

IDEA AND
PERSPECTIVE

A statistical theory for sampling species abundances

Jessica L. Green¹ and
Joshua B. Plotkin^{2*}

¹Center for Ecology and
Evolutionary Biology, University
of Oregon, Eugene, OR, USA

²Department of Biology,
University of Pennsylvania, 433
S. University Ave., Philadelphia,
PA 19104, USA

*Correspondence: E-mail:
jplotkin@sas.upenn.edu

Abstract

The pattern of species abundances is central to ecology. But direct measurements of species abundances at ecologically relevant scales are typically unfeasible. This limitation has motivated a long-standing interest in the relationship between the abundance distribution in a large, regional community and the distribution observed in a small sample from the community. Here, we develop a statistical sampling theory to describe how observed patterns of species abundances are influenced by the spatial distributions of populations. For a wide range of regional-scale abundance distributions we derive exact expressions for the sampled abundance distributions, as a function of sample size and the degree of conspecific spatial aggregation. We show that if populations are randomly distributed in space then the sampled and regional-scale species-abundance distribution typically have the same functional form: sampling can be expressed by a simple scaling relationship. In the case of aggregated spatial distributions, however, the shape of a sampled species-abundance distribution diverges from the regional-scale distribution. Conspecific aggregation results in sampled distributions that are skewed towards both rare and common species. We discuss our findings in light of recent results from neutral community theory, and in the context of estimating biodiversity.

Keywords

Species-abundance distribution, random sampling, negative-binomial sampling, spatial aggregation, biodiversity, community.

Ecology Letters (2007) 10: 1037–1045

INTRODUCTION

The distribution of species abundances is a fundamental topic in ecological research. Species-abundance distributions have been used to examine the influence of niche differentiation, dispersal, density dependence, speciation and extinction on the structure and dynamics of ecological communities (Tokeshi 1993; Hubbell 2001; Chave *et al.* 2002; Magurran 2004; McGill *et al.* in press). In conservation biology, knowledge of the species-abundance distribution helps one to predict the likelihood of population persistence and community stability in face of global change. Despite the theoretical and practical importance of species-abundance distributions, it is difficult to directly measure all species' abundances at ecologically relevant scales. For micro-organisms, this poses a challenge at the scale of a single environmental sample (e.g. < 1 g of soil; Prosser *et al.* 2007). In plant and animal communities as well, the task of exhaustively sampling a full community is typically impossible. Therefore, ecologists have a long-standing interest in the relationship between the underlying species-abundance

distribution of a large, regional community and the observed abundance distribution when sampling a small proportion of the community.

Efforts to develop a sampling theory of species abundances have utilized several approaches. The most widely utilized approach, which dates back to Fisher *et al.* (1943) assumes that individuals are randomly sampled from an ecological community. Fisher *et al.* (1943) sought to understand patterns of species abundance in butterfly, beetle and moth communities sampled throughout the world. They found that species abundances in random samples were well described by a logseries distribution, a distribution they mathematically derived by Poisson sampling from a gamma distribution. Their analyses laid the foundation of biodiversity sampling theory (May 1975; Pielou 1975) and remain at the forefront of literature on species-abundance distributions (Hubbell 2001; Chave 2004; Magurran 2004; McGill *et al.* in press).

Engineer and ornithologist Frank Preston (1948) popularized the lognormal species-abundance distribution in ecology by demonstrating that the lognormal had

scale-invariant properties upon random sampling. He showed by tabulation (Preston 1948) that Poisson sampling individuals from a lognormal species-abundance distribution results in a sample distribution that is approximately lognormal with identical variance. Preston (1962, p. 186) recognized that Poisson sampling individuals was analogous to 'a situation in space and time where the individuals, or pairs, are distributed at random, not clumped on one hand or over-regularized on the other'. He noted that contagion, or conspecific aggregation, would likely to result in a 'somewhat skewed' sample distribution (Preston 1962, p. 203); however, he did not rigorously explore the influence of this contagion on his samples. Biodiversity sampling theory has since predominantly assumed random sampling (e.g. Pielou 1975; Dewdney 1998; Gotelli & Colwell 2001; Chao & Bunge 2002), while less is known about the sampling properties of species-abundance distributions for spatially aggregated populations.

Efforts to understand the sampled abundance distribution of aggregated populations have focused primarily on a specific type of aggregation. The assumption of fractal, or self-similar spatial distributions has been leveraged to explore how species-abundance distributions scale with sampling area (Banavar *et al.* 1999; Harte *et al.* 1999). Recent analyses, however, suggest that such fractal models are biologically unrealistic (Green *et al.* 2003; Pueyo 2006). An alternative statistical model known as the Hypothesis of Equal Allocation Probabilities (HEAP) allocates individuals across a landscape according to a set of assembly-rules, yielding a scale-dependent species-abundance distribution that matches empirical vegetation data relatively well (Harte *et al.* 2005). Although the HEAP framework allows for a range of aggregation patterns, from random to highly clustered, this model has not been utilized to explore how the degree of aggregation influences sampled species-abundance distributions.

Neutral community theory (Hubbell 2001) provides a mechanistic approach for modelling heterogeneity by accounting for the effect of dispersal limitation. Alonso & McKane (2004) applied Poisson sampling to the metacommunity multinomial relative-abundance distribution to derive an analytical expression for the sample distribution under the assumption of zero dispersal limitation. Etienne & Alonso (2005) later invoked dispersal limitation by replacing random sampling with the dispersal-limited binomial (actually, hypergeometric) sampling. This allowed for sampling heterogeneity by modelling the probability for a dispersal-limited species to be present in a sample with a given abundance under the assumptions stipulated by neutral community theory. These recent developments, while powerful, have not explicitly addressed the sampling properties of communities assembled by non-neutral forces.

