CHAPTER 9

Species abundance distributions

Brian J. McGill

9.1 Introduction

A species abundance distribution is a quantitative statement about the difference in abundances between species in a community. In short, it is a statement about the commonness and rarity of species (see Chapter 8). As such, it cuts to the central questions of basic ecological research (the study of the distribution and abundance of organisms: Krebs 1972; Andrewartha & Birch 1984) and of conservation biology (defined by Soule as the science of rarity: Soule 1986). Thus it is not surprising that species abundance distributions are among ecology's oldest patterns, with quantitative measurements as far back as 1909 (Raunkaier 1909) and explanatory models as far back as 1932 (Motomura 1932). Every species abundance distribution ever measured with an adequate number of species (say at least three to five) has shown the same pattern: most species are rare, only a few species are common. In short, 'rare is common, common is rare'.

Perhaps the most useful definition of a species abundance distribution is a vector of comparable abundances representing a community. The idea of 'vector' implies that there is an abundance measure for multiple species. The idea of 'comparable' implies that it is sensible to compare and contrast these numbers. At a minimum comparability requires that the numbers either be true densities or if indices of abundance then that they are estimated in the same fashion across all the species. The idea that they represent a 'community' is the hardest to pin down. Ecologists, despite having community ecology as a central subdiscipline, have never given a truly operational definition to the concept of community. Here we use it to imply a group of organisms that overlap in time and space such that there is a reasonable chance of interaction between the members. The temporal and spatial extent can be quite variable. Spatial extents of as little as $0.25\,m^2$ are common for grasses and milliliters are common for microbial organisms while communities are sometimes taken to cover entire continents. Similarly temporal extent may range from a point in time (in practice, say, a single day) to millions of years in paleontological applications. The taxonomic extent of a community is also quite variable (ranging from a single genus to entire kingdoms, e.g. all plants). The key criteria seems to be that there is an expectation that the members have the potential to interact with each other and the list of species is in some meaningful way exhaustive (species are not omitted arbitrarily). An extraordinary amount of data have been collected that fit this definition of 'a vector of comparable abundances for the species in a community'. Nearly every dataset with multiple species and abundances (density or indices of abundance) measured qualifies.

There is also an issue of what measure of abundance or what 'currency' to measure. Although number of individuals is by far the most common, this approach has limitations when applied to plants that show clonal growth, and it can also distort results in indeterminate growers, such as plants or fish that can vary by orders of magnitude of size within one species. Common currencies include number of individuals, percentage cover, biomass, and energetic use. It has been argued that energy use might be closer to the actual processes regulating community assembly and therefore should be preferred (Tokeshi 1999). Broadly, the distribution observed is the same for all of these currencies, but subtle and possibly important differences do occur (Wilson et al. 1996a; Chiarucci et al. 1999; Connolly et al. 2005; Morlon et al. 2009).

What might one do with a species abundance distribution? A few obvious applications include:

1. using it to test various models or theories of how communities are assembled
2. using it to identify rare and common species for management purposes (which typically focus on rare species and need quantitative assessments of rarity to justify prioritizing some species over others)
3. using it as an indicator of the ecosystem health of a location
4. to compare two or more locations (or one location across time), for example how do the SADs of two or more communities along a productivity gradient or along a human disturbance gradient change? Such comparisons might be done with any of the goals in points 1–3 in this list in mind.

To date, usage 1 has probably been the most common. A great deal of literature has been devoted to testing theoretical models of community processes using species abundance distributions. Thus Motomura (1932) used SADs to suggest niche preemption. Fisher et al. (1943) and Preston (1948) debated whether statistical arguments leading to log-series or log-normal distributions, respectively, were correct. More recently, literally dozens of other theories of community assembly have been tested by species abundance distributions (Hughes 1986; Tokeshi 1996; Dewdney 2000; Hubbell 2001; Pueyo 2006; Harte et al. 2008). Unfortunately, species abundance distributions have shown little power to either reject theories (Popper 1959; Platt 1964) or differentiate between them (Wilson 1993; Chave et al. 2002; McGill 2003a,b). Even when it can be decisively shown that one theory or model fits a particular location better, very similar locations with very similar groups of organisms are usually fit by a different theory best.

For this reason there has been a call to move away from using SADs for point 1 above, and to move towards a more comparative approach (point 4)(McGill et al. 2007). In general, ecology, and indeed all of science, works better with some form of comparison, whether observational or experimental. It is not reasonable to ask physics to predict that a cloud of gas will condense to form a ball of the exact size and location that we now know as Earth (Roughgarden 2009), but it is reasonable to ask physics to predict what will change in the Earth's motion if a given gravitational attraction from the Sun is applied. Similarly, it may be more realistic to expect ecology to predict how the relative proportions of rare and common species will change as productivity is increased or logging occurs than to predict the precise relative proportions of the abundance of each species at a given location and point in time on Earth.

The 1970s saw a burst of research in this direction. A number of authors explored how SADs changed shape along a variety of gradients, such as succession or productivity (Whittaker 1965; Bazzaz 1975; Hubbell et al. 1999). Fairly strong patterns were suggested. For example, SADs become more even with higher productivity. Work on using SADs as ecosystem indicators was also done (Gray 1979), clearly showing identifiable changes to marine communities as a consequence of pollution. At that time all of the comparisons were done graphically (i.e. by eye-ball). Rank abundance distribution (RAD) plots (log abundance vs rank order) were plotted and their shapes were compared. Unfortunately, this research programme died out quickly without significant confirmation of any of the identified patterns (perhaps in part because they were incorporated so quickly into a prominent textbook and were taken as being well established; Whittaker 1975). In any case, relatively little effort has been made since on comparing SADs to inform about underlying processes or management implications (although exceptions exist, e.g. Wilson et al. 1996a).

9.2 State of the field

Whatever one's goals and whether one uses a single SAD or is comparing SADs as advocated above, there are four basic types of tools to use: visual, parametric, non-parametric, and multivariate. These are described in the following sections.

9.2.1 Visual approaches to SADs

The simplest way to use SADs is to plot them graphically. The ease with which this is done is

undoubtedly one of the main reasons for their appeal. There are three main ways to plot the data. The first is to create a histogram. Since the number of unique abundances observed will be a large fraction of the number of species observed, histograms require aggregating or binning the data. There has been great controversy about the best binning method. The simplest way to plot the data is to use an arithmetically scaled axis (Fig. 9.1(a)). Preston (1948) argued for a logarithmically scaled axis (Fig. 9.1(b)) and introduced a method that involved smoothing the data (splitting the number of species found on bin boundaries across two bins). Some have argued for a binning that accounts for the strong skew of the data by lumping the extreme right tail of the data into a single very large bin on an arithmetic axis, which is sometimes referred to as the $N + 1$ method (Nee 2003). Unfortunately, the results and interpretation of the graphs depends heavily on the binning method used (Gray et al. 2006). In particular, the interpretation of the number of rare species depends on whether an arithmetic or logarithmic scale is used and which exact logarithmic scale. More generally, binning represents a loss of data, which is a statistical no-no when avoidable. More modern techniques for estimating probability density functions (PDFs) such as kernel estimation (Martinez & Martinez 2002) could and probably should be used in preference to binning (see the curved line on Fig. 9.1(a) and (b)). As far back as 1957, MacArthur pointed out these problems and suggested an alternative plotting method, later championed by Whittaker (1965), which has become known as the rank abundance

