IDEA AND PERSPECTIVE

The implicit assumption of symmetry and the species abundance distribution

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*
*Correspondence: E-mail: d.alonso@rug.nl

Abstract

Species abundance distributions (SADs) have played a historical role in the development of community ecology. They summarize information about the number and the relative abundance of the species encountered in a sample from a given community. For years ecologists have developed theory to characterize species abundance patterns, and the study of these patterns has received special attention in recent years. In particular, ecologists have developed statistical sampling theories to predict the SAD expected in a sample taken from a region. Here, we emphasize an important limitation of all current sampling theories: they ignore species identity. We present an alternative formulation of statistical sampling theory that incorporates species asymmetries in sampling and dynamics, and relate, in a general way, the community-level SAD to the distribution of population abundances of the species integrating the community. We illustrate the theory on a stochastic community model that can accommodate species asymmetry. Finally, we discuss the potentially important role of species asymmetries in shaping recently observed multi-humped SADs and in comparisons of the relative success of niche and neutral theories at predicting SADs.

Keywords

Diversity patterns, logseries, neutrality, species abundance distribution, stochastic community models, symmetry.

Ecology Letters (2008) **11**: 93–105

INTRODUCTION

Most ecologists agree that species abundance distributions (SADs) represent at least an empirical tool to draw a rough characterization of ecological communities (McGill *et al.* 2007). In recent years, species abundance patterns have played a relevant role in the search for empirical evidence to support particular theories (Hubbell 2001; McGill 2003; Volkov *et al.* 2003, 2005; Azaele *et al.* 2006). It has been often argued that the same patterns of species abundances can arise through different underlying processes (Cohen 1968; Chave *et al.* 2002; Purves & Pacala 2006), and, currently, there seems to be general consensus that SADs provide little information on these processes. However, difficulties in identifying the leading process giving rise to a certain ecological pattern are in part caused by the difficulties in providing sound statistical approaches and alternative models. General stochastic dynamical 'niche' models that are able to make quantitative predictions about the SAD pattern are still lacking (Alonso *et al.* 2006, but see Tilman 2004 and Walker 2007). Therefore, it has been difficult to empirically test such models and compare them

to their neutral counterparts. For instance, there still exists no satisfactory alternative null model to neutrality (Gotelli & McGill 2006) that incorporates asymmetry in dynamics and sampling. Moreover, the common statistical sampling theory that has been used to carry out these analyses makes some implicit assumptions that have been largely overlooked.

Here, we emphasize one such limitation of all current sampling theories (Fisher *et al.* 1943; Pielou 1969; Bulmer 1974; Dewdney 1998; Lande *et al.* 2003; Alonso & McKane 2004; Etienne & Alonso 2005; Green & Plotkin 2007): they ignore species identity, i.e. they assume species symmetry. Evidently, species are different in their life-history traits, such as dispersal and movement. These inherent properties, which are independent from the observer's eye, will determine the distribution of species on a landscape and, hence, the probability of observing a number of individuals of a species from a given sampled area. This distribution arises from the interplay between dispersal (both passive and active), abiotic habitat preferences, and both intra- and inter-specific interactions, processes which all can be very species-specific. Moreover, global processes and environmental variability may constrain species abundances at the

regional level also in a species-specific and diverse manner. Apart from these inherent characteristics, species may simply have different probabilities of being picked up in a sample. They may vary in how easily they are captured, or in size and hence visibility. Therefore, these two sources of species-specificity – inherent and observer-induced differences – warrant an asymmetric treatment of abundances at a given sampling scale.

To our knowledge only one theoretical contribution considers the problem of the distribution of species abundances in the context of a fully asymmetric species assemblage, an *idiosyncratic* community (Pueyo *et al.* 2007). If species are idiosyncratic, i.e. essentially different from each other, these idiosyncrasies somehow balance out and simple descriptions emerge: the most revisited SAD models, the logseries and the lognormal, are shown to arise as the most likely curves under biologically reasonable global constraints acting on the community. However, Pueyo *et al.*'s (2007) approach is fully statistical. It is not meant to study particular mechanisms, and does not focus on the particularities of the sampling process. Although this generality clearly is the main strength of MaxEnt methods (Jaynes 1968), the approach developed by Pueyo *et al.* (2007) is not designed to investigate plausible dynamical processes and alternative models controlling species abundances and community composition in ecological communities.

In this paper we explain the ways in which classical and current sampling theories ignore the possibility of species asymmetry, and discuss the potential for it to lead to multi-humped sample SADs that ecologists have recently observed (Magurran & Henderson 2003; Marquet *et al.* 2004; Gray *et al.* 2005; Labra *et al.* 2005). Furthermore, we show how the community-level SAD emerges as a sum over the distribution of population abundances of the species integrating the community. In this way, the species ensemble can be described by a community-level property: the SAD curve. As different ecological processes, such as competition, mortality, immigration, niche differentiation and species interactions determine species-level dynamics, our approach provides a natural link and a data-driven empirical framework to investigate the relationship between such underlying processes and SAD patterns. We also point out the importance of this issue when measuring the success of niche theory predictions on the SAD against that of neutral theory predictions, a recently popular activity that ecologists have carried out under the assumption of species symmetry (McGill 2003, 2006; Volkov *et al.* 2003).

Our paper is organized in three sections. In the first section, we explain how to modify existing statistical sampling theory to take into account species asymmetries, and how an SAD, as a community-level pattern, can in general be related to the underlying population dynamics of the species constituting the community. In the second section, as an example, we apply the theory to a stochastic model community where species differ and perform an independent birth–death–immigration process with density dependence. We show how asymmetry can arise through species differences in both dynamics and species-specific sampling properties. In the third section, we summarize our results and discuss their relevance in the search for the main determinants of community dynamics and organization.

SAMPLING THEORY FOR SADS

The SAD describes the number (or fraction) of species represented by a given abundance. For a given region, observations of this distribution depends on the spatial scale at which that region is sampled, or the number of individuals collected if sampling is performed randomly in space. Sampling models derive the form of the sample SAD in terms of the regional SAD and the quantitative impacts of sampling on species' abundances (see Table 1).

Historically, SADs have had a dual interpretation either as the expected fraction or the average number of species with a given abundance in a sample given a certain sampling effort, for instance, the total area sampled, the total time invested, the total number of individuals collected, etc. Here, we focus on sample SADs defined as the expected or average number of species with a given number of individuals. We explore equivalent expressions for the fraction or probability of observing a species with a given number of individuals in a sample in Appendix S1.

In this work, we adopt the following definitions: $P(y)$ is the probability for a species to be represented by y individuals in a community. This community can be either a whole region or a locally assembled community, from which samples are taken. Throughout this paper, we will call it 'regional community' or simply 'community'. Particular applications will make it clear what we mean in each case. We further define $E^{(a)}[S_n]$ as the sample SAD, i.e. the expected number of species that are represented by n individuals in a sample from a community or region at a given spatial scale or sampling effort, a. We further define $P^{(a)(i)}(n)$ as the probability of observing a species i with abundance n in a sample at a given spatial scale or sampling effort, a.

