Research Article

An Algebraic Derivation of Chao's Estimator of the Number of Species in a Community Highlights the Condition Allowing Chao to Deliver Centered Estimates

Jean Béguinot^{1,2}

¹ Société d'Histoire Naturelle du Creusot, 12 rue des Pyrénées, 71200 Le Creusot, France

² Biogéosciences, Université de Bourgogne, 21000 Dijon, France

Correspondence should be addressed to Jean Béguinot; jean-beguिनot@orange.fr

Received 30 June 2014; Accepted 19 September 2014; Published 11 November 2014

Academic Editor: Arthur P. Schwab

Copyright © 2014 Jean Béguinot. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Anne Chao proposed a very popular, nonparametric estimator of the species richness of a community, on the basis of a limited size sampling of this community. This expression was originally derived on a statistical basis as a lower-bound estimate of the number of missing species in the sample and provides accordingly a minimal threshold for the estimation of the total species richness of the community. Hereafter, we propose an alternative, algebraic derivation of Chao's estimator, demonstrating thereby that Chao's formulation may also provide *centered* estimates (and not only a lower bound threshold), provided that the sampled communities satisfy a specific type of SAD (species abundance distribution). This particular SAD corresponds to the case when the number of unrecorded species in the sample tends to decrease exponentially with increasing sampling size. It turns out that the shape of this "ideal" SAD often conforms approximately to the usually recorded types in nature, such as "log-normal" or "broken-stick." Accordingly, this may explain why Chao's formulation is generally recognized as a particularly satisfying nonparametric estimator.

1. Introduction

Estimating the total species richness within large communities of species, using only samplings limited in sizes, is a common, long standing challenge which has elicited numerous procedures of estimations.

For a few decades, a series of so-called nonparametric estimators provide elegant and convenient solutions to this question. These new estimators are ordinarily simple format formulations which, moreover, require no specific assumption regarding the statistical distribution of the abundances of species (which, thus, makes these formulations "nonparametric").

Within this category of formulations, Anne Chao proposed a very popular nonparametric expression, which actually stands among the most commonly used estimators of the total species richness of a sampled community.

Let Δ be the number of species that are missed by (i.e., unrecorded within) the limited sampling of a large community of species. Then, according to Chao's formulation [1],

$$\Delta = \frac{f_1^2}{2f_2} \quad (1)$$

with f_1 and f_2 as the numbers of species encountered only *once* and *twice*, respectively, in the sample.

This expression was later generalised [2] as

$$\Delta = \left[\frac{f_1^x}{(x!f_x)} \right]^{(1/(x-1))} \quad (2)$$

with f_x standing for the number of species recorded x times in the sample.

These expressions, derived on a statistical basis, provide a lower-bound estimate of the number Δ of missing species in a

sampled community [1, 2]. That is, Chao's formulation is expected to provide only a minimal threshold (in the statistical sense) for the estimated species richness. Yet, several decades of practice (especially in the field of ecological purposes and biodiversity surveys) call for placing Chao's expression among the most valuable and reliable estimators [3–7] since, in many occasions, this expression may well appear to provide approximately centered rather than only lower-bound estimates. In short, although designed conceptually as a lower-bound evaluation, Chao's formulation may, nonetheless, fairly often provide rather centered estimates in the common practice.

Hereafter, we address this apparent paradox and propose new insights and argumentations, issued from an alternative, algebraic derivation of the originally statistically derived formulation by Chao.

2. The Specific Condition Allowing an Alternative, Algebraic Derivation of the “Chao” Formulation

We demonstrate (see Appendix A for mathematical details) that the general expression of Chao's estimator, established originally on a statistical basis, may also admit, alternatively, an algebraic derivation, leading to exactly the same expression as the statistically derived formulation $\Delta = [f_1^x / (x! f_x)]^{(1/(x-1))}$.

Yet, while the statistical derivation of Chao's formulation requires no particular restriction (“nonparametric”), the algebraic derivation, implies a particular shape for the expected decrease of the proportion Δ/S of missing (i.e., unrecorded) species when the sample size N increases. In fact, as demonstrated in Appendix A, this asymptotic decrease should conform to a negative exponential:

$$\frac{\Delta}{S} = \exp(-kN) \quad (3)$$

with

- (i) “ S ” as the species richness, that is, the total (unknown) number of species of the community,
- (ii) “ k ” as a constant,
- (iii) “ N ” as the sample size, that is, the number of individuals recorded in the sample.

