# Sufficient sampling for asymptotic minimum species richness estimators

Institute of Statistics, National Tsing Hua University, Hsin-Chu, 30043 Taiwan <sup>2</sup> Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut 06269 USA Department of Biology, University of Vermont, Burlington, Vermont 05405 USA

NNE HAO OBERT K. COLWELL <sup>2</sup> HIH-WEI IN AND ICHOLAS J. GOTELLI3,4

Abstract.

 $\mathcal{L}$  all of asymptotic estimates of asymptotic es species richness, and many also generate variances and many also generate variances and  $\bar{H}$ confidence intervals about the estimates. Connolly et al. (2005) provided a formula for estimating how much sampling is required to unveil the parameter  $\mathbf{u}$ abundance distribution. However, no models to date distribution  $\mu_{\omega}$ have included a nonparametric estimate of the sampling  $\mathbb{R}^n$  of individuals or samples or samples of individuals between  $\mathbb{R}^n$ necessary to  $\mathbf{y} = \mathbf{y}$  to reach the asymptote of a species accumu-

 $\mathcal{Q}$  $\Box$  $\overline{\mathcal{Q}}$  and  $\overline{\mathcal{Q}}$  a effective planning of biotic inventories. In this paper, we derive estimators for the sampling  $\overline{B}$  this paper, we derive estimators for the sampling  $\overline{B}$ effort required to reach the asymptotic richness estimation  $\mathbf{u}$ ed by Chao1 and Chao2, two widely used nonparametric

estimators of species richness for abundance and  $\mathcal{L}$  $\overline{u}$  include data, respectively. The relatively simple s  $\mathbf{u}$  are based on a derivation by the founder of  $\mathbf{u}$ modern computer science, Alan Turing, who used it in cryptographic analyses during World War II. We  $\overline{D}$  provide an Excel spreadsheet macro for performing the performi  $\mathbf{u}$  calculations, and we present estimates of complete of co sampling effort for several published biodiversity for several published biodiversity  $\mu$  $\mathbb{D}$  surveys. Simulation results based on data sets from data sets two large biodiversity inventories demonstrate the  $\mu$ robustness of the proposed method to departure from the proposed meth some of the sampling assumptions. The sampling assumptions of the sampling assumptions.

## AMPLE IZES FOR SYMPTOTIC STIMATORS Abundance data Assume that there are S species in a target biological problem in a target biological problem in a target biological problem in  $S$

 $\mathbb{Z}$  random sample of n individuals is selected (with replacement) from the selected  $\mu$  replacement of  $\mu$ community. A lower bound of species richness is species richness is species richness is species richness in the species richness is  $\mathcal{L}$  $\overline{\nu}$ 

#### obs  $\ell \sqrt{1}$  1=  $\frac{2}{1}$  = 2 2  $1 - \frac{1}{2}$

where  $S$  is the number of species observed, and fraction  $\mathbf{f}_r$  is the isomorphism of  $\mathbf{f}_r$  $\tau$  the count (frequency) of species that are observed that are observed that are observed that are observed that  $\mu$ exactly retrieval in the sample (Chao 1989). Thus, for infinite  $\mathbf f$  is is the number of  $\mathbf{00}_{\infty}$  or species represented by species represented by  $\Box$ exactly one individual in the sample; f2 is the sample; f2 is the number of  $\mathbf{f}$  is sample; f2 is the number of  $\mathbf{f}$  $\omega_{\infty}$  or species represented by exactly two species representations  $\omega$  and  $\omega$ individuals in the sample  $\inf$  is the unknown number of  $\mathbf{f}$  is the unknown number of  $\mathbf{f}$ of species that are present in the community but not detected by the sample and the sample and the sample  $\overline{B}$  and the sample  $\overline{B}$  and the sample  $\overline{B}$ individuals in the sample. Because the sample. Because the sample size n is  $\mathbf{n}$ often large, we can ignore the term  $\mathbf n$  in  $\mathbf n$ obtain the following  $S$  is species for species  $S$  $\mathbf{r}$  is a set of  $\mathbf{r}$  .