Symmetry and species independence

The study of the distribution of species abundance in complex communities usually starts with a list of species abundances in observed samples from an ecological survey. This information is fully represented by a sample abundance vector, $\vec{n} \equiv (n_1, n_2, \ldots, n_S)$, where n_i is the abundance of species i in a sample which will have resulted from sampling

Table 1 Sampling theory for SADs. Main examples of statistical sample SADs. Most of these distributions have been introduced as compound statistical distributions without further interpretation

	Reference	Type of sampling, $P^{(a)}(n \lfloor y)$	Regional or community-level SAD, $P(\ y)$
1.	Logseries, Fisher *et al.* (1943)	Poissonian	Gamma distribution
2.	Poisson-lognormal, Bulmer (1974)	Poissonian	Lognormal
3.	Dewdney (1998)	Hypergeometric	Generic
4.	Lande *et al.* (2003)	Poissonian and intraspecific overdispersion	Generic
5.	Neutral theory, Volkov *et al.* (2003)	Dispersal-limited	Gamma distribution
6.	Neutral theory, Alonso & McKane (2004)	Dispersal-limited	One-dimensional representation of Ewens distribution
7.	Neutral theory, Etienne formula, Etienne & Olff (2004), Etienne (2005), Etienne & Alonso (2005)	Dispersal-limited	Multivariate full representation of Ewens distribution
8.	Green & Plotkin (2007)	Negative binomial	Generic

Here, we provide a common description in terms of the type of sampling and the regional community SAD assumed. Note that both Fisher *et al.* (1943) and Volkov *et al.* (2003) considered basically the same distribution for the regional SAD: a continuous representation of a logseries, i.e. a Gamma distribution where one of the parameters is taken to be vanishingly small. While all these expressions consider the sample SAD, $P_S(n)$, to be evaluated for $n = 0, 1, \ldots,$, the abundance distribution at the regional level is assumed to be either a continuous or a discrete function of regional abundances. In essence, all these formulas entail a summation over all possible abundances in the region represented by a certain probability distribution. This summation can be carried out either by an integral, when the distribution at the regional scale is represented by a continuous distribution (1, 2, 4, 5, 6, 8), or by a true discrete sum, when regional abundances are assumed to take discrete values instead (3, 7, see also the example below). In any case, each of these expressions generally assume that the probability for a species to reach certain abundance at the regional scale is the same regardless of species identity. In neutral theory this symmetry is a consequence of the theory's first principles. Unlike the other expressions, the Etienne sampling formula (7) can be seen as a multivariate sample SAD (see eqn 3). These multivariate expressions have been generally called sampling formulas (Etienne *et al.* 2007).

a given community with a given sampling effort. A particular spatial scale defines the community under study. A common way to summarize this information is to count how many species are represented by one individual (singletons), S_1, two individuals, $S_2, \ldots$, until S_m, which is the number of species (usually only one – the most common one) having observed with m individuals, where m is the highest abundance observed in the sample.

Suppose now that we are given an SAD model, such as the logseries (Fisher *et al.* 1943) or the Poisson lognormal (Bulmer 1974). In general, an abundance model provides a probability of observing a species with abundance n is a sample, $P(n|\Theta)$, which depends on some model parameters Θ. A well-established method to fit the model to our abundance data is to use a maximum likelihood approach. All we need is to build a likelihood function which is proportional to the probability of obtaining the observed data set under model assumptions.

In this context, a common likelihood function is given by the multinomial distribution (Chao & Bunge 2002; Alonso & McKane 2004; Etienne *et al.* 2007):

$$\mathscr{P}\{S_1, \ldots, S_m | \Theta\} = \frac{S!}{S_1! \cdots S_m!} P(1|\Theta)^{S_1} \cdots P(m|\Theta)^{S_m},$$

(1)

where S is the total number of species observed in the sample. This expression gives the probability of observing a given data set, $S_1, \ldots, S_m$, when sampling a fixed number of species, S, which can be represented by a conditional probability given S (Etienne & Olff 2005).

As we have observed a singleton species S_1 times, a species with two individuals S_2 times, and so on, and the abundance model provides the probabilities of observing a singleton species, $P(1|\Theta)$, a species with two individuals, $P(2|\Theta)$, and so on, eqn (1) simply provides the probability of obtaining the empirical data, $S_1, \ldots, S_m$, for a given choice of model parameters, Θ. One can show, for instance, that if the abundance model is the logseries, Fisher *et al.*'s (1943) recipe to calculate model parameters is recovered by maximization of this function over parameter values given the data. Therefore, Fisher *et al.*'s estimates are actually the maximum likelihood estimates of model parameters based on the likelihood function eqn (1) (see Appendix S3). Likelihoods are important because when we are studying two alternative abundance models, we can ask: 'How probable are the data given the model?' and use this information as an objective model selection criterion (Hilborn & Mangel 1997, p. 7).

However, an implication of the use of this multinomial likelihood, as well as similar likelihood functions in this

Table 2 Ensemble formulas for SADs. The community-level SAD arises from an ensemble average over the distributions for all species in the system

Species at abundance (n)	Regional	Sample (S is known)
Expected number (average)	$E[S_y] = \sum_{i=1}^{S} P^{(i)}(y)$	$E^{(a)}[S_n] = \sum_{i=1}^{S} P^{(a)(i)}(n)$
Probability (fraction)	$P(y) = \frac{1}{S} \sum_{i=1}^{S} P^{(i)}(y)$	$P^{(a)}(n) = \frac{1}{S} \sum_{i=1}^{S} P^{(a)(i)}(n)$

Here, we summarize the expressions for the ensemble averages of the sample and the region when the total number of species in the region is known. These are the expressions we have used in the analysis of our model community. At the sample level, these expressions become more complicated if we do not know the regional total richness, S, or are dealing with expectations for the fraction of abundances with abundance n in the sample or both (see Appendix S1).

context (Bulmer 1974), is that all rely on the assumption that the species abundances in the sample, $n_1, \ldots, n_S$, are independent identically distributed random variables. In other words, we are assuming that the same underlying abundance distribution holds independently for all sampled species regardless of species identity. These two underlying quite strong assumptions – symmetry and species independence – have been largely overlooked or never clearly stated in common applications (but see Etienne & Olff 2004; Etienne 2005; Etienne & Alonso 2005; Etienne & Olff 2005 and Etienne 2007 with regard to species interdependence and the neutral sampling theory). In the next section we extend current sampling theory to accommodate species asymmetries.

An asymmetric sampling theory

To take into account species asymmetry, we first note that the regional community-level SAD, $P(y)$, can be written as the mean over the species-level abundance distributions $P^{(i)}(y)$ of each of the S species (labelled by the subscript i) present in the region. This was already noted by Volkov et al. (2003) where a non-interactive community was assumed. However, the same holds when species within the community interact and abundances are described by a joint multivariate abundance distribution (see Appendix S1 for a proof of this). We call this expression an 'ensemble formula' (see Table 2) because it summarizes information about the species ensemble defining the community:

$$P(y) = \frac{1}{S} \sum_{i=1}^{S} P^{(i)}(y) \qquad (2)$$

For each species, $P^{(i)}(y)$ is the probability for that species i to have y individuals at the regional community. This probability can be thought of as describing a distribution across regions or over time. The regional abundance distributions (at the either the species- or community-level) are purely theoretical. They are not directly accessible by the observer but only indirectly through sampling.

