)^n\).

Then on average, we have

\[
E\left(\frac{X_i}{n} \mid X_i > 0\right) = E\left(\frac{X_i}{n} \mid X_i > 0\right) P(X_i > 0) + E\left(\frac{X_i}{n} \mid X_i = 0\right) P(X_i = 0) = \frac{\pi_i}{1 - (1 - \pi_i)^n} \times [1 - (1 - \pi_i)^n] + 0 \times (1 - \pi_i)^n = \pi_i.
\]

This (unconditional) expectation, which is valid for all species in the complete assemblage, considers both detection and non-detection and implies unbiasedness.

SIMULATION PART I: THE EMPIRICAL RAD

We use a suite of simple simulations to illustrate the undersampling bias with the empirical RAD when sample size is not large enough to detect all species. We simulated data from two theoretical abundance distributions (the Zipf-Mandelbrot model and the lognormal model) and treated four large empirical diversity surveys as the complete assemblages. For the latter cases, the species-rank abundance distribution from each survey was assumed to be the “true” complete distribution; this true RAD was then compared with the empirical RADs obtained from simulated samples of several sample sizes. Here we report in detail only the simulation results for the Zipf-Mandelbrot model for illustration. See Appendix A for simulation results of other scenarios.

In the Zipf-Mandelbrot model, we fix the number of species at 200 and the true relative abundance takes the form \(p_i = c/(2 + i)\), \(i = 1, 2, \ldots, 200\), where \(c\) is a normalized constant such that the sum of the relative abundances is 1. In Fig. 1a, we compare the true complete RAD of the entire assemblage (light blue line) and the empirical RAD based on 200 simulated data sets of sample sizes 200, 400, and 800 (200 superimposed dark blue lines, each line corresponding to an empirical RAD for each generated data set). When the sample sizes are not large enough \((n = 200\) and 400\) to detect all species, only about half of the complete RAD can be revealed empirically from the simulated data. Even for a large sample size \((n = 800)\), most of the empirical RAD curves still cannot unveil the complete “tail” of the true RAD. Although the observed species (say there are \(K\) of them) in a sample may not correspond to the first \(K\) species in the true RAD, most of the empirical RAD curves lie above the true
RAD, signifying that the positive bias is associated with the empirical RAD for the detected species, as predicted from our theory (Eq. 1) and shown by an example (Table 1).

**GOOD-TURING SAMPLE COVERAGE THEORY AND A GENERALIZATION**

**Sample coverage and coverage deficit**

Let $f_k$ be the number of species represented by exactly $k$ individuals in the sample, $k = 0, 1, \ldots, n$; we refer to $f_k$ as the abundance frequency counts. In particular, $f_1$ is the number of species represented by exactly one individual (singletons) in the sample, and $f_2$ is the number of species represented by exactly two individuals (doubletons). The unobservable frequency $f_0$ denotes the number of species present in the entire assemblage but not detected in the sample. Good and Turing discovered a surprisingly simple estimator for the sample coverage ($C$, a measure of sample completeness, as defined in Introduction) that is a simple function of the number of singletons and the sample size if $f_1 > 0$

$$C = 1 - \frac{f_1}{n}.$$  

(2a)

When $f_1, f_2 > 0$, an improved Turing’s coverage estimator (Chao and Jost 2012) is

$$\hat{C} = 1 - \frac{f_1}{n} \left[ \frac{(n-1)f_1}{(n-1)f_1 + 2f_2} \right].$$  

(2b)

Here, the leading superscript 1 in $\hat{C}$ refers to the first-

---

**FIG. 1.** (a) Comparison of the true species-rank abundance distribution (RAD) of the complete assemblage (light blue line) and the empirical RAD curves (superimposed dark blue lines with 200 replications). (b) Comparison of the true RAD, empirical RAD, and adjusted RAD curves for detected species only (superimposed green lines with 200 replications). (c) Comparison of the true RAD, empirical RAD, and estimated RAD curves for both detected and undetected species (superimposed red lines with 200 replications). For each of the sample sizes 200 (left panels), 400 (middle panels), and 800 (right panels), 200 data sets were generated from the Zipf-Mandelbrot model; thus there are 200 estimated RADs (200 dark blue lines, 200 green lines, and 200 red lines). Note that the $x$-axis is the species list, ranked from most to least abundant, and the $y$-axis (relative abundance) is displayed on a log$_{10}$ scale.
order sample coverage of our generalization (A general-
ization). This improved coverage estimator incorpo-
rates information about the doubletons. It is improved
in the sense that this coverage estimator generally has
smaller mean squared error than Good-Turing’s esti-
mator. In the following derivation, we will adopt this
more accurate estimator. Subtracting the sample cover-
age estimator from unity gives the estimator of coverage
deficit

\[ 1 - \hat{C} = \frac{f_1}{n} \left[ \frac{(n-1)f_1}{(n-1)f_1 + 2f_2} \right]. \]  

(2c)

A tiny percentage of coverage can nevertheless contain
a very large number of rare species. The estimated
coverage deficit is not an estimate of the number or propor-
tion of undetected species, but rather it is an estimate of the proportion of the total number of individuals in the
assemblage that belong to the undetected species. For
this reason, extremely rare, undetected species do not make a significant contribution to that proportion, even
if there are many such species. This distinction intuitively explains why the estimation of species richness in highly
diverse assemblages is so statistically challenging, even
though sample coverage for the same data can be accurately
estimated.

The coverage estimator and its complement make it
possible to adjust the sample relative abundance for
detected species and to infer the relative abundance of
undetected species. This approach allows models with
one parameter, which are useful for assessing sampling
variances in some inference problems (Chao et al. 2013;
Appendix S2; Chao et al. 2014: Appendix G). However,
the one-parameter models are not flexible enough to
provide accurate estimators for the complete RAD. For
this reason, in this study we extend the Good-Turing
concept of coverage for the first time, and develop
improved models for estimating species abundances
beyond the veil line.

A generalization

Lande et al. (2000) commented, “without regard to
the species abundance distribution, the only aspect
of unobserved species that can be accurately extrapo-
lated is their total frequency in a community [i.e., coverage
deficit], using the number of singletons divided by
sample size.” In addition to coverage deficit, however,
there are other aspects of undetected species that we can
measure accurately. To show this, we first generalize
the concept of sample coverage to the rth order sample
coverage \( \hat{C} \) as

\[ \hat{C} = \frac{\sum_{i=1}^{S} p_i^r I(X_i > 0)}{\sum_{i=1}^{S} p_i^r}, \quad r = 1, 2, \ldots, \]

where indicator function \( I(A) = 1 \) if event \( A \) occurs, and

0 otherwise. The coverage \( \hat{C} \) is the fraction of the rth
power of the true relative abundances of those species
detected in sample. For \( r = 1 \), \( \hat{C} \) reduces to Good-
Turing’s sample coverage, and its estimator is given in
Eq. 2b. For \( r = 2 \), \( \hat{2C} \) is the fraction of the squared true
relative abundances of the detected species; it quantifies
the sample completeness for very abundant or dominant
species. When \( f_2, f_1 > 0 \), \( \hat{2C} \) can be accurately estimated
by (see Appendix B for derivation)

\[ 1 - \hat{C} = \frac{2f_2}{\sum_{X_i \geq 2} X_i (X_i - 1) \left[ \frac{(n-2)f_2}{(n-2)f_2 + 3f_3} \right]^2}. \]  

(3a)

We define the rth order coverage deficit as \( \hat{C} = 1 - \hat{C} \). For
\( r = 1 \), \( \hat{1C} \) reduces to the coverage deficit
defined in Chao and Jost (2012), and its estimator is
given in Eq. 2c. For \( r = 2 \), \( \hat{2C} \) can be accurately estimated
by

\[ 1 - \hat{C} = \frac{2f_2}{\sum_{X_i \geq 2} X_i (X_i - 1) \left[ \frac{(n-2)f_2}{(n-2)f_2 + 3f_3} \right]^2}. \]  

(3b)

As we will see, the estimators for the first- and second-
order sample coverages and their deficits make possible
two-parameter models for inferring the complete RAD.
As proved in Appendix B, if the abundance frequency
counts up to \( f_{r+1} \) are all nonzero, then \( \hat{1C}, \hat{2C}, \ldots, \hat{rC} \)
and their deficits can be accurately and efficiently
estimated. Thus, in addition to the coverage deficit, we
have more information (i.e., higher orders of coverage
deficits including \( \hat{1C}, \hat{2C}, \ldots, \hat{rC} \)) about the undetected
species. This information can be used to
help estimate the complete RAD.

Unveiling the Complete RAD

Adjusting the sample relative abundances for
detected species

Based on Eq. 1, we have

\[ p_i = E \left( \frac{X_i}{n} \mid X_i > 0 \right) \left[ 1 - (1 - p_i)^n \right]. \]

If we replace the expected value in this equation with
the observed data, then for \( X_i > 0 \) (i.e., a detected species), we have the following approximation:

\[ p_i \approx \frac{X_i}{n} \left[ 1 - (1 - p_i)^n \right] = \frac{X_i}{n} \left[ 1 - \exp(-np_i) \right]. \]  

(4a)

This formula shows that the approximate adjustment
factor for the sample relative abundance would be
\( 1 - (1 - p_i)^n \approx 1 - \exp(-np_i) \), which depends mainly on the
product of \( n \) and \( p_i \). Note that the (unconditional)
expected abundance of species \( i \) in the sample is \( np_i \), i.e.,
\( E(X_i) = np_i \). However, as we have already argued, the
adjustment factor \( 1 - (1 - p_i)^n \) cannot be estimated
simply by substituting the sample relative abundance \( X_i/n \) for \( p_i \), because the sample relative abundance does not
estimate \( p_i \) well for rare species. Similarly, replacing \( np_i \)
in the adjustment factor \([1 - \exp(-np_i)]\) by the observed abundance \(X_i\) for each individual species \(i\) does not provide a good estimate. Instead, we introduce two parameters, \(\lambda\) and \(\theta\), to the adjustment factor. From Eq. 4a, we assume that parameter \(\lambda > 0\) and parameter \(0 < \theta < 1\), so that for \(X_i > 0\), \(p_i = (X_i/n)(1 - \lambda e^{-\theta X_i})\). Here, parameter \(\theta\) is restricted to be in \([0, 1]\) because \(X_i\) for a detected species overestimates \(np_i\). The special case of \(\theta = 1\) reduces to the one-parameter approach discussed in the following equations: (Eq. 2b) and second-order (Eq. 3a) sample coverage by the following sense: for abundant species, with adjustment scales inversely with the sample abundance \(a\) smaller root mean squared error. The proposed substantial bias inherent in the plug-in estimator and has This is a unified adjustment formula that is valid for all estimated relative abundance of species \(\tilde{X}_i\) with \(\tilde{X}_i > 0\) is 
\[
\tilde{\beta}_i = \frac{X_i}{n} (1 - \hat{\lambda} e^{-\hat{\theta} X_i}).
\] (4d)

This is a unified adjustment formula that is valid for all species abundance distributions. In Appendix A, we show by simulations that the adjusted estimator reduces substantial bias inherent in the plug-in estimator and has a smaller root mean squared error. The proposed adjustment scales inversely with the sample abundance in the following sense: for abundant species, with correspondingly large sample abundance \(X_i\), the adjustment factor \(1 - \lambda e^{(-\hat{\theta} X_i)}\) approaches unity. Thus, virtually no adjustment is needed for abundant species, whereas for rare species, the adjustment factor can be much less than 1. The smaller the abundance \(X_i\), the smaller the adjustment factor and the larger its effect.

