Folia Geobot (2008) 43:259–268
DOI 10.1007/s12224-008-9020-y

# Artifactions in the Log-Transformation of Species Abundance Distributions

**Jeffrey C. Nekola · Arnošt L. Šizling · Alison G. Boyer · David Storch**

**Abstract** One of the most frequently studied pattern in ecology is the Species Abundance Distribution (SAD) that represents the frequency distribution of species abundances in an assemblage. Two main approaches to displaying such information have been employed: histograms constructed using exponentially increasing bin widths as pioneered by Preston (1948), and plots of ranked species abundances. While both techniques have been extensively used in the investigation of community ecology hypotheses, the Preston-style species-abundance histogram has become central to current debates concerning appropriate characterization of the SAD and the processes generating it. Here we point out an important issue in the Preston approach that has profound implications to this debate: by employing bins of exponentially increasing size, the resultant histogram may display a hump-shaped pattern that is not congruent with the shape of the untransformed distribution. Moreover, any distribution constructed from log-transformed abundances will

J. C. Nekola · A. G. Boyer
Biology Department, University of New Mexico, Albuquerque 87131, New Mexico

J. C. Nekola
e-mail: jnekola@unm.edu

A. G. Boyer
e-mail: aboyer@unm.edu

A. L. Šizling
Biodiversity & Macroecology Group, Department of Animal & Plant Sciences,
The University of Sheffield, Sheffield S10 2TN, United Kingdom
e-mail: sizling@cts.cuni.cz

A. L. Šizling · D. Storch (✉)
Center for Theoretical Study, Charles University, Jilská 1, 110 00 Praha 1, Czech Republic
e-mail: storch@cts.cuni.cz

D. Storch
Department of Ecology, Faculty of Science, Charles University, Viničná 7, 128 44 Praha 2,
Czech Republic

necessarily reveal at least one internal mode, even when the non-transformed probability density function is strictly decreasing. We warn against misinterpretation of such transformed datasets, and suggest that rank-abundance plots, which are equivalent to the cumulative distribution functions extensively used in other branches of science, represent a more informative approach as they allow for better discrimination between a number of probability distributions. Ecologists should be aware that logarithmic transformation often generates a log-normal-like shape, and are encouraged to use rank abundance curves to visualize and analyze species-abundance patterns.

**Keywords** Commonness and rarity · Community structure · Diversity · Lognormal distribution · Macroecology · Models · Population size

## Introduction

The analysis of species-abundance patterns within communities has a long and venerable history in ecology. As related by McGill et al. (2007), the first considerations of this distribution may range back over a century and a half to observations made by John James Audubon and Charles Darwin. Empirical analysis of this pattern was initiated by Motomura (1932) and followed by Fisher et al. (1943), who not only plotted a histogram of species abundances but also produced scatter plots of log-transformed species richness within a focal abundance class *vs.* log-transformed abundances. Preston (1948) pioneered the analysis of species abundance distributions through histograms with bins of geometrically increasing width, and this approach remains one of the most common ways that ecological researchers visualize this pattern. Indeed, the shape of Preston-style histograms has become a major criterium used by theorists to attempt differentiation between competing community assembly models (e.g., Harte et al. 1999; Hubbell 2001; Chave et al. 2002; McGill 2003; Volkov et al. 2003; Chave 2004; Tilman 2004; Gaston and Chown 2005; Dewar and Porté 2008). Many have focused on the shape of this histogram, in particular the presence of an internal mode, to test for agreement between data and theory.

While previous investigations have noted that the binning involved in the Preston approach leads to loss of information (McGill et al. 2007) and slight differences in the visualized distributions (Gray et al. 2006), until now none have ever asked whether this method accurately documents the fundamental shape of underlying probability distributions. The impetus for this line of inquiry is based upon investigations of non-ecological abundance distributions reported in Nekola and Brown (2007), in which all analyzed datasets, spanning a large range of physical and human social systems, apparently displayed hump-shaped distributions over log-transformed data.

Here we show that Preston-style binning and any other logarithmic data transformation inevitably produces internal modes, i.e., hump-shaped or multimodal distributions, regardless of the distributions of untransformed abundances. Thus, the apparent ubiquity of hump-shaped species abundance distributions is due to a simple artifaction.