est and obs  $\frac{2}{1}$  = 2 2 :  $\begin{array}{cc} & 2 \\ -1 & 1 \end{array}$  $\begin{array}{ccc} 2 & f \\ 1 & f \end{array}$ number of species present but under the sample. The sample in the sample in the sample. The sample in the sample  $\mu$ The Chao1 estimator represents a universal lower bound  $\mathbf{u}$ 

in the sense that it is valid under all types of species abundance distribution. Thus, all estimated sampling  $\mu$  $\mu$  effort derived in this paper represents minimum effort. If  $\mu$  $\frac{f}{f}$   $\frac{1}{f}$   $\frac{1}{f}$  f f  $\qquad \qquad$   $\qquad$   $\qquad$  $\begin{array}{ccccc} & & & \mathbf{E} & & & \mathbf{E} & &$  $\begin{array}{ccc} \hbox{\LARGE$\mu$} & \hbox{\cr & \ddots \end{array} \qquad \qquad \begin{array}{ccc} \hbox{\LARGE $0$} & \hbox{\cr & \ddots \end{array} \qquad \qquad \begin{array}{ccc} \hbox{\LARGE $\mu$} & \hbox{\cr & \ddots \end{array} \qquad \qquad \begin{array}{ccc} \hbox{\LARGE $\mu$} & \hbox{\cr & \ddots \end{array}$ sampling will be complete with  $\mathbf{v}_i$ 

is represented by at least two individuals (no singletons),  $\mathbf{n}$  and  $\mathbf{n}$  $\nu S \rightarrow S$ effort is needed once this condition is satisfied, as there  $f$  and  $f$  and  $f$ 1  $\sum_{i=1}^{n}$  may not be an integer, the condition  $\sum_{i=1}^{n}$  in data  $\sum_{i=1}^{n}$  in data  $\sum_{i=1}^{n}$  in data  $\sum_{i=1}^{n}$ analysis is modified to , 0.5. That is, when there are fewer than  $\mathbf{r} = \mathbf{r} = \mathbf{r} = \mathbf{r}$  species remaining undetected, the  $\mathbf{r} = \mathbf{r} = \mathbf{r} = \mathbf{r}$  $\mathcal{L}$  is defined and no additional effort a is needed. When  $\geq$  the problem is to estimate the problem in the problem is to esti additional number of individuals needed to observe the  $\mathbf{u}$  and above the above th stopping rule for completeness, sampling should complete  $\mathcal{H}^{\mathcal{C}}_{\mathcal{A}}$  $\mathbf{v} = \mathbf{v}$  until singletons vanish. As we will see, this may will see, the see, this may will see, the see, this may will see, this may will require a very large additional sample size, because by  $\mathbf{r}$  the time the time the time total sampling enough total sampling enough total sampling enough total sampling enough to  $r$  individuals of each species found only on  $\mathbb{R}^n$  $\mathbf{u} = \mathbf{u}$  sample, single individuals of a different individuals of additional  $\mathbf{u} = \mathbf{u}$ species will have increase in the surface inevitably surface in  $\Psi$ communities with a large proportion of very rare species, the challenge of estimating richness from samples  $\mathcal{L}_\mathcal{S}$ data is daunting (Mao and Colwell 2005). The collection  $M$ According to Good (1953, 2000), Alan Turing studied as pects of this problem in the context of deciphering  $\mathcal{L}$ encoded messages intercepted from the German military during World War II. Assume that an original sample of size  $n$  is available. The intervals of the set of that, for the set of the se

the next individual sampled, the probability of encountering each of the frequency class r r  $r_{\rm x}$ . . .

