Three models of ecological community assembly

John Alroy[1]

[1]Macquarie University

May 07, 2024

Abstract

Species abundance distributions, meaning counts of individuals apportioned among species, are fundamental patterns in ecology. Numerous distribution models have been proposed, and most suffer from poor fit to data, complex formulation, excessive parameterisation, or unrealistic modelling of processes. I discuss three that meet all the basic criteria, are easily distinguished, and stem from simple and distinct population dynamics. The log series can be produced by assuming taxonomically and temporally fixed turnover rates. A model derived from scaled odds ratios assumes highly variable dynamics, and one derived from exponential variates assumes taxonomically variable but temporally fixed rates. Mathematical derivations are elementary. Maximum likelihood fits to published empirical data suggest that the two new distributions are more common in nature. Saturated models are rarely better. Ecological communities may be assembled by processes that are easily discerned, instead of being as mysterious as many have thought.

Three models of ecological community assembly

John Alroy[1]

[1]School of Natural Sciences, Macquarie University. Email: john.alroy@mq.edu.au

Running title: Three models of ecological community assembly

Keywords: ecological abundance, exponential-to-e distribution, log series, population dynamics, scaled odds distribution, species richness

Type of article: Letter

Number of words: 143 in the abstract, 4874 in the main text.

References: 60

Figures: 4

Tables: none

Text boxes: none

Statement of authorship: JA designed the study, created the methods, collected the data, analysed the data, and wrote the manuscript.

Data accessibility statement: The data have been archived on Dryad (https://doi.org/10.5061/dryad.brv15dvdc). The source code for the analyses is on Zenodo (https://zenodo.org/records/10156259). The R package is also available on GitHub (https://github.com/johnalroy/richness/tree/main/R).

Correspondence: John Alroy, School of Natural Sciences, Macquarie University, NSW 2109, Australia. Phone +61 (4) 66 819 806. E-mail john.alroy@mq.edu.au

Abstract

Species abundance distributions, meaning counts of individuals apportioned among species, are fundamental patterns in ecology. Numerous distribution models have been proposed, and most suffer from poor fit to data, complex formulation, excessive parameterisation, or unrealistic modelling of processes. I discuss three that meet all the basic criteria, are easily distinguished, and stem from simple and distinct population dynamics. The log series can be produced by assuming taxonomically and temporally fixed turnover rates. A model derived from scaled odds ratios assumes highly variable dynamics, and one derived from exponential variates assumes taxonomically variable but temporally fixed rates. Mathematical derivations are elementary. Maximum likelihood fits to published empirical data suggest that the two new distributions are more common in nature. Saturated models are rarely better. Ecological communities may be assembled by processes that are easily discerned, instead of being as mysterious as many have thought.

INTRODUCTION

Species inventories are the universal currency of community ecology: counting individuals that belong to each of the species in a place is a routine and fundamental practice. These counts are the basis of key community assembly theories (Fisher et al., 1943; MacArthur, 1957; Bulmer, 1974; Caswell, 1976; Hubbell, 2001). It is thought that count distributions usually tail off with an array of rare species (McGill et al., 2007). When this is true, estimating species richness is difficult and dangerous because inventories are likely to be quite incomplete (Colwell & Coddington, 1994). Biodiversity is of deep concern throughout science and society, making it imperative to solve this problem. I focus here on the richness estimation strategy of fitting inventories to mathematical distributions that imply fixed numbers of missing species. In the course of doing so, I show not only that this idea is feasible but that fundamental processes of community assembly can be distinguished using basic inventory data.

Population dynamical models going back to Kendall (1948) have been used before to predict shapes of species abundance distributions (SADs), but the SADs have generally involved multi-parameter equations (Volkov et al., 2005; Jabot & Chave, 2011). The three models considered here all require a single parameter. I put aside other single-parameter models such as the broken stick (MacArthur, 1957), the geometric series as applied to rank-abundance distributions (Motomura, 1932), the logistic-J (Dewdney, 2000), and the Zipf (see Newman, 2005) because they have received little support in comprehensive assessments of distributions (Alroy, 2015; Baldridge et al., 2016) and have not been considered in many studies that have treated two or three distributions at a time (Hughes, 1986; Dewdney, 2000; Connolly et al., 2005; Ulrich et al., 2010; Antão et al., 2021). I do not consider the gambin model (Ugland et al., 2007; Matthews et al., 2014) because it appears only to describe distributions of counts binned into octaves on a log scale (Preston, 1948). Gray et al. (2006) are among several to have pointed out problems with this approach. So like others including Antão et al. (2021), this study is concerned with models such as the log series (Fisher et al., 1943) that predict counts of singletons, doubletons, and so on – i.e., SADs in a restricted sense.

I also do not consider two-parameter distributions such as the classic Poisson log normal (PLN: Bulmer, 1974) and the truncated negative binomial (Connolly et al., 2009; Connolly & Thibaut, 2012). These models have much traction: for example, the PLN has been argued to fit extensive datasets of trees, birds, fishes, and benthic organisms (Antão et al., 2021), not to mention all GBIF occurrence records in the world combined (Callaghan et al., 2023). Meanwhile, the negative binomial has been fit to a vast data set for Amazonian trees (ter Steege et al., 2020). There are two major reasons not to consider these models for the moment. First, they overfit the data, reducing chances of predicting related patterns. Second, one-parameter distributions are often so good that they cannot be rejected by a saturated model. The latter is a highly resolved function that closely mirrors the raw counts instead of following a proper parameterised distribution. This paper shows how to construct a saturated model and how to assess its fit to the data. The upshot is that the three substantive models under consideration perform so well there is little left to explain.