Our adjustment formula (Eq. 4d) also provides a simple nonparametric relationship between the plug-in estimator \((X_i/n)\) calculated for a species in a sample and its estimated true relative abundance in the entire assemblage. Note that the formula is a function of sample abundances of all detected species, not merely the sample abundance of species \(i\). This is because, given the sample coverage estimates \((\hat{\lambda}\hat{C}, \hat{\theta}\hat{C})\), other species also carry information about species \(i\) via \(\hat{\lambda}\) and \(\hat{\theta}\), which are functions of all sample frequencies (by Eqs. 4b and 4c). Thus, our adjustment formula "borrows strength" from the observed abundances of other species. We will discuss how to assess the sampling error of the adjusted estimator after we obtain an estimator for the complete RAD.

**Estimating the relative abundances of undetected species**

As discussed, it is difficult to accurately estimate the number of undetected species in an incomplete sample if there are many, almost-undetectable species in a hyper-diverse assemblage. Practically, an accurate lower bound for species richness is preferable to an inaccurate point estimator. A widely used nonparametric lower bound developed by Chao (1984) uses only the information on rare species (numbers of singletons and doubletons) to estimate the number of undetected species in samples, as rare, detected species contain nearly all information about the number of undetected species. This lower bound for the number of undetected is universally valid for any species abundance distribution and has the following form:

\[
\hat{f}_0 = \begin{cases} 
\frac{(n-1)f_2^2}{2f_2} & \text{if } f_2 > 0 \\
\frac{(n-1)f_1(f_1-1)}{2} & \text{if } f_2 = 0.
\end{cases}
\] (5a)

See Chao and Chiu (2012) for a recent review. Because the number of species must be an integer in later derivations, we define \(\hat{f}_0\) hereafter to be the smallest integer that is greater than or equal to the value computed from Eq. 5a. The empirical RAD ignores the tail, which includes at least \(\hat{f}_0\) species. Although this is a lower bound, when sample size is large enough, this lower bound approaches the true number of undetected species. Based on Eq. 5a, we propose a robust method to estimate the species RAD for the undetected species.

We must assume a functional form for the rank abundances of the undetected tail. There are many options for a functional form, and our method is applicable to any functional form. Here, we adopt more flexible, two-parameter models. Because the method is applied to only the undetected tail part of the true RAD, where all relative abundances are low, a simple functional form with estimable parameters is preferable. A natural assumption is that the abundance distribution of the undetected species is a two-parameter geometric series

\[
p_i = a\beta^i, i = 1, 2, \ldots, \hat{f}_0
\] (5b)

where \(a\) is a normalized constant (see Eq. 6a) and \(\beta\) is a positive decay factor. If all relative abundances for undetected species are approximately equal, then the parameter \(\beta\) is close to 1.

Based on the coverage deficits of the first- and second-order (Eqs. 2c and 3b), we have the following two
of 168 individuals. The nonzero abundance frequency mortality by adelgid infestation. In this experimental were cut and the trees left in place to die, to mimic tree treatment, in which bark and cambium of hemlock trees our method, we use the data of the Hemlock Girdled simulation plots can be better understood. To examine the performance of the estimated RAD form of the relative abundances of undetected species. See (Eq. 4d) and the estimated relative abundances for the undetected species, as will be confirmed by simulations in Appendix C. (4) In (1) through (3), we use bar plots for clearer illustration. Conventionally, only line plots as those plotted in Fig. 1 are sufficient for comparison. In Fig. 2d, we provide the line plots for the empirical RAD and the proposed RAD estimator.

In (1), notice from Fig. 2 that, for abundant species, virtually no adjustment is needed, whereas the adjustment for rare species is substantial and that scales inversely with the sample abundance. In (2), our estimated number of undetected species is only a lower bound, implying that there may have been additional undetected species, but they cannot be statistically estimated from our inference, so they are treated as having negligible abundances. See Discussion for further explanation. In this example (Fig. 2c), the estimated relative abundance for the most abundant of the undetected species is slightly larger than the adjusted species relative abundances of the least-abundant detected species (i.e., singletons). Our analysis shows that the empirical RAD curve differs greatly from the proposed RAD curve in the tail distribution. When there are undetected species, as will be confirmed by simulations in Simulation II: The complete estimated RAD, our proposed approach unveils the tail distribution and provides a more complete picture of the true RAD. In Appendix C, as alternatives to the geometric series, we present a Poisson log-normal and a broken-stick model for the relative abundances of the undetected species. See Discussion for more details.

**Example (Abundance Data)**

Sackett et al. (2011) collected species abundance data for samples of spiders from four experimental forest-canopy-manipulation treatments at the Harvard Forest (Massachusetts, USA). The treatments were established to study the long-term consequences of loss of the dominant forest tree, eastern hemlock (Tsuga canadensis), caused by a nonnative insect, the hemlock woolly adelgid (Adelges tsugae; Ellison et al. 2010). To illustrate our method, we use the data of the Hemlock Girdled treatment, in which bark and cambium of hemlock trees were cut and the trees left in place to die, to mimic tree mortality by adelgid infestation. In this experimental treatment, 26 spider species were represented by a total of 168 individuals. The nonzero abundance frequency counts are \( f_1 = 12, f_2 = 4, f_4 = 1, f_6 = 2, f_8 = f_9 = f_15 = f_23 = f_24 = f_46 = 1 \). The first- and second-order sample coverage estimates are respectively 92.89% and 99.77%; the corresponding coverage deficits are thus respectively 7.11% and 0.23%. Our estimation procedure includes the following four steps (see Fig. 2): (1) Construct the adjusted RAD for the detected species. For the 26 detected species, first plot the empirical RAD, as shown in Fig. 2a (white plus gray bars). Then use Eqs. 4b and c to obtain \( \hat{\lambda} = 0.2980 \) and \( \hat{\theta} = 0.1267 \), and substitute these estimates into Eq. 4d to adjust the sample relative abundances for each detected species downward, as shown in Fig. 2a (white bars). (2) Estimate the RAD for undetected species: based on the Chao1 estimator, which uses the observed numbers of singletons and doubletons, estimate the number of undetected species as \( \hat{f}_0 = 18 \) species (SE = 13.4). The undetected species are labeled Undetected.1 to Undetected.18 in Fig. 2b. For the 18 undetected species, use Eqs. 6a and 6b to obtain \( \hat{\lambda} = 0.0045 \) and \( \hat{\beta} = 0.9865 \), then substitute these two estimates into Eq. 6c to estimate their relative abundances, as shown in Fig. 2b. (3) Combine the adjusted RAD for the detected species in (1) and the estimated RAD for the undetected species in (2) to obtain a complete RAD, as shown in Fig. 2c. A full list of the estimated species relative abundances for the complete RAD is given in Appendix C. (4) For comparing the performance of the estimated RAD with the observed RAD, use bar plots for clearer illustration. Conventionally, only line plots as those plotted in Fig. 1 are sufficient for comparison. In Fig. 2d, we provide the line plots for the empirical RAD and the proposed RAD estimator.

In the estimated complete RAD, there are \( S_{obs} + \hat{f}_0 \) species, where \( S_{obs} \) denotes the number of observed species in the sample. This estimated RAD mimics the profile of the complete assemblage. We can thus assess the sampling error of any estimator of a parameter by bootstrapping or resampling the estimated RAD. For example, we can approximate the sampling variance of the adjusted estimator \( \hat{p}_i \) (Eq. 4d) for any particular detected species \( i \). For each bootstrap replication, we generate a random sample of \( n \) individuals from the estimated RAD, with replacement, yielding a new set of species sample abundances (here we retain only those sets in which species \( i \) is detected, because the estimating target is the relative abundance of a detected species). Based on this new set, we then calculate \( \left( \hat{C}, \hat{\beta} \right) \) to obtain new estimates (\( \hat{\lambda}, \hat{\theta} \)). All these new statistics are
then substituted into Eq. 4d to obtain a bootstrap estimate \( \hat{p}_i \). The procedure is replicated to obtain \( B \) bootstrap estimates \( \{\hat{p}_i^1, \hat{p}_i^2, \ldots, \hat{p}_i^B\} \) (\( B = 1000 \) is suggested in confidence interval construction). The bootstrap variance estimator of our estimator in Eq. 4d is the sample variance of these \( B \) estimates. Moreover, the 2.5\% and 97.5\% percentiles of these \( B \) bootstrap estimates can be used to construct a 95\% confidence interval. See Appendix C: Table C1 for the bootstrap SE of the adjusted estimator \( \hat{p}_i \) for each detected species in the spider example.

Similar procedures can be used to derive variance estimators for any other estimators (e.g., estimators of sample coverages and their deficits) and to construct the
associated confidence intervals. For our proposed estimator \( \hat{\rho}_i \) for undetected species (Eq. 6c), however, sampling variance cannot be assessed because the estimated number of undetected species varies with bootstrap samples and the identities of those undetected species are unknown and thus there is no pre-specified target species.