In this study, we present a general statistical framework for understanding the effect of spatial heterogeneity (or, equivalently, heterogeneity in the sampling scheme) on the species-abundance distribution observed in a sample. Our sampling framework does not assume a particular type of population aggregation (e.g. fractal theory) or community dynamics (e.g. neutral theory). We begin by analysing the simple case in which individuals are randomly sampled from the larger regional community, corresponding to random spatial distributions across the landscape. We then examine a more realistic scenario of negative-binomial sampling, which models spatial clustering of conspecific individuals. We apply our techniques to a wide range of abundance distributions, deriving exact expressions for the sampled abundance distributions as a function of sample size and the degree of conspecific clustering. We also demonstrate two important, generic properties of how spatial aggregation affects the sampled species-abundance distribution.

STATISTICAL FRAMEWORK

Species-abundance distributions are measured in the laboratory or field by counting the number of species in a community represented by n individuals. For the purpose of our analysis, we will characterize species abundances in a large region using a continuous probability density function $\phi(n)$. The expression $\phi(n)dn$ represents the fraction of species whose abundance falls between n and $n + dn$. Proper normalization requires $\int_0^\infty \phi(n)dn = 1$. In reality species' abundances are discrete. We use continuous distributions to provide consistency with a sampling theory of β -diversity (Plotkin & Muller-Landau 2002). Aside from offering analytical tractability, there is a long-standing precedent for using continuous distributions to describe species abundances (Pielou 1975).

Our interest lies in the relationship between the species-abundance distribution at the regional scale, $\phi(n)$, and the abundance distribution observed in a sample that constitutes a proportion a of the larger ambient region, denoted $\phi_a(y)$. Let $\psi_a(y|n)$ denote the probability that a species will be represented by y individuals in the sample, given that it has abundance n in the larger region. Then, the sampled species-abundance distribution may be expressed as:

$$\phi_a(y) = \int_0^\infty \psi_a(y|n)\phi(n)dn \quad (1)$$

Equation 1 provides a general expression for the scaling of the species-abundance distribution with sample size, given an arbitrary sampling scheme $\psi_a(y|n)$. Equation 1 follows directly from the law of total probability. By using different sampling schemes for $\psi_a(y|n)$, we can model different spatial distributions of conspecifics across the landscape.

Plotkin & Muller-Landau (2002) used an analogous approach to derive a sampling theory for β -diversity (the change in species composition between samples). They analysed the probability that a species will be present with any abundance $y \geq 1$ in a sample, given that it has abundance n in the larger region. Equation 1 generalizes their work to describe the full species-abundance distribution in a sample. Below we investigate eqn 1 for a suite of regional-scale species-abundance distributions $\phi(n)$, and for a suite of sampling schemes ranging from random to aggregated.

Random spatial distributions

To examine the scaling of the species-abundance distribution under the assumption of randomly distributed individuals (or, equivalently, randomly sampled individuals), we use the Poisson sampling distribution $\psi_a(y|n) = \frac{e^{-an} (an)^y}{y!}$, which denotes the probability that a species will have y individuals in a sample that constitutes a proportion a of the larger ambient region, given that it has abundance n in the larger region. Inserting the Poisson sampling distribution into eqn 1 yields:

$$\phi_a(y) = \int_0^{\infty} e^{-an} \frac{(an)^y}{y!} \phi(n) dn \quad (2)$$

Equation 2 is equivalent to a mixed Poisson distribution, which has been studied extensively by statisticians (see, for example, Karlis & Xekalaki 2005). Mixed Poisson distributions were first introduced in the ecology literature by Fisher *et al.* (1943) who derived the logseries species-abundance distribution by mixing the Poisson distribution with a gamma distribution (see also Boswell & Patil 1971).

The combination of the Poisson sampling distribution with the continuous abundance distribution, $\phi(n)$, in the integrand of eqn 2 may result in a continuous sampled abundance distribution $\phi_a(y)$ that does not normalize to one, particularly at small sample sizes. To account for this we impose the normalization constraint $\int_0^{\infty} \phi_a(y) dy = 1$, although the normalization factors are typically close to unity. An alternative approach which alleviates the need to normalize is to interpret the sample distribution $\phi_a(y)$ as discrete, but this results in an unwanted inconsistency between the form of the regional-scale and sampled species-abundance distribution. The most precise, yet least tractable approach would entail a fully discrete version of eqn 1, assuming that both the regional-scale and sample species-abundance distributions are discrete, and would model $\psi_a(y|n)$ according to the hypergeometric distribution (Dewdney 1998). Nevertheless, provided regional abundances are large enough to be modelled by a continuous probability density function, eqn 2 provides

a reasonable approximation to discrete, hypergeometric sampling.

Aggregated spatial distributions

We now examine the scaling of species abundances under the assumption that conspecific individuals are autocorrelated in space. This requires us to specify a sampling scheme $\psi_a(y|n)$ that models samples from a spatially aggregated population. Our goal was to understand how the degree of intraspecific aggregation influences the species-abundance distribution observed in the sample, without assuming a particular type of community dynamics such as that invoked in neutral theory (Etienne & Alonso 2005). For this purpose, it is convenient to model the sampling probabilities according to the widely used negative-binomial distribution:

$$\psi_a(y|n) = \frac{\Gamma(k+y)}{y! \Gamma(k)} \left(\frac{an}{an+k} \right)^y \left(\frac{k}{an+k} \right)^k, \quad (3)$$

where the gamma function is defined by $\Gamma(z) = \int_0^{\infty} e^{-t} t^{z-1} dt$. The parameter k reflects the degree of intra-specific spatial clustering, and it can assume values in the intervals $(-\infty, -an)$ and $(0, \infty)$. When k is negative, $\psi_a(y|n)$ describes a regular spatial distribution of individuals; when k is positive, $\psi_a(y|n)$ describes an aggregated spatial distribution of individuals, which is the typical situation for ecological populations (Plotkin 2000; Green *et al.* 2004; Genin *et al.* 2005; Sandin & Pacala 2005). As $k \rightarrow \pm\infty$, the negative-binomial distribution converges to the Poisson distribution, representing random sampling. Aside from providing analytical tractability, the negative-binomial distribution offers a rich historical context as it has been used for decades to model the spatial structure of biological populations (Anscombe 1949; Krebs 1998; He & Gaston 2000) and communities (He & Legendre 2002; Plotkin & Muller-Landau 2002; Green & Ostling 2003).