The effect of correlations between the abundances of different species (species interdependence) is potentially quite important when species differ. In the case of species symmetry, species interdependence arising from zero-sum dynamics leads to a sampling theory which is equivalent to the case of species independence under reasonable assumptions (Etienne et al. 2007). A complete sampling theory that preserves species identity must consider the multivariate probability distribution for species' abundances (Etienne et al. 2007), and rely, accordingly, on a multivariate version (Etienne et al. 2007) of classical sampling distributions (Fisher et al. 1943; Pielou 1969; Bulmer 1974; Dewdney 1998; Lande et al. 2003; Volkov et al. 2003; Alonso & McKane 2004; Etienne & Alonso 2005; Green & Plotkin 2007). This multivariate expression is given by (note that this is an application of the law of total probability):

$$P^{(a)}(\vec{n}) = \int P^{(a)}(\vec{n}|\vec{y}) P(\vec{y}) \, d\vec{y} \qquad (3)$$

where $\vec{n} \equiv (n_1, n_2, \ldots, n_S)$ and $\vec{y} \equiv (y_1, y_2, \ldots, y_S)$ are the abundance vectors in a sample at spatial scale a and in the regional community, respectively. $P(\vec{y})$ is the probability for the community to be described by an abundance vector $\vec{y}$, $P^{(a)}(\vec{n})$ is the probability to obtain an abundance vector $\vec{n}$ in a sample at scale a, and $P^{(a)}(\vec{n}|\vec{y})$ is a sampling transformation distribution which characterizes the probability of obtaining the abundance vector $\vec{n}$ in a sample at scale a given that the regional community abundances are exactly described by the abundance vector $\vec{y}$.

In Appendix S1, we show how to relate a given multivariate sample abundance distribution, $P^{(a)}(\vec{n})$, to a sample SAD. Some progress can be made with this multidimensional approach (Etienne & Alonso 2005; Etienne et al. 2007), but this requires specific assumptions about the sampling process and the spatial dynamics of the community. In this paper, we focus on non-multivariate SADs where species differ both in the sampling process and in the inherent species specific regional abundance distributions. In general terms, an equation analogous to eqn (2),

but at the sampling scale a can also be found (see Table 2 and Appendix S1):

$$E^{(a)}[S_n] = \sum_{i=1}^{S} P^{(a)(i)}(n) \tag{4}$$

where S is the total number of species in the regional community and $P^{(a)(i)}(n)$ is the probability for a species i to have n individuals in a sample taken at scale (or effort) a. If one does not have any knowledge about the total number of species in the region, this sum becomes more complicated (see Appendix S1).

If species are sampled independently, we can take asymmetry into consideration by expressing the sample abundance distribution of each species, appearing in eqn (4) as

$$P^{(a)(i)}(n) = \int P^{(a)(i)}(n|y) P^{(i)}(y)\, \mathrm{d}y \tag{5}$$

In principle, if a multivariate sample distribution, $P^{(a)}(\vec{n})$, can be calculated, we can check to what extent the independent sampling assumption holds. In theory we could also recover $P^{(a)(i)}(n)$ by calculating the corresponding marginal probability distribution from the joint distribution for a sample at sampling effort a, $P^{(a)}(\vec{n})$:

$$P^{(a)(i)}(n) = \sum_{\{n_j\}_{j \neq i}} P^{(a)}(\vec{n}) \tag{6}$$

where the sum is performed over all possible abundance combinations, $\{n_j\}$, with species i being represented by exactly n individuals. In many practical cases, the computation of this sum will be complicated if not impossible.

By substituting eqn (5) into eqn (4), we obtain that, in the case of species asymmetry and independent sampling, the expected sample SAD takes the form:

$$E^{(a)}[S_n] = \sum_{i=1}^{S} \int_0^{\infty} P^{(a)(i)}(n|y) P^{(i)}(y)\, \mathrm{d}y \tag{7}$$

where $P^{(a)(i)}(n|y)$ is the species-specific sampling transformation distribution which, depending on the nature of the sampling, can model both intraspecific aggregation and species-specific detection probabilities.

This expression can be seen as an ensemble sum over species-specific sampling abundance distributions where each term takes into consideration inherent relative differences between species. Importantly, only particular distributions preserve their shape under the assemblage procedure represented by eqn (4) and related ensemble formulas (see Table 2 and Appendix S1). For instance, power laws with the same exponent are invariant under assemblage (Pueyo 2006). Most distributions do not preserve their

shape under assemblage. The assemblage of a number of lognormal distributions with different variances does not produce a lognormal (Allen *et al.* 2001), in general. In fact, we can only be sure that the shape will be preserved if we can assume species symmetry. In this trivial case, the superindex (i) can be dropped in eqn (7), all terms become identical, and we recover a symmetric expression for the sample SAD, $E^{(a)}[S_n]$:

$$E^{(a)}[S_n] = S \int_0^{\infty} P^{(a)}(n|y) P(y)\, \mathrm{d}y \tag{8}$$

Current sampling theories do not consider a sampling transformation distribution that explicitly depends on species identity, and do not take into account species differences in regional abundances either. They predict what we observe in our samples without considering species identity: all of them use the same basic formula for the sample SAD (eqn 8). The expression given in eqn (7) is a simple extension of previous theory. Importantly, it comes from the more general eqn (4) under the assumption of species being detected and sampled independently from one another. In the next section we present a straightforward application of this equation.

A STOCHASTIC MODEL COMMUNITY WITH DENSITY DEPENDENCE

Here, we illustrate the use of ensemble formulas (eqns 2 and 4) to make a theoretical prediction of the community-level sample SAD in terms of species-specific population abundances in the region and sampling transformation distributions at a given sampling effort. Our example considers a community of S non-interacting species undergoing a simple birth–death–immigration process (Alonso & McKane 2002). These stochastic population dynamics are summarized in Box 1.

Our goal in this section is to show that if we assume that species are inherently different either in sampling probabilities or in species-specific dynamics, these differences translate into a community-level sample SAD that can be predicted by using our ensemble formulas (see Table 2). In this way, the exact shape of this curve can be related to the sampling and dynamical specific properties of the species in the community. Therefore, in the context of stochastic community process-based models of the kind presented in Box 1, there is quite a tight correspondence between the underlying process and the shape of the SAD curve. Although two different processes can predict qualitatively very similar shapes (see Figs 1 and 2) for the community-level SAD, as these curves will be essentially different, it is not theoretically impossible to sort out the most likely process at work given some data. In practice, this is only

Box 1: The stochastic community model

Here we consider an ensemble of S species where each of them performs an independent birth–death–immigration process with density dependence (Alonso & McKane 2002). When species independence is assumed (Volkov *et al.* 2005; Azaele *et al.* 2006), the dynamics of species abundance can be described by a one-dimensional master equation for the probability of that species attaining abundance n at a given time t, $P^{(i)}(n,t)$. Usually it is said that population abundances perform a random walk. This

approach has some advantages. First, it is possible to built exact stochastic simulations by using a simple standard algorithm (Gillespie 1976; Renshaw 1991). Second, although the mathematics to treat with these kind of equations can be complicated, model description is actually very simple (see Table A1). For instance, the whole model used to generate the results presented in Fig. A1 (and Figs 1–3 in the main text), is described only by three elemental processes that affect all individuals of all species:

Table A1 Stochastic community models. These are recent and classical community models based on the species independence assumption as described by their total transition rates for either a decrease or an increase of one individual in population abundance