### Artifactions in Logarithmic Transformation

Standard histograms approximate Probability Density Functions (PDF, Appendix 1): when constructed over raw (untransformed) data, all bins have the same width, and the height of each bar is proportional to the number of data points that fall between the upper and lower limits of each respective bin. In many ecological studies, following Preston (1948), histograms of species abundance within a community are based on "octaves". Instead of equally-spaced bins, bin width varies based on a doubling rule (each subsequent bin is twice as wide as its predecessor) and then is presented as a standard histogram with equal-breadth bars. In essence this process is equivalent to first $\log_2$ transforming the data and then constructing a histogram. The resulting bar chart is thus equivalent to a standard histogram over $\log_2$-transformed data (Fig. 1). While binning of any sort always results in a loss of information (Williams 1964; Gray 1987; Magurran and Henderson 2003; Gray et al. 2006), here we additionally consider effects attributable to logarithmic data transformation independent of binning.

Logarithmic transformation of any data necessarily generates an internal mode on any probability density function. Generally speaking, this happens because a logarithm maps the interval from zero to infinity on the interval between minus and plus infinity. Any probability density function is, by definition, finite integrable, which means that the area below the curve is finite (usually set to one) and larger than zero (see Appendix 1). These conditions can be met only if both tails approach zero and if there is at least one peak somewhere between the infinities (Fig. 2a).

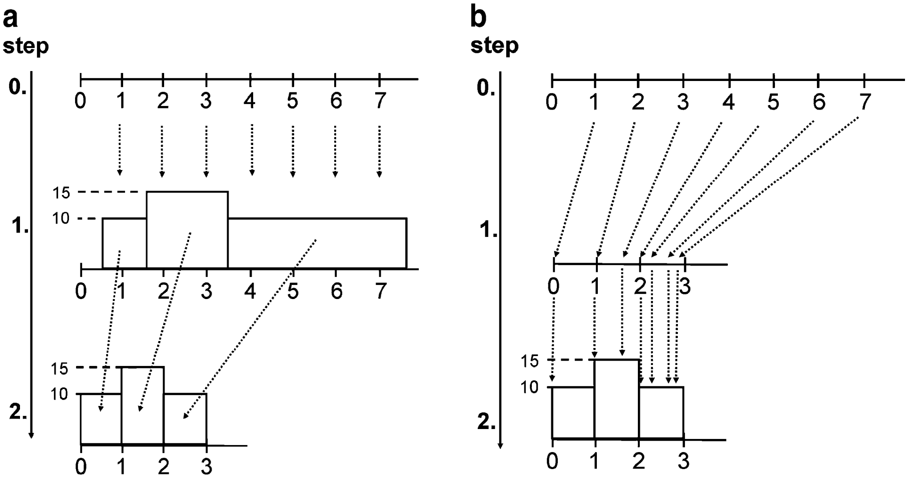

**Fig. 1** Preston-Gray's binning (**a**) and binning of logarithmically transformed abundances (**b**) are equivalent. Having numbers of species 10, 9, 6, 3, 3, 2, 2 within seven abundances classes from 1 to 7, respectively (steps 0), we can either follow Preston *i*) and bin them into exponentially (base 2) widening bins (step 1 in **a**) an then show the bins as equal sized (step 2 in **a**); or *ii*) first logarithmically transform abundances (using base 2 we get 0; 1; 1.58; 2; 2.32; 2.58; 2.81, respectively; step 1 in b) and bin them into bins 0–1 (incl 0 excl 1), 1–2 (incl 1 excl 2) and 2–3 (incl 2 excl 3) (step 2 in **b**). Note that $\log_2 8=3$ and species with abundance of eight would thus fall into the next bin 3–4. Since both processes produce the same results, we can see all Preston-Gray's binning as a histogram constructed over logarithmically transformed abundances

