Ecology, 90(4), 2009, pp. 1125–1133

Sufficient sampling for asymptotic minimum species richness estimators

Anne Chao,[1] Robert K. Colwell,[2] Chih-Wei Lin,[1] and Nicholas J. Gotelli[3,4]

[1]*Institute of Statistics, National Tsing Hua University, Hsin-Chu, 30043 Taiwan*
[2]*Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut 06269 USA*
[3]*Department of Biology, University of Vermont, Burlington, Vermont 05405 USA*

Abstract. Biodiversity sampling is labor intensive, and a substantial fraction of a biota is often represented by species of very low abundance, which typically remain undetected by biodiversity surveys. Statistical methods are widely used to estimate the asymptotic number of species present, including species not yet detected. Additional sampling is required to detect and identify these species, but richness estimators do not indicate how much sampling effort (additional individuals or samples) would be necessary to reach the asymptote of the species accumulation curve. Here we develop the first statistically rigorous nonparametric method for estimating the minimum number of additional individuals, samples, or sampling area required to detect any arbitrary proportion (including 100%) of the estimated asymptotic species richness. The method uses the Chao1 and Chao2 nonparametric estimators of asymptotic richness, which are based on the frequencies of rare species in the original sampling data. To evaluate the performance of the proposed method, we randomly subsampled individuals or quadrats from two large biodiversity inventories (light trap captures of Lepidoptera in Great Britain and censuses of woody plants on Barro Colorado Island [BCI], Panama). The simulation results suggest that the method performs well but is slightly conservative for small sample sizes. Analyses of the BCI results suggest that the method is robust to nonindependence arising from small-scale spatial aggregation of species occurrences. When the method was applied to seven published biodiversity data sets, the additional sampling effort necessary to capture all the estimated species ranged from 1.05 to 10.67 times the original sample (median $\approx$ 2.23). Substantially less effort is needed to detect 90% of the species (0.33–1.10 times the original effort; median $\approx$ 0.80). An Excel spreadsheet tool is provided for calculating necessary sampling effort for either abundance data or replicated incidence data.

Key words: asymptotic species richness estimators; biodiversity sampling; sample size; Turing's frequency formula.

Introduction

Estimating species richness, a central activity in studies of biodiversity (Magurran 2004), presents a statistical challenge because it is rarely possible to collect enough individuals or samples to discover all the species that are present (Gotelli and Colwell 2001). Nevertheless, asymptotic species richness can be estimated statistically from a single random sample of individuals (abundance data) or from a collection of random samples in which only species occurrences are recorded (incidence data). Three methods for estimating species richness are (1) fitting a statistical distribution to rank abundance data, (2) extrapolating a species accumulation curve to its asymptote, and (3) estimating the asymptotic number of species with nonparametric estimators (Longino et al. 2002).

Manuscript received 31 December 2007; revised 12 June 2008; accepted 15 July 2008; final version received 12 August 2008. Corresponding Editor: F. He.

[4] Corresponding author. E-mail: ngotelli@uvm.edu

For individual-based (abundance) data, the area under a fitted, lognormal abundance distribution has been used to estimate the total number of species, including undetected rare species that are hidden by a "veil line" of incomplete sampling (Preston 1962). Other species abundance models such as the log-series, geometric, negative binomial, Zipf-Mandelbrot, and the broken-stick (Magurran 2004) can also be fit to abundance data to estimate asymptotic species richness. Curve-fitting methods, which can be applied to both abundance data and incidence data, extrapolate a fitted function such as the Michaelis-Menten equation or a mixture model out to the asymptote of the species-accumulation graph (Soberón and Llorente 1993, Colwell et al. 2004). A variety of nonparametric estimators can also be used to estimate total species number from either abundance or incidence data (Chao 2005). Nonparametric estimators, which are based on frequency counts, use information on the number of rare or infrequent species in the collection to estimate the number of undetected species.

All of these methods generate estimates of asymptotic species richness, and many also generate variances and confidence intervals about the estimates. Connolly et al. (2005) provided a formula for estimating how much sampling is required to unveil the parametric lognormal abundance distribution. However, no models to date have included a nonparametric estimate of the sampling effort (number of individuals or samples) that would be necessary to reach the asymptote of a species accumulation curve by actually detecting all species present. Because biodiversity sampling is labor intensive (Longino and Colwell 1997, Lawton et al. 1998), such an estimate of sampling effort is of great interest for effective planning of biotic inventories.

In this paper, we derive estimators for the sampling effort required to reach the asymptotic richness estimated by Chao1 and Chao2, two widely used nonparametric estimators of species richness for abundance and incidence data, respectively. The relatively simple solutions are based on a derivation by the founder of modern computer science, Alan Turing, who used it in cryptographic analyses during World War II. We provide an Excel spreadsheet macro for performing the calculations, and we present estimates of complete sampling effort for several published biodiversity surveys. Simulation results based on data sets from two large biodiversity inventories demonstrate the robustness of the proposed method to departures from some of the sampling assumptions.

