The implicit assumption of symmetry and the species abundance distribution

Online Supporting Material

David Alonso[1,2], Annette Ostling[1], and Rampal S. Etienne[2]

[1]Ecology and Evolutionary Biology
University of Michigan
830 North University Av.
Ann Arbor MI 48109-1048. USA.

[2]Community and Conservation Ecology group
University of Groningen
PO Box 14, 9750 AA Haren, The Netherlands

Address of the corresponding author: David Alonso, Community and Conservation Ecology Group, University of Groningen, PO Box 14, 9750 AA Haren, The Netherlands. E-mail: d.alonso@rug.nl

Table of Contents:

S1 Ensemble formulas for SADs

The relationship between a multivariate abundance distribution governing the probability of having the community at a given state, represented by an abundance vector $\vec{n} = (n_1, \ldots, n_S) \equiv \{n_i\}$, at any given time, and the species abundance distribution classically defined as the average number (or fraction) of species presenting a certain abundance has been already highlighted by Etienne and Alonso (2005). Here, we show that this relationship can be written with generality by using Eq. (2) in the main text, provided we know the exact number of species S present at the regional scale.

In other words we derive Eq. (2) beginning with the multivariate probability $P(n_1, n_2, \ldots, n_S)$ for each species i to have n_i individuals. We show that species independence is not a required assumption for Eq. (2) to hold as long as the number of species present is given.

The regional probability for a species to have n individuals, $P(n)$, can be written in terms of the multivariate probability distribution as follows:

$$P(n) = \frac{1}{S} \sum_{n_1} \ldots \sum_{n_S} P(\vec{n})\, S_n(\vec{n}) = \frac{1}{S} \sum_{\{n_j\}} P(\vec{n})\, S_n(\vec{n}) \tag{.1}$$

where the sum is over all possible combinations of abundances of the S species, and $S_n(\vec{n})$ simply counts the number of species with n individuals for a particular combination of abundances $\{n_j\}$. Notice that the sum itself is the expected value, $E[S_n]$, when we average over all possible abundance vectors weighted by the multivariate abundance distribution (Etienne & Alonso, 2005).

Using the fact that $S_n(\vec{n}) = \sum_{i=1}^{S} \delta_{n_i,n}$, where n_i takes the different values observed in the abundance vector $\vec{n}$, and reversing the order of summation, we have:

$$P(n) = \frac{1}{S} \sum_{i=1}^{S} \sum_{\{n_j\}} P(n_1, n_2, \ldots, n_s)\, \delta_{n_i,n} \tag{.2}$$

Finally, using

$$\sum_{\{n_j\}} P(n_1, n_2, \ldots, n_s)\, \delta_{n_i,n} = \sum_{\{n_j\}_{j \neq i}} P(n_1, n_2, \ldots, n, \ldots, n_s) \tag{.3}$$

where the n appears in the ith position. The right-hand side of the last equation is nothing but the marginal probability distribution for the ith species:

$$P^{(i)}(n) = \sum_{\{n_j\}_{j \neq i}} P(n_1, n_2, \ldots, n, \ldots, n_s) \tag{.4}$$

Eq. (2) arises from Eq. (.2).

When we sample at a particular spatial scale a, we can derive a similar expression relating the multivariate probability of observing certain abundance vector (or data set) in a sample at scale a, $P^{(a)}(n_1, \ldots, n_S)$, and the species abundance distribution, $P^{(a)}(n)$. As before, we can write:

$$P^{(a)}(n) = \sum_{\{n_j\}} P^{(a)}(\vec{n})\, F_n(\vec{n}) \tag{.5}$$

where $F_n(\vec{n})$ is the fraction of species with abundance n in a particular data set $\{n_j\}$:

$$F_n(\vec{n}) = \frac{S_n^{(a)}(\vec{n})}{S^{(a)}(\vec{n})} \tag{.6}$$

If we compare Eqs. (.5)-(.6) with its analog given by Eq. (.1), we notice that now the total number of species in the sample, $S^{(a)}(\vec{n})$, depends also on the abundance vector at scale a. However, we can also proceed as before and write:

$$P^{(a)}(n) = \sum_{\{n_j\}} \sum_i P^{(a)}\left(n_1, n_2, \ldots, n_s\right) \frac{\delta_{n_i\, n}}{S^{(a)}(\{n_j\})} \tag{.7}$$

and, by reversing the order in which sums are performed:

$$P^{(a)}(n) = \sum_i \sum_{\{n_j\}_{j\neq i}} P^{(a)}\left(n_1, n_2, \ldots, n, \ldots, n_s\right) \frac{1}{S^{(a)}\left(n_1, n_2, \ldots, n, \ldots, n_s\right)} \tag{.8}$$