In turn, this particular form of the decrease of the number of missing species with enlarging sample sizes constrains the shape of the species abundance distribution (the “SAD,” that is, the distribution of species abundances when species are conventionally ranked by decreasing order of abundance).

According to (3), the number “ r ” of recorded species in the sample is

$$r = S - \Delta = S(1 - \exp(-kN)). \quad (4)$$

Let k' be the number of individuals belonging to the less abundant species among the “ r ” species recorded in the sample (i.e., the species of rank “ r ” when species are ranked by

FIGURE 1: Typical shape of the “SAD” when the number Δ of unrecorded species is decreasing as a negative exponential of the sample size. The sum of abundances is normalised at 100% (open circles: total species richness = 20; black diamonds: total species richness = 50).

decreasing order of abundance). Then, the relative abundance a_r of the species of rank r is expected to be inversely proportional to N , as

$$a_r = \frac{k'}{N}. \quad (5)$$

According to (4) and (5),

$$r = S \left[1 - \exp\left(-\frac{kk'}{a_r}\right) \right]. \quad (6)$$

According to the sample size N , every species may be called to play the role of the less abundant species within the sample (since, by continuously decreasing the size N of the sample, each species of the community [including, at last, the most common] would successively play the role of the least abundant species in the sample).

Therefore, (6) stands the same for any species of any rank “ i ” in the SAD:

$$i = S \left[1 - \exp\left(-\frac{kk'}{a_i}\right) \right]. \quad (7)$$

This equation thus describes the shape of the species abundance distribution ($a_i = f(i)$) when the proportion of unrecorded species in a sample is exponentially decreasing with the sample size, that is, the shape of the species abundance distribution which conditions the validity of the algebraic derivation of Chao estimator. Figure 1 provides examples of the corresponding shapes of the SAD.

3. The Resulting Restrictive Condition Which Allows the “Chao” Formulation to Become a Centered Estimator of Species Richness

As the algebraic derivation is deterministic by essence, it therefore provides a centered estimate of Δ and of the

resulting total species richness S , instead of being only a lower-bound estimate, as is the case in the nonparametric context. As mentioned above, the algebraic derivation of Chao's formulation requires that the sampled community satisfies, at least approximately, the particular type of SAD defined by (7) and illustrated at Figure 1. This restrictive condition is the "price" to be paid for the more accurate estimate, namely, the loss of the strict "nonparametric" character of the statistically based conception.

Yet, this condition assigned to the shape of the distribution of species abundances might not be so restrictive in practice, at least as a first approximation. Reasons for this may be as follows:

- an asymptotic decrease to zero of Δ with N (equation (3)) seems logical and intuitive, because aiming to estimate the total number of species in a community implicitly requires that this number does exist and might be actually reached progressively with sampling size N increasing continuously;
- among the different types of accumulation curves with such an asymptotic evolution, the negative exponential answer of Δ to increasing sampling size is, admittedly, one among the most simple, robust, and seemingly common [8, 9];
- the sigmoidal shape of the prescribed SAD (equation (7) and Figure 1) is not so far from the most classically referred empirical types, broken-stick and log-normal distributions [10]. Yet, a strict conformity is not expected a priori with any empirical models. For example, the equation $i = f(a_i)$ for the SAD corresponding to a broken-stick distribution is $i \approx S \cdot \exp(-S \cdot a_i)$ [10], which is formally different from (7).

Accordingly, it is no real surprise that Chao's formula often approaches a strictly centred estimate, in spite of being only a lower-bound estimate in all generality.

This would explain why, in ecological practice in particular, Chao's formulation is yet considered one of the more accurate and reliable estimators of the total species richness within partially sampled communities.