**Simulation Part II: the Estimated RAD**

The adjusted RAD for detected species (Fig. 1b)

For the scenario considered in Fig. 1a, in addition to the true RAD curve (light blue line) and the empirical RADs (dark blue lines), we now superimpose in Fig. 1b the estimated RADs (green lines) for detected species based on 200 data sets of sample sizes 200, 400, and 800. Thus, there are 200 additional superimposed green lines for each sample size. Most of the green lines are below the empirical RAD, showing the reduction of the positive biases associated with sample relative frequencies for detected species.

The complete estimated RAD (Fig. 1c)

In Fig. 1c, we compare the true RAD curve (light blue line), the empirical RADs (dark blue lines), and the estimated complete RADs including both detected and undetected species (red lines) based on 200 data sets of sample sizes 200, 400, and 800. For sample sizes 200 and 400, the improvement with the estimated RAD is clearly seen: the tail of the true RAD can be revealed, although our estimated tail of RADs for a sample size of 200 unavoidably overestimates the true lines to some extent (i.e., data do not provide sufficient information to accurately infer very small relative abundances; see Appendix A: Fig. A2). When sample size is increased to 400, the proposed RAD curves closely trace the RAD of the complete assemblage; for a sample size of \( n = 800 \), all the proposed RAD curves match closely with the true RAD curve.

**Extension to Incidence Data**

Our statistical framework for abundance data can be extended to incidence data by parallel derivations. Here we only outline the extension; all details are provided in Appendix D. Following the notation and terminology used in Colwell et al. (2012) and Chao et al. (2014), we assume that in the focal assemblage there are \( S \) species indexed by \( 1, 2, \ldots, S \). For any sampling unit, assume that the \( i \)th species has its own unique incidence (or occurrence) probability \( \pi_i \) that is constant for any randomly selected sampling unit. The incidence probability \( \pi_i \) is the probability that species \( i \) is detected in a sampling unit. This incidence probability \( \pi_i \) is analogous to \( \rho_i \), but \( \sum_{i=1}^{S} \pi_i \) may be greater than unity.

As with abundance data, we can similarly define the species incidence distribution (SID) and the corresponding species-rank incidence distribution (RID) for the set \( \{\pi_1, \pi_2, \ldots, \pi_S\} \) of the \( S \) species. Our goal here is to estimate the RID based on incidence data of a set of sampling units. Assume that a set of \( T \) sampling units are randomly selected from the study area, with replacement. The underlying data 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. Under our assumption that the probability of detecting species \( i \) in any sampling unit is a constant \( \pi_i, i = 1, 2, \ldots, S \), the variable \( W_{ij} \) for all \( j \) follows a Bernoulli distribution with parameter \( \pi_i = P(W_{ij} = 1), i = 1, 2, \ldots, S \). 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 and is analogous to \( X_i \) in the abundance data. Species present in the assemblage but not detected in any sampling unit yield \( Y_i = 0 \).

Denote the incidence frequency counts by \( \{Q_0, Q_1, \ldots, Q_T\} \), where \( Q_0 \) is the number of species that are detected in exactly \( k \) sampling units in the data, \( k = 0, 1, \ldots, T \). Here \( Q_k \) is analogous to \( f_k \) in the abundance data. 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. Also, \( Q_1 \) 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 only two sampling units).

Define the sample incidence probability of species \( i \) as \( \hat{\pi}_i = Y_i/T \) (the plug-in estimator); the empirical RID is based on \( \hat{\pi}_i \). Since \( Y_i, i = 1, 2, \ldots, S \) follows a binomial distribution with the total number \( T \) and the detection probability \( \pi_i \), a formula parallel to Eq. 1 can be derived

\[
E(\hat{\pi}_i | Y_i > 0) = E\left(\frac{Y_i}{T} | Y_i > 0 \right) = \frac{\pi_i}{1 - (1 - \pi_i)^T}. \tag{7}
\]

We can similarly define the general \( r \)th sample coverage and its deficits for the incidence probabilities \( (\pi_1, \pi_2, \ldots, \pi_S) \) based on the sample species incidence frequencies \( (Y_1, Y_2, \ldots, Y_S) \). Then, derivation steps parallel to those for abundance data lead to the following adjusted incidence probability for a detected species: \( \hat{\pi}_i = (\hat{Y}_i/T)(1 - \hat{\lambda}e^{-\hat{\theta}T}) \) for \( Y_i > 0 \), where \( \hat{\lambda} \) and \( \hat{\theta} \) are solved from two nonlinear equations involving the estimated sample coverage of the first two orders (see Appendix D).

To estimate the RID for undetected species, we first apply the Chao2 estimator (Chao 1987) to obtain an estimated lower bound on the number of undetected species in \( T \) sampling units

\[
\hat{Q}_0 = \begin{cases} 
\frac{(T - 1) Q_1^2}{2Q_2} & \text{if } Q_2 > 0 \\
\frac{(T - 1) Q_1(Q_1 - 1)}{2T} & \text{if } Q_2 = 0.
\end{cases} \tag{8}
\]
Assuming a geometric series for the incidence probabilities for the undetected species, we can obtain the proposed incidence probabilities for the undetected species:

\[ \tilde{p}_i = \hat{a}^i \hat{b}^i, \quad i = 1, 2, \ldots, \hat{Q}_0. \]

Here \(\hat{a}\) and \(\hat{b}\) are solved from two nonlinear equations involving the estimated sample coverage deficits of the first two orders (see Appendix D). Combining the adjusted incidence probabilities for detected species and the estimated incidence probabilities for undetected species, we can construct a complete RID based on incidence data.

**Example (Incidence Data)**

We use soil ciliate data collected by Foissner et al. (2002) to illustrate our approach for incidence data. A total of 51 soil samples were collected in Namibia and the detection or non-detection of soil ciliate species was recorded in each sample. Detailed sampling locations, procedures, and species identifications are described in Foissner et al. (2002). In total, 331 species were detected in 51 soil samples. The first 14 incidence frequency counts \((Q_1 - Q_{14}) = (150, 53, 42, 18, 12, 10, 7, 6, 1, 0, 2, 3, 2)\), and a full list of the data are given in Appendix D. Combining the adjusted incidence probabilities for detected species and the estimated incidence probabilities for undetected species, we can construct a complete RID based on incidence data.

**Discussion**

We have shown that the empirical RAD using species sample abundances works only when all species are detected in a sample. For an undersampled data set with undetected species, the empirical RAD ignores the set of undetected species, and therefore overestimates the true relative abundances of the set of the detected species (Fig. 1a). We have proposed a general framework to estimate the complete RAD from sample data. Our proposed RAD estimator combines the adjusted RAD (Eq. 4d) for the detected species in samples and the estimated RAD (Eq. 6c) for the undetected species. Both parts are based on Good-Turing’s sample coverage theory and its generalization. For any detected species, we have proposed a novel method to adjust its sample relative abundance to reduce its positive bias (Fig. 1b and Appendix A). For the undetected species (which are assumed to have very low relative abundances), we estimate their relative abundances using an estimator of...
the number of undetected species. See Fig. 2 for an illustrative example to describe our procedures. With our approach, the complete RAD is unveiled if sample size is large enough (Fig. 1c; Appendix A: Fig. A1).

In our inference procedure for undetected species in samples, we use a universal lower bound, i.e., the Chao1 estimator for abundance data and the Chao2 estimator for incidence data; see Chao (1984, 1987). Thus, we essentially assume that there might be additional extremely rare species in the assemblage, but they cannot be statistically estimated, so their relative abundances are estimated to be zero. Our estimator of the number of undetected species could also be replaced by any other reasonable estimator. We also assume that the relative abundances for the set of undetected species follow a simple geometric series model. Nevertheless, our model is not restricted to this distribution. The assumption of a geometric series can be replaced by any other appropriate distribution. There are many other choices, including the commonly used broken-stick model and the Poisson log-normal model, among others. Appendix C provides estimation procedures for these two additional models. For illustration, we also fitted these two models to the data analyzed in Example (abundance data). The Poisson log-normal model and the geometric model yield almost-identical RAD curves. We emphasize that we use these models only for modeling the undetected tail distribution; unless the assemblage is poorly sampled, the relative abundances of those undetected species (i.e., in the tail of the estimated RAD) are typically very small. Thus, the choice of the model for estimating the relative abundances of undetected species is a minor issue in our approach.

Ecologists have recognized that, although an accurate species richness estimator remains beyond our reach, one aspect of undetected species (the coverage deficit) can be accurately estimated; see Eq. 2c. We show that there are other aspects of undetected species (e.g., the deficits of the second- and higher-order sample coverage) that we can also accurately estimate using the information on frequency counts. In this study, we used the first- and second-order sample coverage and their deficits to construct two-parameter models for inferring RAD. In theory, we could have used higher-order (>2) sample coverages and their deficits to build models with more than two parameters. However, the parameter estimates from such models may be too uncertain to be useful, and may be too unstable to estimate properly.

The concept of the SAD/RAD has been also extended in this study to the corresponding SID/RID for incidence data comprising species detection/non-detection records in each sampling unit; a non-detection of a species in a sampling unit may be due to a true absence or an undetected presence, so this model can be applied not only to surveys of sessile plants but also to surveys of mobile animals in which detection probabilities are less than 1.0. If we consider the special case in which a species can always be detected if it is present in a sampling unit, then the detection/non-detection records become presence/absence data and our model can be connected to a special case of occupancy estimation and modeling (e.g., MacKenzie et al. 2006). In this special case, the incidence probability \( \hat{p}_i \) can be interpreted as occupancy rate of species \( i \) in the study area. Our proposed formula \( \hat{p}_i = (Y_i/T)(1 - \lambda e^{-T}) \) for detected species provides a nonparametric adjustment to the sample occupancy rate (i.e., \( Y_i/T \)) and thus can provide a better estimator of the true occupancy rate in the study area.

Our proposed estimator for the RAD/RID is also potentially useful in other inference problems. For example, the proposed RAD can be used for estimating any diversity measure that is a function of species relative abundances \( (p_1, p_2, \ldots, p_j) \). An enormous number of diversity measures have been proposed, not only in ecology but also in other disciplines, e.g., genetics, economics, information sciences, physics, and social sciences, among others; see Magurran and McGill (2011). Hill numbers (including the Shannon diversity and Simpson diversity), originally proposed by Hill (1973) have been increasingly used to quantify species diversity. We specifically discuss (in Appendix E) the use of our estimated RAD in the estimation of diversity profiles based on Hill numbers. The resulting profiles significantly improve over the empirical diversity measures mainly because the relative abundances of undetected species can be incorporated.