However, we can always write:

$$P^{(a)}\left(n_1, n_2, \ldots, n, \ldots, n_s\right) = P^{(a)}\left(n_1, n_2, \ldots, n_s | n_i = n\right)\, P^{(a)}\{n_i = n\} \tag{.9}$$

where, by definition, the probability of observing the species i with abundance n in a sample at scale a, $P(a)\{n_i = n\}$ is given by $P^{(i)(a)}(n)$. In sum, we have:

$$P^{(a)}(n) = \sum_i P^{(a)(i)}(n) \left\langle \frac{1}{S^{(a)}} \right\rangle_{n_i = n} \tag{.10}$$

where $\left\langle \frac{1}{S^{(a)}} \right\rangle_{n_i=n}$ is the expected value of the inverse of the total number of species, $S^{(a)}$ in the sample, weighted by the conditional probability distribution $P^{(a)}\left(n_1, n_2, \ldots, n_s | n_i = n\right)$ and summing over all possible data sets $\{n_j\}$ with the only restriction that the abundance of the ith species is fixed to be n.

If we compare Eq. (.10) with Eq. (4) in the main text we observe that sample SAD is written now as weighted sum rather than a simple sum of marginal probability distribution over the species in the system. These weights are different depending on the particular sampling strategy adopted. In any case, at a particular spatial scale a the sample SAD results from the same assemblage procedure: a sum over species-specific sampling abundance distributions where each factor weights relative differences between species (see also table S for a summary).

S2 Differences in species sampling detectability

The importance of heterogeneity in species detectability (sampling asymmetries) has been explored before in the context of the estimation of species richness (Boulinier *et al.*, 1998). Here we present the simple model to introduce differences in sampling detectability across species that has been used to generate our results (Fig 1, 2, and 3 in the main text).

Species at abundance n	**Regional**	**Sample** (S is unknown)
Average	$E[S_y] = \sum_{i=1}^{S} P^{(i)}(y)$	$E^{(a)}[S_n] = \sum_{i=1}^{S^{(a)}} C^{(i)(a)} P^{(i)(a)}(n)$
Fraction	$P(y) = \frac{1}{S} \sum_{i=1}^{S} P^{(i)}(y)$	$P^{(a)}(n) = \sum_{i=1}^{S^{(a)}} \hat{C}^{(i)(a)} P^{(i)(a)}(n)$

Table S.1 **Ensemble formulas for SADs** Here, we summarize the ensemble formulas at the sample level if we don't know the regional total richness, S, or are dealing with expectations for the fraction of abundances with abundance n in the sample. Compare this table with Table 2 in the main text.

Under the assumption of independent sampling, the general species-specific sampling distribution is given by Eq. (11) in the main text, which we reproduce here once again for the sake of completeness:

$$P^{(a)(i)}(n|y) = \sum_{m=n}^{y} P^{(i)}(n|m)\, P^{(a)}(m|y) \tag{.1}$$

When we sample at spatial scale a and assume equal sampling effort across species the probability of a species contribution with m individuals a this spatial scale given that the species has regional abundance y is given by the sampling distribution $P^{(a)}(m|y)$. Two main models have been used to deal with scale-dependent sampling: random and spatially aggregated sampling. When random sampling can be assumed, either the Poisson or the binomial distribution have been used. While when spatial-aggregation is expected to be important, the negative binomial distribution is used instead. In either case, we require that if the regional abundance of a species is y, then the expected abundance at the sample scale is:

$$E[m] = a\, y \tag{.2}$$

where a defines a common spatial sampling scale across species.

If we now assume different species detectabilities, we mean that we are not detecting in our sample all potential m individuals from a given species, but a species-specific fraction of them, p_i. This situation can be modeled by using the binomial distribution controlled by a species-specific detectability parameter, p_i.

In particular, by assuming spatial aggregation is important, we can rewrite Eq. (.1) by using a compound binomial negative binomial distribution:

$$P^{(i)(a)}(n|y) = \sum_{m=n}^{y} \left(\begin{array}{c} m \\ n \end{array} \right) p_i^n\, (1-p_i)^{m-n}\, \frac{\Gamma(k_i+m)}{\Gamma(k_i)\, m!} \left(\frac{a\, y}{a\, y + k_i} \right)^m \left(\frac{k_i}{a\, y + k_i} \right)^{k_i} \tag{.3}$$

where p_i is the species-specific detectability parameter, and k_i is the clumping parameter of the negative binomial which has a expected value for the number of individuals potentially collected in the sample given by $a\, y$.