	Reference	Birth and death transition rates	
		$n \rightarrow n + 1$	$n \rightarrow n - 1$
1	Kendall (1948)	$b_0 n + m_0$	$d_0 n$
2	Hubbell (2001)	$b_0 \left(1 - \frac{n}{N}\right)(n + Ip)$	$b_0 \frac{n}{N}\left[(N - n) + I(1 - p)\right]$
3	Volkov *et al.* (2005)	$b_0 n + c$	$d_0 n + c$
4	Azaele *et al.* (2006)	$b_0 n + c$	$d_0 n - c$
5	In this article	$b_0 \left(1 - \frac{n}{N}\right)n + m_0(N - n)$	$d_0 n$

Note that Hubbell's standard model for the local community (Hubbell 2001) can be seen as a zero-sum reformulation of Kendall birth–death–immigration classical model and it is equivalent to it at the stationary steady state (Etienne *et al.* 2007). Therefore, models 1 and 2 introduce basically the same kind of immigration-induced density dependence due to an independent term m_0 in the birth rate. Models 3 and 4 introduce effectively also the same kind of density dependence due to the term c in the rates. Unlike the model presented here, none of those is able to produce logistic growth towards a carrying capacity (see Fig. A1). Other ways of introducing density dependence are also possible (Keeling 2000).
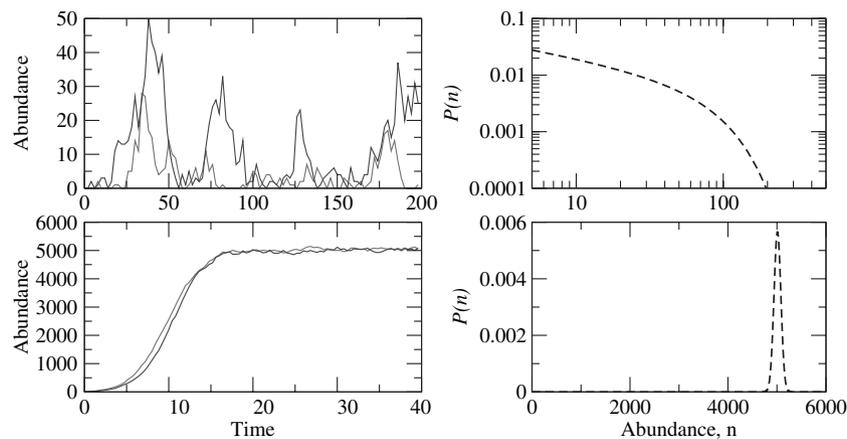

Figure A1 Time evolution of two different species. Upper panels correspond to a low-immigration species invading an area where local growth can barely compensate deaths – a sink. Parameter values are $m_0 = 5.0 \times 10^{-5}$, $b_0 = 0.99$, $d_0 = 1$ and $N = 10^4$. The stochastic trajectories – governed by the three rates given by eqns (A2)–(A4) – of two populations (red and blue) are shown in each case. When immigration is very low, there are constant extinction and reinvasions. Lower panels correspond to a species that thrives in that area reaching carrying capacity after invasion in a logistic almost deterministic fashion. Parameter values are $m_0 = 10^{-3}$, $b_0 = 1$, $d_0 = 0.5$ and $N = 10^4$. On the right the corresponding steady-state probabilities for the species having abundance n at the steady-state are also shown. The logseries shape is characteristic of the low-immigration species, while Gaussian fluctuations around the carrying capacity give rise to the abundance distribution in the second case. These curves are calculated following Alonso & McKane (2002). In both cases, no individual is present in the system at the initial time.

Box 1 (*Continued*)

1. *Local birth*: The percapita birth rate is assumed to be density dependent:

$$b^{(i)} = b_0^{(i)}\left(1 - \frac{n^{(i)}}{N^{(i)}}\right) \qquad (A1)$$

where $N^{(i)}$ is species-specific maximum carrying capacity and $b_0^{(i)}$ is basal percapita rate. Therefore, at any time, the total probability for the species to increase in one individual due to local births per unit time is given by:

$$T_B[n^{(i)} + 1 \mid n^{(i)}] = b_0^{(i)}\left(1 - \frac{n^{(i)}}{N^{(i)}}\right)n^{(i)} \qquad (A2)$$

2. *Local death*: The percapita death rate is assumed to be density independent. Therefore, at any time, there is a total probability for the species to decrease in one individual due to death per unit time given by:

$$T_D[n^{(i)} - 1 \mid n^{(i)}] = d_0^{(i)} n^{(i)} \qquad (A3)$$

3. *Immigration*: Finally, there is also a total probability for the community to receive an immigrant belonging to a given species per unit time that depends on the degree of saturation according to a species-specific maximum carrying capacity, $N^{(i)}$:

$$T_I[n^{(i)} + 1 \mid n^{(i)}] = m_0(N^{(i)} - n^{(i)}) \qquad (A4)$$

These processes allow to define total probability rates of population increase or decrease, $g_n = T_B[n + 1 \mid n] +$

$T_I[n + 1 \mid n]$ and $r_n = T_D[n-1 \mid n]$, respectively (see Table A1). In turn, these total rates are used to build a master equation (McKane *et al*. 2000; Volkov *et al*. 2003) from which it is easy to see that a population governed by these processes tends to a stationary abundance distribution. Thus, each species abundance in the region will tend to a well-defined stationary distribution, $P^{(i)}(n)$, that can be calculated analytically (Alonso & McKane 2002) (see Fig. A1). In addition, it can be also shown that average population abundances are governed by (van Kampen 1992):

$$\frac{d\langle n \rangle}{dt} = \langle g_n \rangle - \langle r_n \rangle \qquad (A5)$$

When N is large, i.e. when $\langle n^2 \rangle \approx \langle n \rangle^2$, eqn (A5) gives rise to logistic growth towards a carrying capacity (see Fig. A1):

$$\frac{d\langle n^{(i)} \rangle}{dt} = b_0^{(i)}\left[1 - \frac{\langle n^{(i)} \rangle}{N^{(i)}}\right]\langle n^{(i)} \rangle + m_0^{(i)}(N^{(i)} - \langle n^{(i)} \rangle) - d_0^{(i)}\langle n^{(i)} \rangle \qquad (A6)$$

Moreover, according to the stochastic rates above, we can compare this process to other stochastic community models from the literature which are also based on species independence (see Table A1). In Fig. A1, we show the time evolution of the abundance of two different species and the corresponding species-level abundance distributions at the stationary steady state.

---

constrained by the statistical power associated with the amount of data and the particular way they were collected. For instance, if we have a single snapshot of the community, it will be in general very difficult to tease apart alternative models (Chave *et al*. 2006), but if we are provided either sequential sampling in time or replication across space our statistical power increases (Etienne 2007).

We have divided our study into two extreme scenarios: (A) Species are symmetric in their dynamics but differ in their sampling probabilities (Fig. 1), and (B) species differ in their dynamics, but not in their inherent sampling probabilities (Fig. 2). In reality, asymmetries arise both in the dynamics and in the sampling probabilities.