In another important application, when diversity measures are complicated functions of species sample abundances, and their variances are therefore difficult to estimate analytically, our proposed RAD estimator can be bootstrapped to assess their sampling variances and to construct the associated confidence intervals. This approach was applied to obtain the variances of the estimator given in Eq. 4d (see Sampling variances of our estimators) and the diversity estimators (see Appendix E). It has many potential applications in the analyses of beta diversity and related similarity (or differentiation) measures based on species relative abundances.

All the estimation procedures and estimators proposed in this study are featured in the freeware application JADE (joint species-rank abundance distribution/estimation; available online).7 The R scripts for JADE are available in the Supplement.

Acknowledgments

The authors thank the subject editor (Brian Inouye), Lou Jost, and an anonymous reviewer for very helpful and thoughtful suggestions and comments. Their comments inspired us to extend our method to deal with incidence data. A. Chao’s work on this project is funded by Taiwan Ministry of Science and Technology under Contract 103-2628-M007-007. T. C. Hsieh is supported by a post-doctoral fellowship, Ministry of Science and Technology, Taiwan. R. L. Chazdon was supported by U.S. NSF DEB 0639393, NSF DEB 1050957, and NSF DEB-1147429. R. K. Colwell was supported by CAPES Ciência sem Fronteiras (Brazil). N. J. Gotelli was

7 http://chao.stat.nthu.edu.tw/blog/software-download/
supported by U.S. NSF DEB 1257625, NSF DEB 1144055, and NSF DEB 1136644.

**LITERATURE CITED**


Foissner, W., S. Agatha, and H. Berger. 2002. Soil ciliates (Protozoa, Ciliophora) from Namibia (Southwest Africa), with emphasis on two contrasting environments, the Etosha region and the Namib Desert. Demnia 5:1–1459.


**SUPPLEMENTAL MATERIAL**

Ecological Archives

Appendices A–E and the Supplement are available online: http://dx.doi.org/10.1890/14-0550.1.sm
APPENDIX A: Simulation results based on two theoretical abundance distributions and four large empirical surveys

Comparisons of the true RAD, the empirical RAD and the proposed RAD for six scenarios

In this appendix, we report the simulation results for six scenarios including the one presented in Fig. 1 of the main text. We simulated data from two theoretical abundance distributions (the Zipf-Mandelbrot model and the log-normal model) and treated four large empirical diversity surveys as the complete assemblages. In the latter case, the species-rank abundance distribution from each survey was assumed to be the “true” complete RAD. The six scenarios are described below. In each case, if \((p_1, p_2, \ldots, p_S)\) are fixed parameters, then we give the true value of the CV (coefficient of variation, which is the ratio of standard deviation and mean) of these probabilities. The CV value quantifies the degree of heterogeneity of the probabilities \((p_1, p_2, \ldots, p_S)\). When all probabilities are equal, CV = 0. A larger value of CV signifies higher degree of heterogeneity among probabilities. If \((p_1, p_2, \ldots, p_S)\) are random variables generated from a distribution (e.g., the log-normal model, as described below), then the average values of CV over 200 simulated data sets are given to approximate the true theoretical value.

**Scenario 1** (The Zipf-Mandelbrot model with 200 species). The relative abundances of the complete assemblage take the general form \(p_i = c / (2 + i)\), \(i = 1, 2, \ldots, 200\), where \(c\) is a normalized constant such that the sum of the relative abundances is unity; see Magurran (2004) for an introduction of the Zipf-Mandelbrot model. CV = 1.751.

**Scenario 2**: (The log-normal model with 200 species). We first generated 200 random variables \((a_1, a_2, \ldots, a_{200})\) from a normal distribution with mean \(\mu = 0\) and standard deviation \(\sigma = 1\). The species relative abundance of species \(i\) of the complete assemblage takes the form \(p_i = c \exp(a_i)\), where \(c\) is a normalized constant. Then all samples were generated from the relative abundances \((p_1, p_2, \ldots, p_{200})\) throughout the simulation. CV = 1.439 for the generated set used in our simulation.

**Scenarios 3, 4, 5**: We treated three tree data sets collected by Chazdon and colleagues as the complete assemblages; see Norden et al. (2009) for details. The three data sets include species abundance survey data taken, respectively, from the LEP old-growth forest site (Scenario 3, 152 species, 943 individuals, CV = 1.545), the LEP older second-growth forest site (Scenario 4, 104 species, 1263 individuals, CV = 2.339), and the LS younger second-growth forest site (Scenario 5, 76 species, 1020 individuals, CV = 2.305). The species abundance frequency counts for the three surveys are given in Table A1.

**Scenario 6**: We treated the census data of Miller and Wiebert (1989) for endangered and rare vascular plant species in the central portion of the southern Appalachian Region (USA) as the true assemblage. The species abundance frequency counts for this survey are given in Table A1; a total of 188 species were represented by 1008 individuals. CV = 1.563.
Given the species relative abundances of each complete assemblage, we generated 200 data sets of sample sizes 200, 400 and 800. Then for each generated data set, we obtained the empirical RAD and the proposed RAD. Comparisons of the true RAD of the complete assemblage, the empirical RAD curves, and the proposed RAD curves are shown in Fig. A1. All the patterns are generally consistent with those presented in Fig. 1 of the main text for Scenario 1. Therefore, the findings and interpretations discussed in the main text for Scenario 1 can be applied to all other scenarios. We thus omit the details.

**Table A1.** Species abundance frequency counts for Scenarios 3–6.

**Scenario 3.** LEP old-growth forest, 152 species, 943 individuals (Norden et al. 2009)

<table>
<thead>
<tr>
<th>i</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>fi</td>
<td>46</td>
<td>30</td>
<td>16</td>
<td>12</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>i</th>
<th>19</th>
<th>20</th>
<th>25</th>
<th>38</th>
<th>39</th>
<th>40</th>
<th>46</th>
<th>52</th>
<th>55</th>
</tr>
</thead>
<tbody>
<tr>
<td>fi</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Scenario 4.** LEP older (29 years) second-growth forest, 104 species, 1263 individuals (Norden et al. 2009)

<table>
<thead>
<tr>
<th>i</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>fi</td>
<td>33</td>
<td>15</td>
<td>13</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>i</th>
<th>16</th>
<th>17</th>
<th>20</th>
<th>22</th>
<th>39</th>
<th>45</th>
<th>57</th>
<th>72</th>
<th>88</th>
<th>132</th>
<th>133</th>
<th>178</th>
</tr>
</thead>
<tbody>
<tr>
<td>fi</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Scenario 5.** LS younger (21 years) second-growth forest, 76 species, 1020 individuals (Norden et al. 2009)

<table>
<thead>
<tr>
<th>i</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>7</th>
<th>8</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>15</th>
<th>31</th>
<th>33</th>
<th>34</th>
</tr>
</thead>
<tbody>
<tr>
<td>fi</td>
<td>29</td>
<td>13</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>i</th>
<th>35</th>
<th>66</th>
<th>72</th>
<th>78</th>
<th>127</th>
<th>131</th>
<th>174</th>
</tr>
</thead>
<tbody>
<tr>
<td>fi</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Scenario 6. The extant rare vascular plant species in the southern Appalachians, 188 species, 1008 individuals (Miller and Wiebert 1989)

<table>
<thead>
<tr>
<th>i</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_i$</td>
<td>61</td>
<td>35</td>
<td>18</td>
<td>12</td>
<td>15</td>
<td>4</td>
<td>8</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>i</th>
<th>16</th>
<th>19</th>
<th>20</th>
<th>22</th>
<th>29</th>
<th>32</th>
<th>40</th>
<th>43</th>
<th>48</th>
<th>67</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_i$</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
**Fig. A1.** Comparison of the true RAD of the complete assemblage (light blue line), the empirical RAD curves (superimposed dark blue lines with 200 replications), and the proposed RAD curves (superimposed red lines with 200 replications) for sample sizes 200 (left panels), 400 (middle panels) and 800 (right panels). Data sets were generated from two theoretical abundance distributions (the Zipf-Mandelbrot model and the log-normal model) and four plant assemblages (Scenarios 3–6, see Table A1 for the species abundance frequency counts). For each assemblage and each sample size, 200 data sets were generated, thus there are 200 estimated RADs (200 red lines and 200 dark blue lines). Note that the X-axis is the species list, ranked from most abundant to least abundant, and the Y-axis (relative abundance) is displayed on a log_{10} scale.
Comparison of the sample relative abundance and the adjusted estimator

In Eq. 4d of the main text, we derive for a detected species \( i \) (i.e., species sample frequency \( X_i > 0 \)) the following adjusted estimator of the plug-in estimator:

\[
\hat{p}_i = \frac{X_i}{n} (1 - \hat{\lambda} e^{-\hat{\lambda} X_i}).
\]

This adjusted estimator also provides a simple nonparametric relationship between the species sample relative abundance (i.e., \( X_i/n \)) and the estimated true relative abundance in the entire assemblage. To examine the relative performance of the two estimators for each species, we conducted simulations by generating data sets from a Zipf-Mandelbrot model with 200 species (Scenario 1 described in Appendix A) and from the assemblage of vascular plants with 188 species (Scenario 6 described in Appendix A).

The species in each assemblage are ordered and indexed from the most abundant (with an index 1) to the least abundant (with an index 200 in the Zipf-Mandelbrot model, and an index 188 in the vascular plant assemblage). Given the species relative abundances, we generated 5000 samples for three sample sizes (200, 400 and 800). For each generated sample, the abundance for each detected species was recorded; the plug-in and the adjusted estimators were calculated. For each particular species, we only considered those samples in which that species was detected because the focus was on estimating the relative abundance of a detected species. The number of times that each species was detected in the 5000 generated samples varies with species, and thus the number of simulated samples used to obtain the averages of the relative biases and RMSEs are different. In Figs. A2 and A3, we respectively present the plots of the average relative bias (with respect to the true relative abundance) and the RMSE for each species indexed from the most abundant to the least abundant.