If we define $\hat{m} \equiv m - n$, a simple rearrangement of the last equation gives rise to:

$$P^{(i)(a)}(n|y) = \left(\frac{k_i}{a\, y + k_i} \right)^{k_i} \left(\frac{p_i\, a\, y}{a\, y + k_i} \right)^n \frac{\Gamma(k_i+n)}{\Gamma(k_i)\, n!} \sum_{\hat{m}=0}^{y-n} \frac{\Gamma(\hat{m}+n+k_i)}{\hat{m}!\, \Gamma(n+k_i)} \left(\frac{(1-p_i)\, a\, y}{a\, y + k_i} \right)^{\hat{m}} \tag{.4}$$

Since the regional abundance y is usually very large (much larger that the observed samples abundances, n), the sum over $\hat{m}$ that appears in the last equation can be approximated with great accuracy by recalling that:

$$\sum_{j=0}^{\infty} \frac{\Gamma(n+j)}{j!\,\Gamma(n)\,F^j} = (1-F)^{-n} \tag{.5}$$

Therefore, we can finally write:

$$P^{(i)(a)}(n|y) = \frac{\Gamma(n+k_i)}{\Gamma(k_i)\,n!} \left(\frac{a_i\,y}{a_i\,y+k_i}\right)^n \left(\frac{k_i}{k_i+a_i\,y}\right)^{k_i} \tag{.6}$$

where $a_i \equiv a\,p_i$. This is the species-specific negative binomial sampling distribution that has been given in Eq. (12) in the main text and used in most part of this work. Species detectability introduces a species-specific shift in the expected value of the negative binomial but leaving k_i, the species-specific clumping parameter, unchanged.

If we assume that individuals are placed at random on the landscape at the spatial scale we are sampling, then $P^{(i)(a)}(m|y)$ can be assumed to be the binomial distribution:

$$P^{(i)(a)}(m|y) = \left(\begin{array}{c} y \\ m \end{array}\right) a^m\,(1-a)^{y-m} \tag{.7}$$

where y is the actual abundance in the regional community and m is the potential abundance we would observe at sampling effort or spatial scale a. As before, the definition of a is such that $E^{(a)}[m] = y\,a$. If there are differences in species detectability we will not be able to count all potential individuals m but a fraction of it that will depend on a species specific detectability factor, p_i. Thus, under random sampling we can write Eq. (.1) as a compound binomial binomial distribution:

$$P^{(i)(a)}(n|y) = \sum_{m=n}^{y} \left(\begin{array}{c} m \\ n \end{array}\right) p_i^n\,(1-p_i)^{m-n} \left(\begin{array}{c} y \\ m \end{array}\right) a^m\,(1-a)^{y-m} \tag{.8}$$

By rearranging the last expression and using $\hat{m} \equiv m-n$, we can write:

$$P^{(i)(a)}(n|y) = \left(\begin{array}{c} y \\ n \end{array}\right) (p_i\,a)^n \sum_{\hat{m}=0}^{y-n} \left(\begin{array}{c} y-n \\ \hat{m} \end{array}\right) [a\,(1-p_i)]^{\hat{m}}\,[1-a]^{y-n-\hat{m}} \tag{.9}$$

where the sum is actually the expansion of the binomial $(1-a\,p_i)^{y-n}$. Notice that this result does not require any approximation. It is equally exact for the whole range of abundance values y at the community or regional level. It has been used to generate Fig 3.

In sum, if we assume that individuals are placed at random on the landscape at the spatial scale we are sampling, a, the species-specific sampling distribution is also binomial, with a species-specific average $E^{(a)}[n] = y\,a_i$, where $a_i = a\,p_i$:

$$P^{(i)(a)}(n|y) = \left(\begin{array}{c} y \\ n \end{array}\right) (a\,p_i)^n\,(1-a\,p_i)^{y-n} \tag{.10}$$

S3 The logseries and the multinomial likelihood

In this appendix we show that logseries parameters, as commonly estimated, are the maximum likelihood estimates (m.l.e) of a multinomial likelihood function that relies on species symmetry (see Eq. (1) in the main text). Chao & Bunge (2002) already applied this likelihood to obtain maximum likelihood estimates of the number of non-observed species. They do explicitly comment that this likelihood is based on independent sampling and on the absence of differences in species detectability, but not present an explicit derivation that classical Fisher parameters are obtained by maximazing this likelihood. Here we add this simple calculation for the sake of completeness.