Substituting eqn 3 into eqn 1 yields the following relationship between the species-abundance distribution at the regional scale, $\phi(n)$, and the distribution observed in a sample of proportion a from the region:

$$\phi_a(y) = \frac{\Gamma(k+y)}{y! \Gamma(k)} \int_0^{\infty} \left(\frac{an}{an+k} \right)^y \left(\frac{k}{an+k} \right)^k \phi(n) dn \quad (4)$$

As before, a continuous distribution $\phi_a(y)$ is obtained from eqn 4 by enforcing the normalization constraint $\int_0^{\infty} \phi_a(y) dy = 1$.

Asymmetric aggregated spatial distributions

In the preceding section we examined the properties of sampled species abundances assuming that all species are symmetric, i.e. that all species have the same sampling

properties regardless of their abundance at the regional scale. But the spatial distributions of species are often not symmetric in this sense. For example, more abundant species typically exhibit less spatial clustering or, equivalently, larger values of the negative-binomial parameter k (Condit *et al.* 2000). How does such asymmetry affect the properties of expected species abundances in a sample?

We can address the asymmetric situation by modifying eqn 4 to account for a clustering parameter $k(n)$ that depends upon a species' abundance, n , at the regional scale:

$$\phi_a(y) = \int_0^\infty \left(\frac{an}{an+k(n)}\right)^y \left(\frac{k(n)}{an+k(n)}\right)^{k(n)} \frac{\Gamma(k(n)+y)}{y!\Gamma(k(n))} \phi(n) dn \tag{5}$$

RESULTS

Random spatial distributions

Table 1 catalogues closed-form solutions to eqn 2 for a range of regional species-abundance distributions. These expressions can be used to quantify the exact relationship between large-scale species-abundance distributions and that observed in a random sample of individuals. Some of the sample distributions in Table 1 have been previously derived as Mixed Poisson distributions in the statistics literature (Johnson *et al.* 2005; Karlis & Xekalaki 2005). The

most important finding, apparent from graphing the expressions in Table 1, is that when individuals are sampled randomly the shape of the sampled abundance distribution $\phi_a(y)$ exhibits a strong resemblance to that of the regional-scale distribution $\phi(n)$. This scaling behaviour was first recognized by Preston (1948) in the special case of the lognormal species-abundance distribution, and later by Dewdney (1998). Here, we mathematically characterize this behaviour by positing a simple relationship:

$$\phi_a(y) \approx \frac{1}{a} \phi\left(\frac{y}{a}\right) \tag{6}$$

We can derive this scaling rule and its associated error expansion by generalizing the approach of Etienne & Alonso (2005). Substituting $z = an$ into Equation 2, we find:

$$\begin{aligned} \phi_a(y) &= \frac{1}{a} \int_0^\infty e^{-z} \frac{z^y}{y!} \phi\left(\frac{z}{a}\right) dz \\ &= \frac{1}{ay} \int e^{-z} \frac{z^{y-1}}{\Gamma(y)} \phi\left(\frac{z}{a}\right) z dz \end{aligned} \tag{7}$$

If we let $f(z) = z\phi(z/a)$, then the expression above is equal to the expected value of $f(Z)$, where Z is a gamma-distributed random variable with shape parameter y and scale 1. Expanding $f(z)$ around the mean of Z leads to the familiar expansion:

$$E[f(z)] = f(\bar{z}) + \frac{1}{2} \text{Var}(Z) f''(\bar{z}) + \dots \tag{8}$$

Distribution name	Regional abundance distribution $[\phi(n)]$	Regional abundance range	Sampled abundance distribution $[\phi_a(y)]$
Continuous logseries	$\frac{x^n}{n\Gamma(0, -\ln x)}$	$n \geq 1$ Otherwise	$\phi_a(y) = -\left(\frac{a}{a-\ln x}\right)^y / yL(x)$
Exponential*	$\lambda e^{-\lambda n}$	$0 \leq n \leq \infty$	$\phi_a(y) = \frac{a^y \lambda}{(a+\lambda)^{y+1}}$
Gamma*	$\frac{\lambda^\beta n^{\beta-1} e^{-\lambda n}}{\Gamma(\beta)}$	$0 \leq n \leq \infty$	$\phi_a(y) = \frac{a^y \lambda^\beta \Gamma(y+\beta)}{y! \Gamma(\beta) (a+\lambda)^{y+\beta}}$
Lognormal*	$\frac{\exp\left[-\frac{(\ln n - \mu)^2}{2\sigma^2}\right]}{n\sigma\sqrt{2\pi}}$	$0 \leq n \leq \infty$	-
Truncated hyperbolic	$\frac{1}{n \ln(M/n)}$ 0	$m \leq n \leq M$ Otherwise	$\phi_a(y) = \frac{\Gamma(y, am) - \Gamma(y, aM)}{y! \ln(M/m)}$

Table 1 Regional and sampled abundance distributions under Poisson sampling

Species-abundance models were chosen to parallel those described in Plotkin & Muller-Landau (2002). Γ denotes the gamma function $\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt$; $\Gamma(\alpha, x)$ denotes the incomplete gamma function $\Gamma(\alpha, x) = \int_x^\infty e^{-t} t^{\alpha-1} dt$; $L(x)$ denotes the logarithmic integral $L(x) = \int_0^x dt / \ln(t)$. Dashes indicate that a closed-form solution is unavailable. Sample distributions marked with an asterisk (*) have been examined previously (see Pielou 1975; Johnson *et al.* 2005 for a comprehensive review). The closed-form solution for the continuous logseries is an approximation which is valid provided $\int_0^1 \psi_a(y|n)\phi(n)dn \ll 1$ (see Appendix A for details on analytical and numerical integration).