MATERIAL AND METHODS

Data

The empirical data consist of 3095 published species inventories from around the world representing terrestrial organisms: trees, arthropods, and tetrapods. They were downloaded from the Ecological Register website (http://ecoregister.org) on 13 November 2022. The 82,870 identifications stem from 2019 publications. A total of 27,045 formal Linnean species are represented by 73,133 of these records, and the rest are specifically indeterminate. Every record is associated with a count of individuals encountered. Major animal groups include ants, butterflies, bats, birds, dung beetles, frogs, large mammals, lizards, mosquitoes, odonates, orthopterans, and small mammals. Earlier versions of the database have featured in several publications (e.g., Alroy 2015, 2017, 2018).

Distribution fitting

Methods of fitting data to abundance models are contentious, with many protocols having been advocated (Matthews & Whittaker, 2014). As mentioned, all of the models considered here seek to explain SADs sensu stricto, which are vectors that record the number of species each sharing a given count of individuals (Fisher et al., 1943).

It is very important to stress that SADs are not equivalent to rank abundance distributions (RADs). These are useful for depicting counts (e.g., Motomura, 1932, MacArthur, 1957) and are commonly used even by some contemporary workers to fit distribution models (e.g., Nekola et al., 2008; Ulrich et al., 2010, 2015). There are at least four major reasons not to fit data to RADs: (1) key theoretical models directly predict SAD shapes, not RADs; (2) most models that do directly predict RADs, such as the geometric series (Motomura, 1932), are no longer considered to be viable descriptors of real-world ecological communities (Alroy, 2015; Baldridge et al., 2016); (3) maximum likelihood methods have been developed to fit models to SADs (e.g., Connolly et al., 2005; Connolly & Thibaut, 2012) and are generally advocated over the many alternatives (Gray et al. 2006; Whittaker & Matthews, 2014; Antão et al., 2021), but RADs are generally fit using frequentist methods; and (4) it is difficult to model the error in species ranks because any variation in the count of a species could also change its rank, so the x- and y-axes in an RAD are not statistically independent.

A third approach is also worth mentioning: to bin the counts into classes on a log scale, equivalent to a histogram where the boxes show the number of species in classes 1, 2, 3 – 4, 5 – 8, 9 – 16, etc. (Preston, 1948). This strategy is still used (e.g., Matthews et al., 2014), but it has rightly been rejected because it loses too much information and can introduce artefacts (Gray et al., 2006; Nekola et al., 2008).

Here I use a fast and reliable maximum likelihood computation for fitting. It is the most obvious approach: define the likelihood by multiplying the probabilities of the individual counts based on the SAD. Specifically, if there are s_1 singletons and s_2 doubletons out of S species and if the hypothesised PMF is $p_1, p_2, p_3 \ldots$, then the joint likelihood is $p_1^{s_1} \times p_2^{s_2} \ldots$.

This calculation works as well in practice as any other I have investigated, surpassing rivals in a suite of tests that I do not have space to detail. It has a very interesting property: exactly the same solution is always found by fitting a given set of counts to a multinomial model. The reason is that the combinatorial terms distinguishing multinomial distributions from simple products of probabilities are constant across all possible parameter values, so they wash out of any optimisation.

All models considered here use just one free parameter. However, many of the remaining models in the literature assume two parameters (such as the PLN). For comparison across models in general, the corrected Akaike information criterion (Hurvich and Tsai, 1993) is recommended. It has been used previously in this context (Antão et al., 2021).

In practice, analysing a large data set requires examining a limited set of classes. Here, the computational limit is treated as $2^{14} = 16{,}384$. Imposing this cutoff makes hardly any difference because just 45 out of 82,870 species counts in the database (0.05%) exceed it.

The log series

Also called the logarithmic series, the log series is the oldest description of SADs sensu stricto in the eco-

logical literature (Fisher et al., 1943). The log series has been flagged as fitting tropical tree inventory data sets particularly well (Ulrich et al., 2015) – but see below. The neutral model of biodiversity (Hubbell, 2001) was developed in part to justify this belief, and it predicts the log series as a result of steady immigration, continuous point speciation, and exactly balanced birth and death: each individual lost is replaced immediately. The assumption of constant turnover was used in earlier population dynamical models that predicted the log series (e.g., Kendall, 1948; Caswell, 1976).

A particular algorithm that produces the series (Figs. 1A, B) is as follows: (1) the total population size for a community of 10,000 species is fixed at some value N (here, 100 x 10,000); (2) at each time step, all N individuals are removed; (3) and the Nreplacements are drawn from a probability distribution that is a linear function of the preceding population sizes of individual species, so the relative probability of drawing an individual of a species with $n$individuals (i.e., its weight) is just n . The prevent permanent loss of species, at the beginning of each step immigration is simulated. Whenever z species have zero counts, z individuals are randomly assigned to increase the values for those species. So if$z$ is two, a missing species could end up with 0, 1, or 2 individuals.