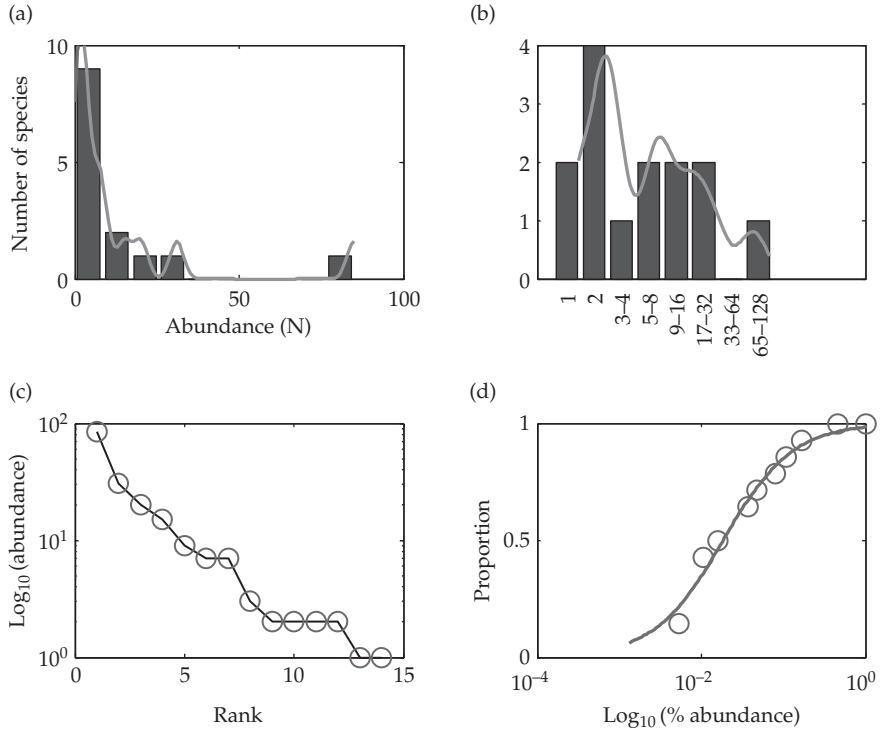

Figure 9.1 Four different ways to plot species abundance data: (a) a histogram (binned) plot on an arithmetic scale; (b) a histogram plot on a log scale (in both (a) and (b) the curved line represents a kernel-smoothed representation of the data); (c) a classic rank abundance distribution; (d) a scaled empirical cumulative distribution function plot (see Box 9.1). The curved line represents a logistic curve fit to the data. This data and the data used in Figure 9.3(a) and (b) are based on trees >1 cm dbh on transects run at Mt St Hilaire nature preserve operated by McGill University as a field station. The data were collected by students in the 2006 Field Methods in Ecology and Behavior course.

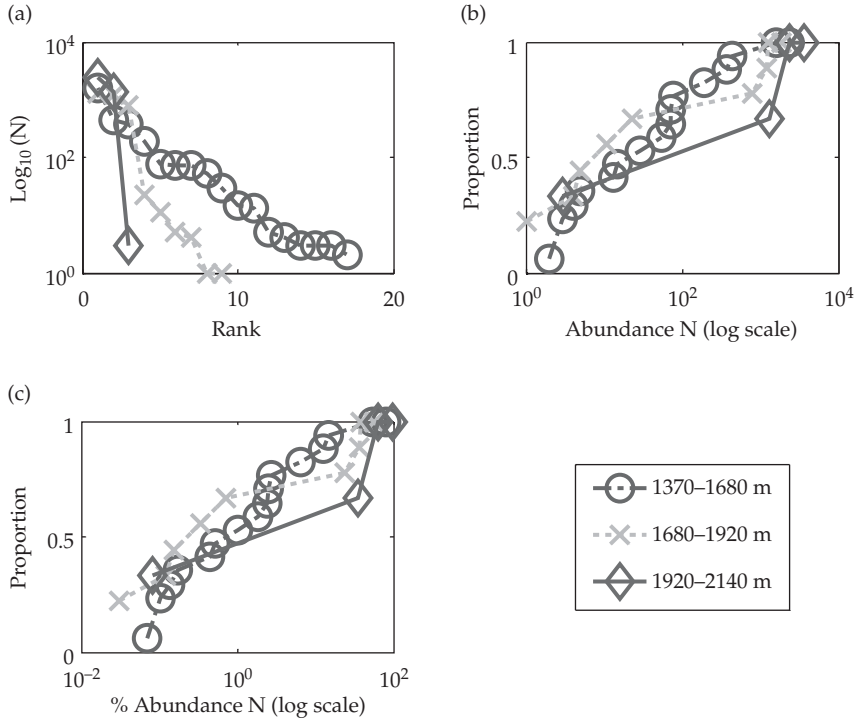

Figure 9.2 Different ways to plot an SAD. This figure is a plot of the data reported in Whittaker 1960. (a) Classic rank abundance distribution (RAD) plot with log N on the y-axis and rank ($1 \ldots S$) on the x-axis. (b) The empirical cumulative distribution function (ECDF) with log N on the x-axis and proportion of all species on the y-axis. (c) ECDF but with the abundances transformed into percentages of all individuals observed for the given community (here N = 2949, 3342, and 3741, respectively).

plot (Fig. 9.1(c)). In this plot the individual species abundances are sorted highest to lowest. The y-axis is log-abundance and the x-axis is rank (1 = most abundant down to S = number of species for the least abundant). As MacArthur pointed out, this plot has a close relationship to the cumulative distribution function (CDF) used in probability theory to characterize a probability distribution. RADs appear to be easy to interpret. Whittaker (1965) and others have interpreted RADs such as those shown in Fig. 9.2(a) to indicate that low diversity communities are geometric (Motomura 1932), medium diversity communities are log-series (Fisher et al. 1943) and high diversity communities are log-normal (Preston 1948).

Two problems occur with RAD plots. The first is mathematical. Rank is mathematically difficult to deal with (it invokes the field of order statistics), but the use of a traditional CDF plot would allow for the invocation of a vast array of mathematical machinery developed for CDFs. Secondly, RAD plots suffer from a severe confoundment—by definition the bottom right corner of a RAD occurs at S, the number of species in the community. Thus species with higher diversity will have RADs that are more stretched out along the x-axis (rank) and appear more even. This makes it very difficult to use RADs to provide meaningful comparisons of species distributions across communities with even moderate variations in species richness as the differences in richness quickly dominate all other patterns. A solution to these two problems, called the empirical cumulative density function (ECDF), is my recommended way to plot SADs (Box 9.1 and Fig. 9.2).

Box 9.1 Empirical cumulative distribution function plot

A simple and quick method for correcting for different species diversities between communities is to use a new type of plot (Figs 9.1(d) and 9.2). Specifically, the cumulative distribution function is a plot of the proportion of points with a value (here abundance) less than a given value (CDF(X) = P{$x \leq X$}). This is a standard way to analyse probability distributions. When the CDF is calculated from data rather than from a formula, it is called an empirical cumulative distribution function (ECDF). The Kolmogorov–Smirnov statistic calculates the difference between the ECDF and the theoretical CDF (or between two ECDFs).

The CDF is also the integral of the PDF. The PDFs are often plotted for SADs. This means peaks in the PDF correspond to steep upward slopes in the CDF. For example, the Preston plot (1948) is just a PDF on a $\log_2$ scale with binning (i.e. a histogram) used to estimate the PDF from empirical data. However, in comparison to the PDF, plotting the CDF has two distinct advantages:

1. The ECDF is easily estimated. Sort the abundances into increasing order. Use these as the values on the x-axis. Place the ranks (here from lowest to highest abundance, the reverse of RAD) divided by the total number of species on the y-axis. This will give values ranging from $1/S$ to 1. Normally the point (0, 0) is also included to make the graph complete.
2. Because it sums across data, the ECDF is usually more robust (less sensitive to outliers) than the empirical PDF.

The ECDF is also closely related to the RAD (Fig. 9.2). Specifically, the ECDF is generated by flipping the RAD about the $y = x$ line, then rescaling the rank by dividing by S, then inverting the rank axis (smallest to largest instead of largest to smallest). The ECDF has two distinct advantages over the RAD:

1. By rescaling by S the inordinate sensitivity to differences in S allows for useful comparisons to be made between sites.
2. By tapping into traditional mathematical machinery, it should be possible to import new tools into the analysis of SADs.