As the mean of variance of Z both equal to y , this yields

$$\begin{aligned} \phi_a(y) &= \frac{1}{ay} \left(f(y) + \frac{1}{2} y f''(y) + \dots \right) \\ &= \frac{1}{ay} \left(\phi\left(\frac{y}{a}\right) \cdot y + \frac{1}{2} y f''(y) + \dots \right) \\ &= \frac{1}{ay} \left(\phi\left(\frac{y}{a}\right) \cdot y + \frac{y}{2a^2} \left[2a\phi'\left(\frac{y}{a}\right) + y\phi''\left(\frac{y}{a}\right) \right] + \dots \right) \\ &= \frac{1}{a} \phi\left(\frac{y}{a}\right) + O\left(\frac{1}{a^2}\right) \end{aligned} \tag{9}$$

This scaling relationship, or ‘rule of thumb’, is intuitive because it suggests that under random sampling the abundance distribution scales according to a change of variables by proportion a . Figure 1 illustrates a comparison between the Poisson-sampled species-abundance distribution (eqn 2) and the rule of thumb (eqn 6). The sampled abundance distribution and the rule of thumb are indistinguishable except in the limit of small sample sizes. In general, given the error term derived in eqn 9, we expect the rule of thumb scaling relationship to approximate eqn 2 accurately except when the proportion of area sampled is very small. In this regime, combining Poisson sampling with a continuous abundance distribution yields inaccurate results, and a fully discrete approach is required.

Our proposed rule of thumb (eqn 6) is also supported by results from applied statistics stating that the probability

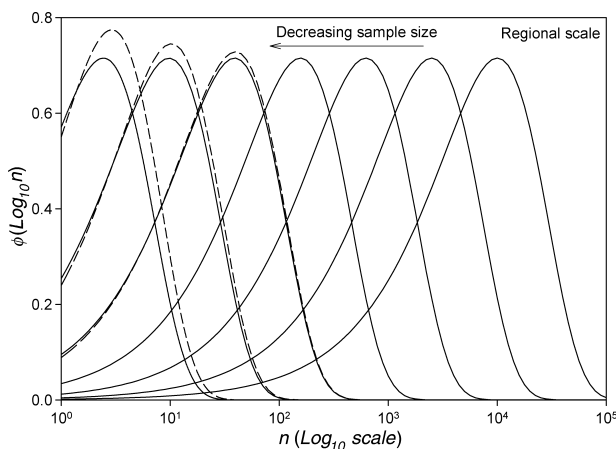


Figure 1 Scaling of relative distributions in the case of random sampling. This example assumes a gamma distribution at the regional scale with parameters $\lambda = 0.000075$, $\beta = 0.75$. Sampling proportion a decreases from right to left as $a = 1$ (regional scale), $a = 2^{-2}$, $a = 2^{-4}$, $a = 2^{-6}$, $a = 2^{-8}$, $a = 2^{-10}$, $a = 2^{-12}$. The scaling relationship is represented by the solid lines, and the solution to eqn 2 assuming Poisson sampling is represented by the dashed lines. The scaling relationship $\phi_a(y) \approx \phi(y/a)/a$ and eqn 2 are indistinguishable except in the limit of small sample size and abundance, where the Poisson approximation breaks down. All distributions are plotted as probability density $f(y)$ vs. y , where $y = \log_{10} n$.

function of a mixed Poisson distribution often resembles that of its mixing distribution (Johnson *et al.* 2005; Karlis & Xekalaki 2005). In particular, the asymptotic tails of mixed Poisson distributions agree with the tails of their mixing distributions provided $\phi(n)$ satisfies $\phi(n) = C(n)n^\alpha e^{-\beta n}$, as $n \rightarrow \infty$, where $C(n)$ is locally bounded and varies slowly at infinity (Willmot 1990).

Aggregated spatial distributions

Figure 2 illustrates the relationship between the sampled distribution $\phi_a(y)$ and the regional-scale distribution $\phi(n)$, when conspecific individuals are aggregated in space. Unlike the case for randomly sampled individuals, the shapes of the sampled species-abundance distributions significantly differ from the regional-scale distribution. In particular, intraspecific aggregation skews the abundance distribution towards both more rare and more abundant species in the sample, compared to random sampling.

Table 2 catalogues closed-form analytical solutions to eqn 4 for the regional species-abundance distributions listed in Table 1. Table 2 provides a means of quantifying the relationship between the regional-scale distribution and the sample distribution for any degree of sampling heterogeneity as specified by the clustering parameter k . These sample distributions may be readily utilized in future theoretical work aimed at linking small-scale sample distributions to large-scale biodiversity patterns.

Despite the complexity of the derived sample distributions, they each exhibit a similar behaviour under clustered sampling. The generality of this behaviour is illustrated in Fig. 3: clustered sampling results in a greater proportion of

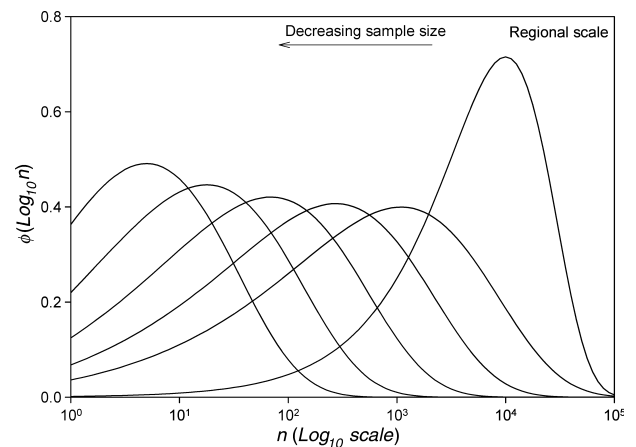


Figure 2 Scaling of relative-abundance distributions in the case of clustered sampling. This example assumes a gamma distribution of abundances at the region scale with parameters $\lambda = 0.000075$, $\beta = 0.75$, and a clustering parameter $k = 0.5$. Sampling proportion a decreases from right to left as $a = 1$ (regional scale), $a = 2^{-2}$, $a = 2^{-4}$, $a = 2^{-6}$, $a = 2^{-8}$, $a = 2^{-10}$.