Based on Figs. A2 and A3, we have the following findings:

(1) Both figures reveal that for those relatively abundant species the plug-in estimator works well and is nearly unbiased, and thus no adjustment is actually needed. This is predicted from our theory discussed in the main text. Due to downward bias-correction, the adjusted estimator is nevertheless subject to slight negative bias for relatively abundant species, but the difference in bias between the two estimators for relatively abundant species is generally limited.

(2) For those relatively rare species, the plug-in estimator exhibits positive bias, reiterating our theory (Eq. 1 of the main text). The degree of the positive bias increases with increasing relative rarity of each species, as shown by the increasing trend of the relative bias when species abundance is decreased. Fig. A2 shows that our adjusted estimator for the relatively rare species can reduce the positive bias inherent in the plug-in estimator; the reduction is substantial for smaller sample sizes (\( n = 200 \) and 400).

(3) Fig. A3 further demonstrates that the adjusted estimator is generally more accurate in terms of smaller RMSE. The improvement over the plug-in estimator is clearly seen for smaller sample sizes (\( n = 200 \) and 400) and also increases with increasing relative rarity of each species.
Fig. A2. The average relative bias of the plug-in estimator $\hat{p}_i = X_i / n$ and the proposed adjusted estimator $\tilde{p}_i = (X_i / n)(1 - \hat{\lambda}e^{-\hat{\theta}X_i})$ (Eq. A.1 or Eq. 4d of the main text) for species ordered and indexed from the most abundant to the least abundant under the Zipf-Mandelbrot model (left panels) and the assemblage of vascular plants (right panels) for sample sizes 200 (upper panels), 400 (middle panels) and 800 (lower panels). The horizontal dotted line in each panel represents the relative bias = 0 line.
Fig. A3. The root mean squared error (RMSE) of the plug-in estimator $\hat{p}_i = X_i / n$ and the proposed adjusted estimator $\tilde{p}_i = (X_i / n)(1 - \hat{\lambda}e^{-\hat{\lambda}X_i})$ (Eq. A.1 or Eq. 4d of the main text) for species ordered and indexed from the most abundant to the least abundant under the Zipf-Mandelbrot model (left panels) and the assemblage of vascular plants (right panels) for sample sizes 200 (upper panels), 400 (middle panels) and 800 (lower panels).
LITERATURE CITED

APPENDIX B: A generalization of Good-Turing sample coverage theory.

Assume that there are $S$ species in the assemblage and that the true species relative abundances are $(p_1, p_2, \ldots, p_S)$, $\sum_{i=1}^{S} p_i = 1$. Suppose a random sample of $n$ individuals is selected with replacement. Let $X_i$ denote the abundance of the $i$-th species in the sample, $i = 1, 2, \ldots, S$. The coverage of the sample originally developed by Turing and Good (Good 1953) is expressed as

$$1^C = \sum_{i \text{ detected}} p_i = \sum_{i=1}^{S} p_i I(X_i > 0), \quad (B.1)$$

where $I(A)$ is an indicator function that equals 1 when $A$ is true and 0 otherwise. The leading superscript “1” in the notation $1^C$ signifies that this is the first-order of our generalization of sample coverage. This sample coverage can be regarded as one measure of sample completeness, giving the fraction of the true relative abundances of those species detected in the sample (or equivalently, the proportion of the total number of individuals in an assemblage that belong to the species represented in the sample). Subtracting the sample coverage from unity gives the “coverage deficit” (Chao and Jost 2012), $1^C_{\text{def}}$, the proportion of the assemblage belonging to undetected species:

$$1^C_{\text{def}} = 1 - 1^C = \sum_{i=1}^{S} p_i I(X_i = 0). \quad (B.2)$$

The second-order sample coverage is defined as the fraction of the squared true relative abundances of those species detected in the sample. It can be expressed as

$$2^C = \frac{\sum_{i \text{ detected}} p_i^2}{\sum_{i=1}^{S} p_i^2} = \frac{\sum_{i=1}^{S} p_i^2 I(X_i > 0)}{\sum_{i=1}^{S} p_i^2}. \quad (B.3)$$

The second-order sample coverage quantifies the sample completeness for very abundant or dominant species. The expected second-order sample coverage is

$$E(2^C) = 1 - \frac{\sum_{i=1}^{S} p_i^2 (1 - p_i)^n}{\sum_{i=1}^{S} p_i^2}. \quad (B.4)$$

The denominator and numerator for the right-most term in the above formula can be accurately estimated. The minimum variance unbiased estimator of $\sum_{i=1}^{S} p_i^2$ is $\sum_{X_i \geq 2} X_i (X_i - 1)/[n(n - 1)]$ (Good 1953). Using a similar derivation as in Chao et al. (2009),
we can obtain an estimator for \( \sum_{i=1}^{S} p_i^r (1 - p_i)^n \) as \( \frac{2f_2}{n(n-1)} \left[ \frac{(n-2)f_2}{(n-2)f_2 + 3f_3} \right] \) if \( f_2, f_3 > 0 \); details are provided below for a general case. We have the following estimator for the second-order sample coverage:

\[
2\hat{C} = 1 - \frac{2f_2}{\sum_{X_i \geq 2} X_i(X_i - 1)} \left[ \frac{(n-2)f_2}{(n-2)f_2 + 3f_3} \right]^2. \tag{B.4}
\]

The deficit of the second-order sample coverage is defined as \( 2C_{\text{def}} = 1 - 2\hat{C} \), leading to the corresponding estimator:

\[
2\hat{C}_{\text{def}} = \frac{2f_2}{\sum_{X_i \geq 2} X_i(X_i - 1)} \left[ \frac{(n-2)f_2}{(n-2)f_2 + 3f_3} \right]^2. \tag{B.5}
\]

We can extend the above approach to define the \( r \)-th order sample coverage as

\[
\hat{C} = \frac{\sum_{i \text{ detected}} p_i^r}{\sum_{i=1}^{S} p_i^r}, \quad r = 1, 2, \ldots,
\]

with the expected value

\[
E(\hat{C}) = 1 - \frac{\sum_{i=1}^{S} p_i^r (1 - p_i)^n}{\sum_{i=1}^{S} p_i^r}.
\]

We separately estimate the denominator and numerator for the right-most term in the above formula. The minimum variance unbiased estimator of \( \sum_{i=1}^{S} p_i^r \) for \( r \leq n \) is \( \sum_{X_i \geq r} X_i^{(r)} / n^{(r)} \) (Good, 1953), where \( X_i^{(r)} = X_i(X_i - 1)\ldots(X_i - r + 1) \) and \( n^{(r)} = n(n-1)\ldots(n-r+1) \) denote the descending factorial. We now derive an estimator for the numerator \( \sum_{i=1}^{S} p_i^r (1 - p_i)^n \) (provided \( f_r, f_{r+1} > 0 \)) following Chao et al. (2009) approach. Under the model that the species sample frequencies \( (X_1, X_2, \ldots, X_S) \) follow a multinomial distribution with parameters \( (p_1, p_2, \ldots, p_S) \), we have

\[
E(f_k) = E\left( \sum_{i=1}^{S} I(X_i = k) \right) = \sum_{i=1}^{S} \frac{n^{(k)}}{k!} p_i^k (1 - p_i)^{n-k}, \quad k = 0, 1, \ldots, n.
\]

Then we can write

\[
\sum_{i=1}^{S} p_i^r (1 - p_i)^n = \sum_{i=1}^{S} \frac{r!}{n^{(r)}} (1 - p_i)^{r} \left[ \frac{n^{(r)}}{r!} p_i^r (1 - p_i)^{n-r} \right] = \sum_{i=1}^{S} \frac{r!}{n^{(r)}} (1 - p_i)^{r} P(X_i = r) = \frac{r!}{n^{(r)}} E\left( \sum_{i=1}^{S} (1 - p_i)^{r} I[X_i = r] \right).
\]
Note that in the sum $\sum_{i=1}^{S} (1 - p_i)^{r} I[X_i = r]$, only species with sample frequency $r$ would contribute a term $(1 - p_i)^{r}$, and other species do not contribute. Denote the mean relative abundance for all species with sample frequency $r$ by $\overline{p}_{(r)}$. We have an approximation formula:

$$\left\{ \sum_{i=1}^{S} (1 - p_i)^{r} I[X_i = r] \right\} \approx f_r \times (1 - \overline{p}_{(r)})^{r},$$

implying

$$\sum_{i=1}^{S} p_{(r)} (1 - p_i)^{r} \approx \frac{r! E(f_r)}{n} (1 - \overline{p}_{(r)})^{r}.$$ \hspace{1cm} (B.6)

To estimate $\overline{p}_{(r)}$, we consider the following expected sum:

$$E\sum_{i=1}^{S} \frac{p_i}{1 - p_i} I[X_i = r] = \sum_{i=1}^{S} \frac{p_i}{1 - p_i} \frac{n^{(r)}}{r!} p_{(r)}^{r} (1 - p_i)^{n-r}$$

$$= \frac{r+1}{n-r} \sum_{i=1}^{S} \frac{n^{(r+1)}}{(r+1)!} p_{(r+1)}^{r+1} (1 - p_i)^{n-(r+1)} = \frac{r+1}{n-r} E(f_{r+1}).$$ \hspace{1cm} (B.7)

Applying the same approach used in the derivation of Eq. B.6, we have

$$E\sum_{i=1}^{S} \frac{p_i}{1 - p_i} I[X_i = r] \approx \frac{\overline{p}_{(r)}}{1 - \overline{p}_{(r)}} E(f_r).$$ \hspace{1cm} (B.8)

Substituting the expected frequency counts involved in Eqs. B.7 and B.8 by sample data, we obtain

$$\frac{\overline{p}_{(r)}}{1 - \overline{p}_{(r)}} f_r \approx \frac{r+1}{n-r} f_{r+1}.$$ \hspace{1cm} (B.9)

Therefore, we can solve $\overline{p}_{(r)}$ in the above equation to get

$$\overline{p}_{(r)} \approx \frac{(r+1)f_{r+1}}{(n-r)f_r + (r+1)f_{r+1}}.$$ \hspace{1cm} (B.10)

Substituting the resulting $\overline{p}_{(r)}$ into Eq. B.6, we obtain the following estimator of

$$\sum_{i=1}^{S} p_{(r)} (1 - p_i)^{r} :$$
These lead to the following coverage estimator:

\[
\hat{C}_r = 1 - \frac{r! f_r}{\sum X_i^{(r)}} \left[ \frac{(n-r)f_r}{(n-r)f_r + (r+1)f_{r+1}} \right].
\]

For \( r = 1 \), it reduces to Eq. 2b of the main text, and for \( r = 2 \), it reduces to Eq. B.4 and Eq. 3a of the main text. The corresponding estimator for the deficit of the \( r \)-th order sample coverage is

\[
\hat{C}_{def} = \frac{r! f_r}{\sum X_i^{(r)}} \left[ \frac{(n-r)f_r}{(n-r)f_r + (r+1)f_{r+1}} \right].
\]

For \( r = 1 \), it reduces to Eq. 2c of the main text, and for \( r = 2 \), it reduces to Eq. B.5 and Eq. 3b of the main text.