The improved comparative power of an ECDF over a RAD is demonstrated in Fig. 9.2. The data are from vegetation at different elevations in the Siskiyou mountains (Whittaker 1960). As noted in the main text, this and related data plotted as in Fig. 9.2(a) were used to argue that SADs are fundamentally different at different elevations,

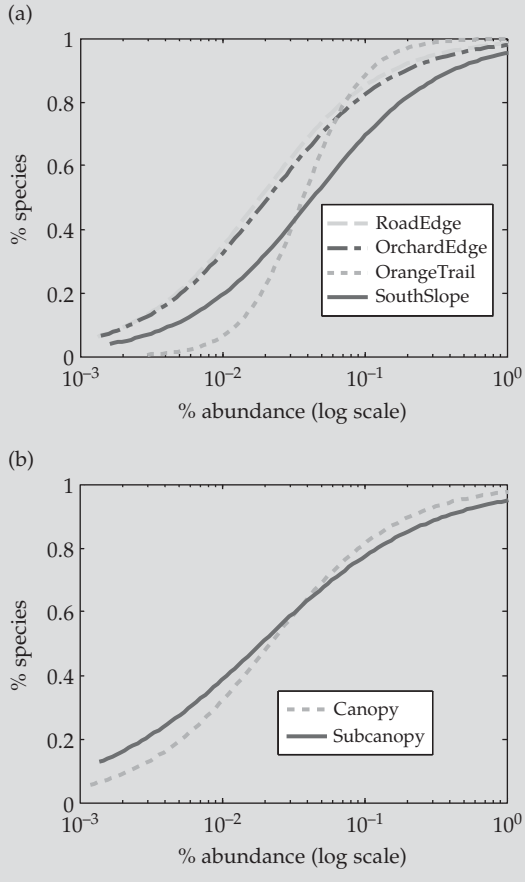

Figure 9.3 Comparisons of SADs using the ECDF plot. (a) Four different plots compared. Data as in Fig. 9.1. We can see that the site at Orange Trail is very different in evenness (the evenness is much higher since the line is more sharply vertical, implying that most species have an abundance of about 4% of the total population). The other sites are broadly similar in this aspect of evenness with species spread out between 0.1 and 10%, but SouthSlope overall has a lower proportion of rare species (the line is shifted down, implying that a given percentile—say 50% of species read on the vertical axis—occurs at a higher proportional abundance, roughly 4% instead of 1%). (b) Two groups of species compared. Here data are summed across all four sites and one SAD is plotted for species primarily found in the canopy, and one SAD is plotted for species primarily found in the subcanopy (shade or early successional). Here it is clear that the subcanopy has more rare species (higher on the left side of the graph).

Continued

Box 9.1 (*Continued*)
productivities, and diversities (Whittaker 1965). However, the ECDF in Fig. 9.2(b) gives a much more realistic depiction of how different the SADs are due directly to elevation and productivity rather than the indirect effect these factors have on species richness (S). Fig. 9.2(b) is also suggestive of greater evenness in the lowest site (perfect evenness shows in the ECDF as a step function or equivalently in the PDF as a delta function with all abundances concentrated at a single value). However, this pattern is nowhere near as strong once species richness is adjusted for. Indeed, to our knowledge, Fig. 9.2(a) has never invoked a call for the need to place confidence intervals around the SADs since the effects are so strong, while Fig. 9.2(b) clearly demands an assessment of significance. On the other hand, the ECDF can clearly discriminate between communities. In Fig. 9.3 the raw data (with its noise) is replaced by a logistic function fit to the data. Note that fitting such a function is tantamount to modelling the data as sampled from a log-logistic probability distribution.

9.2.2 Parametric approaches to SADs

The use of the probability-based ECDF immediately suggests that viewing a SAD as a sample from a probability distribution might be useful. To have a reasonable chance of fitting the data, the probability distribution needs to be defined only for positive numbers and to be strongly skewed with a long right tail on an arithmetic scale. The 1940s saw two suggested probability distributions that met these requirements: Fisher's log-series (1943) and Preston's log-normal (1948). Both fit the data reasonably well with the main difference between the two occurring in the prediction for extremely rare species (species with abundances of 1 or 2) (Chapter 10). Preston and Fisher both argued that their approach was superior and this argument has continued unresolved to the present day, with the added complexity of literally dozens of different probability distributions proposed (for extensive lists see Marquet et al. 2003 and McGill et al. 2007). It seems likely that SADs do not contain enough data to decisively decide which one probability distribution is the best fit (if indeed there is such a thing as one distribution to describe all communities) (McGill et al. 2007).

Some probability distributions commonly applied to species abundance distributions are described in Box 9.2. Each distribution will have several (usually one to three) parameters that can be chosen to make the general probability distribution fit a particular dataset (see Chapter 10). These parameters can be interpreted in one of two ways. First, it can be assumed that the underlying probability distribution and any model that produced this distribution is correct. In this case the parameters can be interpreted in this context. For example, the migration parameter, m, and the speciation parameter, θ, in neutral theory can be estimated to provide the best fit to the neutral theories predicted probability distribution (the zero-sum multinomial or ZSM) and then interpreted according to their mechanistic definitions. Alternatively, these parameters can simply be seen as geometric descriptors of the shape of the species abundance distribution without assuming the underlying model is necessarily true or applicable. In this scenario, for the ZSM, m can be seen as a descriptor of the shape of the left tail of the SAD and the number of rare species, while θ can be seen as a scaling parameter that broadly affects the overall shape and the right tail (McGill et al. 2006). Under either interpretation, these metrics can broadly be called 'parametric' as they are derived as the parameters of a mathematical probability distribution that summarizes the shape of the SAD.

9.2.3 Non-parametric approaches to SADs

Many other numbers or metrics can be used to describe the shape of a SAD that do not derive from a probability distribution. Some of these are intended to measure a specific property of the SAD, such as the diversity or the evenness. Others are just geometric descriptors of the shape of the SAD. An example of the latter is the slope of the RAD (Nee et al. 1992). A number of these measures are summarized in Box 5.1.

Box 9.2 Parametric measure of SAD shape

Parametric measures are the parameters of a mathematically described probability distribution. Such parameters may be interpreted either under the assumption that the probability distribution applies to the data or merely as a useful descriptor of the shape of the data. In the following list, the name of the distribution is given with its parameters following in parenthesis. After some brief comments, a best-practice method for estimation is referenced and the formula for the distribution is given. Some common and useful distributions include the following:

- **Pareto/power (*c*):** The Pareto or power distribution is based on $p(n) \propto 1/n^c$. The Pareto can be justified as the limit of the log-normal distribution with high variance (Montroll & Shlesinger 1982), as the sole distribution surviving basic transformations (Mandelbrot 1963), as a sign of fracticality (Mandelbrot 1982), or when $c = 1$ as a sign of self-organized criticality (Pueyo 2006). Most SADs fail to be well fit by the Pareto. There are several variants of this distribution and parameter estimation can be tricky (White et al. 2008). I recommend the Discrete Pareto which runs from 1 to N. A simple analytical maximum likelihood estimator exists (if using software that can calculate the harmonic function, $\zeta(c)$). $p(n) = n^{-c}/\zeta(c)$
- **PowBend (β, ω):** Pueyo (2006) suggested that the Pareto distribution should apply to SADs but failed because of sample size limitations, in particular the very common species are not accurately represented in their true proportions due to finite sample size. He derived a modified power distribution (the power-bended distribution or PowBend) that bends the distribution down to have fewer large species to represent this. Nominally, the parameter β is the same as the Pareto c, while ω represents the degree of bending. In practice ω ensures a fit to common species, which allows β to fit the rare species portion of the SAD well. Pueyo recommends a four-step programme:
  1. Fit a $1/x$ distribution (i.e. power distribution with $c = 1$ or, equivalently, the powbend with $\omega = 0$ and $\beta = 1$).
  2. If the fit is poor, allow c to vary in the power distribution (or equivalently vary β with $\omega = 0$).
  3. If the fit is still poor, switch to the PowBend and allow ω to also vary.
  4. If this still fails, a log-normal can be fit.