Sample Sizes for Asymptotic Estimators

Abundance data

Assume that there are S species in a target biological community or assemblage. A random sample of n individuals is selected (*with* replacement) from the community. A lower bound of species richness is obtained as

$$S_{\text{obs}} + (1 - 1/n)f_1^2/(2f_2) \quad (1)$$

where S_{obs} is the number of species observed, and f_r is the count (frequency) of species that are observed exactly r times in the sample (Chao 1989). Thus, f_1 is the number of "singletons," or species represented by exactly one individual in the sample; f_2 is the number of "doubletons," or species represented by exactly two individuals in the sample; and f_0 is the unknown number of species that are present in the community but not detected by the sample and therefore each have zero individuals in the sample. Because the sample size n is often large, we can ignore the term $(1 - 1/n)$ in Eq. 1 and obtain the following Chao1 estimator (S_{est}) for species richness if $f_2 > 0$:

$$S_{\text{est}} = S_{\text{obs}} + f_1^2/(2f_2). \quad (2)$$

In this equation, $\hat{f}_0 = f_1^2/(2f_2)$ is an estimator for the number of species present but undetected in the sample. The Chao1 estimator represents a universal lower bound in the sense that it is valid under all types of species abundance distribution. Thus, all estimated sampling effort derived in this paper represents *minimum* effort. If $f_2 = 0$, the Chao1 estimator is replaced by $S_{\text{est}} = S_{\text{obs}} + f_1(f_1 - 1)/[2(f_2 + 1)]$ (Chao 2005).

Even before we derive a formal result, Eq. 2 already provides a heuristic "stopping rule" for biodiversity sampling: Sampling will be complete when every species is represented by at least two individuals (no singletons), so that $\hat{f}_0 = 0$ and $S_{\text{est}} = S_{\text{obs}}$. No additional sampling effort is needed once this condition is satisfied, as there are no additional undetected species. Because $\hat{f}_0 = f_1^2/(2f_2)$ may not be an integer, the condition $\hat{f}_0 = 0$ in data analysis is modified to $\hat{f}_0 < 0.5$. That is, when there are fewer than 0.5 species remaining undetected, the sampling is deemed complete and no additional effort is needed. When $\hat{f}_0 \geq 0.5$, the problem is to estimate the additional number of individuals needed to observe the remaining, undetected species. Applying the above stopping rule for completeness, sampling should continue until singletons vanish. As we will see, this may require a very large additional sample size, because by the time the total sampling effort is extensive enough to reveal two individuals of each species found only once in the original sample, single individuals of additional species will have inevitably surfaced. For hyperdiverse communities with a large proportion of very rare species, the challenge of estimating richness from sample data is daunting (Mao and Colwell 2005).

According to Good (1953, 2000), Alan Turing studied aspects of this problem in the context of deciphering encoded messages intercepted from the German military during World War II. Assume that an original sample of size n is available. Turing (and others) proved that, for the next individual sampled, the probability of encountering each of the f_r species in frequency class r, $r = 0, 1, \ldots$ is approximately

$$(r + 1)f_{r+1}/(nf_r). \quad (3)$$

As a special case, the probability of encountering *each* of the undetected species ($r = 0$) is thus $f_1/(nf_0)$. Because there are f_0 species in the frequency class $r = 0$, the probability q_0 that the next individual sampled represents a previously undetected species can be estimated by

$$q_0 = f_0 \times f_1/(nf_0) = f_1/n. \quad (4)$$

Good (1953, 2000) interpreted Eq. 3 in the following way: the relative abundance (or discovery probability) of any species in the frequency class r is approximately $(r + 1)f_{r+1}/(nf_r)$. A remarkable implication for $r = 0$ is that the relative abundance for each undetected species is roughly $f_1/(nf_0)$, and the total relative abundances of the undetected species can thus be accurately estimated by $q_0 = f_1/n$. The usual maximum likelihood estimate (MLE) for the relative abundance of any undetected species is 0, but this estimate is obviously not reasonable when sampling is incomplete. Eq. 4 is a special case of the nonparametric empirical Bayes method, in which the

estimator is "updated" by information contained in the sample (the proportion of singletons), but the functional form of the prior distribution is not assumed (Good 2000).

For our purposes, the information in the original sample will be used to predict the minimum number m of additional observations (individuals) needed to achieve the following version of the stopping rule: There are no singletons in the enlarged sample of size $n + m$. Equivalently, the *expected* number of singletons in the enlarged sample of size $n+m$ is less than 0.5 (because the theoretical expected value may not be an integer).

Note that the singletons in the enlarged sample of size $n + m$ include two groups of species: (1) any species represented by only a singleton in the original sample for which no additional individuals are detected by the enlarged sample, and (2) any species not detected in the original sample for which exactly one individual is observed in the additional sampling. Let p_i be the relative abundance (or discovery probability) of the ith species in the community and X_i be the number of individuals of the ith species observed in the original sample. Then the expected number of species in the first group is

$$\sum_{i=1}^{S}(1-p_i)^m I(X_i=1) \tag{5}$$

where $I(\cdot)$ is an indicator function that equals 1 when true and 0 otherwise, meaning in this case that only singletons ($X_i = 1$) in the original sample contribute to the sum. The term $(1 - p_i)^m$ denotes the probability that the ith species is not observed in any of the m additional observations. Similarly, the expected number of species in the second group is

$$\sum_{i=1}^{S} mp_i(1-p_i)^{m-1} I(X_i=0) \tag{6}$$

where the term $mp_i(1 - p_i)^{m-1}$ denotes the probability that the ith species is observed only once in the m additional observations. In this case, only undetected species ($X_i = 0$) in the original sample contribute to the sum. Based on the information in the original sample data, the expected number of singletons in the enlarged sample with size $n + m$ is

$$\sum_{i=1}^{S}(1-p_i)^m I(X_i=1)+\sum_{i=1}^{S} mp_i(1-p_i)^{m-1} I(X_i=0). \tag{7}$$