For the sake of simplicity, independence in dynamics as well as sampling is assumed. By assuming also that the number of species in the region is known, we can calculate the sample SAD or average number of species in a sample with abundance n, $E[S_n]$ (see eqn 4). Given that the regional abundance distribution is discrete (see Box 1), the proba-

bility of contributing with n individuals to the sample for each species can be written as:

$$P^{(a)(i)}(n) = \sum_{y=n}^{N} P^{(a)(i)}(n \mid y)P^{(i)}(y) \qquad (9)$$

where $P^{(a)(i)}(n \mid y)$ is a species-specific sampling distribution. By substituting eqn (9) into eqn (4), we obtain:

$$E^{(a)}[S_n] = \sum_{i=1}^{S} \sum_{y=n}^{N} P^{(a)(i)}(n \mid y)P^{(i)}(y) \qquad (10)$$

There is a potential number of individuals m out of the total regional abundance y that will be collected in the sample according to some probability distribution. This potential number is determined by an overall sampling effort, a. However, the actual number n will be usually lower than m and, in general, will depend on a species-specific detectability factor, p_i. If $p_i = 1$, then the potential number of individuals of that species at that sampling
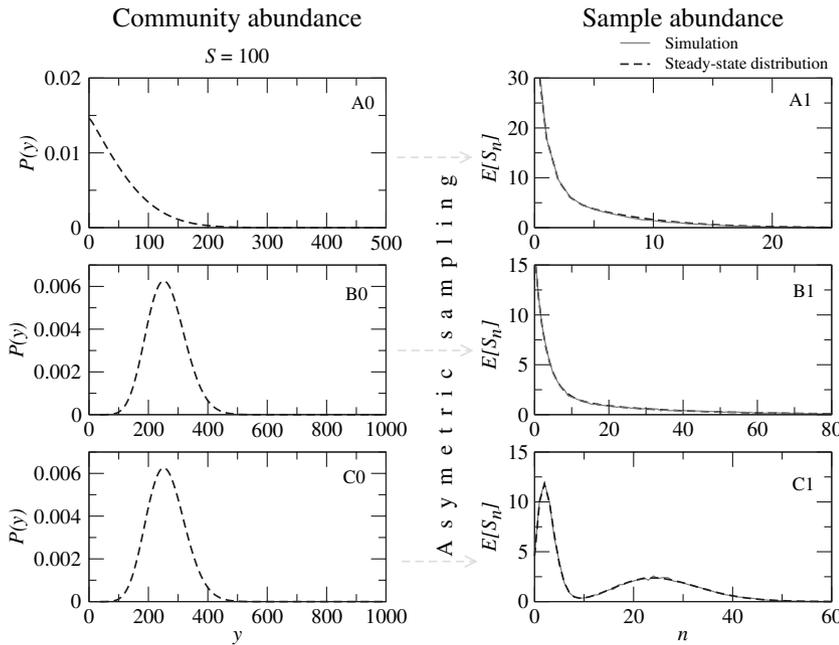
Figure 1 *Scenario A*. Here we show how asymmetric negative binomial sampling modifies the shape of an underlying regional distribution. The regional community is an ensemble of 100 species performing a birth–death dynamics (see Box 1) with identical parameters (A) $m_0 = 10^{-4}$, $b_0 = 0.99$ and $d_0 = 1$; (B, C) $m_0 = 10^{-3}$, $b_0 = 0.99$ and $d_0 = 1$. $N = 10^4$ in all three cases. In the three cases, we consider two sampling-induced groups of 50 species each ($a = 0.1$, $p_1 = 1$ and $a = 0.1$ and $p_2 = 0.1$), which can be detected in the sample SAD only in case C. The clumping parameter of the negative binomial is 100, 1 and 100 in the cases A, B and C, respectively. The curves resulting from simulation when the community is sampled at $t = 100$ time units are show in thin lines (panels A1, B1 and C1).

effort, m, can be fully detected ($n = m$). In general, this can be summarized as a compound distribution (see Appendix S2):

$$P^{(a)(i)}(n \mid y) = \sum_{m=n}^{y} P^{(i)}(n \mid m) P^{(a)}(m \mid y) \quad (11)$$

If we assume that each individual enters the sample according to a species-specific probability, p_i, and we assume that spatial aggregation is important, the compound sampling distribution in eqn (11) can be written as a single negative-binomial distribution with species-specific parameters $a p_i y$ and clumping parameter k_i (see Appendix S2 for further details):

$$P^{(a)(i)}(n \mid y) = \frac{\Gamma(n + k_i)}{n! \Gamma(k_i)} \left(\frac{a_i y}{a_i y + k_i} \right)^n \left(\frac{k_i}{a_i y + k_i} \right)^{k_i} \quad (12)$$

where $a_i = a p_i$.

In sum, a community-level sample SAD is fully determined by the abundance distributions of the species in the region, $P^{(i)}(y)$, through eqn (10), and by a particular type of sampling (for instance, eqns 11 and 12). These sampling distributions assume a species-specific differential detectability (under either random or negative binomial sampling assumption, see Appendix S2), and have been used to generate the predicted SADs for each scenario (right column panels of Figs 1–3).

Scenario A

This scenario is an example of genuine asymmetric sampling. We assume a symmetric community where all

species are described by the same regional abundance distribution. However, we assume that the community can be split into two groups of S_1 and S_2 species according to a different detectability factor. By using again the general eqn (10) and (9), we observe that the SAD can be written under this scenario as:

$$E^{(a)}[S_n] = \sum_{y=n}^{N} \left(S_1 P^{(a)(p_1)}(n \mid y) + S_2 P^{(a)(p_2)}(n \mid y) \right) P(y \mid \Theta) \quad (13)$$

where Θ stands for model parameter values, b_0, d_0, m_0, N (see Box 1) describing the species at the region R. The sampling distributions $P^{(a)(p_i)}(n \mid y)$ are given by eqn (12).

In Fig. 1 we show the results when the detectability of one half of the species is 90% less than the detectability of the other half. Note that panels B0 and C0 in Fig. 1 are identical, but sampling differences determine different shapes for the assembled sample SAD, panels B1 and C1. The only difference between these two plots is the clumping parameter of the negative binomial. The sample SAD for case B has a logseries-like shape, while for the case C the sample SAD is bimodal. In contrast, panels A0 and B0 are very different, but because they also differ in their sampling properties (the clumping parameter), they actually show very similar shapes in the sample, panels A1 and B1. See Box 1, the caption of Fig. 1, and Appendix S2 for further details.