  These four steps correspond to using the first to fourth terms in a Taylor power series, and can be used to detect the complexity of the community. Parameters can be estimated by numerical maximization of the likelihood surface. $p(n) = Cn^{-\beta}\exp(-\omega n)$
- **Zipf–Mandelbrot (*b*, *c*):** Another modification of the power distribution is the Zipf–Mandelbrot distribution, which contains the power distribution as a special case when $b = 0$. It was first applied to word frequencies and in ecology is interpreted as a sort of successional model where later species have increasingly complex ecological requirements and as a result are extra rare (Frontier 1985). In practice, unlike the Pareto with its excessively rigid one-parameter shape, but like the PowBend, the Zipf–Mandelbrot has a greatly improved fit with two parameters, allowing one parameter (c) to describe fit to common species and the other (b) to describe fit to rare species. Fitting the Zipf–Mandelbrot can be tricky (Izsak 2006). I recommend minimizing a χ^2 goodness-of-fit only on abundances >5 with a careful choice of initial values to speed convergence. $p(n) = (b + n)^c/H(N,b,c)$ where $H(N,b,c)$ is a normalizing constant and is an extension of the ζ function where $H(N, b, c) = \sum(i + b)^{-c}$ and the sum goes from $i = 1 \ldots N$.
- **Log-normal (μ, $\sigma\mu^*$, σ^*, **cv**, **cv***):** The log-normal distribution arises from the central limit theorem on a product of random variables (Limpert et al. 2001). A variety of biological interpretations can fit this description (McGill 2003a; McGill & Nekola 2010). The two parameters are the mean, μ, and the standard deviation, σ, of the (usually natural) log-transformed abundances. However, $\mu^* = \exp(\mu)$ is often reported because μ^* is in units of abundance. Some also transform $\sigma^* = \exp(\sigma)$ to return σ to units of abundance, but this is less common. Clearly μ is a scale parameter (position on the x-axis of abundance), but shape depends on the interaction between μ and σ. The coefficient of variation ($cv = \sigma/\mu$ or $cv^* = \sigma^*/\mu^*$) serves as a shape descriptor, with $cv \ll 1$ indicating a nearly normal shape (even on an arithmetic scale) and the skewness increasing as the cv increases. Parameter estimation is trivial (simply log transform and take the mean and standard deviation).
  $p(n) = \exp[-\log(n/\mu)^2/(2\sigma^2)]/[n\sigma(2\pi)^{1/2}]$

Continued

Box 9.2 (*Continued*)

- **Poisson log-normal (μ, σ, μ^*, σ^*, cv, cv*):** The main problem with the log-normal is that it is a continuous distribution allowing for fractional abundances. One approach is to develop a sampling model (Pielou 1975; Dewdney 1998; Green & Plotkin 2007) (see Chapter 10) where the abstract abundance of a species is given by λ sampled from a log-normal, but the observed abundance is sampled from the Poisson distribution with the one parameter (the average value) given by λ. Parameter estimation requires numerical maximization of the likelihood surface and can be computationally intense as well as requiring some good numerical approximations (Chapter 10, Bulmer 1974). $p(n) = \int_0^\infty \frac{\lambda^n e^{-\lambda}}{n!} p_{LN}(\lambda) d\lambda$ where $p_{LN}(\lambda)$ is the log-normal probability (see above)
- **Gamma (scale, shape):** The gamma distribution is another continuous probability distribution on the range $(0, \infty)$. It is broadly used in mathematics as a general, flexible distribution. Unlike the log-normal, its two parameters are clearly divided into a scale parameter (position on the *x*-axis) and the shape parameter, with shape again varying from modal to a single mode near $N = 0$. Both the gamma and the Weibull distribution (next) are mathematically more tractable than the log-normal. The gamma distribution is not frequently used for SADs, but occasionally it is used and in some cases found to be the best-fitting distribution (Plotkin & Muller-Landau 2002). There is a simple formula for maximum likelihood estimation (Evans et al. 1993). $p(n) = (n/\text{scale})^{\text{shape-1}} \exp(-n/\text{scale})/(\Gamma(n)\text{scale})$, where $\Gamma(n)$ is the gamma function.
- **Weibull (scale, shape):** The Weibull is another widely used continuous probability distribution defined on $(0, \infty)$ and therefore sensible for modelling abundances. Again there is a simple MLE formula for estimation (Evans et al. 1993). $p(n) = \text{shape} \times (n/\text{scale})^{\text{shape-1}} \exp[-(n/\text{scale})^{\text{shape}}]/\text{scale}$
- **NegBin (*P*,*K*):** Just as one can make a Poisson log-normal to represent integer sampling from an abstract continuous distribution, one can make a Poisson gamma distribution. This is called the negative binomial distribution. Traditionally, the distribution is zero truncated (the probability of sampling abundance 0 is removed by rescaling). The negative binomial has several different parameterizations and can be confusing. I recommend following the one given by Pielou (Pielou 1975, p. 35–38) along with her estimation method. In this case, *P* is the scale parameter of the gamma distribution and *k* is the shape parameter. One interpretation is that *k* measures inverse distance from the Poisson distribution ($k = \infty$ gives the Poisson while $k = 0$ is far from Poisson). Thus *k* has been used in studies of abundances in quadrats to indicate the degree of clumping, but it is unclear how this relates to SADs. $p(n) = \frac{\Gamma(k+n)}{n!\Gamma(k)} \left(\frac{P}{1+P}\right)^n \frac{1}{(1+P)^k - 1}$
- **Log-series (*c*):** The log-series is the limit of the negative binomial as *k* goes to zero (with some mathematical details added) and the zero abundances are truncated (Fisher et al. 1943). It has a single parameter, *c*. Fisher's α is technically not a parameter of the distribution but is instead an implicit function of the parameters, *S* and *N* (specifically $S = \alpha \ln[1 + N/\alpha]$). Fisher's α is generally recognized as a measure analogous to species richness, although in different units. Estimation of *c* can be done by numerically solving for *c* in the equation $\overline{n_i} = \frac{C}{-(1-C)\log(1-C)}$. $p(n) = kc^n/n$, where $k = -1/\ln(1 - c)$
- **ZSM (m, θ):** The birth–death–immigration equations of neutral theory can be solved to produce what is known as the zero-sum multinomial (ZSM) distribution (Hubbell 2001; Etienne 2005), although the zero-sum assumption is not critical (Etienne et al. 2007). It has two parameters. Nominally θ is a speciation rate and *m* is the percentage of individuals migrating from outside the local community. In practice these values are impossible to measure and θ and *m* are fit by maximum likelihood estimation with the likelihood calculated using complex combinatoric methods (Etienne 2005). This optimization is made even more tricky because there is a long ridge of nearly equal likelihood (Etienne et al. 2006). The parameter θ is strongly correlated with species richness (and is equal to Fisher's α when $m = 0$, in which case the ZSM is the same as the log-series). $p(n)$ has a complex representation (Etienne & Olff 2004; Etienne 2005).
- **Geometric (*r*):** The first used probability distribution for SADs is the geometric, but ironically it is not actually a probability distribution but a series. The proportion of abundance in each species is a function of its rank, not a random sample from a distribution, and is different from the geometric probability distribution (May 1975). However, it is often still treated as a probability distribution. It has a single parameter, often called *k* or *r*. This represents the fraction of the remaining community captured by each successive species. Several approaches for estimation are available (He & Tang 2008) of which I find He and Tang's new method the best. $n_i = C(k)Nk(1 - k)^{i-1}$, where $C(k)$ is a

normalizing constant $C(k) = 1/[1 - (1 - k)^S]$, N is total abundance and S is species richness.

In summary two groups of distributions emerge. One group contains the Pareto distribution with extensions to create the PowerBend, Zipf–Mandelbrot and the conceptually distinct but mathematically related log-series. The second group contains three continuous functions (log-normal, γ and Weibull) and their Poisson sampling extensions (Poission–Lognormel, Poisson–Weibull, not discussed here, and Poisson Gamma = negative binomial). The ZSM and geometric stand alone. There are numerous (probably dozens) of other probability distributions that are not mentioned here (Marquet et al. 2003; McGill et al. 2007). The power fraction extensions of the geometric (Tokeshi 1996) are interesting, but currently have no analytical approaches and must be solved by repeated simulation. The log-logistic distribution (the logistic distribution on log abundances) is a little known but valid distribution and is essentially the distribution one assumes when fitting a logistic curve to the ECDF (Fig. 9.1 and section 9.2.1). Since it is primarily a descriptor of visual shape I treat it as non-parametric (Box 5.1), but it has sometimes been argued as a probability distribution outright relevant to SADs (Williamson & Gaston 2005).