The basic form of the log series, meaning the probability mass function (PMF), is a proportion p between 0 and 1 raised to an integer series k and then divided by the same series:

$$\mathrm{P}(X = x\ ) = -1/\ln(1 - p\ )p^k\ /k\ (1)$$

where the left-hand term is a scaling constant that causes the sum of probabilities to be 1 and p is a constant fitted by a standard recursive equation (Chatfield, 1969; May, 1975). This form is a special case of the negative binomial, but otherwise unique. It implies semi-log plotted RADs that each start with a rapid drop in the counts of common species and then asymptote quickly on a straight falling line (Figs. 1A, B).

The scaled odds distribution

The second model assumes high variance across species and through time in counts (Figs. 1C, D). In this case, not all individuals are replaced at each time step, so there is a non-trivial death process. Both births and deaths depend on a random uniform variate τ that is reset at each time step. The number of deaths per time step is a random binomial draw from the n individuals of each species based on its own τ probability of death. The birth process is independent of population size, and random draws for each species are geometrically distributed with a success probability of exactly τ. Thus, birth and death rates are positively correlated, and an equilibrium is maintained because birth rates are steady and not per-capita.

The key biological assumption is that replacement probabilities at a given time are fixed for each given species but highly variable across species. Because the probabilities change completely at each time step, species have no innate properties and the model is neutral.

The PMF is derived as follows. First, relative abundances on a continuous scale are defined as tracking an odds distribution. The odds are multiplied by a constant μ and specifically taken to predict abundance values x on a continuous scale:

$$x = \mu\ (1 - U\ )/U\ (2)$$

U is solved for by rearranging the odds:

$$x = \mu/U - \mu\ (3)$$

$$U = \mu/(x + \mu)\ (4)$$

Note that μ/U – μ still an odds ratio, just a scaled one.

Equation 4 defines 1 minus the cumulative distribution function (CDF) of the scaled odds distribution. Any CDF starts at zero and reaches 1 asymptotically, and eqn. 4 does in fact start with a value of 1 where$x$ is 0 and decline to zero where x reaches infinity. A monotonically trending function that has limits of 1 and 0 also has an integral of 1, as in this case. Any PMF can be derived from a monotonic CDF by taking first

differences and rounding down to integer values. This is the exact procedure used to derive the PMF of the geometric series from the CDF of the exponential distribution (see Cohen, 1968 for an early example of using the exponential in the context of constructing SADs). Thus, the scaled odds PMF is:

$$P(X = x\ ) = \mu/(x_i + \mu) - \mu/(x_{i\ +\ 1} + \mu) \quad (5)$$

which can be rearranged algebraically as:

$$P(X = x\ ) = \mu/[(\xi_\iota + \mu)\ (x_{i\ +\ 1} + \mu)] \quad (6)$$

The PMF's value for the zero class is just:

$$P(X = 0) = \mu/[\mu\ (1 + \mu)] = 1/(1 + \mu) \quad (7)$$

The sum of the values for the non-zero classes is therefore:

$$1 - 1/(1 + \mu) = \mu/(1 + \mu) \quad (8)$$

The PMF discounting the zero class is then easily defined by dividing eqn. 6 by eqn. 8:

$$P(X = x\ ,\ X > 0) = (1 + \mu)/[(x_i + \mu)\ (x_{i+\ 1} + \mu)] \quad (9)$$

A valid estimator of the total species richness of the community$R$ is then trivially derived by dividing observed richness$S$ by eqn. 8:

$$R = S\ (1 + \mu)/\mu \quad (10)$$

The exponential-to-*e* distribution

This quite different but equally simple distribution can be motivated by a very similar population dynamics model (Figs. 1E, F). Here, the turnover of each species is fixed and intrinsic, not changing through time. It is based on a random number ε with an exponential distribution. In the illustrated trial, the distribution's rate λ is 0.5. Variable ε is converted to a proportion π using the ratio $1/(\varepsilon + 1)$, which scales π between 0 and 1. The binomial death probability is $\pi^2$and the geometric success probability is π.

A similar set of manipulations can be used to derive the PMF and richness estimator of this exponential-to-e($\exp^e$) distribution. The x values on a continuous scale are assumed to come not from an odds ratio but from an exponential random variate with a rate of λ that is raised to the power of e , keeping in mind that the negative logarithm of a uniform variate is exponentially distributed:

$$x = [-\log(U\ )/\lambda]^{\epsilon} \quad (11)$$

Τηε ΠΜΦ φολλοωσ χυιςκλψ:

$$\Upsilon = \varepsilon\xi\pi(-\lambda\ x^{\ 1/e}\ ) \quad (12)$$

$$P(X = x\ ) = \exp(-\lambda x_i^{\ 1/e}\ ) - \exp(-\lambda x_{i\ +\ 1}^{\ 1/e}\ ) \quad (13)$$

which does not reduce algebraically into anything more simple. However, the size of the zero class, size of the non-zero class, and richness estimator are all trivial:

$$P(X = 0) = \exp(-\lambda\ 0^{1/e}\ ) - \exp(-\lambda\ 1^{1/e}\ ) = 1 - \exp(-\lambda) \quad (14)$$

$$P(X = x\ ,\ X > 0) = \exp(-\lambda) \quad (15)$$

$$R = S\ /\exp(-\lambda) \quad (16)$$

The $\exp^e$ distribution is deeply related to the Weibull distribution, which has been of interest to ecologists seeking to model SADs (Ulrich et al., 2018). Specifically, the e power term in eqn. 11 is a direct function of the k variable in the Weibull probability density function. As a result, eqn. 13 is a special case of the discrete Weibull distribution (Nakagawa & Osaki, 1975). The fact that eqn. 13 exactly follows from a basic population dynamics model (Figs. 1E, F) justifies the choice of k used here.