Scenario B

In this scenario we have assumed a heterogeneous community where species belong to distinct guilds. We
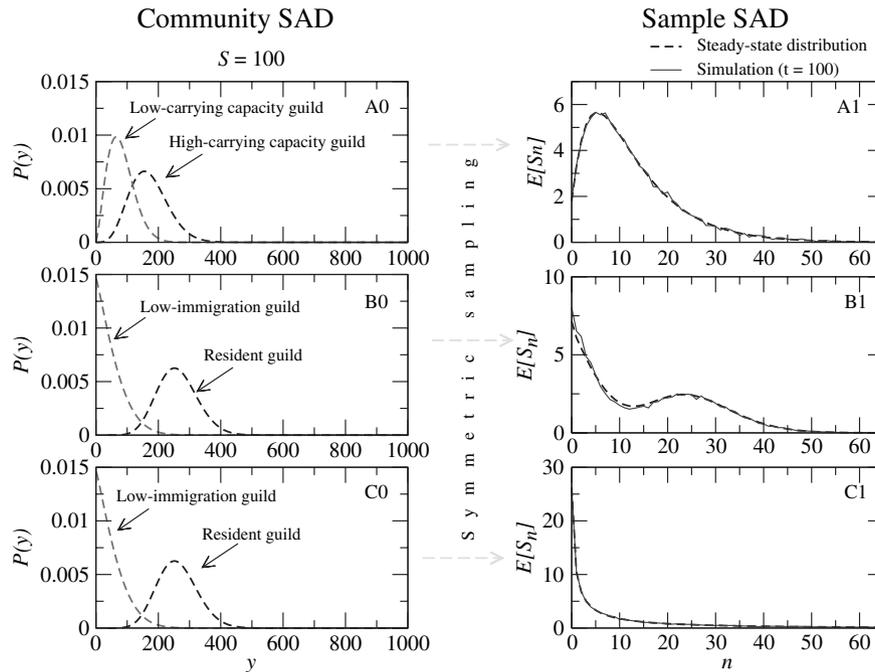
Figure 2 *Scenario B.* Here we show how symmetric sampling – i.e. same sampling parameters across species – modifies a non-symmetric regional distribution in different ways. By using the ensemble formula (eqns (9) and (10)), and assuming also a type of sampling, the steady-state stationary sample SAD (thick broken lines) can be predicted analytically. The curves resulting from simulation when the community is sampled at $t = 100$ time units are show in thin lines (panels A1, B1 and C1). Negative binomial sampling has been assumed with $a = 0.1$, $p_i = 1$ and clumping parameter of 5, 100 and 0.5 in the sampling of A, B and C communities, respectively. The two guilds present in the regional community (panel A0) are not recovered in the sample SAD (panel A1) where a common lognormal-like curve for the sample SAD is observed instead. When we compare B and C cases, we observe that the two immigration guilds in the regional community (panels B0 and C0) are perfectly distinguishable in the sample SAD of panel B1, but not in the sample SAD of panel C1. When species aggregation is important (low clumping parameter, panel C1), it is more difficult to separate the two guilds present in the regional community. Parameter values are (B, C) $N = 10^4$, $m_0^{(1)} = 10^{-4}$, $m_0^{(2)} = 10^{-3}$, $b_0 = 0.99$ and $d_0 = 1$, and (A) $b_0 = 0.99$, $d_0 = 1$, $m_0 = 5 \times 10^{-4}$, $N_1 = 10^4$ and $N_2 = 5.0 \times 10^3$.
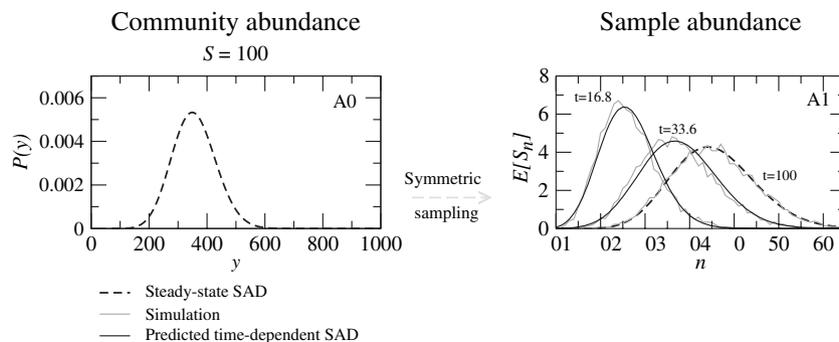

Figure 3 *Time evolution.* Here we show the time evolution of sample abundance distribution until a steady-state stationary sample SAD arises (thick lines). Both the time-dependent SAD (thin lines at three different sampling times, see panel A1) and the stationary distribution can be predicted analytically. Binomial symmetric sampling has been used in this example, with $a = 0.1$, $p_i = 1$ for all species in the system. Parameter values are $N = 10^4$, $m_0 = 10^{-3}$, $b_0 = 1$ and $d_0 = 0.99$.

consider a resident community made up of S_R species that are constantly reinvading the community and relatively occasional S_O species that migrate into the community at a

much lower rate. Hence, immigration rate characterizes two subguilds within the community. This framework has been shown to be relevant both in theoretical and empirical

studies (Magurran & Henderson 2003; Schwilk & Ackerly 2005). The presence of a mixed community made up of resident and occasional species has been found in some fish communities (Magurran & Henderson 2003). The community-level abundance distribution has two clear different components: a logseries-like curve (red) and lognormal-like curve (black) (Magurran & Henderson 2003; Magurran 2007) (see Fig. 2). As there are no differences regarding sampling, by using eqn (10), we can write the sample SAD as a mixture of two different distributions:

$$E^{(a)}[S_n] = \sum_{y=n}^{N} P^{(a)}(n\,|\,y)(S_O P(y\,|\,\Theta_O) + S_R P(y\,|\,\Theta_R))$$

(14)

where Θ_R and Θ_O stand for model parameter values to the resident and accidental community, respectively, which describe the community at the regional level, and the sampling distribution $P^{(a)}(n\,|\,y)$ is given by eqn (12), when negative binomial sampling is performed, or a corresponding binomial distribution, when random sampling is assumed. See also Appendix S2 and figure captions for further details. Panels B0 and B1 in Fig. 2 show that the difference in immigration rates is picked up as dual modes in the sample SAD.

We now consider again a community consisting of two guilds. For the first group of species the sampled locality is assumed to be more advantageous than for the second one. So, the first group of species is controlled by a higher carrying capacity and reaches higher abundance levels. An analogous equation to eqn (14) gives also the sample ensemble SAD in this case. Panels A0 and A1 of Fig. 2 show the results for a twofold difference in the carrying capacity between the guilds. In this case, the inherent difference between the guilds does not translate into dual modes, and by looking at this SAD, one may be tempted to think that the community consists of only a single guild.

In Fig. 3 the time evolution of the sample SAD is shown when the community is assembled from scratch, so all abundances are set to zero at the initial time. Then the birth–death–immigration process (see Box 1) models successional dynamics assembling the community through constant species-specific immigration. As the time evolution of each species in the system can be predicted analytically by using a Gaussian approximation (see Alonso & McKane (2002) for details), and our ensemble formulas for the community-level SADs apply along this temporal process, we can also calculate a time-dependent community-level SAD analytically. The community has been sampled at time $t = 16.7$, $t = 33.3$ and $t = 100$. At this last time stationarity is almost reached in all simulations. It is interesting to note that although the time evolution of species-level abundance distributions are approximated by Gaussian curves (Alonso

& McKane 2002), the assembled community level sample SAD is not Gaussian in general, but lognormally shaped or more skewed (see thin black lines in Fig. 3A1). This is an example of an ensemble sum of Gaussian curves (see eqn 4) producing a community-level sample SAD which is no longer Gaussian (Allen *et al.* 2001). Finally, a steady-state distribution describing the community-sample SAD emerges (thick black lines). We note the excellent match between the theoretical predictions of the community-level SAD for each scenario and the results from simulations. Simulated SADs have been calculated by averaging over 250–500 replicates. The process was simulated with the Gillespie algorithm (Gillespie 1976).

CONCLUSIONS

Species symmetry, either by implicitly assuming that all species are governed by the same abundance distribution at the regional level, or as a form of symmetric sampling underlies all previous expressions of sample SADs (Fisher *et al.* 1943; Pielou 1969; Bulmer 1974; Dewdney 1998; Lande *et al.* 2003; Volkov *et al.* 2003; Alonso & McKane 2004; Etienne & Alonso 2005; Green & Plotkin 2007, but see Walker 2007 and Pueyo *et al.* 2007). The estimation of species abundances has long tradition in ecology (Seber 2002). In particular, 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). In this paper, we have shown how a community-level sample SAD arises under species asymmetry through an ensemble sum that can be linked to a process-based stochastic community model. The way this ensemble sum is actually evaluated will depend on how exactly the sampling process is defined. We have first derived a general expression for the expected sample SAD under asymmetric and generic sampling from a species ensemble undergoing generically coupled community dynamics (see eqn 4, supplemented by eqns 6 and 3). Then, we have given a much simpler expression under the assumption of asymmetric and independent sampling (eqn 7). When independent sampling and species symmetry are assumed, the classical expression for a statistical sample SAD arises (see eqn 8). In general, when these assumptions no longer hold, it is necessary to solve first a multivariate sampling community model and use the resulting multivariate sample abundance distribution, $P^{(a)}(\vec{n})$, to generate a prediction for the sample SAD (see eqn 4). This has been achieved in neutral theory, where species interdependence at the sample level is modelled as a hypergeometric or dispersal-limited sampling distribution (Etienne & Alonso 2005, 2007; Etienne *et al.* 2007).