9.2.4 Multivariate approaches to SADs

The final approach to dealing with the vector of numbers measured in an SAD is to treat it in its full multivariate nature. This is rarely done with a single SAD (but see, for example, Alonso 2005). With multiple SADs it becomes possible to define measures of dissimilarity distance between communities (over 70 different distance measures have been defined—Legendre & Legendre 1998; Chapter 6—and perform cluster analysis, ordination and other techniques- Clarke & Warwick 2001; Manly 2004). For an example see Figure 17.4. This approach shares with the visual approach the benefit that no data are thrown away. The use of one or a few metrics (parametric or non-parametric) as described in sections 9.2.2. and 9.2.3 involves data reduction, and the multivariate approach has the benefit over the visual approach of being quantitatively rigorous. Yet the multivariate approach has had limited uptake. This is probably largely due to the complexity and training required to correctly perform multivariate statistics. One of the major benefits and reasons for frequent use of SADs is their easy-to-grasp nature. Moreover, our intuition starts to fail in a multivariate world, for example the definition of order falls apart in a multivariate world, but the first three approaches allow for unambiguous statements like there are more rare species in data set X or data set Y is more even.

9.3 Identifying a useful, parsimonious subset of SAD metrics

If one wishes to use the visual or multivariate approaches, then the large number of metrics of SADs is not a problem. Otherwise, the proliferation of dozens of metrics to describe SADs is problematic. Indeed, for SADs with only a few (5–15) species, there are more metrics that can be calculated on it than there are raw data points. Thus there must be redundancy among the diverse metrics. There have been some attempts to identify some metrics as superior to others (Hurlbert 1971; Kempton & Taylor 1974; Smith & Wilson 1996; Wilsey et al. 2005). However, a truly systematic evaluation of both parametric and non-parametric metrics on SADs is lacking. I seek to identify a subset of all metrics that can usefully but parsimoniously describe the essential features of an SAD.

I suggest that there are four properties that a good set of metrics should have:

1. **Efficiency:** No metric will work perfectly on small samples (the shape of the SAD is meaningless with one individual and is nearly as meaningless with 10). However, good metrics should quickly converge to their true values as sample size increases and the standard error should likewise quickly decrease as sample size increases.
2. **Unbiased:** A good metric will not consistently undershot or overshoot the true value when the sample size is small. In other words, the metric

should be on average equal to the true value when many different samples are compared.

3. **Interpretability:** A good metric should be intuitive. It should describe geometric aspects of the SAD that can be easily visualized or that describe factors of the SAD which are biologically relevant or ideally both. Evenness, in its qualitative sense, is an example that has a clear geometric interpretation (close to the horizontal line in RAD or close to the step function ECDF) as well as a clear biological relevance.
4. **Orthogonality:** The first three are properties of individual metrics. Orthogonality is a property of sets of metrics. A set of metrics should contain little overlap or redundancy in what the member metrics measure. In short, the collinearity or correlation between the different metrics should be low. Thus species richness, S, and Shannon diversity, H, have a lot of redundancy, as do H and Shannon evenness, J, but S and J are moderately orthogonal (although it has been shown that even J depends on S and evenness measures other than J show greater orthogonality to S, e.g. Smith & Wilson 1996). I suggest that SAD metrics are currently in the state that landscape metrics were in before Riiters and colleagues (1995) performed their analysis. Over 50 landscape metrics had been suggested, but a factor analysis showed that all the metrics were well described by only six underlying factors.

Factors such as dependency on sample size and orthogonality can be calculated analytically for specific individual metrics or pairs of metrics (Hurlbert 1971; May 1975; Smith & Wilson 1996), but such analytic results are not generally available for most of our proposed metrics, thereby necessitating an empirical approach. To assess which subset(s) of metrics out of all parametric (Box 9.2) and non-parametric (Box 5.1) meet these goals, I performed two simple analyses.

9.3.1 Efficiency and bias

The first analysis assessed bias and efficiency. I took five empirical datasets with more than 10 000 individuals (N): Barro Colorado Island tropical trees (Condit et al. 1996), Winemiller's Venezuelan fish (Winemiller 1990), Fisher's Malaysian moth light trap data (Fisher et al. 1943), Hick's copepod data (Hicks 1980) and one site from Pearson's data on macrobenthos in Loch Linhe (Pearson 1975). Ten replicate subcommunities each of sizes $N = 30, 100, 1000$ and 10 000 were randomly sampled from the original community. All 52 metrics (32 non-parametric, 20 parametric) were calculated on each replicate subcommunity and also on the original full community. The latter was assumed to be the true value of the metric for each dataset. Estimated metrics were averaged across the 10 replicates and then calculated as a ratio to the true metric for the full community (so a value < 1 indicates smaller than the true value, > 1 indicates greater than the true value). The results are found in Figs 9.4 and 9.5 and summarized in Tables 9.1 and 9.2.

Overall, it can be seen that small sample sizes give very misleading estimates. Most metrics only began to come close to their true value (even to within $\pm 50\%$) with at least 1000 individuals sampled and in many cases only with 10 000 individuals sampled. This is problematic as a vast majority of studies of SADs are done with at most a few hundred individuals. Most metrics are also biased. Among the parametric measures, those related to the power distribution (Pareto/Power, Zipf–Mandelbrot and especially the log-series) do the best. Among the non-parametric measures, the classic diversity measures (the Shannon diversity and the closely related pair of Simpson diversity and Hurlbert Probability of Interspecific Encounter diversity measures) have nearly perfect performances converging to the true answer with only a few dozen individuals and being unbiased. The equivalent evenness measures do not fare as well, although the Shannon evenness remains one of the best evenness measures. The non-absolute dominance measures (relative dominance and McNaughton dominance), the percentage rare measures and the intercept from the logit on the ECDF (i_{logit}) all do fairly well (although to have any species that are rare at the 1% level several hundred individuals are needed). Overall the message of this analysis is a sobering one—large samples in the order of 1000 individuals are needed to truly assess species abundance distributions. With