Applying Turing's formula to the first sum in Eq. 7 and using Eq. 3 with $r = 1$, we have that the relative abundance p_i for a singleton (i.e., $X_i = 1$) is approximately equal to $2f_2/(nf_1)$. Thus, a first-order approximation for the number of singletons represented by first sum in Eq. 7 can be found by substituting p_i in this sum by $2f_2/(nf_1)$, noting that there are f_1 of them:

$$\sum_{i=1}^{S}(1-p_i)^m I(X_i=1) \approx f_1\left(1-\frac{2f_2}{nf_1}\right)^m \approx f_1\exp\left[-\frac{m}{n}\left(\frac{2f_2}{f_1}\right)\right]. \tag{8}$$

Similarly, for the second sum in Eq. 7, the relative abundance for an undetected species is approximately $f_1/(nf_0)$, using Eq. 3 with $r = 0$. Thus the second sum in Eq. 7 can be approximated by

$$\sum_{i=1}^{S} mp_i(1-p_i)^{m-1} I(X_i=0) \approx f_0 m\left(\frac{f_1}{nf_0}\right)\left(1-\frac{f_1}{nf_0}\right)^{m-1} \approx f_1\frac{m}{n}\exp\left[-\frac{m}{n}\left(\frac{f_1}{f_0}\right)\right]. \tag{9}$$

Note that f_0 appears in Eq. 9, indicating we have to provide an estimate for the number of undetected species. From Eq. 2, we substitute $\hat{f}_0 = f_1^2/(2f_2)$ into Eq. 9 as an estimate of the number of undetected species and obtain

$$f_1\frac{m}{n}\exp\left[-\frac{m}{n}\left(\frac{f_1}{\hat{f}_0}\right)\right] = f_1\frac{m}{n}\exp\left[-\frac{m}{n}\left(\frac{2f_2}{f_1}\right)\right]. \tag{9a}$$

Combining Eq. 8 and Eq. 9a, the approximate number of singletons in the enlarged sample with size $n + m$ is (letting $x = m/n$, the ratio between the additional and original sample sizes) $f_1(1 + x)\exp[-x(2f_2/f_1)]$, which is less than 0.5 if and only if

$$2f_1(1+x) < \exp\left[x\left(\frac{2f_2}{f_1}\right)\right]. \tag{10}$$

The function $h(x) = 2\ f_1(1 + x)$ for $x > 0$ is a linear function of x, whereas $v(x) = \exp[x(2f_2/f_1)]$ is an exponentially increasing function of x (see the figure in the Excel spreadsheet calculator in the Appendix). To estimate m, if we first find the solution x^* for the equation $2f_1(1 + x) = \exp[x(2f_2/f_1)]$, then the minimum required additional number of *individuals* is $m = nx^*$. A bootstrap percentile method (Efron and Tibshirani 1993:170) described in the Supplement can be used to construct a lower confidence limit.

In many cases, the sampling effort required to reach the asymptote may be prohibitively large (as we later show for empirical data sets). However, a large fraction of the S_{est} may be reached with considerably less sampling. If g is the fraction of S_{est} that is desired ($0 < g < 1$), then the objective is to find the additional sample size m_g such that the number of species reaches the target value gS_{est}, i.e., the number of previously undetected species discovered in the additional sample is $gS_{\text{est}} - S_{\text{obs}}$. In order to make the requirement sensible, the target number must be greater than the observed number in the current sample, so we must require that $g > S_{\text{obs}}/S_{\text{est}}$. We can apply Eq. 10 in Shen et al. (2003) to predict that the number of previously undetected

species in the additional sample of size m_g approximately equals

$$\hat{f}_0\left\{1-\exp\left[-\frac{m_g}{n}\left(\frac{2f_2}{f_1}\right)\right]\right\}. \quad (11)$$

From Eq. 11, the additional number of individuals needed to detect a fraction g of S_{est} is obtained from the following equation:

$$\hat{f}_0\left\{1-\exp\left[-\frac{m_g}{n}\left(\frac{2f_2}{f_1}\right)\right]\right\}=gS_{\text{est}}-S_{\text{obs}}.$$

This gives the following solution:

$$m_g \approx \frac{nf_1}{2f_2}\log\left[\frac{\hat{f}_0}{(1-g)S_{\text{est}}}\right]. \quad (12)$$

Incidence data

In most biodiversity studies, individual organisms are not sampled randomly and independently, as required by our sampling model and by most statistical models for biodiversity estimation. Instead, multiple individuals are collected or censused in traps, baits, quadrats, plots, or timed surveys. It is these sampling units, and not the individual organisms, that are actually sampled randomly and independently. For very abundant organisms (such as microbes), or taxa with clonal growth forms (such as many plants and invertebrates), it may not even be possible to count individuals within each sampling unit, and only their presence or incidence can be recorded. However, estimation is still possible for a set of replicated samples in which the incidence of each species is recorded in the sample.

When applied to incidence data based on t replicated samples, let Q_1 and Q_2 represent the number of species that occur in exactly one sample ("uniques") or in exactly two samples ("duplicates"), respectively (Colwell and Coddington 1994). For replicated incidence data, the estimator of species richness, known as Chao2, incorporates a correction for small sample size

$$S_{\text{est}}=S_{\text{obs}}+(1-1/t)Q_1^2/(2Q_2). \quad (13)$$

Parallel arguments and results allow estimation of m (for incidence data, m is the number of *samples* needed to achieve $S_{\text{est}}=S_{\text{obs}}$, and m_g is the number of *samples* needed to achieve gS_{est}). (Details of the derivation are provided in the Supplement.) For replicated incidence data, the probability q_0 that the next *incidence* (the next species collected, regardless of its abundance) represents a previously undetected species is $q_0=Q_1/T$, where $T=\sum_{i=1}^{t} iQ_i$ denotes the total number of incidences in t samples. Thus, q_0 also represents the proportion of previously undetected species in an additional sample. The additional number of samples needed to reach the asymptotic Chao2 estimate is equal to $m=tx^*$, where x^* is the solution of the following equation:

$$2Q_1(1+x)=\exp\left[x\frac{2Q_2}{(1-1/t)Q_1+2Q_2/t}\right]. \quad (14)$$

To reach a fraction g of S_{est} for sample-based data, the required number of additional *samples* is

$$m_g \approx \frac{\log\left[1-\frac{t}{(t-1)}\frac{2Q_2}{Q_1^2}(gS_{\text{est}}-S_{\text{obs}})\right]}{\log\left[1-\frac{2Q_2}{(t-1)Q_1+2Q_2}\right]}. \quad (15)$$