As mentioned in particular by Gotelli (personal communication), a trend would remain for Chao's estimates to increase somehow when a series of sampling of growing sizes are extracted from the same community instead of remaining ideally stable on average. This, however, is not necessarily contradictory to preceding arguments but should certainly result from residual discrepancy between the real SAD and the ideal model described by (7) and exemplified at Figure 1.

Appendix

A. An Alternative, Algebraic Derivation of the Chao 1 Estimator

Consider a community of species, actually containing an unknown total number " S " of species.

Let p_i be the probability of occurrence (assimilated to the relative abundance) of species " i " within this community and

let N be the sample size, that is, the number of individuals recorded in the sample.

The estimated number of species that escape recording during sampling of the community is $\Delta_{(N)}$:

$$\Delta_{(N)} = \sum_i (1 - p_i)^N \quad (\text{A.1})$$

with $\sum_i$ as the operation summation extended to the totality of the " S " species " i " (either recorded or not).

The number f_x of species recorded x times in the sample is then, according to the binomial distribution:

$$\begin{aligned} f_x &= \left[\frac{N!}{(N-x)!} \right] \sum_i [(1 - p_i)^{N-x} p_i^x] \\ &= C_{N,x} \sum_i (1 - p_i)^{N-x} p_i^x. \end{aligned} \quad (\text{A.2})$$

A.1. Deriving f_x as a Function of $\Delta_{(N)}$. The number f_x of species recorded x times in the sample may be derived as a function of $\Delta_{(N)}$.

(i) Consider

$$\begin{aligned} f_1 &= N \sum_i [(1 - p_i)^{N-1} p_i] \\ &= N \sum_i [(1 - p_i)^{N-1} (1 - (1 - p_i))] \quad \text{according to (A.2)} \\ &= N \sum_i [(1 - p_i)^{N-1}] \\ &\quad - N \sum_i [(1 - p_i)^{N-1} (1 - p_i)] \\ &= N \sum_i [(1 - p_i)^{N-1}] - N \sum_i [(1 - p_i)^N]. \end{aligned} \quad (\text{A.3})$$

According to (A.1), it becomes as follows: $f_1 = N(\Delta_{(N-1)} - \Delta_{(N)}) = -N(\Delta_{(N)} - \Delta_{(N-1)})$.

As appropriate sampling requires a size $N \gg 1$, the difference $(\Delta_{(N)} - \Delta_{(N-1)})$ can be likened to the corresponding derivative $(\partial \Delta_{(N)} / \partial N)$. Then,

$$f_1 = -N \left(\frac{\partial \Delta_{(N)}}{\partial N} \right) = -N \Delta'_{(N)}, \quad (\text{A.4})$$

where $\Delta'_{(N)}$ is the first derivative of $\Delta_{(N)}$ with respect to N . Thus,

$$f_1 = -N \Delta'_{(N)} (= -C_{N,1} \Delta'_{(N)}). \quad (\text{A.5})$$

Similarly,

(ii) consider

$$\begin{aligned} f_2 &= C_{N,2} \sum_i [(1 - p_i)^{N-2} p_i^2] \quad \text{according to (A.2)} \\ &= C_{N,2} \sum_i [(1 - p_i)^{N-2} (1 - (1 - p_i^2))] \end{aligned}$$

$$\begin{aligned}
&= C_{N,2} \left[\sum_i \left[(1-p_i)^{N-2} \right] - \sum_i \left[(1-p_i)^{N-2} (1-p_i^2) \right] \right] \\
&= C_{N,2} \left[\sum_i \left[(1-p_i)^{N-2} \right] \right. \\
&\quad \left. - \sum_i \left[(1-p_i)^{N-2} (1-p_i) (1+p_i) \right] \right] \\
&= C_{N,2} \left[\sum_i \left[(1-p_i)^{N-2} \right] - \sum_i \left[(1-p_i)^{N-1} (1+p_i) \right] \right] \\
&= C_{N,2} \left[(\Delta_{(N-2)} - \Delta_{(N-1)}) - \frac{f_1}{N} \right]
\end{aligned}$$

according to (A.1) and (A.2)

$$= C_{N,2} \left[-\Delta'_{(N-1)} - \frac{f_1}{N} \right] = C_{N,2} \left[-\Delta'_{(N-1)} + \Delta'_{(N)} \right]$$

since $f_1 = -N\Delta'_{(N)}$ (cf. equation (A.5)).