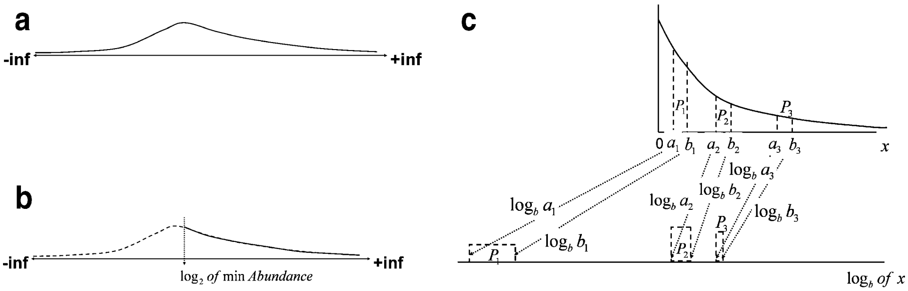

**Fig. 2** Probability density function (PDF) over data that range from minus infinity to plus infinity has always at least one peak (**a**) unless there is a lower boundary (log of minimum abundance), which can veil the peak (**b**). It is because the area below PDF is, by definition, always finite. The peak originates (**c**) even when PDF is strictly decreasing. The area between points $a_i$ and $b_i$ (labeled as $P_i$) is the probability that an $x$ falls between those two points. So, transforming data logarithmically (dotted arrows), the focal $P_i$ must be retained for the area between log $a_i$ and log $b_i$. More formally, $P_i(a_i \leq x < b_i) = P_i(\log a_i \leq \log x < \log b_i)$ (for details see Appendix 2). Now, as $a_i$ approaches zero, the distance between log $a_i$ and log $b_i$ increases. Thus, the lower is $a_i$, the lower is the value of the PDF over the transformed data. This mechanism deforms the left side of the PDF toward lower values and creates a mode as well as lifts the right tail and makes it heavier

However, this peak may not always be observable, particularly in situations when minimum abundance is constrained (Fig. 2**b**). Note that this minimum abundance value is not necessarily comparable with a "veil line" (see Discussion).

Given that PDFs over log-transformed data must possess an internal mode, we now turn our attention to the underlying mathematical mechanism. Probability density is proportional to the occurrence frequency of an observed value: higher probability densities correspond to higher occurrence frequencies ($\pi \cong f/(n+1)$, where $\pi$ is probability density, $f$ is frequency, and $n$ is the number of observations). The frequency can be estimated as a reciprocal value of a distance between two neighboring observations ranked along the rare-common gradient ($frequency \cong (a_{i+1} - a_i)^{-1}$). Since *i*) any PDF over values between zero and infinity becomes sooner or later decreasing and *ii*) any logarithm shortens the distance between two high neighboring abundances but enlarges distance between two low neighboring abundances, these two counteracting forces must generate at least one peak in a PDF over log-transformed data. Expressing this mathematically (Fig. 2**c**; Appendix 2) we can show that

$$g(y) = \beta^y f(\beta^y) \ln \beta \qquad \text{(Eqn.1)}$$

where g($y$) is a PDF over log-transformed data, $y$ is the log-transformed abundance of $a$ ($y = \log_\beta a$), $f(a)$ is a PDF over non-transformed data, $\beta$ is the base of the chosen logarithm and *ln* is the natural logarithm. The term $\beta^y$ represents the logarithmic force that creates the internal mode. Note that this force usually cancels all peaks on the original distribution (*f*($a$)), so the generation of internal modes in a log-transformed PDF is the simple consequence of logarithmic transformation. Since for this reasoning we do not need the assumption of binning, this artifaction is not limited to histograms.

The hump-shaped PDF over logarithmically transformed abundances thus says nothing about the existence of an internal mode in a species abundance distribution, and the presence of an internal mode can be considered a typical example of artifaction *sensu* Palmer (2008, this issue).

### *An Alternative Approach: Rank-Abundance Plots*

MacArthur (1957) and Whittaker (1965) initiated a different approach to visualizing the abundance distribution whereby the proportion of each species contribution to the total assemblage was plotted against the rank order of that species from the most to least common. Whittaker (1965) termed this plot a “dominance-diversity” curve (“rank-abundance” plot is a widely used, and actually more appropriate synonym), and it represents the other common way that ecologists have pictured the species abundance distribution (e.g. Hubbell 2001; McGill et al. 2007). This visualization technique does not result in a loss of detail or information because it does not gather abundance classes into bins and thereby shows every single abundance. Moreover, because proportional rank (i.e. normalized to the interval 0–1) estimates the probability that a randomly drawn abundance is less than or equal to that of the focal abundance, rank-abundance plots with untransformed axes are identical to the empirical cumulative distribution functions that have long been used in statistics.