Empirical Examples

Table 1 illustrates the calculation of m and m_g ($g=0.95$ and 0.90) for four examples from the literature for abundance-based data. Table 2 shows three examples of incidence-based data. The estimates for m (for $g=1.00$) vary considerably, ranging from 1.05 times more data than the original sample (tropical rain forest tree seedlings, Butler and Chazdon 1998) to 10.67 times (forest Lepidoptera; Fisher et al. 1943). Substantially less effort is needed to detect 95% or 90% of the species. The probability that the next individual discovered represents a previously undetected species (for abundance data) or that the next sample includes a previously undetected species (for replicated incidence data) varies accordingly from 0.0004 to 0.0556, although these values are a nonlinear function of sampling effort.

Although these examples include both abundance-based and incidence-based data, in reality all of the data sets in Tables 1 and 2 represent some form of sample-based collection, because individual organisms are not randomly sampled. For example, the Fisher et al. (1943) light trap data consist of pooled records of light traps taken over four years, and the Dahlberg and Odum (1970) fish data represent multiple trawls, pooled within a season and across habitats. Both of these data sets contain a large number of singletons, and it is possible that some of this rarity reflects the pooling of heterogeneous samples (Longino et al. 2002). The Cunningham et al. (2002) data are also pooled pitfall collections, although these were taken within a single habitat type. The robustness of our method to the dependence of sampled individuals will be shown in *Simulation analyses*.

Similarly, two of the incidence data sets (Maudsley et al. 2002, Ellison et al. 2007) are themselves pooled collections of pitfall traps, litter samples, and other methods that were used within a single plot, which is considered the sampling unit for these analyses. Each sample in the seedbank data of Butler and Chazdon (1998), in contrast, represented a single soil sample.

When incidence-based methods are used with sample plots of fixed area, we can estimate the total area that would be needed to collect all of the species. Ellison et al. (2007) used a variety of standardized sampling methods (hand collecting, litter samples, pitfall traps, and baits) to sample ant occurrences in 18 75 × 75-m plots of red oak forest in Blackrock, New York, USA. Combining all sampling methods, 33 species were collected. At the

TABLE 1. Examples of estimated sampling effort for abundance-based data.

Habitat, taxon, and locale	n	S_{obs}	S_{est}	f_1	f_2	q_0	$g = 1$	$g = 0.95$	$g = 0.90$	Source
Forest Lepidoptera at light traps (UK)	15 609	240	296	35	11	0.0022	166 509	32 930	15 718	Table 3 in Fisher et al. (1943)
Estuarine fish in trawls (USA)	31 637	70	90	14	5	0.0004	243 369	65 371	34 670	Table 1 in Dahlberg and Odum (1970) (summarized in Magurran [2004:220])
Forest lizards in pitfall traps (USA)	161	9	11	3	2	0.0186	358	167	84	Table 5 in Cunningham et al. (2002)
Tropical rain forest tree seedlings (Costa Rica)	952	34	35	2	2	0.0021	1002	†	†	Butler and Chazdon (1998)

Notes: Each row represents a different data set from the literature. Abbreviations are: n, number of individuals collected; S_{obs}, observed species richness; S_{est}, estimated asymptotic species richness, based on the Chao1 estimator; f_1, the number of species represented by exactly one individual ("singletons"); f_2, the number of species represented by exactly two individuals ("doubletons"); q_0, the probability that the next individual sampled represents a previously undetected species, estimated as f_1/n; g, target fraction of S_{est} that is to be reached. The entries in each "g" column represent the number of additional individuals needed to reach 100% ($g = 1$), 95% ($g = 0.95$), and 90% ($g = 0.90$), respectively, of S_{est}.

† For this case, g must be greater than 0.971 because of the restriction $gS_{est} > S_{obs}$.

plot level, the data set contains eight uniques (species with one occurrence) and five duplicates (species with two occurrences). The estimated total number of species is 39, which would require an additional 62 randomly selected plots to achieve (Table 2). Because the area of each plot is 5625 m^2, we estimate a total of $(18 + 62) \times 5625 = 393\,750$ $m^2 = 39.4$ ha as a minimum area that would be needed to detect all of the ant species present. In contrast, the total area of comparable habitat in the Blackrock reserve is 1092 ha (B. Schuster, *personal communication*), so the necessary sampling represents ~3.6% of the total habitat area.

SIMULATION ANALYSES

We conducted a simulation experiment to examine the performance of the proposed methods. We treated the data from each of several large biodiversity surveys/censuses as the true community, generated subsamples from it, and then used our method to estimate how much sampling would be needed to reach a fraction (including 100%) of the estimated asymptote. Here we report only three representative cases because the findings were consistent with those from other data sets. For abundance data, we analyzed Lepidoptera records at light traps in Britain (Fisher et al. 1943; the first example in Table 1). This data set included 15 609 individuals representing 240 species. For incidence data, we analyzed two quadrat sizes (50 × 50 m and 25 × 25 m) from the 50-ha Barro Colorado Island (BCI), Panama, 1985 tree census (Hubbell et al. 2005). This data set included 238 018 individual trees and shrubs (≥1 cm in diameter at breast height) representing 299 species. Because the spatial location of each stem in the BCI census was recorded, we could test the robustness of our method to spatial aggregation of individuals within and among species, which is known to be present in these data.