$$\begin{aligned}
&= C_{N,2} \left(\frac{\partial \Delta'_{(N)}}{\partial N} \right) \\
&= \left[\frac{N(N-1)}{2} \right] \left(\frac{\partial^2 \Delta_{(N)}}{\partial N^2} \right) = \left[\frac{N(N-1)}{2} \right] \Delta''_{(N)},
\end{aligned} \tag{A.6}$$

where $\Delta''_{(N)}$ is the second derivative of $\Delta_{(N)}$ with respect to N . Thus,

$$f_2 = \left[\frac{N(N-1)}{2} \right] \Delta''_{(N)} = C_{N,2} \Delta''_{(N)}. \tag{A.7}$$

(iii) Consider

$$f_3 = C_{N,3} \left[\sum_i (1-p_i)^{N-3} p_i^3 \right]$$

which, by the same process, yields :

$$\begin{aligned}
&= C_{N,3} \left[\sum_i (1-p_i)^{N-3} - \sum_i (1-p_i)^{N-2} \right. \\
&\quad \left. - \sum_i \left[(1-p_i)^{N-2} p_i \right] \right. \\
&\quad \left. - \sum_i \left[(1-p_i)^{N-2} p_i^2 \right] \right] \\
&= C_{N,3} \left[(\Delta_{(N-3)} - \Delta_{(N-2)}) - \frac{f_1^*}{N-1} - \frac{2f_2}{N(N-1)} \right]
\end{aligned}$$

according to (A.1) and (A.2),

(A.8)

where f_1^* is the number of singletons that would be recorded in a sample of size $(N-1)$ instead of N .

According to (A.5) and (A.7),

$$\begin{aligned}
f_1^* &= -(N-1) \Delta'_{(N-1)} = -C_{N-1,1} \Delta'_{(N-1)}, \\
f_2 &= \left[\frac{N(N-1)}{2} \right] \Delta''_{(N)} = C_{N-1,2} \Delta''_{(N)},
\end{aligned} \tag{A.9}$$

where $\Delta'_{(N-1)}$ is the k th derivate of $\Delta_{(N)}$ with respect to N , at point $(N-1)$. Then,

$$\begin{aligned}
f_3 &= C_{N,3} \left[(\Delta_{(N-3)} - \Delta_{(N-2)}) + \Delta'_{(N-1)} - \Delta''_{(N)} \right] \\
&= C_{N,3} \left[-\Delta'_{(N-2)} + \Delta'_{(N-1)} - \Delta''_{(N)} \right] \\
&= C_{N,3} \left[\Delta''_{(N-1)} - \Delta''_{(N)} \right] = C_{N,3} \left[-\frac{\partial \Delta''_{(N)}}{\partial N} \right] \\
&= C_{N,3} \left[-\frac{\partial^3 \Delta_{(N)}}{\partial N^3} \right] = C_{N,3} \Delta'''_{(N)},
\end{aligned} \tag{A.10}$$

where $\Delta'''_{(N)}$ is the third derivative of $\Delta_{(N)}$ with respect to N . Thus,

$$f_3 = -C_{N,3} \Delta'''_{(N)}. \tag{A.11}$$

(iv) Generalising this approach to calculate the number f_x of species recorded x times in the sample:

$$\begin{aligned}
f_x &= C_{N,x} \sum_i \left[(1-p_i)^{N-x} p_i^x \right] \quad (\text{according to (A.2)}) \\
&= C_{N,x} \sum_i \left[(1-p_i)^{N-x} (1 - (1-p_i^x)) \right] \\
&= C_{N,x} \left[\sum_i (1-p_i)^{N-x} - \sum_i \left[(1-p_i)^{N-x} (1-p_i^x) \right] \right] \\
&= C_{N,x} \left[\sum_i (1-p_i)^{N-x} \right. \\
&\quad \left. - \sum_i \left[(1-p_i)^{N-x} (1-p_i) (\sum_j p_i^j) \right] \right]
\end{aligned} \tag{A.12}$$