The analysis of our simple model community shows that the claim that SADs are not informative, because different

mechanistic models can predict essentially the same SAD (Chave *et al.* 2002; Purves & Pacala 2006), is too crude (see also Walker 2007 and Etienne 2007). Evidently, it is a trivial truism that the consistency of a single empirical pattern with a quantitative prediction from a single model is insufficient to accept that the basic processes generating this pattern in nature are precisely those assumed by the model (McGill 2006). However, a scientific theory is meant to be only an approximation to reality. This consideration becomes relevant when judging ecological theories (Ginzburg & Jensen 2004). In general, science is not based on absolute truth, but on relative evidence (Taper & Lele 2004). Therefore, if we are able to relate empirical patterns to several alternative process-based models and generate quantitative predictions, model selection techniques will allow to compare these theoretical predictions against data. By performing these analyses across as many systems and ecological situations as possible, we will accumulate relative evidence for the various alternatives models and thereby gain insight into the importance of the underlying mechanisms they represent. In relation to SADs, in the context of the recent niche-neutrality debate, such an approach has remained elusive (Volkov *et al.* 2005; Chave *et al.* 2006) (but see also Etienne 2007). We hope that the clear link we have established between community-level SADs, sampling theory and species-level abundance distributions will help to provide new and better tools to analyse this type of data.

Dispersal limitation tends to produce a single mode in the community-level sample SAD as a result of neutral dispersal limitation (Hubbell 2001). A one-humped abundance distribution, i.e. a relative depression of the most rare and the most common species, has been shown to arise in local samples as a consequence of con-specific clumping (Green & Plotkin 2007). Species aggregation can be caused by dispersal limitation or by any other biotic or abiotic process. In fact, many factors can contribute to this 'one-hump' effect. For instance, Zillio & Condit (2007) argue that SADs are largely determined by the introduction process. Recently, lognormality has been shown to arise statistically under general global community constraints by applying the MaxEnt approach (Jaynes 1968), i.e. without assuming symmetry, explicit species-specific sampling properties or any particular mechanism controlling species abundances (Pueyo *et al.* 2007). The extent to which the species belonging to a given community can differ and still produce a single mode in the community-level sample SAD will depend on the sampling process as well as on the particularities of community dynamics (compare panels A0–A1 and B0–B1 in Fig. 2, see also Pueyo *et al.* 2007). In practice, very often empirical abundance distributions are not smooth one-hump curves. Instead, they show the clear presence of two or more distinguishable humps (Gray *et al.* 2005). This is so precisely because the symmetry assumption

is violated. Through eqns (4) and (7), we can see that when we are sampling a heterogeneous community, a multimodal distribution could arise from substantial differences either in species-specific abundance distributions at the regional level (Scenario B), or in species-specific sampling probabilities (Scenario A) or both. Our ability to tease apart these two possibilities will increase if we use spatial or temporal replication. In the presence of differential processes controlling species regional abundances, we usually observe an SAD pattern that reveals the presence of different guilds, as, for instance, Magurran and Henderson's finding (Magurran & Henderson 2003; Magurran 2007) that a resident community is governed by the lognormal and an accidental community is governed by the logseries. In general, the SAD curve results from the aggregation of different species. If we deconstruct it into different guilds, we may be able to identify that different processes determine the assembly of each particular guild (Magurran & Henderson 2003; Marquet *et al.* 2004; Labra *et al.* 2005).

As a consequence, the common oversimplifying assumption that the lognormal (Preston 1948; Bulmer 1974) is the natural output of a niche-assembled community (which emphasizes species differences), and the use of this distribution as an alternative model to test neutrality (McGill 2003) is unjustified (Etienne 2007), because this pattern may arise both in a completely symmetric community (Engen & Lande 1996) and in a completely *idiosyncratic* community (Pueyo *et al.* 2007). If in some particular settings these two models have been teased apart from each other, it is because the kind of symmetry they assume is different. Neutrality assumes effective interchangeability of individuals (*hard symmetry*) or fitness equalization on a per capita basis (*ecological equivalence*) in a dispersal-limited community (Hubbell 2001), while lognormality only requires that the community is a homogeneous species ensemble (*soft symmetry*), so that species are similar enough to produce lognormality under assemblage (see eqns 4 and 6, Pueyo 2006; Pueyo *et al.* 2007). Ironically, as it requires weaker assumptions, the initial intuition that lognormality was the simplest null hypothesis to test the neutral model may arguably be right (McGill 2003). Here we disagree however with the conclusion that most ecologists have drawn from such a test. If we want to test for the importance of niche-assembly and species differences, we should move from a description where species identity is lost (the classical SAD) to a multivariate description of species abundances where species identity is taken into account explicitly (Etienne *et al.* 2007; Etienne 2007) and can be linked to different process-based assumptions of community assembly (Gotelli & McGill 2006). This is not to say that symmetrical classical SADs are useless. Classical SADs present a simple and fast empirical description of community complexity. For instance, they have been used successfully as indicators of

ecosystem health (Gray 1979). It is in this empirical context that the classical statistical theory for sampling SADs (Pielou 1969; Green & Plotkin 2007; McGill *et al.* 2007) is most valuable.

REFERENCES

Allen, A.P., Li, B. & Charnov, E.L. (2001). Population fluctuations, power laws and mixtures of lognormal distributions. *Ecol. Lett.*, 4, 1–3.

Alonso, D. & McKane, A.J. (2002). Mainland-island metapopulations: an *N*-patch stochastic approach. *Bull. Math. Biol.*, 64, 913–958.

Alonso, D. & McKane, A.J. (2004). Sampling Hubbell's neutral theory of biodiversity. *Ecol. Lett.*, 7, 901–910.

Alonso, D., Etienne, R.S. & McKane, A.J. (2006). The merits of neutral theory. *Trends Ecol. Evol.*, 21, 451–457.

Azaele, S., Pigolotti, S., Banavar, J.R. & Maritan, A. (2006). Dynamical evolution of ecosystems. *Nature*, 444, 926–928.

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.

Bulmer, M.G. (1974). On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics*, 30, 101–110.

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

Chave, J., Muller-Landau, H.C. & Levin, S.A. (2002). Comparing classical community models: theoretical consequences for patterns of diversity. *Am. Nat.*, 159, 1–23.

Chave, J., Alonso, D. & Etienne, R.S. (2006). Comparing models of species abundance. *Nature*, 441, E1–E2.

Cohen, J.E. (1968). Alternate derivations of a species abundance relation. *Am. Nat.*, 152, 165–172.

Dewdney, A.K. (1998). A general theory of the sampling process with applications to the veil line. *Theor. Popul. Biol.*, 54, 294–302.