For each data set, we considered a range of 14 subsample sizes. We generated 1000 random subsamples for each fixed subsample size. Note here that some subsample sizes were allowed to exceed the total number of individuals in the inventory or census because our sampling was conducted *with* replacement. For each subsampling level in each data set, we first calculated S_{est}, the estimated asymptotic richness, using Chao1 (for abundance data) or Chao2 (for incidence data), and then computed the required additional sample size to reach

TABLE 2. Examples of estimated sampling effort for incidence-based data.

Habitat, taxon, and locale	t	T	S_{obs}	S_{est}	Q_1	Q_2	q_0	$g = 1$	$g = 0.95$	$g = 0.90$	Source
Hedgerow carabid beetles in soil, litter, and vegetation samples (UK)	16	72	20	22	4	5	0.0556	20	2	†	Appendix A in Maudsley et al. (2002)
Temperate forest ants in pitfall traps, bait, litter, and hand collections (USA)	18	208	33	39	8	5	0.0385	62	16	6	Ellison et al. (2007)
Tropical rain forest tree seedlings (Costa Rica)	121	461	34	36	3	2	0.0065	270	19	‡	Butler and Chazdon (1998)

Notes: Abbreviations are: t, number of samples collected; T, $\Sigma_{i=1}^{t} iQ_i$ = total number of incidences; S_{obs}, observed species richness; S_{est}, estimated asymptotic species richness, based on the Chao2 estimator; Q_1, the number of species represented by exactly one sample ("uniques"); Q_2, the number of species represented by exactly two samples ("duplicates"); q_0, the probability that the next observed sample contains a species new to the survey (i.e., the proportion of species in the next sample that are new to the survey), estimated as Q_1/T; g, target fraction of S_{est} that is to be reached. The entries in each "g" column represent the number of additional samples needed to reach 100% ($g = 1$), 95% ($g = 0.95$), and 90% ($g = 0.90$), respectively, of S_{est}.

† For this case, g must be greater than 0.930 because of the restriction $gS_{est} > S_{obs}$.

‡ For this case, g must be greater than 0.938 because of the restriction $gS_{est} > S_{obs}$.

TABLE 3. Simulation results based on the Fisher et al. (1943) Lepidoptera abundance data set (15 609 individuals, 240 species).

		$g = 1$				$g = 0.95$			
n	S_{obs}	Target S_{est}	Estimated m	Achieved S_{est}	Achieved g	Target gS_{est}	Estimated m_g	Achieved gS_{est}	Achieved g
750	119.3	165.6	5378	198.7	1.21	157.3	1415	161.1	0.98
1500	148.3	188.4	9236	214.6	1.15	179.0	2098	180.9	0.96
3000	175.4	209.4	16 593	227.4	1.09	199.0	3166	199.1	0.95
5000	193.8	222.3	24 888	233.4	1.05	211.1	3935	209.7	0.95
6000	199.7	225.4	28 017	234.9	1.04	214.1	3966	212.5	0.95
7000	204.4	227.7	30 980	235.9	1.04	216.3	3924	215.1	0.95
8000	208.6	229.8	33 266	236.6	1.03	218.3	3710	217.2	0.95
10 000	214.4	233.1	39 637	237.8	1.02	221.5	3681	220.5	0.95
15 000	223.7	237.6	51 914	238.9	1.01	225.7	2944	226.2	0.95
20 000	228.9	239.5	60 372	239.3	1.00	227.5	†	†	†
50 000	238.6	240.6	57 931	239.7	1.00	228.6	†	†	†
100 000	239.9	240.1	15 209	240	1.00	228.1	†	†	†
200 000	240.0	240.0	0	240	1.00	228.0	†	†	†
300 000	240.0	240.0	0	240	1.00	228.0	†	†	†

Notes: Each row represents the average of 1000 simulation runs. Abbreviations are: n, size of the random subsample; S_{obs}, average number of species in the subsample; target S_{est}, average estimated number of species present in the assemblage, based on the subsample; g, target fraction of S_{est} that is to be reached; estimated m, average estimated additional number of individuals needed to be sampled to reach the target S_{est}; estimated m_g, average estimated additional number of individuals needed to be sampled to reach the target gS_{est} (calculated from Eq. 12); achieved S_{est}, the average number of species obtained when additional sampling effort is simulated; achieved g, achieved S_{est}/target S_{est}.

† In this case the average $gS_{est} < S_{obs}$, so no simulation was performed for this subsample size.

100% ($g = 1$) and 95% ($g = 0.95$) of S_{est}. Table 3 shows the average results (over 1000 subsample runs) for abundance data (Fisher's Lepidoptera data set) for $g = 1$ and $g = 0.95$. Table 4 shows the results for incidence data, based on the BCI tree data set, for $g = 1$ and $g = 0.95$, for 50×50 m quadrats. In Table 5, we compare analyses based on both abundance data and the corresponding incidence data for the BCI tree data set (for $g = 1$), for 25×25 m quadrats, in order to investigate the sensitivity of our method to spatial aggregation of individuals.

Ideally, we would compare the estimated additional sample size (as calculated from our equations) with the simulated sample size (which is obtained by continuing our simulated process until we reach the target). However, in some data sets, the estimate S_{est} may exceed the observed number of species in the full data set, so that the simulated size is not attainable (because we can never reach a species richness higher than the full observed species richness in the inventory or in the census). Therefore, we used an alternative metric: the achieved number of species (or equivalently, the

TABLE 4. Simulation results based on 50-ha Barro Colorado Island (BCI), Panama, incidence data (Hubbell et al. 2005; 200 50×50 m quadrats, 238 018 individuals, 299 species).