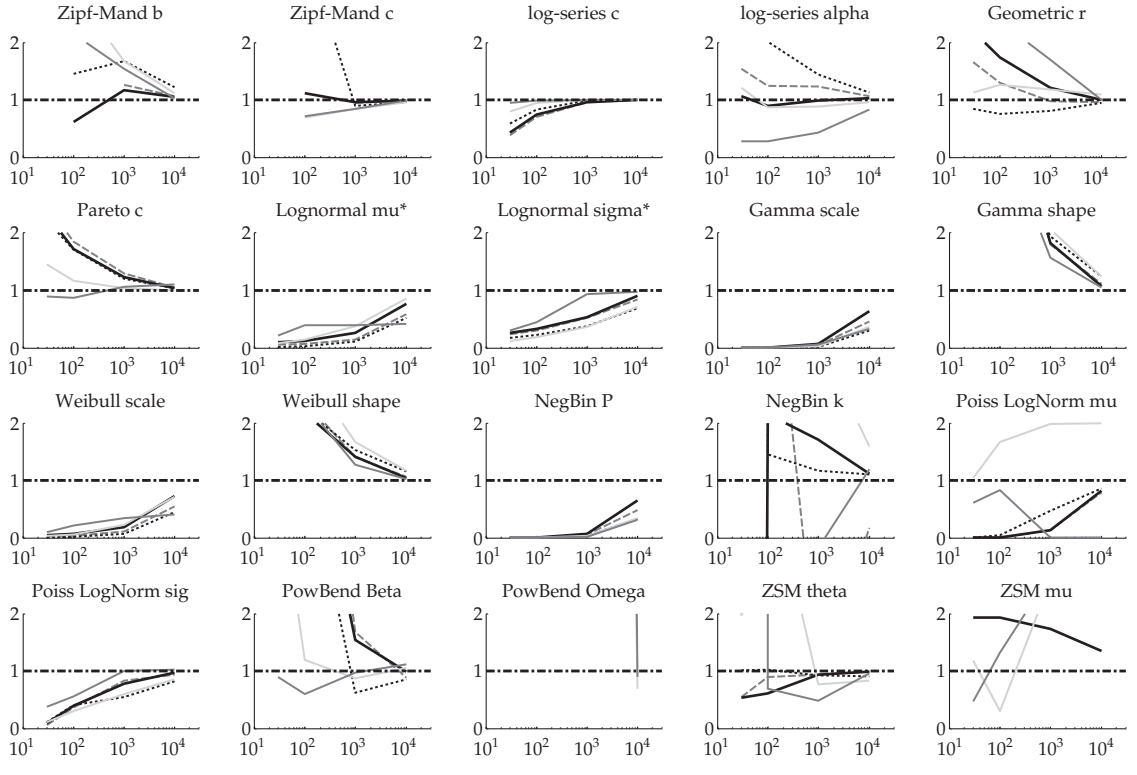

Figure 9.4 Efficiency and bias of parametric metrics. Each line represents the average across 10 replicate draws from a single dataset. The lines give the ratio to the true value, with 1 (the dashed line) representing a perfect match. Values greater than 1 represent overestimates. The *x*-axis represents sample size, *N*.

smaller samples, one is best to limit analysis to the Simpson and Shannon diversity measures, the dominance and percentage rare measures and the log-series *c* and Zipf–Mandelbrot *c* measures.

9.3.2 Independence of measures

The second analysis seeks to reduce the dimensionality and find orthogonal metrics. Datasets representing 91 communities were analysed, including 21 from Morlaix (different points in time before and after a disturbance: Dauvin 1984), the aforementioned Hicks copepod data, the aforementioned BCI data divided into 10 strips of 5 ha each, 13 communities of macrobenthos from a succession experiment (Arntz & Rumohr 1982), 30 communities of macrobenthos from Loch Linhe and Loch Eil (Pearson 1975), 20 randomly chosen routes from the North American Breeding Bird Survey with at least 400 individuals (Robbins et al. 1986; Patuxent Wildlife Research Center 2001), and the aforementioned Winemiller fish and Fisher moths. The last three datasets (22 communities) were not exhaustively sampled but instead depended on individuals that approached the counting device or observer. The first five datasets (69 communities) were all exhaustively sampled (every individual in a well-defined area was identified and counted). In addition 55 randomly generated communities (samples from a probability distribution) were created, composed of five replicates of 11 different distribution/parameter combinations (four sets of parameters on the ZSM, three sets of parameters on the log-series and four sets of parameters on the Poisson log-normal). These communities were analysed separately (exhaustively sampled, non-exhaustively sampled and random) and collectively by principal component analysis and *k*

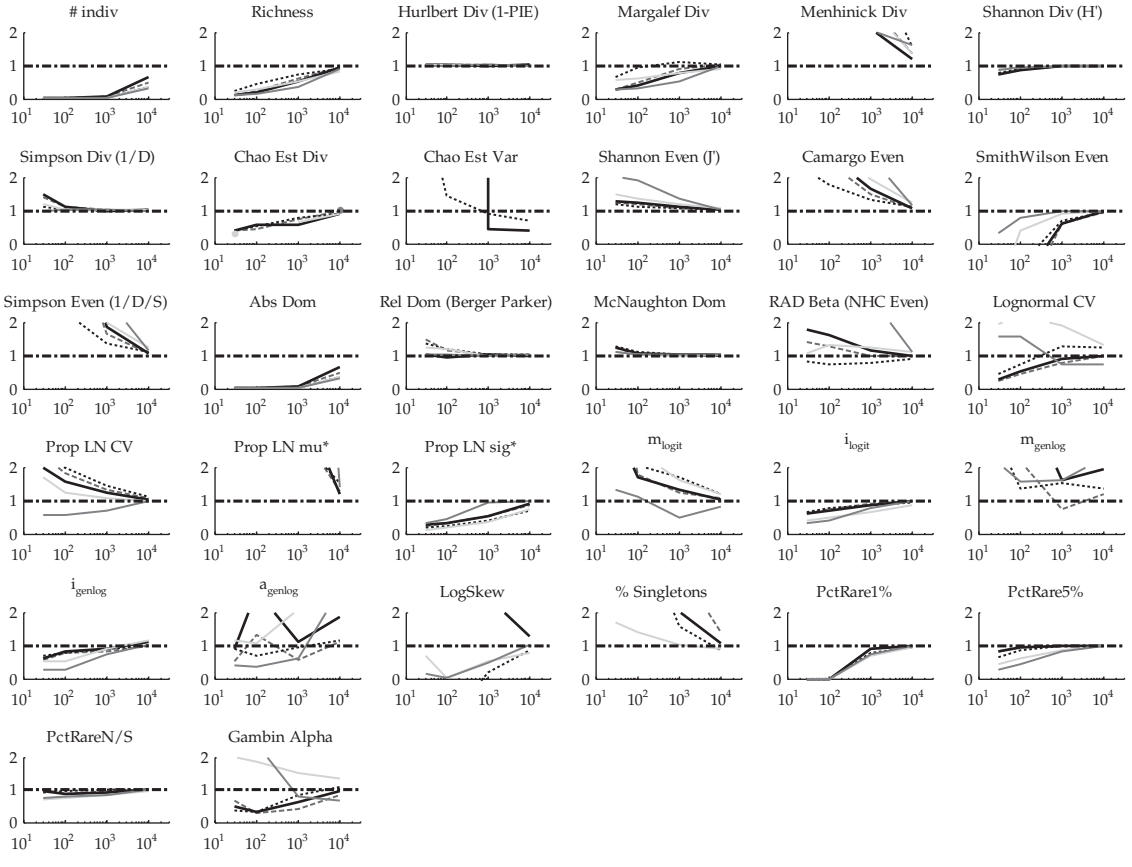

Figure 9.5 Efficiency and bias of non-parametric metrics. Interpretation as in Fig. 9.4.

means and UPGMA clustering where the 52 metrics were treated as measures and the 146 communities were treated as observations. The results are shown in Fig. 9.6 and summarized in Tables 9.1 and 9.2. Simple univariate correlation of all measures with S and N were also calculated (Table 9.3). In general, results from the non-exhaustively sampled communities were less robust than those from sampled communities. It would appear that it is highly desirable to use exhaustive sampling methods whenever possible.

Many measures (Table 9.3) were strongly correlated with S or N or both (echoing the results of the previous section on effects of sample size). However, several measures were largely independent of both S and N, indicating that there is indeed additional information contained in SADs. The metrics most independent of S and N include the slope of the logistic intercept (both the standard, m_{logit}, and the generalized logistic, m_{genlog}), the migration parameter, m, of the ZSM and log-skewness. The logistic slopes can be seen as a measure of evenness while m and skewness more specifically describe proportions of rare species.

The clustering (see Fig. 9.6) produced several useful results. First, three distinct groups of metrics emerged. These groups were robust across all subsets of data, different measures of distance and methods of clustering (hierarchical or K means). One group had measures broadly associated with species richness, including richness per se and various estimators of it as well as diversity measures and Zipf–Mandelbrot b. Another group included evenness measures, including traditional evenness measures, shape parameters of the log-normal, Weibull, gamma and negative binomial distribution, and the slope of the ECDF logit (m_{logit}) and Gambin α. The third group included

Table 9.1 Scorecard of performance for non-parametric metrics.