Engen, S. & Lande, R. (1996). Population dynamics models generating the lognormal species abundance distribution. *Math. Biosci.*, 132, 169–183.

Etienne, R.S. (2005). A new sampling formula for neutral biodiversity. *Ecol. Lett.*, 8, 253–260.

Etienne, R.S. (2007). A neutral sampling formula for multiple samples an 'exact' test for neutrality. *Ecol. Lett.*, 10, 608–618.

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

Etienne, R.S. & Alonso, D. (2007). Neutral community theory: how stochasticity and dispersal-limitation can explain species coexistence. *J. Statist. Phys.*, 28, 485–510.

Etienne, R.S. & Olff, H. (2004). A novel genealogical approach to neutral biodiversity theory. *Ecol. Lett.*, 7, 170–175.

Etienne, R.S. & Olff, H. (2005). Bayesian analysis of species abundance data: assessing the relative importance of dispersal and niche-partitioning for the maintenance of biodiversity. *Ecol. Lett.*, 8, 493–504.

Etienne, R.S., Alonso, D. & McKane, A.J. (2007). The zero-sum assumption in neutral theory of biodiversity. *J. Theor. Biol.*, 248, 522–536.

Fisher, R., Corbet, A. & Williams, C. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.*, 12, 42–58.

Gillespie, D.T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.*, 22, 403–434.

Ginzburg, L.R. & Jensen, C.X.J. (2004). Rules of thumb for judging ecological theories. *Trends Ecol. Evol.*, 19, 121–126.

Gotelli, N.J. & McGill, B.J. (2006). Null *versus* neutral models: what's the difference. *Ecography*, 29, 793–800.

Gray, J.S. (1979). Pollution induced changes in populations. *Philos. Trans. R. Soc. Lond. B*, 286, 545–557.

Gray, J.S., Bjørgesaeter, A. & Ugland, K.I. (2005). The impact of rare species on natural assemblages. *J. Anim. Ecol.*, 74, 1131–1139.

Green, J. & Plotkin, J. (2007). A statistical theory for sampling species abundances. *Ecol. Lett.*, 10, 1037–1045.

Hilborn, R. & Mangel, M. (1997). *The Ecological Detective. Confronting Models with Data*. Princeton University Press, Princeton.

Hubbell, S.P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, Princeton.

Jaynes, E.T. (1968). Information theory and statistical mechanics. *Phys. Rev.*, 106, 620–630.

van Kampen, N.G. (1992). *Stochastic Processes in Physics and Chemistry*. Elsevier, Amsterdam.

Keeling, M.J. (2000). Simple stochastic models and their power-law type behaviour. *Theor. Popul. Biol.*, 58, 21–31.

Kendall, D.G. (1948). On some modes of population growth leading to R. A. Fisher's logarithmic series distribution. *Biometrika*, 35, 6–15.

Labra, F.S., Abades, S. & Marquet, P.A. (2005). Distribution and abundance. Scaling patterns in exotic and native bird species. In: *Species Invasions. Insights into Ecology, Evolution and Biogeography* (eds Sax, D.F., Stachowicz, J.J. & Gaines, S.D.). Sinauer Associates Inc., Sunderland, MA, pp. 421–446.

Lande, R., Engen, S. & Saether, B.E. (2003). *Stochastic Population Dynamics in Ecology and Conservation*. Oxford Series in Ecology and Evolution. Oxford University Press, Oxford.

Magurran, A.E. (2007). Species abundance distributions over time. *Ecol. Lett.*, 10, 347–354.

Magurran, A.E. & Henderson, P.A. (2003). Explaining the excess of rare species in natural species abundance distributions. *Nature*, 422, 714–716.

Marquet, P.A., Fernandez, M., Navarrete, S.A. & Valdivinos, C. (2004). Diversity emerging: towards a deconstruction of biodiversity patterns. In: *Frontiers of Biogeography: New Directions in the Geography of Nature* (eds Lomolino, M. & Heaney, L.R.). Cambridge University Press, Cambridge, pp. 192–209.

McGill, B.J. (2003). A test of the unified theory of biodiversity. *Nature*, 422, 881–885.

McGill, B.J. (2006). Empirical evaluation of neutral theory. *Ecology*, 87, 1411–1423.

McGill, B.J., Etienne, R.S., Gray, J.S., Alonso, D., Anderson, M.J., Benecha, H.K. *et al.* (2007). Species abundance distribution: moving beyond single prediction theories to integration within an ecological framework. *Ecol. Lett.*, 10, 995–1015.

McKane, A.J., Alonso, D. & Solé, R.V. (2000). A mean field stochastic theory for species rich assembled communities. *Phys. Rev. E*, 62, 8466–8484.

Pielou, E.C. (1969). *An Introduction to Mathematical Ecology*. Wiley, New York.

Preston, F.W. (1948). The commonness and rarity of species. *Ecology*, 29, 254–283.

Pueyo, S. (2006). Diversity: between neutrality and structure. *Oikos*, 112, 392–405.

Pueyo, S., He, F. & Zillio, T. (2007). The maximum entropy formalism and the idiosyncratic theory of biodiversity. *Ecol. Lett.*, 10, 1017–1028.

Purves, D.W. & Pacala, S.W. (2006). Ecological drift in niche-structured communities: neutral pattern does not imply neutral process. In: *Biotic Interactions in the Tropics* (eds Burslem, D., Pinard, M. & Hartley, S.). Cambridge University Press, Cambridge, pp. 107–138.

Renshaw, E. (1991). *Modelling Biological Populations in Space and Time. Vol. 11 of Cambridge Studies in Mathematical Biology*. Cambridge University Press, Cambridge.

Schwilk, D.W. & Ackerly, D.D. (2005). Limiting similarity and functional diversity along environmental gradients. *Ecol. Lett.*, 8, 272–281.

Seber, G.A.F. (2002). *The Estimation of Animal Abundance and Related Parameters*. The Blackburn press, Caldwell, New Jersey. Reprint of 2nd edition, 1982.

Taper, M.L. & Lele, S.R. (2004). *The nature of scientific evidence: philosophical and empirical considerations*. Cambridge University Press, London.

Tilman, D. (2004). Niche tradeoffs, neutrality, and community structure: a stochastic theory of resource competition, invasion, and community assembly. *Proc. Natl. Acad. Sci. USA*, 101, 10854–10861.

Volkov, I., Banavar, J.R., Hubbell, S.P. & Maritan, A. (2003). Neutral theory and relative species abundance in ecology. *Nature*, 424, 1035–1037.

Volkov, I., Banavar, J.R., He, F., Hubbell, S.P. & Maritan, A. (2005). Density dependence explains tree species abundance and diversity in tropical forests. *Nature*, 438, 658–660.

Walker, S.C. (2007). When and why do non-neutral metacommunities appear neutral? *Theor. Popul. Biol.*, 71, 318–331.

Zillio, T. & Condit, R. (2007). The impact of neutrality, niche differentiation and species input on diversity and abundance distributions. *Oikos*, 116, 931–940.

SUPPLEMENTARY MATERIAL

The following supplementary material is available for this article:

Appendix S1 Ensamble formulas for SADs.
Appendix S2 Differences in species sampling detectability.
Appendix S3 The logseries and the multinomial likelihood.
References

This material is available as part of the online article from: http://www.blackwell-synergy.com/doi/full/10.1111/j.1461-0248.2007.01127.x.

Please note: Blackwell publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.