		$g = 1$				$g = 0.95$			
t	S_{obs}	Target S_{est}	Estimated m	Achieved S_{est}	Achieved g	Target gS_{est}	Estimated m_g	Achieved gS_{est}	Achieved g
5	191.2	219.2	25	248.8	1.14	208.2	3	208.7	0.95
7	204.1	230.7	33	256.4	1.11	219.2	4	219.3	0.95
10	216.8	241.5	46	264.3	1.10	229.5	5	229.3	0.95
20	238.1	262.5	102	280.0	1.07	249.4	11	248.4	0.95
40	256.7	279.3	206	290.3	1.04	265.4	19	264.4	0.95
60	266.7	287.1	291	294.0	1.03	272.8	23	272.1	0.95
80	272.8	290.7	358	295.5	1.02	276.1	22	276.6	0.95
100	277.9	294.1	423	296.7	1.01	279.4	21	280.6	0.95
200	289.5	298.8	625	298.2	1.00	283.9	†	†	†
300	293.9	299.4	687	298.4	1.00	284.4	†	†	†
500	297.3	299.7	750	298.5	1.00	284.8	†	†	†
1000	298.9	299.2	296	298.9	1.00	284.2	†	†	†
2000	299.0	299.0	4	299.0	1.00	284.1	†	†	†
3000	299.0	299.0	0	299.0	1.00	284.1	†	†	†

Notes: Each row represents the average of 1000 simulation runs. Abbreviations are: t, number of quadrats randomly selected; S_{obs}, average number of species in the subsample; target S_{est}, average estimated number of species present in the assemblage, based on the subsample; g, target fraction of S_{est} that is to be reached; estimated m, average estimated additional number of quadrats needed to be sampled to reach the target S_{est}; estimated m_g, average estimated additional number of quadrats (Eq. 15) needed to be sampled to reach the target gS_{est}; achieved S_{est}, the average number of species obtained when additional sampling effort is simulated; achieved g = achieved S_{est}/target S_{est}.

† In this case the average $gS_{est} < S_{obs}$, so no simulation was performed for this subsample size.

TABLE 5. Simulation results of random quadrat selection based on 50-ha BCI incidence data (Hubbell et al. 2005; 800 25 × 25 m quadrats, 238,018 individuals, 299 species).

		$g = 1$ (replicated incidence data)				$g = 1$ (abundance data, nonrandom sampling)				
t	S_{obs}	Target S_{est}	Estimated m (quadrats)	Achieved S_{est}	Achieved g	n (individuals)	Target S_{est}	Estimated m (individuals)	Achieved S_{est}	Achieved g
5	137.8	175.5	30	216.4	1.24	1488	175.2	9338	222.3	1.28
7	152.3	189.9	41	225.8	1.20	2081	190.2	12526	231.2	1.22
25	205.2	235.2	128	257.1	1.10	7432	230.6	35723	258.9	1.13
50	228.1	253.7	249	271.8	1.07	14 878	250.2	71 685	273.6	1.10
100	247.4	272.1	539	285.4	1.05	29 771	268.2	150 624	285.4	1.07
150	257.4	280.4	785	290.0	1.04	44 638	277.3	221 801	290.2	1.05
200	264.5	285.5	997	292.6	1.03	59 479	282.4	281 031	292.8	1.04
300	273.4	291.1	1365	295.4	1.02	89 242	288.7	383 625	295.1	1.02
500	282.7	296.3	2004	297.4	1.00	148 766	294.1	535 595	297.1	1.01
1000	292.2	299.5	2778	298.3	1.00	297 493	298.6	747 673	298.2	1.00
2000	297.4	299.7	2901	298.5	1.00	595 100	299.6	879 848	298.5	1.00
4000	298.9	299.2	1173	298.9	1.00	1 190 146	299.2	417 107	298.9	1.00
8000	299.0	299.0	0	299.0	1.00	2 380 191	299.0	1368	299	1.00
10 000	299.0	299.0	0	299.0	1.00	2 974 992	299.0	0	299	1.00

Notes: Each row represents the average of 1000 simulation runs. In the data, t quadrats were randomly selected from the whole area. Records in the selected quadrats were treated and analyzed either as replicated incidence data (left half of the table) or as the underlying abundance data (right half of the table). Terms are as defined in Tables 3 and 4.

achieved g) when the estimated additional sampling has been carried out in the simulation. Thus, if the estimator is performing well, for any fixed value of g (including $g = 1$), we should find the achieved species richness is very close to our target gS_{est} with an observed value of g very close to the anticipated g value.

We found that, as subsample size is increased and more information is collected, the estimated asymptotic target gS_{est} (including $g = 1$) increases accordingly (Tables 3, 4, and 5). Thus, the estimated additional sampling effort needed to reach the target gS_{est} initially increases with subsample size. This result reflects a general property of species accumulation curves: They typically have initially steep slopes because common species are quickly sampled, but their slopes decrease at large sample sizes because much greater effort is needed to sample the remaining rare species (Gotelli and Colwell 2001). For the target of complete sampling ($g = 1$), up to a critical point, as S_{est} is approaching the true species richness, the estimated effort starts to decline and eventually falls to 0. In all of our analyses, these critical points correspond to very large subsample sizes, implying that the search for rarest species requires substantial effort. However, if the target is set to be 95% of S_{est} ($g = 0.95$), then the additional effort needed is much less than the level required for complete sampling.