As noted, the odds and $\exp^e$ population models differ only in how year-to-year variation works. This fact raises the question of whether the two might be unified by adding a single parameter to express the rate at which relative abundances are scrambled through time: rapidly in an odds world, and not at all in an $\exp^e$ world. The answer is no. In a potential unified model, a scaling variable $\varkappa$ could be introduced. Variable μ in the odds eqn. 3 is then multiplied by $\varkappa$, and U is raised to the power $1/\varkappa$. As $\varkappa$ increases, simple calculations show that eqn. 3 rapidly converges on a scaled exponential function:

$$\varkappa\,\mu/U^{\,1/\varkappa} - \varkappa\,\mu \sim -\ln(U)/\mu \quad (17)$$

The right-hand side is just a root of the function defining $\exp^e$. Therefore, bridging the gap between the distributions would require defining a model with at least three parameters: the equivalents of $\varkappa$ and μ plus a power term.

Saturated models

Intuitively, certain SAD shapes might be so unusual that shoehorning them into any of the models under consideration would be uninformative. A good test is to show whether a distribution generated directly from the raw data fits better. By this I mean what might be expected to be found in a new data set highly resembling the one under consideration.

Specifically, I assume that the individual counts are random outcomes of a geometric sampling process. The maximum likelihood estimate of the governing parameter of the geometric series p is just $1/(n+ 1)$ where n is an individual count. I propose finding $p$for each n , computing the PMF of the geometric series based on each p , and averaging the PMFs that result. A smooth PMF lacking gaps results. It is necessary to assume there are no zero counts for the calculation to replicate observed richness. Thus, each geometric series PMF needs to be rescaled before summation, specifically by multipling the entire range of probabilities by n /(n + 1).

The alternative would be to assume a Poisson process (Bulmer, 1974). But Poisson variation around high integers is small, making it difficult to fill the gaps at the high end of a predicted SAD shape. Furthermore, a direct comparison of the two possible saturated models, which I lack space to illustrate, shows that the one based on the geometric series yields better log likelihoods (LLs) for the current empirical data set.

Assessing LLs here is non-trivial because any saturated model overfits the data by preserving too much fine-scale detail. A good solution is to randomly vary the species list across many trials. During each one, the counts are sampled with replacement (i.e., bootstrapped), the averaged PMF is recomputed, and the LL is found. The LLs are then averaged across trials. Consistent with the logic of all bootstrapping methods, this procedure renders the PMF more of a prediction of future data then a description of existing data, which is appropriate.

Empirical test of relative support

A ternary diagram is used to illustrate which distributions are most consistent with the tree and animal species inventories. The saturated model is used as a baseline. For each inventory, the LL of each distribution is subtracted from the LL for the saturated model and the difference is exponentiated to produce a relative weight. For example, if the saturated model LL is 10 and the log series LL is 8, the weight is $\exp(2) = 7.39$. The weights are used to generate x -$y$coordinates in a ternary diagram with the standard equation: $x = 0.5\ (2\ w_o + w_l)/(w_o + w_l + w_e)$ where w_o, w_l, and w_e are the weights for the odds, log series, and $\exp^e$ distributions; and$y = 3^{0.5}/2\ w_l/(w_o + w_l + w_e)$.

Empirical test of sampling bias

Unlike conventional lower-bound richness extrapolators such as Chao 1 (Chao, 1984), the species richness estimators proposed here (eqns. 10 and 16) are free of substantial sample size bias. This fact is demonstrated with a subsample-and-extrapolate protocol. First, richness estimates are obtained for each of the species inventories. Second, each one is randomly degraded by removing half of the individuals, and the estimates

are recomputed. Consistency is assessed by computing the median offset on a log scale between the two sets of estimates.

RESULTS

Relative to a saturated model, the terrestrial species inventory data show split support for the three distributions (Fig. 2A). However, there is an interesting trend as more and more non-singleton species are picked up by the sampling process. When samples are poor (blue points in Fig. 2A), support is ambiguous but slightly shifted toward the axis connecting the log series and odds models. The reason is that small samples are often either flat at random (resembling the log series) or heavily dominated by a single species at random (resembling the odds distribution). The more heavily sampled and rich an inventory, the closer it moves toward the $\exp^e$ corner of the graph (red points in Fig. 2A). There are exceptions: some of these points are on edges of the triangle leading to the other distributions.

However, it seems fair to argue that although different communities follow different rules, a plurality of diverse ones most closely adhere to the $\exp^e$ population dynamics system. Thus, it may be common for turnover rates to vary among species but not through time. This fact is striking because the log series has been heralded as a key feature of tropical tree communities (Ulrich et al., 2015), but it makes little sense unless species instead behave almost identically.