Table 2 Regional and sampled abundance distributions under negative-binomial sampling

Name	Sampled abundance distribution
Continuous logseries	$\phi_a(y) = \frac{\Gamma(1-k)\Gamma(k+y) {}_1F_1(k+y, 1+k, -k \ln(x)/a) (-k \ln(x))^k}{a^k \Gamma(1+y)L(x)} - \frac{{}_1F_1(y, 1-k, k \ln(x)/a)}{yL(x)}$
Exponential	$\phi_a(y) = \frac{\lambda k {}_1F_1(1+y, 2-k, k\lambda/a)}{a(k-1)} + \frac{(k\lambda)^k \pi \text{CSC}(k\pi) \Gamma(k+y) {}_1F_1(k+y, k, k\lambda/a)}{a^k \Gamma(k)^2 \Gamma(y+1)}$
Gamma	$\phi_a(y) = \frac{\pi \binom{k+y-1}{k-1} \text{csc}(\pi(\beta-k))}{\Gamma(\beta)\Gamma(k+y)} \left[\left(\frac{k\lambda}{a}\right)^k \Gamma(k+y) {}_1F_1(k+y, 1-\beta+k, \frac{k\lambda}{a}) - \left(\frac{k\lambda}{a}\right)^\beta \Gamma(\beta+y) {}_1F_1(\beta+y, 1+\beta-k, \frac{k\lambda}{a}) \right]$
Lognormal	–
Truncated hyperbolic	$\phi_a(y) = \frac{1}{y \ln(M/m)} \left(\frac{k^k \Gamma(k+x)}{\Gamma(x+1)(k-1)} \left(\frac{m {}_2F_1(k-1, k+x, k, -k/am)}{(am)^k} - \frac{M {}_2F_1(k-1, k+x, k, -k/aM)}{(aM)^k} \right) \right)$

Regional-scale distributions are defined in Table 1. $F_1(a, b, z)$ denotes the Kummer confluent hypergeometric function, and $F_2(a, b, z)$ denotes the Gauss hypergeometric function. Dashes indicate that a closed-form solution is unavailable. The closed-form solution for the continuous logseries is an approximation which is valid provided $\int_0^1 \psi_a(y|n)\phi(n)dn \ll 1$ (see Appendix A for details on analytical and numerical integration).

both rare and abundant species, compared to random sampling or compared to the regional abundance distribution. This phenomenon is more exaggerated as the degree of conspecific aggregation increases. Although a surplus of both rare and common species may seem surprising, there is an intuitive explanation for this fundamental result: when conspecific individuals are aggregated, there is a greater chance of either ‘hitting’ or ‘missing’ a cluster of individuals in a sample, resulting in a sampled abundance distribution that is skewed towards both rare and common species.

Asymmetric aggregated spatial distributions

Figure 4 illustrates how asymmetry in clustering properties (eqn 5) can affect the species-abundance distribution in a sample. In the figure we compare the regional species-abundance distribution to the abundance distribution under Poisson sampling, negative-binomial sampling with fixed k , and negative-binomial sampling with varying $k(n)$. The sampled abundance distribution is clearly influenced by asymmetries in species’ clustering properties, but the resulting sampled abundance distribution is very similar to the situation in which all species have the mean clustering parameter $\bar{k} = \int_0^\infty k(n)\phi(n)dn$. In other words, the symmetric and asymmetric models yield qualitatively similar results, for k equal to the average aggregation tendency of species.

DISCUSSION

The distribution of species abundances is a long-standing topic of research in community ecology, informing basic ecological theory and applied conservation management. Most species-abundance data are based on limited samples, and it is important to understand how sampled abundance patterns relate to the underlying species-abundance distribution in the regional community. In this study, we have addressed this need by developing a sampling theory of

species abundances for random and autocorrelated populations, based on widely used statistical sampling schemes. Under Poisson sampling, the sampled and regional-scale distributions have nearly the same shape, and the relationship between the two can be expressed as a simple scaling function with an error term. In the presence of conspecific aggregation, however, the sampled abundance distribution is markedly different from the underlying, regional-scale distribution: spatial clustering increases the frequency of both rare and common species in the sample distribution.

The scaling rule we have derived for Poisson sampling (eqn 6) is in close agreement with the empirical observations made by Preston (1948, 1962) in the special case of the lognormal distribution. Preston found that when plotting species-abundance distributions as the fraction of species within the abundance interval $[\log(y), \log(y)+d\log(y)]$, the sample distribution had nearly identical shape as the regional-scale lognormal species-abundance distribution, but shifted to the left by a factor of $\log(y/a)$ – in accordance with our ‘rule of thumb’ (eqn 6). Our results also resolve Dewdney’s (1998) conjecture that there exists a large family of species-abundance distributions whose random sampling properties are approximated by a simple scaling behaviour (Figs 1 and 3). Although the Poisson-sampled abundance distribution can change shape at very small abundances (cf. McGill 2003), this phenomenon only occurs in parameter regimes and at abundances for which the use of a continuous species-abundance distribution is not well-justified to begin with.