		Metric	Traditional symbol	Proposed Symbol	Bias (+/0/−)	Efficiency	Tie to *S* and *N*	Group	Complexity
Non-parametric	Abundance ()	Number of individuals	*N*	N/A	−−	10 000	N/A	N	**L**
	Richness (*S*)	Richness	*S*	*S*	−−	10 000	N/A	S	**L**
		Margalef diversity		$S_{Margalef}$	−	1000	VH	S	**L**
		Menhinick diversity		$S_{Menhinick}$	++	10 000	VH	S	**L**
		Chao diversity estimator		S_{Chao}	−	1000		S	**L**
		Chao diversity variance		$S_{ChaoVar}$	**0**	10 000		?	**L**
	Diversity (*D*)	Shannon diversity	H'	$D_{Shannon}$	**0**	**30**	VH	D	**L**
		Simpson diversity	$1/D$	$D_{Simposon}$	**0**	**30**	H	D	**L**
		Hurlbert diversity	$1-D$, PIE	$D_{Hurlbert}$	**0**	**30**	H	D	**L**
	Evenness (*E*)	Shannon evenness	J'	$E_{Shannon}$	+	1000	H	E	**L**
		Simpson evenness	$1/D/S$	$E_{Simpson}$	++	10 000	**L**	E	**L**
		Camargo evenness	$E_{Camargo}$	$E_{Camargo}$	++	10 000	**L**	E	M
		Smith–Wilson evenness	E_{Var}	$E_{SmithWilson}$	−	10 000	**L**	N	M
		RAD β or NHC	E_{NHC}	E_{NHC}	+	10 000	M	S	M
	Dominance or commonness (*C*)	Absolute dominance		C_{Abs}	−−	10 000	H	N	**L**
		Relative dominance		C_{Rel}	**0**	**100**	H	C	**L**
		McNaughton dominance		C_{McInt}	**0**	**100**	H	C	**L**
	Rarity (*R*)	LogSkew		$R_{LogSkew}$	**0**	10 000	**VL**	R	**L**
		% Singletons		$R_{\%Single}$	+	10 000	M	R	**L**
		Percentage rare 1%		$R_{1\%}$	−	1000	VH	R	**L**
		Percentage rare 5%		$R_{5\%}$	**0**	1000	H	S	**L**
		Percentage rare *N*/*S*		$R_{N/S}$	**0**	**100**	M	R	**L**

Bias is indicated by sign (+ indicates consistent overestimation) with 0 bias being best and in bold. Efficiency represents the approximate sample size required to get an estimate close to the true, full community value; lower values are best and in bold. Metrics that are independent of both *S* and *N* are ranked as VL (= very low), low (L), medium (M), high (H) and VH (= very high) with VL and L in bold. Group indicates which of the seven basic categories of measurement the metric is found in. Complexity indicates how difficult the metric is to calculate (L = low, M = medium, H = high). This takes into account the complexity of the formula, the computer skills required and the computational time required. PIE, RAD, NHC

Table 9.2 Scorecard of performance for semi-parametric and parametric metrics.

	Metric	Traditional symbol	Proposed symbol	Bias (+/0/−)	Efficiency	Tie to *S* and *N*	Group	Complexity
Semi-parametric	Fisher's α		α	**0**	10 000	VH	S	M
	Log-normal CV		LNCV	**0**	1000	**L**	E	**L**
	Proportional LN μ^*		%LNμ^*	++	10 000	H	E50	**L**
	Proportional LN σ^*		%LNσ^*	−	1000	H	N	**L**
	Proportional LN CV		%LNCV	**0**	1000	H	N	**L**
	Gambin α		α_{GamBin}	**0**	1000	M	E	H
	Slope of logit at inflection		m_{logit}	+	10 000	**VL**	E	L/H
	Intercept of logit at inflection		i_{logit}	**0**	1000	VH	E50	L/H
	Slope of generalized logit at inflection		m_{genlog}	+	10 000	**VL**	?	L/H
	Intercept of generalized logit at inflection		i_{genlog}	**0**	1000	H	E50	L/H
	Assymetry of generalized logit at inflection		a_{genlog}	**0**	10 000	**L**	?	L/H
	Pareto power *c*	*c*	*c*	++	1000	M	R	M
	PowerBend β	β	B	++	10 000	M	R	H
	PowerBend ω	ω	ω	++	10 000	VL	D	H
	Zipf–Mandelbrot *b*	*b*	*b*	++	10 000	H	D	H
	Zipf–Mandelbrot *c*	*c*	c_{ZM}	**0**	1,000	H	C	H
	Log-series *c*	*x*, *c*	c_{ls}	−	**100**	M	N	M
	Log-normal μ^*	μ^*, scale	M^*	−−	100 000	M	N	**L**
	Log-normal σ	σ, shape	σ	−−	10 000	H	N	**L**
Parametric	Poisson log-normal μ	Scale	μ_{PLN}	−−	10 000	**L**	E	H
	Poisson log-normal σ	Shape	σ_{PLN}	−	100 000	**L**	N	H
	γ scale	*b*	$Scale_{gam}$	−−	100 000	VH	N	**L**
	γ shape	*c*	$Shape_{gam}$	++	10 000	**L**	E	**L**
	Weibull scale	*b*	$Scale_{weib}$	−−	100 000	H	N	**L**
	Weibull shape	*c*	$Shape_{weib}$	++	10 000	M	E	**L**
	Negative binomial *P*	*P*, scale	*P*	**0**	100 000	H	N	M
	Negative binomial *k*	*k*, shape	*k*	**0**	100 000	M	E	M
	ZSM *m*	*m*	*m*	++	100 000	**VL**	R	H
	ZSM θ	θ	Θ	−	10 000	VH	S	H
	Geometric *r*	*r*, *k*	*r*	++	10 000	H	N	**L**

Same notations as Table 2. A complexity of L/H means that it can be easily assessed visually (L) but to compute an actual metric is of high (H) complexity. CV, LNCV, ZSM,