There is clear variation in support across ecological groups (Figs. 2B – D). Tree inventories occupy much of the $\exp^e$corner and the edge of the central arch on that side (Fig. 2B). Bird inventories also take up the outer edge of this arch, but mammal inventories cluster on the inner edge. Insect samples are widely scattered, but ant, butterfly, dung beetle, and mosquito communities are all routinely found in the $\exp^e$ area. Finally, frog and lizard samples are concentrated in the main arch. They tend to approach the odds corner when sampling is strong.

In principle, saturated models might be hard to beat because any extreme outlier count or suggestion of bimodality could challenge the one-parameter models. This is rarely true in practice: the saturated model is best in only 859 cases (27.8%). It is notable that $\exp^e$ cannot be rejected most of the time: it fits 1749 of 3095 SADs (56.6%) better than the saturated model, as opposed to 1231 cases and coincidentally also 1749 cases (39.8 and 56.6%) for the odds and log series. Regardless, support for saturated models should never be mistaken as potential support for multi-parameter distributions like the PLN because outliers, bimodal patterns, and the like aren't naturally predicted by those distributions either.

The SADs best fitting the four models are all diverse (Fig. 3). Of particular note, the Barro Colorado Island tree counts (Fig. 2C: Condit et al., 1996) have been the subject of a high-profile dispute: Hubbell (2001) argued that they follow the zero-sum multinomial distribution generated by his neutral theory, McGill (2003b) that they are closer to the log normal, and Gray et al. (2006) that nothing could be concluded because of difficulties with the octave plotting method of Preston (1948) used by the other authors. So what may well be the most famous SAD in ecology is very closely matched by one of two models presented here ($\exp^e$).

The scaled odds and $\exp^e$ species richness estimators are highly consistent (Figs. 4B, C). The median estimates for half-samples are respectively 0.093 and 0.020 log units below the values for full samples. Fisher's α is equally consistent (Fig. 4A: median 0.055), but Chao 1 is greatly biased (Fig. 4D: 0.204). It is meaningful that when assumptions are met, the estimates are almost exactly right: the median offset is 0.012 for samples best matching $\exp^e$ and 0.003 in the other direction for those best matching the odds distribution.

DISCUSSION

It is extremely important to emphasise that the two new distribution models disagree with the log series because they explicitly assume a hard limit to species richness. No conflict among theories could be more basic. Indeed, each novel distribution comes with an elementary species richness estimator that follows directly from its single governing parameter (eqns. 10 and 16).

Therefore, we now have within reach a real solution to the problem of estimating diversity that plagues the ecological literature (Colwell & Coddington, 1994; Bunge et al., 2014). One could argue nonetheless that a distribution-free approach to estimating richness is preferable. This is beside the point because fitting distributions is a straightforward process, so identifying the right one is often easy.

Regardless, the most common so-called non-parametric richness estimators are all designed to produce lower bounds instead of accurate, unbiased values. Examples include the jackknife estimators (Burnham & Overton, 1978), the bootstrap estimator (Smith & van Belle, 1984), Chao 1 (Chao, 1984), the abundance coverage estimator (Chao & Lee, 1992), and interpolation and extrapolation based on sample coverage (iNEXT: Hsieh et al., 2016).

The reason for their systematic error is that they implicitly assume uniform abundance distributions, as I have briefly noted before (Alroy, 2017). Otherwise, for example, the Poisson sampling theory that can be used to easily derive Chao 1 (Alroy, 2017) would make no sense. The fact that real-world distributions are virtually never uniform renders all of the many lower-bound estimators non-starters.

Computing joint evenness-richness measures such as Hill numbers (Hill, 1973; Chao et al., 2014) is a widespread alternative approach to the diversity problem, and it is motivated by the difficulty of extrapolating richness from incomplete survey data (Chao et al., 2014). After all, the key Hill number called the Simpson index (Simpson, 1949) has no sampling bias if properly formulated (Hurlbert, 1971), and the classic Shannon index (Shannon, 1948) is also robust. However, the odds and $\exp^e$ models produce no sample size bias – even when their assumptions aren't met (Figs. 4B, C). Thus, the undersampling problem that drives the Hill number and extrapolation literature is now moot.

As for evenness, it does not exist as a separate concept in cases where any of the three one-parameter models hold: there is no role for distribution "shape" in these scenarios, so evenness is always fixed. Specifically, the x parameter of the log series, the μ parameter of the odds distribution, and the λ parameter of $\exp^e$ are all instead sampling intensity measures. The evenness concept is fundamental to ecology (Pielou, 1966; Hill, 1973; Tuomisto, 2012) and is routinely taught to undergraduate students as a core part of the discipline's theory. Ironically, the existence of "evenness" as an attribute of real-world communities can now be called into question. In cases where evenness is a non-concept, the Hill number approach is even less relevant: differences amongst Hill numbers reduce to differences in how richness and evenness are balanced (Hill, 1973). The only Hill number entirely representing richness is richness itself.

The simulation models undergirding the three models make completely conflicting assumptions about birth and death processes: they are either entirely invariant (log series), variable both taxonomically and temporally (odds), or variable taxonomically but not temporally ($\exp^e$). The latter two distributions both follow from assuming that birth and death rates jointly track random uniform variates. The only difference is that $\exp^e$ holds the rates constant through time, so they result from the invariant ecological traits of individual species.