**LITERATURE CITED**


APPENDIX C: Other models for the relative abundances of the undetected species.

In our estimation procedures, we adopt a nonparametric method for adjusting sample relative abundances for species detected in a sample (Eq. 4d of the main text). The adjustment method for detected species is valid for all species abundance distributions. For estimating the species relative abundances for undetected species, a functional form of parametric model is needed. In the main text, we assume a geometric series model based on an estimated number of undetected species \( \hat{f}_0 \) (see Eq. 5a of the main text). There are other models or distributions that could be applied to the set of undetected species indexed by 1, 2, \( \ldots, \hat{f}_0 \).

**The broken-stick model**

Under a broken-stick model with \( \hat{f}_0 \) species, Pielou (1975) and Baczkowski (2000) showed that the \( i \)-th ordered species relative abundance of the set \( ( p_1 > p_2 > \ldots > p_{\hat{f}_0} ) \) takes the form:

\[
p_i = \frac{K}{\hat{f}_0} \left( \frac{1}{i} + \frac{1}{i+1} + \ldots + \frac{1}{\hat{f}_0} \right) = \frac{K}{\hat{f}_0} \sum_{j=0}^{\hat{f}_0-i} \frac{1}{\hat{f}_0-j}, \quad i = 1, 2, \ldots, \hat{f}_0,
\]

where \( K \) is a normalized constant such that the sum of the relative abundances is equal to the coverage deficit estimator \( \hat{C}_{def} \) (Eq. 2c in the main text), i.e.,

\[
\sum_{i=1}^{\hat{f}_0} p_i = \hat{C}_{def}.
\]

Because the expected relative abundances have an analytic form as indicated above, the RAD is fully determined by the number of undetected species. The model is thus not flexible.

**The Poisson log-normal model**

As indicated in Magurran (2004, p. 39), the continuous log-normal model should only be applied to continuous abundance data, such as biomass or cover measures, rather than to discrete data. For individual-based data, Bulmer (1974) presented a Poisson log-normal model. When Bulmer’s model is used to the set of undetected species, the species relative abundances are:

\[
p_i(\bar{M},V) = \frac{1}{\sqrt{2\pi V}} \int_0^\infty \frac{2^{i-2} e^{-\lambda} e^{\frac{(\log \lambda - \bar{M})^2}{2V}}}{(i-1)!} d\lambda, \quad i = 1, 2, \ldots, \hat{f}_0,
\]

where \( \bar{M} \) and \( V \) denote respectively the mean and variance of the log-normal distribution. Based on the Eqs. 6a and 6b in the main text, we have the following two equations for the
undetected species in terms of two parameters $M$ and $V$:

$$\sum_{i \in \text{undetected}} p_i \approx \sum_{i=1}^{\hat{N}} p_i(M, V) = \hat{C}_{\text{def}};$$

$$\sum_{i \in \text{undetected}} p_i^2 \approx \sum_{i=1}^{\hat{N}} \left[p_i(M, V)\right]^2 = \hat{C}_{\text{def}} \times \frac{\sum_{i \geq 2} X_i(X_i - 1)}{n(n-1)}.$$

**Comparison of models using spider data**

For the spider data (with 26 detected species and an estimate of 18 undetected species) discussed in the main text, we compare in Fig. C1 the empirical RAD and three fitted RADs including the proposed RAD curve based on a geometric series model, the broken-stick model, and the Poisson log-normal model. These three models were only fitted to the undetected species; for the 26 detected species, Eq. 4d of the main text is used for adjusting the plug-in estimator for all three models. So there are no differences in the adjusted relative abundances for the 26 detected species, as shown in Table C1 where the estimated complete RAD under each of the three models are listed.

For the proposed geometric series model, we also obtained the bootstrap s.e. based on 5000 replications. See the main text for the description of the bootstrap method. From Fig. C1 and Table C1, the Poisson log-normal and our proposed geometric series model yield almost identical RADs for this example. Here the parameters for the Poisson log-normal model are estimated to be $\hat{M} = 6.089$ and $\hat{V} = 4.779$ by the above system of nonlinear equations. (For some other data sets, solutions may not exist.) Except for a few species in the broken-stick model, the empirical RAD which ignores the set of the undetected species gives higher relative abundances than those in the other three models. The broken-stick model yields substantial low abundances for the tail of the fitted RAD.

Our simulation studies have suggested that the iterative steps for the Poisson log-normal models in many cases failed to converge to proper solutions. This is the main drawback to fit a Poisson log-normal model. For the simple geometric series model we used, all the iterative steps converged quickly. As discussed in the main text, we use these models only for modeling the undetected tail distribution; unless the assemblage is poorly sampled, the relative abundances of those undetected species (i.e., in the “tail” of the estimated RAD) are typically very small. Thus, the choice of the model for estimating the relative abundances of undetected species is a minor issue in our approach.
**Fig. C1.** Comparison of the empirical RAD curve (blue solid line) and three fitted RAD curves for the spider data discussed in the main text: the proposed RAD curve based on a geometric series model (red solid line), the broken-stick model (fuchsia dotted line), and the Poisson log-normal model (green dotted line). All these three models were fitted only to the set of undetected species. The adjustment method (Eq. 4d of the main text) for the 26 detected species is independent of species abundance distributions or models.
Table C1. Comparison of the estimated species-rank abundances of the empirical RAD and the estimated RADs based on three models (geometric series, Poisson log-normal, and broken-stick) fitted to the set of undetected species for the spider data discussed in the main text. The estimated abundances for the detected species are in black print; the estimated abundances for the undetected species are in red print. The empirical RAD applies only to 26 observed species whereas the other three models additionally consider 18 undetected species. The three models differ only in the estimated abundances for undetected species as the estimated abundances of the 26 detected species for all three models are based on the same adjusted formula (Eq. 4d of the main text). The bootstrap s.e. of the adjusted estimator for each detected species under the geometric series model is based on 5000 bootstrap replications.

<table>
<thead>
<tr>
<th>Species rank</th>
<th>Empirical RAD</th>
<th>Geometric series model (s.e.)</th>
<th>Poisson log-normal</th>
<th>Broken-stick model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2738</td>
<td>0.2735 (0.0349)</td>
<td>0.2735</td>
<td>0.2735</td>
</tr>
<tr>
<td>2</td>
<td>0.1310</td>
<td>0.1282 (0.261)</td>
<td>0.1282</td>
<td>0.1282</td>
</tr>
<tr>
<td>3</td>
<td>0.1012</td>
<td>0.0973 (0.0228)</td>
<td>0.0973</td>
<td>0.0973</td>
</tr>
<tr>
<td>4</td>
<td>0.0893</td>
<td>0.0849 (0.0213)</td>
<td>0.0849</td>
<td>0.0849</td>
</tr>
<tr>
<td>5</td>
<td>0.0893</td>
<td>0.0849 (0.0215)</td>
<td>0.0849</td>
<td>0.0849</td>
</tr>
<tr>
<td>6</td>
<td>0.0536</td>
<td>0.0482 (0.0162)</td>
<td>0.0482</td>
<td>0.0482</td>
</tr>
<tr>
<td>7</td>
<td>0.0476</td>
<td>0.0422 (0.0151)</td>
<td>0.0422</td>
<td>0.0422</td>
</tr>
<tr>
<td>8</td>
<td>0.0357</td>
<td>0.0306 (0.0130)</td>
<td>0.0306</td>
<td>0.0306</td>
</tr>
<tr>
<td>9</td>
<td>0.0357</td>
<td>0.0306 (0.0128)</td>
<td>0.0306</td>
<td>0.0306</td>
</tr>
<tr>
<td>10</td>
<td>0.0238</td>
<td>0.0194 (0.0100)</td>
<td>0.0194</td>
<td>0.0194</td>
</tr>
<tr>
<td>11</td>
<td>0.0119</td>
<td>0.0091 (0.0060)</td>
<td>0.0091</td>
<td>0.0138</td>
</tr>
<tr>
<td>12</td>
<td>0.0119</td>
<td>0.0091 (0.0060)</td>
<td>0.0091</td>
<td>0.0099</td>
</tr>
<tr>
<td>13</td>
<td>0.0119</td>
<td>0.0091 (0.0061)</td>
<td>0.0091</td>
<td>0.0091</td>
</tr>
<tr>
<td>14</td>
<td>0.0119</td>
<td>0.0091 (0.0060)</td>
<td>0.0091</td>
<td>0.0091</td>
</tr>
<tr>
<td>15</td>
<td>0.0060</td>
<td>0.0044 (0.0039)</td>
<td>0.0044</td>
<td>0.0091</td>
</tr>
<tr>
<td>16</td>
<td>0.0060</td>
<td>0.0044 (0.0039)</td>
<td>0.0044</td>
<td>0.0091</td>
</tr>
<tr>
<td>17</td>
<td>0.0060</td>
<td>0.0044 (0.0039)</td>
<td>0.0044</td>
<td>0.0079</td>
</tr>
<tr>
<td>18</td>
<td>0.0060</td>
<td>0.0044 (0.0039)</td>
<td>0.0044</td>
<td>0.0066</td>
</tr>
<tr>
<td>19</td>
<td>0.0060</td>
<td>0.0044 (0.0038)</td>
<td>0.0044</td>
<td>0.0056</td>
</tr>
<tr>
<td>20</td>
<td>0.0060</td>
<td>0.0044 (0.0039)</td>
<td>0.0044</td>
<td>0.0048</td>
</tr>
<tr>
<td>21</td>
<td>0.0060</td>
<td>0.0044 (0.0039)</td>
<td>0.0044</td>
<td>0.0044</td>
</tr>
<tr>
<td>22</td>
<td>0.0060</td>
<td>0.0044 (0.0039)</td>
<td>0.0044</td>
<td>0.0044</td>
</tr>
<tr>
<td>23</td>
<td>0.0060</td>
<td>0.0044 (0.0039)</td>
<td>0.0044</td>
<td>0.0044</td>
</tr>
<tr>
<td>24</td>
<td>0.0060</td>
<td>0.0044 (0.0037)</td>
<td>0.0044</td>
<td>0.0044</td>
</tr>
<tr>
<td>25</td>
<td>0.0060</td>
<td>0.0044 (0.0039)</td>
<td>0.0044</td>
<td>0.0044</td>
</tr>
<tr>
<td>26</td>
<td>0.0060</td>
<td>0.0044 (0.0038)</td>
<td>0.0044</td>
<td>0.0044</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>0.0044 (0.0039)</td>
<td>0.0044</td>
<td>0.0044</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>0.0044</td>
<td>0.0044</td>
<td>0.0044</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>0.0043</td>
<td>0.0044</td>
<td>0.0044</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>0.0042</td>
<td>0.0043</td>
<td>0.0044</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>0.0042</td>
<td>0.0043</td>
<td>0.0044</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>0.0041</td>
<td>0.0042</td>
<td>0.0044</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>0.0041</td>
<td>0.0042</td>
<td>0.0041</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>0.0040</td>
<td>0.0041</td>
<td>0.0036</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>0.0040</td>
<td>0.0040</td>
<td>0.0031</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>0.0039</td>
<td>0.0040</td>
<td>0.0026</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>0.0039</td>
<td>0.0039</td>
<td>0.0022</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>0.0038</td>
<td>0.0038</td>
<td>0.0019</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>0.0038</td>
<td>0.0038</td>
<td>0.0015</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>0.0037</td>
<td>0.0037</td>
<td>0.0012</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>0.0037</td>
<td>0.0036</td>
<td>0.0010</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>0.0036</td>
<td>0.0036</td>
<td>0.0007</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>0.0036</td>
<td>0.0035</td>
<td>0.0005</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>0.0035</td>
<td>0.0031</td>
<td>0.0002</td>
<td></td>
</tr>
</tbody>
</table>