The generality of our rule of thumb is evidenced by recent findings in the neutral theory of community assembly. Alonso & McKane (2004) have shown that under the neutral model, when there is no dispersal limitation in the regional species pool, the local community species-abundance distribution may be approximated by Poisson sampling from the metacommunity’s multinomial relative-abundance distribution, $f_M(y)$. For a large enough local community, the sampled

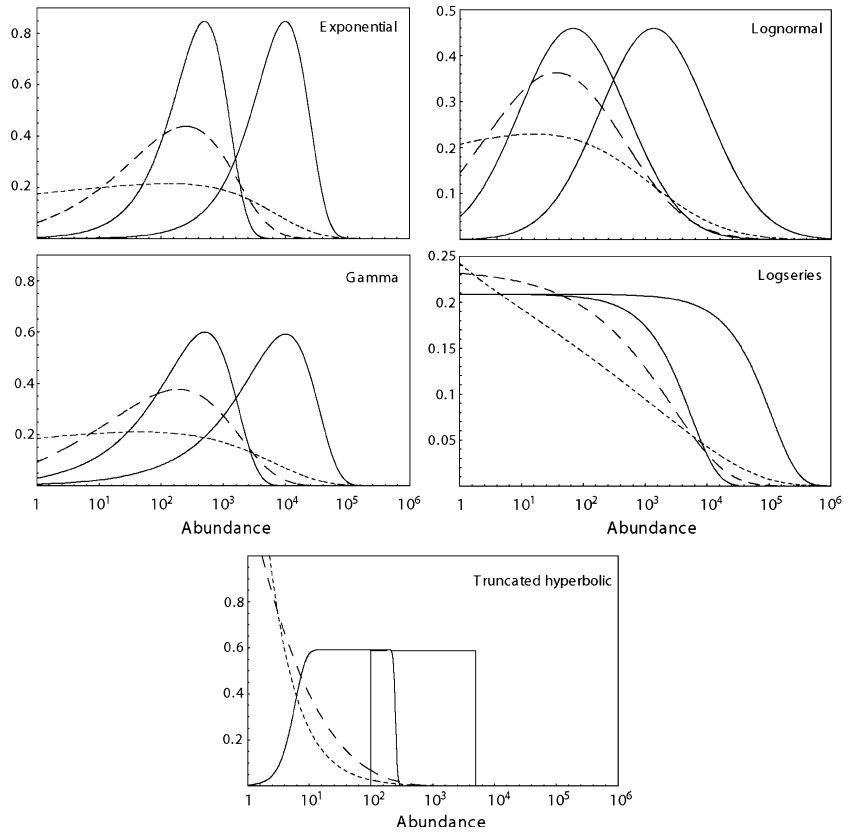


Figure 3 The scaling of five different regional abundance distributions (solid) under Poisson sampling (solid), aggregated sampling (dashed, negative-binomial parameter $k = 0.5$) and highly aggregated sampling (dotted, negative-binomial parameter $k = 0.05$). In all cases, 5% of the regional community is sampled. The shape of the abundance distribution under random sampling is similar to that of the regional abundance distribution. The abundance distribution under negative-binomial sampling shows an excess of both rare and common species, especially for highly aggregated distributions (dotted). Distribution parameters are as follows: continuous logseries $x = 0.99999093735$; exponential $\lambda = 0.0001$; gamma $\beta = 0.55, \lambda = 0.000055$; lognormal $\mu = 7.2103, \sigma = 2$; truncated hyperbolic $m = 100, M = 5000$.

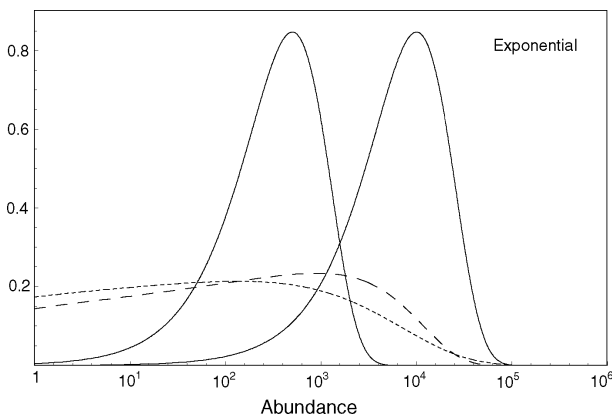


Figure 4 The scaling of the regional abundance distribution (solid) under Poisson sampling (solid), symmetric aggregated sampling (dashed, negative-binomial parameter $k = 0.05$), and asymmetric aggregated sampling (dotted). The sampled area represents 5% of the regional area, and the regional abundances are exponentially distributed with $\lambda = 10^{-4}$. In the asymmetric case, the clustering parameter depends upon abundance according to the formula $k(n) = n/200\,000$. The sampled abundance distribution in the asymmetric case is very similar to the symmetric sampled distribution, for k equal to the average clustering parameter across species: $\bar{k} = \int_0^\infty k(n)\phi(n)dn = 0.05$. Alternative choices for $k(n)$ and $\phi(n)$ yield similar results.

distribution $S_f(n)$, defined as the expected number of species with abundance n represented in a sample of J individuals, is closely approximated by the scaling relationship $S_f(n) = f_M(n/J)/J$ (*Ecol. Lett.*, 8; eqn 13). This scaling relationship parallels our rule of thumb (eqn 6), but is expressed in the context of sampling from the distribution of relative (rather than absolute) abundances. Thus, eqn 6 is applicable beyond the family of species-abundance distributions examined in this paper. Indeed, according to our derivation, we can expect eqn 6 to provide an accurate approximation for the randomly sampled abundance distribution, to first order in $1/a$, provided the derivatives of the regional-scale abundance distribution are all bounded.

Our analysis of spatially autocorrelated populations also parallels recent results for the neutral theory in the case of a dispersal-limited local community (Hubbell 2001; Vallade & Houchmandzadeh 2003; Volkov *et al.* 2003; Chave 2004). Neutral theory predicts that the shape of the local species-abundance distribution sampled from the metacommunity multinomial distribution will be a function of the immigration rate. Increased dispersal limitation leads to an increasingly left-skewed sample distribution. Dispersal limitation is one obvious mechanistic force that induces aggregation, although habitat preferences will have similar effects. Our results suggest that any mechanism leading to autocorrelated

samples of conspecific individuals will result in a sampled abundance distribution that is skewed towards both rare and abundant species, compared to the metacommunity. We have shown that this phenomenon is not limited to the multinomial distribution discussed in neutral theory, but rather it applies across a wide range of regional-scale abundance distributions.