Rank-abundance plots are easy to interpret in terms of presence or absence of an internal mode and symmetry and shape of the corresponding histogram constructed over the data. A clear logical correspondence exists between the rank-abundance plots and histograms generated from a given dataset (Fig. **3d**). The internal mode of the histogram is easy to see as an inflection point on the rank plot (the point when concavity becomes convexity; triangle in Fig. **3d**), and the potential symmetry of the distribution is represented by the symmetry around this inflection. Even if the inflection point does not exist in raw data, logarithmic transformation produces it — and thus the internal mode in the probability density function.

In the abundance distribution literature the rank-abundance plot is often presented with a logarithmically transformed abundance axis, which emphasizes the rare species part of the distribution. Zipf (1949) pioneered rank plots with logarithmic transformation applied to both axes. Astronomers, geologists, economists, sociologists, statistical physicists, linguists and complexity scientists frequently use this expression (Newman 2005).

### *Case Study*

We demonstrate this theory in the case of an exponential distribution ($f(x) = 0.1e^{-0.1x}, x \geq 0$), the probability density function of which is strictly decreasing (Fig. **3**). Neither the raw histogram (Fig. **3a**) nor rank plot (Fig. **3c**) shows any internal mode. The rank plot is strictly convex — and thus without any inflection. However, when data is log$_2$-transformed, the analytical form of the PDF becomes $g(y) = 0.1 \cdot 2^{y}e^{-0.1 \cdot 2^{y}} \ln 2$ ($y$ is any real number; see eqn 1, Fig. **2**c and Appendix **2**), which is a unimodal curve approaching zero in minus and plus infinity. The mode is at abundance of 10 individuals, i.e., at approximately $2^{3.32}$. The Preston-style histogram accurately represents the log$_2$-transformed PDF, as does the