*An undetected species, so its bootstrap s.e. is not obtainable; see the subsection Sampling variances of our estimators of the main text for explanation.

LITERATURE CITED


Appendix D: Unveiling the species-rank incidence distribution for incidence data.

In the main text, we introduced the basic model, data format and terminology for incidence data. In order for this presentation to be self-contained, we duplicate the introductory text here; see Colwell et al. (2012) and Chao et al. (2014) for the necessary backgrounds. Assume that in the focal assemblage there are $S$ different species indexed by $1, 2, \ldots, S$. For any sampling unit, assume that the $i$th species has its own unique incidence (or occurrence) probability $\pi_i$ that is constant for any randomly selected sampling unit. The incidence probability $\pi_i$ is the probability that species $i$ is detected in a sampling unit. The species incidence distribution (SID) and the corresponding species-rank incidence distribution (RID) refer to the set $(\pi_1, \pi_2, \ldots, \pi_S)$ of the $S$ species.

Assume that a set of $T$ sampling units are randomly selected from the study area with replacement. The underlying data consist of an $S \times T$ detection/non-detection matrix $\{W_{ij}; i = 1, 2, \ldots, S, j = 1, 2, \ldots, T\}$; here $W_{ij} = 1$ if species $i$ is detected in sampling unit $j$, and $W_{ij} = 0$ otherwise, $i = 1, 2, \ldots, S, j = 1, 2, \ldots, T$. Let $Y_i = \sum_{j=1}^{T} W_{ij}$ denotes the sample species incidence frequency of species $i$. Denote the incidence frequency counts by $(Q_0, Q_1, \ldots, Q_T)$, where $Q_k$ is the number of species that are detected in exactly $k$ sampling units in the data, $k = 0, 1, \ldots, T$. 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. Also, $Q_1$ 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 only two sampling units). Define the sample incidence probability as $\hat{\pi}_i = Y_i / T$, which is the conventional plug-in estimator. The empirical RID is based on this plug-in estimator.

All derivations below are based on that $Y_i, i = 1, 2, \ldots, S$ follows a binomial distribution with the total number $T$ and the detection probability $\pi_i$:

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

Chao et al. (1992) first defined the sample coverage for a set of $T$ sampling units as the following expression:

$$^1C^* = \frac{\sum_{i \text{ detected}} \pi_i}{\sum_{i=1}^{S} \pi_i} = \frac{\sum_{i=1}^{S} \pi_i I(Y_i > 0)}{\sum_{i=1}^{S} \pi_i}, \quad (D.1)$$

where $I(A)$ is an indicator function that equals 1 when $A$ is true and 0 otherwise. Here we use a superscript “*” to refer to the sample coverage of incidence data, and use a leading superscript “1” to signify that this is the first-order of our generalization of sample coverage. The sample coverage quantifies the fraction of the total incidence probabilities of the
discovered species in the $T$ sampling units. It is an objective measure of sample completeness for incidence data. Subtracting the sample coverage from unity gives the fraction of the assemblage belonging to the undetected species. As with the abundance data, we denote the first-order coverage deficit by $1^{1}C_{\text{def}}^{*}$, which can be expressed as (Chao and Jost 2012):

$$1^{1}C_{\text{def}}^{*} = 1 - C^{*} = \frac{\sum_{i} \pi_{i} I(Y_{i} = 0)}{\sum_{i} \pi_{i}}. \quad (D.2)$$

Following exactly the same approach for abundance data, we can generalize the sample coverage theory to incidence data by defining the $r$-th order sample coverage as

$$r C^{*} = \frac{\sum_{i} \pi_{i}^{r} I(Y_{i} > 0)}{\sum_{i} \pi_{i}^{r}}, \quad r = 1, 2, \ldots \quad (D.3)$$

with the expected value

$$E(r C^{*}) = 1 - \frac{\sum_{i} \pi_{i}^{r} (1 - \pi_{i})^{r}}{\sum_{i} \pi_{i}^{r}}.$$

If the incidence frequency counts $Q_{r}$ and $Q_{r+1}$ are both non-zero, then we have the following estimator

$$r \hat{C}^{*} = 1 - \frac{r^{1}Q_{r}}{\sum_{Y \geq r} Y_{i}^{(r)}} \left[ \frac{(T - r)Q_{r}}{(T - r)Q_{r} + (r + 1)Q_{r+1}} \right]^{r}, \quad r = 1, 2, \ldots \quad (D.4)$$

where $Y_{i}^{(r)} = Y_{i} (Y_{i} - 1) \ldots (Y_{i} - r + 1)$ denotes the descending factorial. The estimator in Eq. D.4 is derived by noting that the minimum variance unbiased estimator of $\sum_{i=1}^{S} \pi_{i}^{r}$ is $\sum_{Y \geq r} Y_{i}^{(r)}/T^{(r)}$ (Good, 1953), and an estimator for $\sum_{i=1}^{S} \pi_{i}^{r} (1 - \pi_{i})^{r}$ is (if $Q_{r} , Q_{r+1} > 0$)

$$\frac{r^{1}Q_{r}}{T^{(r)}} \left[ \frac{(T - r)Q_{r}}{(T - r)Q_{r} + (r + 1)Q_{r+1}} \right]^{r}.$$

The derivation is parallel to that given in Appendix B for abundance data. The corresponding estimator for the deficit of the $r$-th order coverage is $r \hat{C}_{\text{def}}^{*} = 1 - r \hat{C}^{*}$, i.e.,

$$r \hat{C}_{\text{def}}^{*} = \frac{r^{1}Q_{r}}{\sum_{Y \geq r} Y_{i}^{(r)}} \left[ \frac{(T - r)Q_{r}}{(T - r)Q_{r} + (r + 1)Q_{r+1}} \right]^{r}, \quad r = 1, 2, \ldots$$

*Estimated the RID for incidence data*

Similar derivation steps as those for abundance data lead to the following adjusted estimator for a detected species:
\[ \widetilde{\pi}_i = \frac{Y_i}{T} (1 - \lambda e^{-\theta Y_i}), \quad Y_i > 0. \]  
(D.5)

Since an unbiased estimator for \( \sum_{i=1}^{S} \pi_i \) (the denominator of \( ^1C^* \)) is \( \sum_{Y_{i>1}} Y_i / T \), and an unbiased estimator for \( \sum_{i=1}^{S} \pi_i^2 \) (the denominator of \( ^2C^* \)) is \( \sum_{Y_{i>2}} Y_i(Y_i - 1) / [T(T - 1)] \), the two parameters \( \lambda \) and \( \theta \) satisfy the following two equations:

\[ \sum_{i \in \text{detected}} \pi_i \approx \sum_{Y_{i>1}} (Y_i / T)(1 - \lambda e^{-\theta Y_i}) = \hat{\lambda}^* \sum_{Y_{i>1}} Y_i / T; \]  
(D.6)

\[ \sum_{i \in \text{detected}} \pi_i^2 \approx \sum_{Y_{i>2}} [(Y_i / T)(1 - \lambda e^{-\theta Y_i})]^2 = \hat{\beta}^* \sum_{Y_{i>2}} Y_i(Y_i - 1) / T(T - 1). \]  
(D.7)

Any software can be readily used to obtain the solution \( \hat{\lambda} \) and \( \hat{\beta} \) in the above system of nonlinear equations. If the solution \( \hat{\beta} \) is out of the range of \( [0, 1] \), we replace it by 1 so that the model reduces to the one-parameter case. A bootstrap method as the one we proposed for abundance data in the main text can be similarly applied to assess the sampling variance of the estimator \( \widetilde{\pi}_i \) and construct the associated confidence interval of \( \pi_i \).