Testing the models with empirical data could be straightforward. For example, a study of tree community ecology could examine counts of seedlings or saplings sorted by species. The counts might follow a geometric distribution because this is assumed of birth rates in both of the new models. Just as interestingly, repeated censuses across several years could be used to show whether varation is temporally consistent: do species with high counts in a given year still have high counts in following years? If so, then the $\exp^e$ model should hold for counts of adult individuals – meaning that traits matter. If not, then the odds model might be sufficient.

The published empirical data suggest that the $\exp^e$ dynamic is actually the most common when sampling is intense, richness is high, and a coherent, non-saturated pattern is present (Fig. 2A). The database is so highly eclectic that this unlikely to reflect a bias in the primary literature or sampling of that literature – trees (Fig. 2B), birds (Fig. 2C), and certain insect groups (Fig. 2D) have little in common, but all of them often drift into the $\exp^e$ range of the diagram. There is some additional interesting variation across ecological categories, but either $\exp^e$ or odds is often the strongest model. This fact is not very surprising

considering that their assumptions are biologically basic: there is every reason to think that reproductive rates do vary amongst species in the same communities, regardless of whether that variation is consistent through time.

By contrast, the log series is usually not the best model for well-sampled terrestrial data (Fig. 2A; Antão et al., 2021). It is widely believed that most empirical SADs do have a "hollow curve" pattern, with long tails of rare species that loosely fit the log series (McGill et al. 2007). But such RADs very often fit $\exp^e$.

The fact that the log series isn't dominant has very broad implications. The description of the model by Fisher et al. (1943) is a classic in the field, and the deeply related neutral model of biodiversity (Hubbell, 2001) has been profoundly influential (Rosindell et al., 2011). Ironically, a large body of literature has been devoted to altering Hubbell's original model by adding yet more biological assumptions: variation through dispersal limitation and habitat preference (Zillio & Condit, 2007), conspecific frequency dependence (Jabot & Chave, 2011), and much more (e.g., Al Hammal et al., 2015). The neutral model was complex from the start, explicitly assuming a major role for varying immigration rates and steady speciation rates in addition to assuming complete ecological equivalence among species. Its variants are even more complex, and elaborate models are unlikely to have fidelity to the world (Finocchiaro, 2021).

The odds and $\exp^e$ models succeed in capturing large differences among communities that one might sensibly expect to exist, but they sidestep all the complexity. It is literally not possible for simpler models to exist. As a result, ecologists may now have the tools to discern biologically important patterns by contrasting highly testable and distinct theories of community assembly – and to finally estimate species richness in a robust and justifiable manner.

FUNDING INFORMATION

The author is the recipient of a Discovery Project Award (project number DP210101324) funded by the Australian Government.

ACKNOWLEDGEMENTS

I thank many colleagues, including Mark Westoby and particularly Michael Foote, for comments on the current research.

ORCID

John Alroy https://orcid.org/0000-0002-9882-2111

REFERENCES

Al Hammal, O., Alonso, D., Etienne, R.S., & Cornell, S.J. 2015. When can species abundance data reveal non-neutrality? *PLoS Computation Biology,* **11** (3), e1004134.

Alroy, J. 2015. The shape of terrestrial abundance distributions.*Science Advances,* **1** , e1500082.

Alroy, J. 2017. Effects of habitat disturbance on tropical forest biodiversity. *Proceedings of the National Academy of Sciences, USA,* **114** (23), 6056-6061.

Alroy, J. 2018. Limits to species richness in terrestrial communities.*Ecology Letters,* **21** (12), 1781-1789.

Antão, L.H., Magurran, A.E., & Dornelas, M. 2021. The shape of abundance distributions across spatial scales. *Frontiers in Ecology and Evolution,* **9** , 626730.

Baldridge, E., Harris, D.J., Xiao, X., & White, E.P. 2016. An extensive comparison of species-abundance models. *PeerJ,* **4,** e2823.

Belshaw, R., & Bolton, B. 1994.. A survey of the leaf litter ant fauna in Ghana, West Africa (Hymenoptera: Formicidae). *Journal of Hymenoptera Research,* **3** , 5-16.

Bulmer, M.G. 1974. On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics,* **30** (1), 101-110.

Bunge , J., Willis, A., & Walsh, F. 2014. Estimating the number of species in microbial diversity studies. *Annual Review of Statistics and Its Application,* **1** , 427-445.

Burnham, K.P., & Overton, W.S. 1978. Estimation of the size of a closed population when capture probabilities vary among animals.*Biometrika,* **65** (3), 625-633.

Caswell, H. 1976. Community structure: a neutarl model analysis.*Ecological Monographs,* **46** (3), 327-354.

Chao, A. 1984. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics,* **11** (4), 265-270.

Chao, A., Gotelli, N.J., Hsieh, T.C., Sander, E.L., Ma, K.H., Colwell, R.K., & Ellison, A.M. 2014. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs,* **84** (1), 45-67.

Chao, A., & Lee, S.-M. 1992. Estimating the number of classes via sample coverage. *Journal of the American Statistical Association,***87** (417), 210-217.

Chatfield, C. 1969. On estimating the parameters of the logarithmic series and negative binomial distributions. *Biometrika,***56** (2), 411-414.