## ð 1  $n \geq 3$  $\overline{a}$  special case, the probability of each of  $\overline{a}$  $\sum_{\alpha}$  undetected under  $\sum_{\alpha}$  is the  $\sum_{\alpha}$ there are follows in the frequency class r  $\mathbf{r}_{\mathbf{y}}$  , the frequency  $\mathbf{q}$  that the next individual sampled representation  $\mathbf{u}$  $\begin{array}{ccc}\n u & u & \nu \\
 v & 3 & 1 = 0 \end{array}$  , 1= :  $0 \rightarrow 0^3 1^2 0 \rightarrow 1^2$

 $\overline{\omega}$  interpreted Eq. 3 in the following Eq. 3 in the fol way: the relation  $\mathcal{L}$  abundance  $\mathcal{L}$  abundance  $\mathcal{L}$  $r = \frac{r}{r}$  $f_p$   $nf_r$   $\qquad \qquad$   $r_{\gamma}$   $\qquad \qquad$   $\qquad$  $r$  abundance for each undetected speciesf $\mu$ 

$$
\begin{array}{ccccccccc}\n & 1 & \exp & \stackrel{\prime}{\longrightarrow} & \frac{22}{1} & \cdots & \cdots & \cdots & 11\n\end{array}
$$

 $\mathbf{F} = \mathbf{F} \mathbf{H}$  individuals number of individuals  $\mathbf{F} = \mathbf{F} \mathbf{H}$ needed to detect a fraction g of Sestema fraction  $\mathbf{g}$  of Sestema from the sestema from the sestema from the set

$$
\begin{array}{ccccccccc}\n & 1 & \exp & \stackrel{\prime}{\longleftarrow} & \frac{2}{1} & \cdots & \text{est} & \text{obs} \\
 & & \searrow & & & \\
\hline\n\end{array}
$$
\n
$$
\begin{array}{ccccccccc}\n & & & & & \\
\hline\n\end{array}
$$
\n
$$
\begin{array}{ccccccccc}\n & & & & & \\
\hline\n\end{array}
$$
\n
$$
\begin{array}{ccccccccc}\n & & & & & \\
\hline\n\end{array}
$$
\n
$$
\begin{array}{ccccccccc}\n & & & & & \\
\hline\n\end{array}
$$
\n
$$
\begin{array}{ccccccccc}\n & & & & & \\
\hline\n\end{array}
$$
\n
$$
\begin{array}{ccccccccc}\n & & & & & \\
\hline\n\end{array}
$$
\n
$$
\begin{array}{ccccccccc}\n & & & & & \\
\hline\n\end{array}
$$
\n
$$
\begin{array}{ccccccccc}\n & & & & & \\
\hline\n\end{array}
$$
\n
$$
\begin{array}{ccccccccc}\n & & & & & \\
\hline\n\end{array}
$$

### Incidence data

In most biodiversity studies, individual organisms are  $\begin{array}{ccc} \hbox{\it II} & \hbox{\it II} & \hbox{\it II} & \hbox{\it II} & \hbox{\it II} \end{array}$  $\mathbb D$  our sampling model and by model and by most statistical models with  $\mathbb D$ for biodiversity estimation. In the individual of  $\mathbb{R}$  in the individual of  $\mathbb{R}$  individuals in the individual of  $\mathbb{R}$ are collected or collected or census experimental in traps, plots, plots, plots, plots,  $p$  $\overline{u}$  is the surveys. It is the sampling units, and not the sampling units, an individual organisms, that are actually sampled ranged range domly and independently. For very abundant organisms  $\mathcal{L}_{\mathbf{S}}$  as microbes), or taxa with clonal growth forms  $\mathcal{L}_{\mathbf{S}}$ (such as many plants and invertebrates), it may not even  $\mathbf{b}$  be possible to count individuals within each sampling  $\mathbf{b}$  $\mathcal{U}$  their presence or incidence can be  $\mathcal{U}$ recorded. However, estimation is still possible for a set  $\overline{D}$  replication in which the incidence of each the incidence of each of species is recorded in the sample. When applied to incidence data based on t replicated samples and  $\mathbf{Q}$  represent the number of species  $\mathbf{Q}$  $\frac{1}{\sqrt{2}}$  $\omega_{\rm m}$ and the code of  $\omega^{00}$ . For  $\omega^{00}$  $\mathbf{y} = \mathbf{y}$  , as  $\mathbf{y} = \mathbf{y}$  , as  $\mathbf{y} = \mathbf{y}$  ,  $\mathbf{y} = \mathbf{y}$  , as  $\mathbf{y} = \mathbf{y}$ 