When subsample sizes are relatively small, our estimates are conservative in the sense that the required additional sampling effort is slightly overestimated, up to a maximum of 20%, as shown for the smallest samples in Tables 3, 4, and 5. When sample size is increased, based on the achieved values of g in these tables, our method is generally satisfactory because most values of g are close to their nominal levels, especially for the case of $g = 0.95$. Our method performed similarly well for both abundance (Table 3) and incidence data (Table 4, and the left half of Table 5).

For the BCI forest census, in which there is spatial aggregation in the occurrence of some tree species (Table 4 and the left half of Table 5), our method was quite robust to spatial aggregation if the data were analyzed as incidences in quadrats.

Does our method work for nonrandom sampling? To answer this question, we randomly drew quadrats from the BCI data and then counted all individuals within the selected quadrats. Because individual species occurrences are to some extent aggregated within a quadrat, the sampled individuals are not statistically independent and may not satisfy the assumption of random sampling.

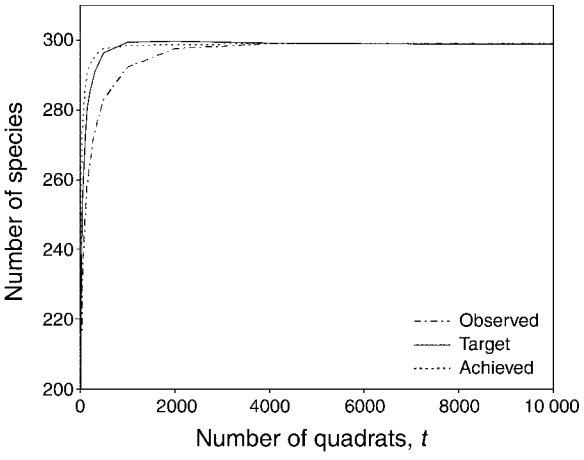

FIG. 1. Average of the observed, target, and achieved number of species as a function of the number of 25 × 25 m quadrats subsampled from the 50-ha Barro Colorado Island (BCI), Panama, incidence data (Hubbell et al. 2005). "Observed" is the mean observed species richness; "Target" is the mean estimated species richness for a given number of quadrats; and "Achieved" is the mean richness obtained when the prescribed additional sampling effort is simulated. See Table 5 for details.

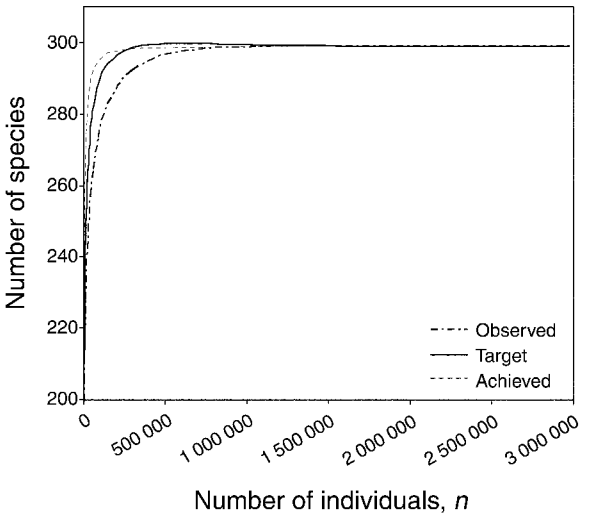

FIG. 2. Average of the observed, target, and achieved numbers of species as a function of the number of individuals subsampled from the 50-ha BCI incidence data (Hubbell et al. 2005). "Observed" is the mean observed species richness; "Target" is the mean estimated species richness for a given number of individuals; and "Achieved" is the mean richness obtained when the prescribed additional sampling effort is simulated. See Table 5 for details.

However, the achieved g values performed adequately and differ little from those based on incidence sampling (Figs. 1 and 2).

DISCUSSION

Biodiversity sampling is an important, but labor-intensive activity, and the sampling effort may have to be increased several-fold in order to detect all of the rare species in the tail of the rank abundance distribution, as we have shown in the simulation study. The methods presented here can provide guidance regarding how much additional sampling would be minimally required to detect all of the species (or a specified target proportion) present in an area.

We applied our method to two ecological sampling protocols: (1) Individuals are sampled independently from the study community and abundance data are recorded (abundance data); and (2) the community is sampled multiple times and incidences are recorded within each sample (replicated incidence data). For the first protocol, our method has some degree of robustness to the spatial aggregation of individuals (Table 5 for BCI data). For the second protocol, our method performed well with quadrat sampling (Tables 4 and 5).

Of course, spatial aggregation affects all species diversity estimators and statistical inferences about biodiversity patterns. In theory, if the functional form of the aggregation were known (e.g., negative binomial), our method could be modified to take this into account (e.g., Kobayashi 1982, 1983, Smith et al. 1985), but in practice the functional form cannot be reasonably inferred. When strong aggregation is present, we suggest that investigators should not sample individuals, but instead should sample quadrats or other sampling units, so that aggregation is no longer present at the larger spatial scale. Then our methods for replicated incidence data can be applied (Table 4 and the left half of Table 5). Chazdon et al. (1998) showed a similar advantage of incidence sampling for richness estimation when aggregation is present.

If the goal is to detect all estimated species, it is inevitable that the estimated effort required will often be prohibitively large (the case of $g = 1$ in all tables). We suggest using a slightly smaller fraction ($g = 0.95$) for more realistic sampling objectives. An alternative method is to establish a low target for q_0 in Eq. 4, the probability of finding a new additional species, and use this as a guide for additional sampling. This alternative criterion is based on a statistical optimal stopping rule and has some good statistical properties (Rasmussen and Starr 1979).