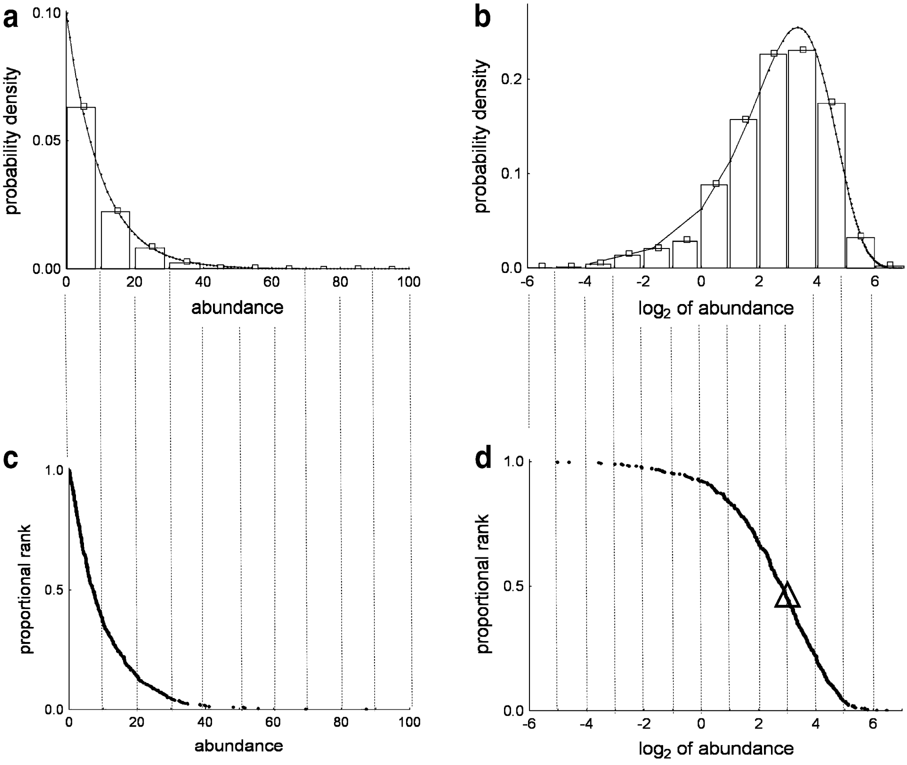

**Fig. 3** Logarithmic transformation of data which follow an exponential distribution (**a, c**) creates hump shaped distribution (**b, d**). **a** Histogram (squares and columns) constructed over 500 abundances randomly drawn (Appendix 2) from the exponential distribution $f(x) = 0.1e^{-0.1x}, x \geq 0$ (*full line*). The randomly drawn abundances are plotted in a rank plot (dots in **c**). (Note that the axes are replaced with each other in comparison to the usual rank-abundance plots in order to match the histogram above. This position of the axes also conforms to the customary orientation of a cumulative distribution function.) The number of dots within the corresponding bin (*dotted lines*) is proportional to the height of the corresponding column of the histogram in '**a**'. Thus, strictly convex rank plot means strictly decreasing histogram. **b** The exponential distribution from '**a**' for log$_2$-transformed data (i.e., $g(\log_2 x) = 0.1xe^{-0.1x}\ln 2$; *full line*; for explanation see Fig. 2 and Appendix 1) and histogram over the same sample (following methods in Gray 1987). **d** The rank plot '**c**' with a log$_2$-transformed abundance axis. The highest probability density is reflected by steepest slope of the rank plot. The modal bin in **b** corresponds to the inflection point on the rank plot (*triangle*)

rank-abundance plot, which shows an inflection point (i.e., mode) at abundance of about $2^3$. Data for all histograms and rank plots (Fig. 3) were randomly drawn (500 data points) from the exponential distribution (Appendix 3).

## Discussion

The theory presented here clearly documents that logarithmic transformation usually generates an artificial internal mode. If the internal mode does not occur, it only means that *i*) the sample was too small to sample a sufficient number of low abundance events; *ii*) the modal abundance occurred below the minimum possible abundance (usually one); or *iii*) there is a veil line, which prevents low abundances from being observed. The apparent ubiquity of hump-shaped abundance distributions over log-transformed data thus seems likely artifaction. Such artifactions are not limited to the analysis of species abundances, but are possible in any study analyzing log-transformed data, such as the analysis of species range sizes reported in Gaston and He (2002). In fact, this simple mathematical process is likely the general mechanism that underlies the omnipresence of hump-shaped frequency histograms reported by Nekola and Brown (2007). It is beyond the purview of this paper to interpret this finding in terms of current debates regarding models of community assembly. However, the shape of the many published examples of lognormal-like SADs in the ecological literature over the last 60 years may well tell us more about the process of log-transformation than it does about the actual shape of these distributions or the ecological processes governing them.

However, to dismiss all use of logarithmic binning as a visualization technique would be a mistake, for every kind of visualization has its own strengths and weaknesses. Histograms and rank plots of raw data properly demonstrate the proportional changes across various abundance classes, but in so doing tend to emphasize statistical fluctuations within high abundance classes while at the same time obscuring fine-scale variation at low abundances. This is particularly important given that assemblages are typically dominated by low-abundance species. While Preston-style histograms and rank plots of log-transformed data generally create artificial internal modes, they can also *i*) reveal whether data point densities change at the same rate to the left and right of this artificial mode (thus the symmetry of distributions over log-transformed data); and *ii*) emphasize abundance variation in lower abundance classes. The observed enrichment of rare species in many Preston-style histograms is thus neither an artifact nor artifaction *sensu* Palmer (2008), but reflects a genuine pattern. Differences between log-transformed distributions also describe real discrepancies between datasets. An example of this is provided by McClain and Nekola (2008) who used log-linear modeling to compare log-binned (i.e., with exponentially increasing bin widths) abundances of species and individuals across body size in Gastropods.