Based on empirical data and simulations, McGill (2003) recognized the tendency of sample autocorrelation to fatten the rare tail of the lognormal species-abundance distribution. Here, we have used explicit sampling formulae and demonstrated that autocorrelation produces sampled distributions that are skewed towards both rare *and* common species. In addition, we have shown that this phenomenon holds over a wide range of abundance distributions. In the case of the lognormal distribution, the sampled distributions are skewed towards both rarity and commonness, but the rare skew is more pronounced than the common skew (Fig. 3, red curve). These results suggest that the apparent rare-skewed lognormal abundance distribution observed in censused tropical forest plots (Hubbell 1997, 2001) is consistent with the common assumption of a lognormal regional abundance distribution (Preston 1948, 1962; May 1975) combined with observed, local aggregation of conspecific trees (He & Legendre 1996; Condit *et al.* 2000; Plotkin & Muller-Landau 2002).

Our analysis of the sampled species-abundance distribution sheds light on a topic of practical interest in applied ecology – namely, inferring the properties of a community from one or more samples. One common approach, which has recently gained popularity in microbial ecology (Curtis *et al.* 2002; Finlay 2002; Bohannan & Hughes 2003; Venter *et al.* 2004), uses parametric distributions to estimate the number of species in a community. Sample data are fit to models of relative abundance (or assumed on theoretical grounds), and the sample frequency distribution is projected to estimate the number of unobserved species in the community (Pielou 1975; Bunge & Fitzpatrick 1993). This approach is grounded in the assumption that the sample frequency distribution is a truncated version of the community-level distribution when individuals that are randomly sampled from the community (eqn 6). In many ecological contexts, however, this assumption may be seriously violated. Microbial communities, for example, are commonly investigated by identifying individuals from soil or sediment cores across a landscape. Even if the cores are randomly distributed in space, spatial aggregation in microbial populations would result in a non-random sample of individuals from the community. In this case, the sampled distribution would be characterized by a relative-abundance distribution that is markedly different in shape than the community-level distribution, complicating the problem of estimating community-level species richness.

Estimating species richness in a single environmental sample from clone libraries using the random sampling assumption may also be inaccurate, because PCR-based methods may yield a biased, non-random subsample of individuals (Wintzingerode *et al.* 1997; Lueders & Friedrich 2003). Thus, parametric-based distribution approaches that ignore conspecific aggregation for estimating species richness are likely to yield unreliable predictions.

Our results on sampling may inform future research aimed at leveraging abundance and aggregation patterns measured at local scales to predict biodiversity patterns at larger, regional scales.

ACKNOWLEDGEMENTS

We thank Steve Hubbell for useful discussions. We thank J. Chave and two anonymous referees for substantial input. J. L. G. acknowledges support from the National Science Foundation (MCB 0500124). J. B. P. acknowledges support from the Burroughs Wellcome Fund.

REFERENCES

- Alonso, D. & McKane, A.J. (2004). Sampling Hubbell's neutral theory of biodiversity. *Ecol. Lett.*, 7, 901–910.
- Anscombe, F.J. (1949). The statistical analysis of insect counts based on the negative binomial distribution. *Biometrika*, 5, 165–173.
- Banavar, J.R., Green, J.L., Harte, J. & Maritan, A. (1999). Finite size scaling in ecology. *Phys. Rev. Lett.*, 83, 4212–4214.
- Bohannan, B.J.M. & Hughes, J. (2003). New approaches to analyzing microbial biodiversity data. *Curr. Opin. Microbiol.*, 6, 282–287.
- Boswell, M.T. & Patil, G.P. (1971). Chance mechanisms generating the logarithmic series distribution used in the analysis of number of species and individuals. In: *Statistical Ecology: Spatial Patterns and Statistical Distributions*. (eds Patil, G.P., Pielou, E.C., Waters, W.E.) Pennsylvania State University Press, University Park, pp. 99–130.
- Bunge, J. & Fitzpatrick, M. (1993). Estimating the number of species: a review. *J. Am. Stat. Assoc.*, 88, 364–373.
- Chao, A. & Bunge, J. (2002). Estimating the number of species in a stochastic abundance model. *Biometrics*, 58, 531–539.
- Chave, J. (2004). Neutral theory and community ecology. *Ecol. Lett.*, 7, 241–253.
- Chave, J., Muller-Landau, H.C. & Levin, S. (2002). Comparing classical community models: theoretical consequences for patterns of diversity. *Am. Nat.*, 159, 1–23.
- Condit, R., Ashton Peter, S., Baker, P., Bunyavejchewin, S., Gunatilleke, S., Gunatilleke, N. *et al.* (2000). Spatial patterns in the distribution of tropical tree species. *Science*, 288, 1414–1418.
- Curtis, T.P., Sloan, W.T. & Scannell, J.W. (2002). Estimating prokaryotic diversity and its limits. *Proc. Natl Acad. Sci. U.S.A.*, 99, 10494–10499.
- Dewdney, A.K. (1998). A general theory of the sampling process with application to the “veil line”. *Theor. Popul. Biol.*, 54, 294–302.
- Etienne, R.S. & Alonso, D. (2005). Dispersal-limited sampling theory for species and alleles. *Ecol. Lett.*, 8, 1147–1156.
- Finlay, B.J. (2002). Global dispersal of free-living microbial eukaryote species. *Science*, 296, 1061–1063.