Fisher's classical recipe to estimate the logseries parameters, α and x, from a given abundance data set with N individuals and S species is given by:

$$S = -\alpha \log(1-x) \tag{.1}$$

$$N = -\alpha \frac{x}{1-x} \tag{.2}$$

This is a non-linear system of two equations and two parameters that can be readily solved numerically (Rosenzweig, 1995).

We, first, recall that, as a probability distribution, Fisher logseries is a uni-parametric distribution:

$$P(n) = \gamma(x)\,\frac{x^n}{n} \tag{.3}$$

The normalizing condition, $\sum_{n=1}^{\infty} P(n) = 1$ requires the normalization factor γ to be:

$$\gamma(x) = \frac{1}{\left[\log\left(\frac{1}{1-x}\right)\right]} \tag{.4}$$

In fact, the first equation in Fisher's recipe is nothing but a straightforward consequence of this normalization. When the whole probability is rescaled by a factor S, the total number of species in the sample, we obtain that Fisher α is equal to $\gamma\, S$.

Let us write the abundance model that gives the probability of observing a species with abundance n in the sample as:

$$P(n|x) = \gamma(x)\,\frac{x^n}{n} \tag{.5}$$

Now, we simply re-write the multinomial probability as a loglikelihood function, where the constant prefactor in Eq. (1) given by the factorials can be dropped because it has no effect on the optimization procedure to calculate the m.l.e:

$$\mathcal{L}(x|S_1,\ldots,S_m) = \log \sum_{n=1}^{m} S_n \,\log P(n|x) \tag{.6}$$

where m is the maximum abundance in the sample. Notice that this function depends on the single parameter x given some data, summarized as $S_1,\ldots,S_m$, where S_i is the number of species observed with abundance i in the sample.

As usual, in order x to be a maximum value, we require that:

$$\frac{d\mathcal{L}(x|S_1,\ldots,S_m)}{dx} = 0 \tag{.7}$$

$$\frac{d^2\mathcal{L}(x|S_1,\ldots,S_m)}{dx^2} < 0 \tag{.8}$$

The first derivative (Eq. (.7)) can be written as, after some algebra:

$$\frac{d\mathcal{L}(x|S_1,\ldots,S_m)}{dx} = -\sum_{n=1}^{m} S_n \frac{\gamma(x)}{(1-x)} + \sum_{n=1}^{m} n\, S_n\, x \tag{.9}$$

Since $\sum_{n=1}^{m} S_n = S$ and $\sum_{n=1}^{m} n\, S_n = N$, where S and N are the total number of species and individuals collected, Eq. (.7) and Eq. (.9) result in:

$$\frac{N}{S} = \frac{x}{1-x}\,\frac{1}{\left[\log\left(\frac{1}{1-x}\right)\right]} \tag{.10}$$

Notice that the average number of collected individuals per species is all we need to fit Fisher probability model, given by Eq. .5, to abundance data. Now, we observe that the normalization condition, $\alpha = \gamma(x)\, S$, and the last equation make up a non-linear system which is equivalent to Fisher's classical recipe in Eqs. (.1)-(.2), as we announced.

The second derivative (Eq. (.8)) can be shown to be negative. In fact, when evaluated on the extreme value for x, which satisfies Eq. (.10), the inequality (.8) boils down to the condition $x\,(1-x) < 0$, which is always true since we require x to be strictly less than 1 for probabilities to be well-defined. In conclusion, the estimate for x (and, therefore, α) is a maximum of the function $\mathcal{L}(x|S_1,\ldots,S_m)$, and so it is of the initial multinomial likelihood —which is based on species sampling independence and symmetry as given by Eq. (1) in the main text.

References

Boulinier, T., Nichols, J. D., Sauer, J. R., Hines, J. E. & Pollock, K. H. (1998). Estimating species richness: the importance of heterogeneity in species detectability. *Ecology* **79**:1018–1028.

Chao, A. & Bunge, J. (2002). Estimating the number of species in a stochastic abundance model. *Biometrics* **58**:531–539.

Etienne, R. S. & Alonso, D. (2005). A dispersal-limited sampling theory for species and alleles. *Ecol. Lett.* **8**:1147–1156.

Rosenzweig, M. L. (1995). *Species Diversity in Space and Time*. Cambridge University Press, Cambridge.