Based on the Chao2 lower bound or estimator (Chao, 1987), we have the following estimator for the number of the undetected species in \( T \) sampling units:

\[
\hat{Q}_0 = \begin{cases} 
\frac{(T - 1) Q_1^2}{T}, & \text{if } Q_2 > 0; \\
\frac{(T - 1) Q_1 (Q_1 - 1)}{2}, & \text{if } Q_2 = 0.
\end{cases}
\]  
(D.8)

Assuming a geometric series for the incidence probabilities for the undetected species, i.e., \( \pi_i = \alpha \beta_i^i, \ i = 1, 2, ..., \hat{Q}_0 \), we obtain the following two equations in terms of two parameters \( \alpha \) and \( \beta \) for the undetected species:

\[ \sum_{i \in \text{detected}} \pi_i \approx \sum_{i=1}^{\hat{Q}_0} \alpha \beta_i^i = \hat{\alpha}^* \sum_{Y_{i>1}} Y_i / T; \]  
(D.9)

\[ \sum_{i \in \text{detected}} \pi_i^2 \approx \sum_{i=1}^{\hat{Q}_0} (\alpha \beta_i^i)^2 = \hat{\beta}^* \sum_{Y_{i>2}} Y_i(Y_i - 1) / T(T - 1). \]  
(D.10)

We can solve \( \hat{\alpha} \) and \( \hat{\beta} \) by the above system of nonlinear equations. Therefore, the proposed estimated relative abundances for the undetected species are

\[ \widetilde{\pi}_i = \hat{\alpha} \beta_i^i, \ i = 1, 2, ..., \hat{Q}_0. \]  
(D.11)

Combining the adjustment method for detected species in Eq. D.5 and the estimated relative abundances for undetected species in Eq. D.11, we can construct a complete RAD based on \( T \) sampling units.

**Soil ciliates example**
The incidence frequency counts for the ciliates data described in the main text are summarized in the following table (Foissner et al. 2005):

<table>
<thead>
<tr>
<th>k</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q_k</td>
<td>150</td>
<td>53</td>
<td>42</td>
<td>18</td>
<td>12</td>
<td>9</td>
<td>10</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>k</th>
<th>17</th>
<th>19</th>
<th>20</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>26</th>
<th>27</th>
<th>29</th>
<th>32</th>
<th>33</th>
<th>34</th>
<th>35</th>
<th>37</th>
<th>39</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q_k</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The estimation procedures are generally parallel to those presented for the abundance data, as illustrated in Fig. 2 of the main text. We omit the details. The first- and second-order sample coverage estimates are respectively 88.45% and 99.38%; the corresponding coverage deficit estimates are thus respectively 11.55% and 0.62%. Also, we have \( \sum_{i=1}^{T} Y_i / T = 1281/51 = 25.12 \) (the average number of incidences per sampling unit) and \( \sum_{i=2}^{T} Y_i (Y_i - 1) / [T(T - 1)] = 6.1082 \).

Based on these statistics and data, the numerical solution for Eqs. D.6 and D.7 are \( \hat{\lambda} = 0.3264 \) and \( \hat{\theta} = 0.1528 \); and the numerical solution for Eqs. D.9 and D.10 are \( \hat{\alpha} = 0.0139 \), and \( \hat{\beta} = 0.999934 \); see the main text for more analysis. The estimated decay factor \( \hat{\beta} \) is close to unity, so the estimated relative abundances for the 209 undetected species (Fig. 3 in the main text) differ little.

LITERATURE CITED


Foissner, W., Agatha, S., and Berger, H. 2002. Soil ciliates (Protozoa, Ciliophora) from Namibia (Southwest Africa), with emphasis on two contrasting environments, the Etosha region and the Namib Desert. Denisia 5:1-1459.
APPENDIX E: Estimating diversity profile based on the proposed RAD estimator (for abundance data).

With the estimated RAD, we can infer many parameters of the focal assemblage. Here we present the estimation of diversity (Hill numbers) profile. Hill (1973) integrated species richness and species abundances into a class of diversity measures later called Hill numbers, or the effective numbers of species, defined as

\[ q D = \left( \sum_{i=1}^{S} p_i^q \right)^{1/(1-q)} , \quad q \neq 1. \]  

(E.1)

The parameter \( q \) determines the sensitivity of the measure to the relative abundances. When \( q = 0 \), \( qD \) is simply species richness. For \( q = 1 \), Eq. E.1 is undefined, but its limit as \( q \) tends to 1 is the exponential of the familiar Shannon index, referred to as Shannon diversity (Chao et al. 2014). When \( q = 2 \), Eq. E.1 yields Simpson diversity, the inverse of the Simpson concentration. A complete characterization of the abundance-based species diversity of an assemblage is conveyed by a diversity profile – a plot of \( qD \) versus \( q \) – from \( q = 0 \) to \( q = 3 \) or 4 (beyond 3 or 4 it generally changes little). The diversity profile also completely characterizes the RAD itself.

It is well known that the empirical diversity profile obtained by substituting sample relative abundances (the “plug-in estimator”) into Eq. E.1 underestimates the true profile especially for \( q \leq 1 \). Previous approaches to correcting for this negative bias were proposed only for non-negative integers \( q = 0, 1 \) and 2 (see Gotelli and Chao 2013 for a review). Now, with our proposed RAD estimator, we can provide a bias-corrected diversity profile for all \( q \geq 0 \), including non-integer values of \( q \), by Eqs. 4d and 6c of the main text.

The variance of the resulting diversity estimator and the associated confidence intervals for the diversity of order \( q \) can also be constructed by a bootstrap method based on sampling with replacement from the estimated RAD. The bootstrap procedures are similar to those presented in the main text in the section Sampling variances of our estimators. First, a random sample of \( n \) individuals is generated from the estimated RAD with replacement to obtain a new set of species sample abundances. Based on this new set, we then calculate new coverage estimates (\( \hat{\lambda}, \hat{\theta} \)) and their deficits, a new estimated number of undetected species, and new estimates (\( \hat{\alpha}, \hat{\beta} \)); all these new statistics are then substituted into Eqs. 4d and 6c to obtain a new RAD estimator. That is, all statistics in our RAD estimator are replaced by those computed from the new generated set of species sample abundances. Then a bootstrap diversity estimate \( \hat{q} \hat{D} \), based on the estimated RAD is calculated. The procedure is replicated \( B \) times to obtain \( B \) bootstrap diversity estimates \( \{\hat{q} \hat{D}^1, \hat{q} \hat{D}^2, ..., \hat{q} \hat{D}^B\} (B = 1,000 \) is suggested in confidence interval construction). The bootstrap variance estimator of \( \hat{q} \hat{D} \) is the sample variance of these \( B \) estimates. Moreover, the 2.5% and 97.5% percentiles of these \( B \) bootstrap estimates can be used to construct a 95% confidence interval. Similar procedures can be used to derive variance estimators for any other estimator (e.g., empirical diversity)
and their associated confidence intervals (see the example below).

**Example (abundance data)**

In Fig. E1, we show the diversity profile based on the empirical RAD and the estimated RAD respectively for the spider data discussed in the main text (26 detected species and an estimate of 18 undetected species, Sackett et al. 2011). The associated 95% confidence intervals using 1,000 bootstrap replications are also shown. For any order $q$, $0 \leq q \leq 3$, the estimated diversity of order $q$ is higher than the corresponding empirical diversity. When $q \leq 1$, substantial differences exist between the two profiles; the difference arises because in our approach the contribution from undetected species is included whereas the empirical estimates ignore the undetected species. We later use simulation to show the estimated diversity profile based on the proposed RAD estimator eliminates most of the negative bias associated with the empirical diversity profile.

![Comparison of the empirical diversity profile (blue dashed line) and estimated diversity profile (red solid line) as a function of order $q$, $0 \leq q \leq 3$, for the abundance data of forest spiders (Sackett et al. 2011). The shaded areas denote the 95% confidence intervals based on 1000 bootstrap replications from the estimated RAD.](image)

**Fig. E1.** Comparison of the empirical diversity profile (blue dashed line) and estimated diversity profile (red solid line) as a function of order $q$, $0 \leq q \leq 3$, for the abundance data of forest spiders (Sackett et al. 2011). The shaded areas denote the 95% confidence intervals based on 1000 bootstrap replications from the estimated RAD.

**Simulation comparisons of empirical diversity profiles and the proposed diversity profiles**

Based on the same scenarios used in Appendix A, we examine in Fig. E2 the diversity profiles based on the estimated RAD. For each generated data set under a particular scenario, we compare the true diversity profile of the complete assemblage, the empirical diversity profile, and the estimated diversity profile for $0 \leq q \leq 3$. All empirical diversity profiles are much below the true curves and thus exhibit negative biases, especially for $q \leq 1$. The biases
are substantial for \( n = 200 \). Our estimated profile eliminates much of the bias and produces nearly unbiased diversity estimates (which fluctuate below and above the true profile) when \( q \) is not close to 0. Note that when \( q \) is equal to 0, our estimate is the Chao1 estimator, which is an estimated lower bound of the true species richness.

In the literature, the most widely used measures under the framework of Hill numbers are the species richness (\( q = 0 \)), Shannon diversity (\( q = 1 \)) and Simpson diversity (\( q = 2 \)). For \( q = 0 \), the empirical diversity is the observed species richness; its bias is thus the number of undetected species in the sample. This bias-correction issue is a well known subject in many disciplines; see Chao and Chiu (2012) for a review. For \( q = 1 \), the empirical diversity is the observed exponential of Shannon entropy, which also exhibits substantial negative bias in hyper-diverse assemblages; see Chao et al. (2013). Our simulation results (Fig. E2) reveal that the diversity estimates based on the proposed RAD perform much better than the empirical diversity measures. However, our estimation procedures require solving nonlinear equations (Eqs. 4b, 4c and Eqs. 6a, 6b) and occasionally the solutions may not lie in the expected range, as discussed after Eq. 4c in the main text. We are currently working on analytic estimators of diversity and entropy profiles.
**Fig. E2.** Comparison of the true diversity profile of the complete assemblage (light blue line in each panel), the empirical diversity profiles (superimposed dark blue lines with 200 replications) and the estimated diversity profiles based on the proposed RAD estimator (superimposed red lines with 200 replications) for sample size 200 (left panels), 400 (middle panels) and 800 (right panels). Data sets were generated from two theoretical abundance distributions (the Zipf-Mandelbrot model and the log-normal model) and four plant assemblages (see Appendix A for data details). For each assemblage and each sample size, 200 data sets were generated, thus there are 200 estimated profiles (200 red lines and 200 dark blue lines). Note that the X-axis is the order of Hill numbers, while the Y-axis is in units of “effective number of species”.

---

**Table 5.** The number of species in each sample set.
LITERATURE CITED