- Fisher, R.A., Corbet, S.A. & Williams, C.B. (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.
- Genin, A., Jaffe, J.S., Reef, R., Richter, C. & Franks, P.J.S. (2005). Swimming against the flow: a mechanism of zooplankton aggregation. *Science*, 308, 860–862.
- Gotelli, N.J. & Colwell, R.K. (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.*, 4, 379–391.
- Gradshteyn, I.S. & Ryzhik, I.M. (2000). *Table of Integrals, Series and Products*, 6th edn. Academic Press, London.
- Green, J. & Ostling, A. (2003). Endemics-area relationships: the influence of species dominance and spatial aggregation. *Ecology*, 84, 3090–3097.
- Green, J.L., Harte, J. & Ostling, A. (2003). Species richness, scaling and abundance patterns: tests of two fractal models in a serpentine grassland. *Ecology Lett.*, 6, 919–928.
- Green, J.L., Holmes, A.J., Westoby, M., Oliver, I., Briscoe, D., Dangerfield, M. *et al.* (2004). Spatial scaling of microbial eukaryote diversity. *Nature*, 432, 747–750.
- Harte, J., Kinzig, A. & Green, J. (1999). Self-similarity in the distribution and abundance of species. *Science*, 284, 334–336.
- Harte, J., Conlisk, E., Ostling, A., Green, J.L. & Smith, A.B. (2005). A theory of spatial-abundance and species-abundance distributions. *Ecol. Monogr.*, 75, 179–197.
- He, F. & Gaston, K.J. (2000). Estimating species abundance from occurrence. *Am. Nat.*, 156, 553–559.
- He, F. & Legendre, P. (1996). On species-area relations. *Am. Nat.*, 148, 719–737.
- He, F. & Legendre, P. (2002). Species diversity patterns derived from species-area models. *Ecology*, 83, 1185–1198.
- Hubbell, S.P. (1997). A unified theory of biogeography and relative species abundance and its application to tropical rain forests and coral reefs. *Coral Reefs*, 16, S9–S21.
- Hubbell, S.P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, Princeton.
- Johnson, J.L., Adrienne, W.K. & Kotz, S. (2005). *Univariate and Discrete Distributions*, 3rd edn. John Wiley and Sons, New York.
- Karlis, D. & Xekalaki, E. (2005). Mixed Poisson distributions. *Int. Stat. Rev.*, 73, 35–58.
- Krebs, C.J. (1998). *Ecological Methodology*, 2nd edn. Harper Collins Publishers, New York.
- Lueders, T. & Friedrich, M.W. (2003). Evaluation of PCR amplification bias by terminal restriction fragment polymorphism analysis of small-subunit rRNA and mcrA genes using defined template mixtures of methanogenic pure cultures and soil DNA extracts. *Appl. Environ. Microbiol.*, 69, 320–326.
- Magurran, A.E. (2004). *Measuring Biological Diversity*. Blackwell Science Ltd, Malden, MA.
- May, R.M. (1975). Patterns of species abundance and diversity. In: *Ecology and Evolution of Communities* (eds Cody, M.L. & Diamond, J.M.), Harvard University Press, Cambridge, pp. 81–120.
- McGill, B.J. (2003). Does mother nature really prefer rare species or are log-left-skewed SADs a sampling artifact? *Ecol. Lett.*, 6, 766–773.
- McGill, B.J., Etienne, R.S., Gray, J.S., Alonso, D., Anderson, M.J., Benecha, H.K. *et al.* (in press) Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecol. Lett.*, 10, doi: 10.1111/j.1461-0248.2007.01092.x.
- Pielou, E.C. (1975). *Ecological Diversity*. John Wiley and Sons Inc., New York.
- Plotkin, J.B. (2000) Species-area curves, spatial aggregation, and habitat specialization in tropical forests. *J. Theor. Biol.*, 207, 81–99.
- Plotkin, J.B. & Muller-Landau, H.C. (2002). Sampling the species composition of a landscape. *Ecology*, 83, 3344–3356.
- Preston, F.W. (1948). The commonness, and rarity, of species. *Ecology*, 29, 254–283.
- Preston, F.W. (1962). The canonical distribution of commonness and rarity. *Ecology*, 43, 182–215.
- Prosser, J.I., Bohannan, B.J.M., Curtis, T.P., Ellis, R.J., Firestone, M.K., Freckleton, R.P. *et al.* (2007). The role of ecological theory in microbial ecology. *Nat. Rev. Microbiol.*, 5, 384–392.
- Pueyo, S. (2006). Self-similarity in species-area relationship and in species abundance distribution. *Oikos*, 112, 156–162.
- Sandin, S.A., Pacala, S.W. (2005). Fish aggregation results in inversely density-dependent predation on continuous coral reefs. *Ecology*, 86, 2716–2725.
- Tokeshi, M. (1993). Species abundance patterns and community structure. *Adv. Ecol. Res.*, 24, 112–186.
- Vallade, M. & Houchmandzadeh, B. (2003). Analytical solution of a neutral model of biodiversity. *Phys. Rev. E*, 68, 061902.
- Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A. *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 1093857, 304, 66–74.
- Volkov, I., Banavar, J.R., Hubbell, S.P. & Maritan, A. (2003). Neutral theory and relative species abundance in ecology. *Nature*, 424, 1035–1037.
- Willmot, G.E. (1990). Asymptotic tail behaviour of Poisson mixtures with applications. *Adv. Appl. Probab.*, 22, 147–159.
- Wintzingerode, F.V., Göbel, U.B. & Stackebrandt, E. (1997). Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS. Microbiol. Rev.*, 21, 213–229.

SUPPLEMENTARY MATERIAL

The following supplementary material is available for this article:

Appendix S1 Deriving the Sampled Abundance Distributions.

This material is available as part of the online article from: <http://www.blackwell-synergy.com/doi/full/10.1111/j.1461-0248.2007.01101.x>.

Please note: Blackwell Publishing is 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.

Editor, Jerome Chave

Manuscript received 22 June 2007

First decision made 5 July 2007

Manuscript accepted 30 July 2007