Cohen, J.E. 1968. Alternate derivations of a species-abundance relation.*The American Naturalist,* **102** (924), 165-172.

Colwell, R.K., & Coddington, J.A. 1994. Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society B,* **345** (1311), 101-118.

Condit, R., Hubbell, S.P., & Foster, R.B. 1996. Changes in tree species abundance in a Neotropical forest: impact of climate change.*Journal of Tropical Ecology,* **12** , 231-256.

Connolly, S.R., Dornelas, M., Bellwood, D.R., & Hughes, T.P. 2009. Testing species abundance models, a new bootstrap approach applied to Indo-Pacific coral reefs. *Ecology,* **90** (11), 3138-3149.

Connolly, S.R., Hughes, T.P., Bellwood, D.R., & Karlson, R.H. 2005. Community structure of corals and reef fish at multiple scales.*Science,* **309** , 1363–1365.

Connolly, S.R., & Thibaut, L.M. 2012. A comparative analysis of alternative approaches to fitting species-abundance models.*Journal of Plant Ecology,* **5** (1), 32-45.

Dewdney, A.K. 2000. A dynamical model of communities and a new species-abundance distribution. *The Biological Bulletin,***198** (1), 152-165.

Dowdy, W.W. 1947. An ecological study of the Arthropoda of an oak-hickory forest with reference to stratification. *Ecology,***28** (4), 418-439.

Finocchiaro, P. 2021. High-fidelity metaphysics: ideological parsimsony in theory choice. *Pacific Philosophical Quarterly,***102** (4), 613-632.

Fisher, R.A., Corbet, A.S., & Williams, C.B. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology,***12** (1), 42-58.

Gray, J.S., Bjorgesaeter, A., & Ugland, K.I. 2006. On plotting species abundance distributions. *Journal of Animal Ecology,***75** (3), 752-756.

Hill, M.O. 1973. Diversity and evenness: a unifying notation and its consequences. *Ecology,* **54** (2), 627-639.

Hsieh, T.C., Ma, K.H., & Chao, A. 2016. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers).*Methods in Ecology and Evolution,* **7** (12), 1451-1456.

Hubbell, S.P. 2001. *The unified neutral theory of biodiversity and biogeography.* Princeton University Press, Princeton, NJ.

Hughes, R.G. 1986. Theories and models of species abundance. *The American Naturalist,* **128** (6), 879-899.

Hurlbert, S.H. 1971. The nonconcept of species diversity: a critique and alternative parameters. *Ecology,* **52** (4), 577-586.

Hurvich, C.M., & Tsai, C.L. 1993. The corrected Akaike information criterion for vector autoregressive model selection. *Journal of Time Series Analysis,* **14** (3), 271-279.

Jabot, F., & Chave, J. 2011. Analyzing tropical forest tree species abundance distributions using a nonneutral model and through approximate Bayesian inference. *The American Naturalist,* **178** (2), E37-E47.

Kendall, D.G. 1948. On some modes of population growth leading to R. A. Fisher's logarithmic series distribution. *Biometrika,***35** (1-2), 6-15.

MacArthur, R.H. 1957. On the relative abundance of bird species.*Proceedings of the National Academy of Sciences, USA,***43** (3), 293-295.

Matthews, T.J., Borregaard, M.K., Ugland, K.I., Borges, P.A.V., Rigal, F., Cardoso, P., & Whittaker, R.J. 2014. The gambin model provides a superior fit to species abundance distributions with a single free parameter: implementation and interpretation. *Ecography,***37** (10), 1002-1011.

Matthews, T.J., & Whittaker, R.J. 2014. Fitting and comparing competing models of the species abundance distribution: assessment and prospect.*Frontiers of Biogeography,* **6** (2), 67-82.

May, R.M. 1975. Patterns of species abundance and diversity. In Cody, M.L., & Diamond, J.M. (Eds.), *Ecology and evolution of communities.* Belknap, Cambridge, MA.

McGill, B.J. 2003a. Does Mother Nature really prefer rare species or are log-left-skewed SADs a sampling artefact? *Ecology Letters,***6** (8), 766-773.

McGill, B.J. 2003b. A test of the unified theory of biodiversity.*Nature,* **422** , 881-885.

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

Mishra, B.P., Tripathi, O.P., & Laloo, R.C. 2005. Community characteristics of a climax subtropical humid forest of Meghalaya and population structure of ten important tree species. *Tropical Ecology,* **46** (2), 241-251.

Motomura, I. 1932. A statistical treatment of associations.*Japanese Journal of Zoology,* **44** , 379-383.

Nakagawa, T., & Osaki, S. 1975. The discrete Weibull distribution.*IEEE Transactions on Reliability,* **24** (5), 300-301.

Nekola, J.C., Sizling, A.L., Boyer, A.G., & Storch, D. 2008. Artifactions in the log-transformation of species abundance distributions. *Folia Geobotanica,* **43** , 259-268.

Newman, M.E.J. 2005. Power laws, Pareto distributions and Zipf's law.*Contemporary Physics,* **46** (5), 323-351.

Pielou, E.C. 1966. Species-diversity and pattern-diversity in the study of ecological succession. *Journal of Theoretical Biology,***10** (2), 370-383.

Preston, F.W. 1948. The commonness, and rarity, of species.*Ecology,* **29** (3), 254-283.