with $\sum_j$ as the summation from $j = 0$ to $j = x-1$. It becomes as follows:

$$\begin{aligned}
f_x &= C_{N,x} \left[\sum_i (1-p_i)^{N-x} - \sum_i \left[(1-p_i)^{N-x+1} (\sum_j p_i^j) \right] \right] \\
&= C_{N,x} \left[\sum_i (1-p_i)^{N-x} \right. \\
&\quad \left. - \sum_i (1-p_i)^{N-x+1} \right. \\
&\quad \left. - \sum_k \left[(\sum_i (1-p_i)^{N-x+1} p_i^k) \right] \right]
\end{aligned} \tag{A.13}$$

with $\sum_k$ as the summation from $k = 1$ to $k = x-1$; that is,

$$\begin{aligned}
f_x &= C_{N,x} \left[(\Delta_{(N-x)} - \Delta_{(N-x+1)}) - \sum_k \left(\frac{f_k^*}{C_{(N-x+1+k),k}} \right) \right] \\
&\quad (\text{according to (A.1) and (A.2)}),
\end{aligned} \tag{A.14}$$

where $C_{(N-x+1+k),k} = (N-x+1+k)!/k!/(N-x+1)!$ and f_k^* is the number of species recorded k times during a sampling of size $(N-x+1+k)$ (instead of size N).

The same demonstration, which yields previously the expression of f_1^* above (see (A.9)), applies for the f_k^* (with k up to $x-1$) and gives

$$f_k^* = (-1)^k (C_{(N-x+1+k),k}) \Delta_{(N-x+1+k)}^{(k)}, \tag{A.15}$$

where $\Delta_{(N-x+1+k)}^{(k)}$ is the k th derivate of $\Delta_{(N)}$ with respect to N , at point $(N - x + 1 + k)$. Then,

$$f_x = \left[\frac{N!/x!}{(N-x)!} \right] \times \left[(\Delta_{(N-x)} - \Delta_{(N-x+1)}) - \sum_k \left((-1)^k \Delta_{(N-x+1+k)}^{(k)} \right) \right], \quad (\text{A.16})$$

which finally yields

$$\begin{aligned} f_x &= \left[\frac{N!/x!}{(N-x)!} \right] \left[\frac{(-1)^x \partial \Delta_{(N)}^{(x-1)}}{\partial N} \right] \\ &= \left[\frac{N!/x!}{(N-x)!} \right] \left[\frac{(-1)^x \partial^x \Delta_{(N)}}{\partial N^x} \right]. \end{aligned} \quad (\text{A.17})$$

That is,

$$f_x = (-1)^x C_{N,x} \Delta_{(N)}^{(x)}, \quad (\text{A.18})$$

where $\Delta_{(N)}^{(x)}$ is the x th derivative of $\Delta_{(N)}$ with respect to N , at point N .

A relationship is thus derived between the series of the numbers f_x of species recorded x times and the series of derivatives $\Delta_{(N)}^{(x)}$ at order x for $\Delta_{(N)}$:

$$\begin{aligned} f_0 &= \Delta_{(N)} = C_{N,0} \Delta_{(N)}, \\ f_1 &= -C_{N,1} \Delta'_{(N)}, \\ f_2 &= C_{N,2} \Delta''_{(N)}, \\ f_3 &= -C_{N,3} \Delta'''_{(N)}, \\ f_x &= (-1)^x C_{N,x} \Delta_{(N)}^{(x)}. \end{aligned} \quad (\text{A.19})$$

These relations will serve as the basis for deriving the formulation of Δ as a function of the f_x .