While log-transformation and Preston-style binning can be used to accurately document asymmetries within and differences between distributions, it is vital to remember that such transformed data cannot properly characterize the shape of the raw probability distribution function. Other analytical approaches thus must be used to achieve this goal. Rank-abundance plots, either with transformed or untransformed axes, allow all degrees of freedom in the original dataset to be utilized (Newman 2005), and a large number of distributions to be included within a single graph, aiding in the easy comparison of observed or simulated datasets. When untransformed, such plots properly show the proportion of species richness of neighboring abundance classes as well as their rates of change. Doubly log-transformed rank-abundance plots can be used to reveal deviations from a power-law distribution, while semi-log plots can be used to reveal deviations from a log-normal distribution.

## Appendix 1: Probability Density Functions (PDF)

Let us start with definition of probability density: “probability density equals the probability that a species of given abundance falls within a given bin of unit width”. Since a bin of unit width is often so wide that it veils variation in probability densities of various abundances, we define the probability density, more precisely, as a probability over an interval (the probability is normalized by the interval size) where interval size approaches zero. We usually estimate the density as the probability over some small reasonably observed interval divided by the size of the interval. Note that probability over an interval is proportional to number of species which fall into the interval. The sense of probability density is to get as fine detail on probability of occurrence of a focal abundance class within an assemblage as possible; in other words, the sense is to have a variable whose integral (i.e., infinitesimal summing) across an interval equals the probability over the interval. Integrating the probability density across all possible abundance classes thus gives one (i.e. all species of an assemblage normalized by all species of the assemblage). Changes of probability density with abundance classes show how probability of occurrence of particular abundances vary within the focal assemblage. A histogram is then a rough approximation to the probability density function and requires bars with equal widths. If bars are not equal, their height must be normalized by their width according to the definition of the probability density (see above and also Newman 2005).

## Appendix 2: Probability Density Function Over Log-Transformed Data

*Aim* To derive the relationship between a probability density function (*f*(*x*)) and the corresponding probability density function for logarithmically transformed axis ($g(\log_\beta x)$, where $\beta$ is a base).

*Solution* Since $a \le x < b \Rightarrow \log_\beta a \le \log_\beta x < \log_\beta b$, the probability that an $x$ falls between any a,b equals the probability that $\log_\beta x$ falls between $\log_\beta a$ and $\log_\beta b$ (i.e., $P(a \le x < b) = P(\log_\beta a \le \log_\beta x < \log_\beta b)$; see also Fig. 2). Since $(b-a)f(x) \overset{a\to b}{\to} P(a \le x < b)$ and $(\log_\beta b - \log_\beta a)g(\log_\beta x) \overset{a\to b}{\to} P(\log_\beta a \le \log_\beta x < \log_\beta b)$, then $(b-a)f(x) \overset{a\to b}{\to} (\log_\beta b - \log_\beta a)g(\log_\beta x)$. Replacing $a$ with $x$ and $b$ with $x$+Δ$x$, we get

$$\frac{\ln(x+\Delta x) - \ln x}{\ln\beta} g(\log_\beta x) \xrightarrow{\Delta x \to 0} f_{(x)}((x+\Delta x) - x). \tag{Eqn.S1}$$

(Note that $\log_\beta x \equiv \ln x / \ln \beta$.)
Therefore,

$$g(\log_\beta x) = f(x) \ln\beta \lim_{\Delta x \to 0} \frac{\Delta x}{\ln(1 + \Delta x / x)} \tag{Eqn.S2}$$

which, using the l’Hospital theorem, is

$$g(\log_\beta x) = x f(x) \ln\beta, \qquad \text{where } x \ge 0. \tag{Eqn.S3}$$

Or, replacing $\log_\beta x$ with $y$,

$$g(y) = \beta^y f(\beta^y)\ln\beta, \qquad \text{where } y \in \Re \text{ (i.e. } y \text{ is a real number).} \qquad \text{(Eqn.S3)}$$

## Appendix 3: Random Sample from an Exponential Distribution

The probability density function of an exponential distribution follows $f(x) = be^{-bx}$, where $b>0$. Cumulative distribution function then follows $F(\xi < x) = b\int_0^x e^{-b\xi}d\xi = 1 - e^{-bx}$. Distribution function always maps regular distribution onto the distribution in question. Thus any $x = -\ln(1-y)/b$ where $y$ is sampled from regular distribution between 0 and 1 is a sample drawn from the exponential distribution. In short, this transformation maps a point $\{x; f(x)\}$ into the point $\{\log_\beta x;\ xf(x)\ln\beta\}$.