Rosindell, J., Hubbell, S.P., & Etienne, R.S. 2011. The unified neutral theory of biodiversity and biogeography at age ten. *Trends in Ecology and Evolution,* **26** (7), 340-348.

Shannon, C.E. 1948. A mathematical theory of communication. *Bell System Technical Journal,* **27** (3), 379-423.

Simpson, E.H. 1949. Measurement of diversity. *Nature,***163** , 688.

Smith, E.P., & van Belle, G. 1984. Nonparameteric estimation of species richness. *Biometrics,* **40** (1), 119-129.

ter Steege, H., Prado, P.I., de Lima, R.A.F., Pos, E., de Souza Coelho, L., de Andrade Lima Filho, D., et al. 2020. Biased-corrected richness estimates for the Amazonian tree flora. *Scientific Reports,***10** , 10130.

Tuomisto, J. 2012. An updated consumer's guide to evenness and related indices. *Oikos,* **121** (8), 1203-128.

Ugland, K.I., Lambshead, P.J.D., McGill, B., Gray, J.S., O'Dea, N., Ladle, R.J., & Whittaker, R.J. 2007. Modelling dimensionality in species abundance distributions: description and evaluation of the Gambin model. *Evolutionary Ecology Research,* **9** (2), 313-324.

Ulrich, W., Ollik, M., & Ugland, K.I. 2010. A meta-analysis of species-abundance distributions. *Oikos,* **119** (7), 1149-1155.

Ulrich, W., Kusumoto, B., Shiono, T., & Kubota, Y. 2015. Climatic and geographic correlates of global forest tree species–abundance distributions and community evenness. *Journal of Vegetation Science,* **27** (2), 295-305.

Ulrich, W., Nakadai, R., Matthews, T.J., & Kubota, Y. 2018. The two-parameter Weibull distribution as a universal tool to model the variation in species relative abundances. *Ecological Complexity,***36** , 110-116.

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** , 658-661.

Zillio, T., Condit, R. 2007. The impact of neutrality, niche differentation and species input on diversity and abundance distributions. *Oikos,* **116** (6), 931-940.

Hosted file

`image1.emf` available at https://authorea.com/users/363764/articles/918868-three-models-of-ecological-community-assembly

Figure 1. Simulation analyses based on three models of population dynamics. See the text for details of the models. Species pools of 10,0000 species are assumed. Each illustrated pattern is the final set of counts obtained after 1000 iterations of the relevant death and birth algorithm. Black lines = simuated data; red lines = fits of the log series; blue lines = fits of the scaled odds distribution; green lines = fits of the $\exp^e$ distribution. (A) A routine model producing a log series-shaped pattern, which assumes no variation among species or through time in turnover rates (Kendall, 1948; Caswell, 1976; Hubbell, 2001). (B) The same data with the x-axis on a square root scale. (C) A model producing the scaled odds distribution, which assumes high variability among species and through time. (D) Square root scale. (E) A model producing the $\exp^e$ distribution, which assumes high variability among species and no variation through time. (F) Square root scale.

Hosted file

`image2.emf` available at https://authorea.com/users/363764/articles/918868-three-models-of-ecological-community-assembly

Figure 2. Relative support for the three models. Cartesian coordinates are a function of log likelihood scores relative to a saturated model (see text). Points near a corner are much better supported by the noted distribution; points at a centre are equally well-described by all three distributions. (A) Performance as a function of sampling intensity and species richness. Redder and larger points include more non-singleton

species, characteristic of large sample sizes. (B) Tree inventories. (C) Mammal inventories (gold) and bird inventories (turquoise). (D) Frog and lizard samples (green) and insect inventories (violet).

Hosted file

`image3.emf` available at https://authorea.com/users/363764/articles/918868-three-models-of-ecological-community-assembly

Figure 3. Empirical distributions best fitting four species abundance distributions. Each one is favoured over the next-best supported model by the largest margin observed across the entire Ecological Register data set. Line colours are as in Fig. 1. (A) The log series is most strongly supported by an ant survey from southern Ghana (Belshaw & Bolton, 1994). (B) The odds distribution is best supported by an arthropod inventory from Jefferson City, Missouri (Dowdy, 1947). (C) The $\exp^e$ distribution is best supported by a tree inventory from Barro Colorado Island, Panama (Condit et al., 1996). (D) A saturated model (line not shown) is most strongly needed to account for a tree inventory from Mawnai, India (Mishra et al., 2005).

Hosted file

`image4.emf` available at https://authorea.com/users/363764/articles/918868-three-models-of-ecological-community-assembly

Figure 4. A test of accuracy applied to four species diversity estimators. Each point represents a species inventory of terrestrial trees, arthropods, or tetrapods drawn from the Ecological Register (Alroy, 2015). The logged x-axis shows estimates based on the original, complete inventories; the logged y-axis shows estimates based on estimates recomputed after randomly excluding half of the individuals in each inventory. Diagonal lines are lines of unity, so scatters centered below the lines indicate bias. (A) Fisher's α (Fisher et al., 1943). (B) Estimates based on fitting the scaled odds distribution (eqn. 10). (C) Estimates based on fitting the $\exp^e$distribution (eqn. 16). (D) Estimates based on the Chao 1 index (Chao, 1984).