A limitation of this study is that our method assumes sampling *with* replacement, because the asymptotic species richness estimators (Chao1 and Chao 2), as well as Turing's frequency formulas were all derived from such sampling schemes. However, for most quadrat sampling in plant and animal surveys, quadrats/plots/transects/traps are selected *without* replacement. We are currently investigating modifications to our method for sampling without replacement.

ACKNOWLEDGMENTS

We thank two anonymous reviewers for thoughtful and helpful comments and suggestions that greatly improved the paper. We also thank the Center for Tropical Forest Science for making the BCI data available on the internet. N. J. Gotelli was supported by NSF (DEB-0541936). A. Chao and C. W. Lin were supported by Taiwan National Science Council under Contract NSC-95-2118-M007-003. R. K. Colwell was supported by NSF (DEB-0639979). This work was conducted as a part of the Synthetic Macroecological Models of Species Diversity Working Group supported by the National Center for Ecological Analysis and Synthesis, a Center funded by NSF (Grant number DEB-0553768), the University of California–Santa Barbara, and the State of California.

LITERATURE CITED

Butler, B. J., and R. L. Chazdon. 1998. Species richness, spatial variation, and abundance of the soil seed bank of a secondary tropical rain forest. Biotropica 30:214–222.

Chao, A. 1989. Estimating population size for sparse data in capture–recapture experiments. Biometrics 45:427–438.

Chao, A. 2005. Species estimation and applications. Pages 7907–7916 *in* N. Balakrishnan, C. B. Read, and B. Vidakovic, editors. Encyclopedia of statistical sciences. Second edition, volume 12. Wiley, New York, New York, USA.

Chazdon, R. L., R. K. Colwell, J. S. Denslow, and M. R. Guariguata. 1998. Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of NE Costa Rica. Pages 285–309 *in* F. Dallmeier and J. A. Comiskey, editors. Forest biodiversity research, monitoring and modeling: conceptual background and Old World case studies. Parthenon Publishing, Paris, France.

Colwell, R. K., and J. A. Coddington. 1994. Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society of London B 345:101–118.

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

Connolly, S. R., T. P. Hughes, D. R. Bellwood, and R. H. Karlson. 2005. Community structure of corals and reef fishes at multiple scales. Science 309:1363–1365.

Cunningham, S. C., R. D. Babb, T. R. Jones, B. D. Taubert, and R. Vega. 2002. Reaction of lizard populations to a catastrophic wildfire in a central Arizona mountain range. Biological Conservation 107:193–201.

Dahlberg, M. D., and E. P. Odum. 1970. Annual cycles of species occurrence, abundance, and diversity in Georgia estuarine fish populations. American Midland Naturalist 83: 382–392.

Efron, B., and R. J. Tibshirani. 1993. An introduction to the bootstrap. Chapman and Hall, New York, New York, USA.

Ellison, A. M., S. Record, A. Arguello, and N. J. Gotelli. 2007. Rapid inventory of the ant assemblage in a temperate hardwood forest: species composition and assessment of sampling methods. Environmental Entomology 36:766–775.

Fisher, R. A., A. Steven-Corbet, and C. B. Williams. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. Journal of Animal Ecology 12:42–58.

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

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

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

Hubbell, S. P., R. Condit, and R. B. Foster. 2005. Barro Colorado forest census plot data. ⟨http://ctfs.si.edu/datasets/bci⟩

Kobayashi, S. 1982. The rarefaction diversity measurement and the spatial distribution of individuals. Japanese Journal of Ecology 32:255–258.

Kobayashi, S. 1983. Another calculation for the rarefaction diversity measurement for different spatial distributions. Japanese Journal of Ecology 33:101–102.

Lawton, J. H., D. E. Bignell, and B. Bolton. 1998. Biodiversity inventories, indicator taxa, and effects of habitati modification in tropical forest. Nature 391:72–76.

Longino, J. T., J. A. Coddington, and R. K. Colwell. 2002. The ant fauna of a tropical rain forest: estimating species richness three different ways. Ecology 83:689–702.

Longino, J. T., and R. K. Colwell. 1997. Biodiversity assessment using structured inventory: Capturing the ant fauna of a lowland tropical rain forest. Ecological Applications 7:1263–1277.

Magurran, A. E. 2004. Measuring biological diversity. Blackwell, Malden, Massachusetts, USA.

Mao, C. X., and R. K. Colwell. 2005. Estimation of species richness: mixture models, the role of rare species, and inferential challenges. Ecology 86:1143–1153.

Maudsley, M., B. Seeley, and O. Lewis. 2002. Spatial distribution patterns of predatory arthropods within an English hedgerowin early winter in relation to habitat variables. Agriculture, Ecosystem, and Environment 89:77–89.

Preston, F. W. 1962. The canonical distribution of commonness and rarity: Parts 1 and 2. Ecology 43:185–215, 410–432.

Rasmussen, S. L., and N. Starr. 1979. Optimal and adaptive stopping in the search for new species. Journal of the American Statistical Association 74:661–667.

Shen, T.-J., A. Chao, and J.-F. Lin. 2003. Predicting the number of new species in a further taxonomic sampling. Ecology 84:798–804.

Smith, E. P., P. M. Stewart, and J. Cairns, Jr. 1985. Similarities between rarefaction methods. Hydrobiologia 120:167–169.

Soberón, J., and J. Llorente. 1993. The use of species accumulation functions for the prediction of species richness. Conservation Biology 7:480–488.

APPENDIX

Statistical derivations (*Ecological Archives* E090-073-A1).

SUPPLEMENT

Excel-sheet calculator and calculator instructions (*Ecological Archives* E090-073-S1).