est , obs  $\ell \sqrt{1}$  1 =  $\frac{2}{1}$  =  $\sqrt{2}$  2 :  $\sqrt{13}$ 



 $\frac{1}{\sqrt{2}}$ 

 $\Box$ 

 $\sim$ 



 $10\%$  g  $10\%$  g  $15$  $\mathbf{r} = \mathbf{r} \cdot \mathbf{r}$  average results (over  $\mathbf{r} = \mathbf{r} \cdot \mathbf{r}$  and  $\mathbf{r} = \mathbf{r} \cdot \mathbf{r}$  and  $\mathbf{r} = \mathbf{r} \cdot \mathbf{r}$ abundance data (Fishert data set) for  $g_{\infty}$  $\begin{array}{ccc} \hline \textbf{u} & \textbf{g} \\ \textbf{g} & \textbf{u} \end{array}$  is the results for incidence incidence incidence incidence incidence incidence in  $\textbf{u}$ data, based on the BCI tree data set, for  $\mathbf{g}_{\infty}$  and  $\mathbf{g}_{\infty}$  $\overline{3}$  and  $\overline{4}$  such that  $\overline{5}$ analyses based on both abundance data and the both abundance data and the both abundance data and the  $\alpha$  $\underline{D}$  incidence data for the BCI tree data for the BCI tree data for the BCI tree data set of the BCI tree data set of the BCI tree data for the BCI tree data set of the BCI tree data set of the BCI tree data set of th  $g_{\text{in}}$ , in order to  $3$  , in order to  $\mu$ investigate the sensitivity of our method to spatial the sensitivity of our method to spatial the spatial value of our method to spatial value of  $\mathbf{u}$  $\overline{u}$  in  $\overline{u}$ 

Ideally, we would compare the estimated additional sample size (as calculated from our equations) with the  $\mathcal{S}$  $\overline{M}$ simulated sample size (which is obtained by continuing by continuing by continuing  $\overline{M}$  $\overline{B}$  simulated process until we reach the target). We reach the target  $\overline{y}$  $\mathbb{H}$  in some data sets, the estimate Sest may be  $\mathbb{S}$ exceed the observed number of species in the full data  $s$  that the size is not attained size is not attained size is not attained size is not attained because is not attained by  $\overline{B}$ we can never reach a species richness higher than the full species richness higher than the full species  $\mathcal{A}$ observed species richness in the inventory or in the census). Therefore, we use used an alternative metric metric: the second and alternative metric: the second and alter achieved number of species (or equivalently, the species (or equivalently, the species (or equivalently, the species of  $\mathbf{v}_i$ )

 $\frac{1}{x}$  $\frac{d}{dt} \left( \frac{d}{dt} \right) = \frac{d}{dt} \left( \frac{d}{dt} \right)$  $\mathcal{P}=\mathbf{g}$  $\bar{\mathcal{L}}_{\mu\nu}$  $\bar{B}=B\bar{B}$   $\bar{\omega}$  $\sim$  $\sim$   $\sim$  $\bar{\omega}$  $\rightarrow$  ${\bf g}$  .  $\overline{u}$   $g_{\gamma}$  $\ddot{\phantom{0}}$  $\bar{U}$ ÷ L.  $_{\star}$  $\bar{\gamma}$  $\bar{r}_{\rm esc}$  $\boldsymbol{\varPsi}$  $\overline{g}$  $\overline{\phantom{a}}$  $\bar{\bar{z}}$  $\omega$  $\mathbf{v}^{(i)}$  $\qquad \qquad \text{if} \quad \mathbf{g}$  $\bar{\omega}$  $\bar{u}$  $\omega$  $\omega$  $g\mathcal{S}$  $\bar{U}$  $g<sub>x</sub>$  $\omega$  $\bar{B}$  and  $\bar{B}$  $\boldsymbol{D}$  $\downarrow$  $gS$  $\bar{B}$  –  $\bar{B}$ ç,  $\ddot{\phantom{0}}$  $\frac{1}{\sqrt{2}}$  $\ddot{\phantom{0}}$  $\omega$  $\bar{D}$  $\sim$  $\bar{D}$  –  $\bar{D}$  $\ddot{\phantom{1}}$  $\bar{U}$ l.  $\boldsymbol{g}$  $\frac{1}{\sqrt{2}}$ Ļ.  $\mathbf S$  $_{\star}$ Ń  $\bar{B}$  $\mu$  $\bar{U}$  $\overline{\phantom{a}}$  .  $\overline{\phantom{a}}$ ç.  $\%$ Ş.  $g_{\gamma}$  $\bar{D}$  –  $\bar{D}$  $\ddot{\phantom{1}}$  $\bar{\omega}$  $\sim$  $\bar{\mathcal{L}}$  $\bar{U}$  $\frac{1}{\sqrt{2}}$  $\sim$  $\ddot{\phantom{0}}$  $\omega$  $\overline{\phantom{a}}$  $_{\scriptscriptstyle U\!U}$  $\overline{\phantom{a}}$  $\omega$  $\downarrow$  $\%$  $\frac{1}{x}$  $\bar{B}$ 