**Acknowledgments** We thank Tomáš Herben for inviting this contribution, and Jiří Reif and two anonymous referees for theirs comments. A.L.Š. was supported by the Marie Curie Fellowship nr. 039576-RTBP - EIF. A.B. was partially supported by the Program in Interdisciplinary Biological and Biomedical Sciences at the University of New Mexico. This work was supported by the Czech Ministry of Education (MSM0021620845, LC06073) and the Grant Agency of the Academy of Sciences of the Czech Republic (IAA601970801).

## References

Chave J (2004) Neutral theory and community ecology. *Ecol Lett* 7:241–253

Chave J, Muller-Landau HC, Levin SA (2002) Comparing classical community models: theoretical consequences for patterns of diversity. *Amer Naturalist* 159:1–23

Dewar RC, Porté A (2008) Statistical mechanics unifies different ecological patterns. *J Theor Biol* 251:389–403

Fisher RA, Corbet AS, Williams CB (1943) The relation between the number of species and the number of individuals in a random sample of an animal species. *J Anim Ecol* 12:42–58

Gaston KJ, He F (2002) The distribution of species range size: a stochastic process. *Proc Roy Soc London* B 269:1079–1086

Gaston KJ, Chown SL (2005) Neutrality and the niche. *Funct Ecol* 19:1–6

Gray JS (1987) Species-abundance patterns. *Symp Brit Ecol Soc* 27:42–58

Gray JS, Bjørgesǽter A, Ugland KI (2006) On plotting species abundance distributions. *J Anim Ecol* 75:752–756

Harte J, Kinzig A, Green JL (1999) Self-similarity in the distribution and abundance of 634 species. *Science* 284:334–346

Hubbell SP (2001) *The unified neutral theory of biodiversity and biogeography. Monographs in Population Biology # 32*. Princeton University Press, Princeton, New Jersey

MacArthur RH (1957) On the relative abundance of bird species. *Proc Natl Acad USA* 43:293–295

Magurran AE, Henderson PA (2003) Explaining the excess of rare species in natural species abundance distributions. *Nature* 422:714–716

McClain C, Nekola JC (2008) The role of local-scale on terrestrial and deep-sea Gastropod body size distributions across multiple scales. *Evol Ecol Res* 10:129–146

McGill BJ (2003) A test of the unified neutral theory of biodiversity. *Nature* 422:881–885

McGill BJ, Etienne RS, Gray JS, Alonso A, Anderson MJ, Benecha HK, Enquist BJ, Green JL, He F, Hurlbert AH, Magurran AE, Marquet PA, Maurer BA, Ostling A, Soykan CU, Ugland KI, White EP (2007) Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecol Lett* 10:995–1015

Motomura I (1932) A statistical treatment of associations (in Japanese). *Zool Mag Tokyo* 44:379–383

Nekola JC, Brown JH (2007) The wealth of species: ecological communities, complex systems, and the legacy of Frank Preston. *Ecol Lett* 10:188–196

Newman MEJ (2005) Power laws, Pareto distributions, and Zipf's law. *Contemp Phys* 46:323–351

Palmer MW, McGlinn DJ, Fridley JD (2008) Artifacts and artifictions in biodiversity research. *Folia Geobot* 43(3):245–257

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

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

Volkov I, Banavar JR, Hubbell SP, Maritan A (2003) Neutral theory and relative species abundance in ecology. *Nature* 424:1035–1037

Whittaker RH (1965) Dominance and diversity in land plant communities. *Science* 147:250–260

Williams CB (1964) *Patterns in the balance of nature and related problems in quantitative ecology.* Academic Press, London

Zipf GK (1949) *Human behavior and the principle of least effort*. Addison–Wesley, Reading, Massachusetts

Received: 31 March 2008 / Revised: 19 May 2008 / Accepted: 27 June 2008 /
Published online: 5 November 2008