A.2. Deriving a Generalised Expression for the “Chao” Formulation. Let us consider now the case when the number of missing species $\Delta_{(N)}$ conforms (or is close) to a negative exponential with respect to the sampling size N . Accordingly, $\Delta_{(N)}$ would thus verify the following series of differential equations:

$$\frac{\Delta_{(N)}}{\Delta'_{(N)}} = \frac{\Delta'_{(N)}}{\Delta''_{(N)}} = \frac{\Delta''_{(N)}}{\Delta'''_{(N)}} = \dots = \frac{\Delta_{(N)}^{(x-1)}}{\Delta_{(N)}^{(x)}} \quad (\text{A.20})$$

that yields $\Delta_{(N)}/\Delta_{(N)}^{(x)} = (\Delta_{(N)}/\Delta'_{(N)})^x$ and then $\Delta_{(N)} = \left[(\Delta'_{(N)})^x / \Delta_{(N)}^{(x)} \right]^{1/(x-1)}$, where $\Delta'_{(N)}$, $\Delta''_{(N)}$, $\Delta'''_{(N)}$, and $\Delta_{(N)}^{(x)}$ are the first, second, third, and x° derivatives of $\Delta_{(N)}$ with respect to N .

As $\Delta'_{(N)} = -(f_1/N)$ and $\Delta_{(N)}^{(x)} = (-1)^x f_x / C_{N,x}$ (see (A.5) and (A.18)), it follows for $\Delta_{(N)}$ that

$$\Delta_{(N)} = \left(C_{N,x} \left[\frac{f_1^x}{N^x f_x} \right] \right)^{(1/(x-1))}. \quad (\text{A.21})$$

Note that $C_{N,x}/N^x$ is $\sim 1/x!$ since, in practice, x remains, by far, quite smaller than N . Accordingly,

$$\Delta_{(N)} = \left[\frac{f_1^x}{(x! f_x)} \right]^{(1/(x-1))}. \quad (\text{A.22})$$

In particular,

$$\begin{aligned} \Delta_{(N)} &= f_0 = \frac{f_1^2}{(2f_2)} \quad [\text{the genuine Chao's formulation}] \\ &= \left[\frac{f_1^3}{(6f_3)} \right]^{0.5} = \left[\frac{f_1^4}{(24f_4)} \right]^{0.33} \dots \end{aligned} \quad (\text{A.23})$$

Conflict of Interests

The author declares that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

Two anonymous reviewers are gratefully acknowledged for their useful suggestions on a previous version of the paper. Also, Anne Chao and Nick Gotelli provide encouraging comments on the original version of the paper.

References

- [1] A. Chao, “Nonparametric estimation of the number of classes in a population,” *Scandinavian Journal of Statistics*, vol. 11, no. 4, pp. 265–270, 1984.
- [2] A. Chao and T.-J. Shen, “Nonparametric prediction in species sampling,” *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 9, no. 3, pp. 253–269, 2004.
- [3] J. A. Coddington, L. H. Young, and F. A. Coyle, “Estimating spider species richness in a Southern Appalachian cove hardwood forest,” *Journal of Arachnology*, vol. 24, no. 2, pp. 111–128, 1996.
- [4] P. T. Hrabar, *Discovering Molecular Mechanisms of Mutualism with Computational Approaches to Endosymbiosis*, University of New Mexico, Santa Fe, NM, USA, 2001.
- [5] S. Aubry, F. Magnin, V. Bonnet, and R. C. Preece, “Multi-scale altitudinal patterns in species richness of land snail communities in South-Eastern France,” *Journal of Biogeography*, vol. 32, no. 6, pp. 985–998, 2005.
- [6] S. Brittain and D. Böhning, “Estimators in capture-recapture studies with two sources,” *Advances in Statistical Analysis*, vol. 93, no. 1, pp. 23–47, 2009.
- [7] A. Chao, R. K. Colwell, C.-W. Lin, and N. J. Gotelli, “Sufficient sampling for asymptotic minimum species richness estimators,” *Ecology*, vol. 90, no. 4, pp. 1125–1133, 2009.
- [8] L. R. Holdridge, W. G. Grenke, W. H. Hatheway, T. Liang, and J. A. Tosi, *Forest Environments in Tropical Life Zones*, Pergamon Press, Oxford, UK, 1971.
- [9] M. J. Soberon and B. J. Llorente, “The use of species accumulation functions for the prediction of species richness,” *Conservation Biology*, vol. 7, no. 3, pp. 480–488, 1993.

- [10] R. M. May, "Patterns of species abundance and diversity," in *Ecology and Evolution of Communities*, M. L. Cody and J. M. Diamond, Eds., The Belknap Press of Harvard University, Cambridge, Mass, USA, 1975.