 $\boldsymbol{\omega}$ 

 $\sim$  $\mathbf{y} = \mathbf{y} + \mathbf{y}$  and a chieved adequately performed adequately perfo and different little from the incident little from incident little from incidence sampling incidence sampling  $\mathbf{r}$ (Figs. 1 and 2).

 $\mathcal{L}(\mathcal{L})$  is a set of  $\mathcal{L}(\mathcal{L})$  $B$ iodiversity sampling is an important, but laborintensive activity, and the sampling effort may have to sampling effort may have to sampling effort may have to  $\mathcal{L}_{\mathcal{A}}$  $\mu$  in order to detect all of the rare  $\omega$ species in the tail of the rank abundance distribution, as  $\omega$ we have shown in the simulation study. The simulation study is shown in the methods of  $\mathbf{u}$ presented here can present for  $p$  and  $p$  an much additional sampling would be minimally required be  $m$  $\overline{u}$  detect all of the specified target all of the specified target all or a specified target all or a specified target all of the sp proportion) present in an area.  $\Box$  we applied to two equations are method to two equations of two equations  $\Box$ protocols: (1) individuals are sampled in the sampled independent independent in  $\mathcal{D}$  $f(x) = \frac{1}{2} \int_0^x \frac{dx}{(x^2 + y^2)^2} dx$ 

 $\begin{array}{ccc} \begin{array}{ccc} \text{$\mu$} & \text{$\mu$} \end{array} & \text{if} & \text{if}$  $\Box$ 

- $\hat{H} = \hat{H} \hat{H} + \hat{H} \hat{H}$  and  $\hat{H} = \hat{H} \hat{H}$  and  $\hat{H} = \hat{H} \hat{H}$  and  $\hat{H} = \hat{H} \hat{H}$ terrestrial biodiversity through extrapolation. Philosophical biodiversity through extrapolation. Philosophical biodiversity of  $\mathcal{P}_\text{max}$ Transactions of the Royal Society of London B  $345$  $\mathbf{E} = \mathbf{E} \times \mathbf{E}$  $\begin{array}{ccc} \hbox{\it II} & & \hbox{\it II} & & \hbox{\it III} \end{array}$
- mulation curves. Ecology 85:2717–2727. Connolly, S. R., T. P. Hughes, D. R. Bellwood, and R. H.
- $\overline{B}$  and reef fishes and at multiple scales. Science 309:1363–1365. Cunningham, S. C., R. D. Babb, T. R. Jones, B. D. Taubert,
- $\overline{u}$  and R. Vega. 2002. Reaction of lizard population of lizard population of lizard populations to a set of lizard population of lizard populations to a set of lizard population of lizard populations to a set of liza catastrophic wildfire in a central Arizona mountain range. The strophic mountain range  $\mu$
- Biological Conservation 107:193–201.
- $\mathcal{D}$  and  $\mathcal{D}$  and  $\mathcal{D}$  and  $\mathcal{D}$
- species occurrence, abundance, abundance, abundance, and diversity in Georgia estuarine fish populations. American Midland Naturalist  $M$  and  $M$
- 382–392.
- $E = \frac{1}{2}$  and  $E = \frac{1}{2}$  a bootstrap. Chapman and Hall, New York, New York, USA.