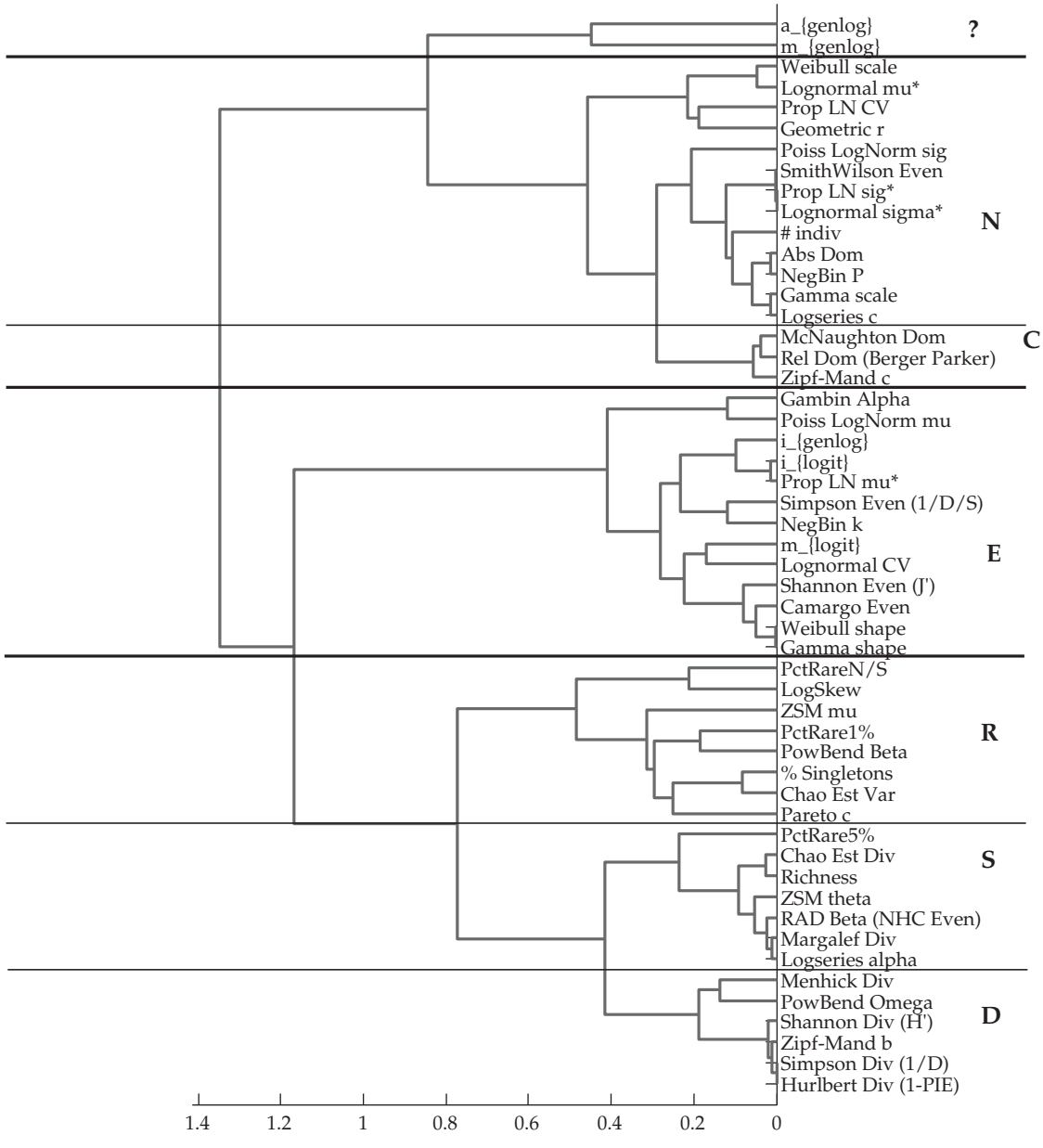

Figure 9.6 Hierarchical tree of 52 metrics (parametric and non-parametric) to describe species abundance distributions. UPGMA clustering based on Spearman Rank correlation distances shown here. Three major groups N/C, E, R/S/D emerge and are repeatable across many datasets and methods. The subgroups (N vs C and R vs S vs D) are less consistent but are useful divisions.

measures broadly related to total community size N, including N itself, scale parameters for the log-normal, Weibull, gamma and negative binomial, the standard deviations (σ) of the various log-normal distributions, the dominance measures, Zipf–Mandelbrot c and, surprisingly, Smith–Wilson evenness. The existence of these three groups was robust. The relationship between the three groups (which two were most closely related) was not robust, varying with data and methods. Similarly, several subgroups within the three groups emerged (Fig. 9.6). These subgroupings were moderately robust across data and methods, although individual measures did switch between subgroups, and the relationships between subgroups often varied. However, because of the overall intuitive sensibility of these subgroupings I am inclined to regard them as real pending further studies. Specifically, three measures of dominance (abundance of common species and here grouped under the label C) always were grouped together within the larger group N. These measures were relative dominance (Berger–Parker), McIntyre dominance and the Zipf–Mandelbrot c. The only consistent subgroup within evenness is three measures (i_{logit}, i_{genlog} and proportionate log-normal μ) that all directly measure the abundance of the 50th percentile species, which might be denoted E50. The richness group had three subgroups, one directly measuring richness and denoted S (such as the Chao estimator, ZSM θ, Margalef diversity, Fisher's α), one closely related group (sometimes intermingling in some analyses) on diversity (D), including the Shannon diversity, the Simpson and Hurlbert diversities and the Zipf–Mandelbrot b. The third subgroup was directly related to numbers of rare species (R), including log skewness, ZSM m, pareto c, PowerBend β, percentage singletons, and PctRare1% and PctRareN/S. Thus in summary, three groups were strongly supported: the N or NC group, the E group (including E50), and the S or RSD group (Fig. 9.6). The subgroups (N vs C and R vs S vs D) were usually found and are intuitive and therefore probably useful.

Principal component analysis was also performed on the same data as the cluster analysis. Results were difficult to interpret and showed variability across analysis methods. However,

Table 9.3 Correlations between different metrics and S (species richness) and N (number of individuals) in 91 communities. Dif = $|N| + |S|$ as a measure of independence of S and N.

	N	S	Dif
m_{genlogit}	−0.08	−0.16	0.24
Zero Sum Multinimial m	−0.14	0.25	0.39
m_{logit}	−0.38	0.03	0.41
PowBend Omega	−0.26	0.16	0.42
Log-Skew	0.276	0.16	0.44
Smith–Wilson Evenness	0.43	−0.08	0.51
Poisson Log-Normal μ	−0.07	−0.46	0.53
gamma shape	−0.50	0.05	0.55
Carmago Evenness	−0.54	−0.02	0.56
Poisson Log-Normal σ	0.53	−0.03	0.56
a_{genlog}	0.20	−0.38	0.58
Simpson Evenness (1/D/S) (1	−0.43	−0.15	0.58
Lognormal CV	−0.39	−0.21	0.59
% Singletons	0.05	0.56	0.60
Weibull shape	−0.54	0.08	0.62
PctRare N/S	0.51	0.13	0.64
NegBin k	−0.25	−0.39	0.64
PowBend β	0.306	0.34	0.65
RAD B (NHC Evenness)	−0.05	0.62	0.67
Lognormal μ^*	0.326	−0.39	0.72
Gambin alpha	−0.23	−0.50	0.73
Pareto c	−0.06	0.67	0.73
Log-series c	0.40	−0.35	0.74
Weibull scale	0.42	−0.38	0.80
Prop LN CV	0.13	−0.71	0.84
Lognormal σ	0.58	−0.27	0.85
Proportionate Log-normal σ^*	0.58	−0.27	0.85
Proportionate Log-normal μ^*	−0.39	−0.47	0.86
Shannon Evenness (J)	−0.63	0.23	0.86
Geometric r	0.13	−0.73	0.86
PctRare5%	0.19	0.69	0.87
Negative Binomial P	0.88	0.00	0.89
Zipf-Mandelbrot c	0.59	−0.31	0.90
i_{genlogit}	−0.62	−0.30	0.93
Absolute Dominance	0.92	−0.03	0.95
Relative Dominance (Berger	0.53	−0.43	0.96
Hurlbert Diversity (1-PIE)	−0.58	0.39	0.97
Zipf-Mandelbrot b	−0.32	0.65	0.97
Simpson Diversity (1/D)	−0.27	0.70	0.97
McNaughton Dominance	0.50	−0.48	0.99
gamma scale	0.82	−0.20	1.02
PctRare 1%	0.40	0.63	1.02
ZSM θ	−0.16	0.89	1.05
N (# of individuals)	1.00	0.05	1.05
S (Richness)	0.05	1.00	1.05
i_{logit}	−0.64	−0.43	1.07
Margalef Diversity	−0.10	0.98	1.08
Logseries α	−0.15	0.95	1.10
Shannon Diversity (H)	−0.46	0.65	1.12
Menhenick Diversity	0.40	0.75	1.15

very roughly, the first axis explained about 40% of the variance and captured large numbers of intermediate abundance species vs large numbers of rare and common species. The second axis captured about 25% of the variance and could perhaps be interpreted as how like the geometric series the data were (with one end of the axis having a steep RAD β, low richness and lots of rare species). The third axis captured about 10% of the variance and represented having high numbers of rare species vs. high numbers of common species.

9.3.3 Overall assessment of useful, parsimonious metrics of SADs

Tables 9.1 and 9.2 summarize these findings and add a column indicating how complex it is to calculate the various metrics. Weighing all of these factors (unbiased and efficient, independent and consistent in meaning, easy to calculate), some recommendations emerge. The best measures depend on your goals. Three scenarios follow.

- **Comparing equal-sized samples from closely related communities:** If the sample size is constant and the communities are directly comparable (e.g. herbaceous plants from a series of 10 × 10 m quadrats) then it is meaningful to directly compare S and N (see Chapter 4). With these simple measures available, it would appear from the three clusters identified that a measure of evenness would add the most and the Shannon evenness or m_{logit} (the latter especially as sample sizes get into the hundreds) would appear best. It is not clear that the Smith–Wilson evenness truly captures evenness since it is clustered in group N and the Camargo and Simpson evenness measures have poor convergence behaviour. One can even imagine doing a simplistic three-dimensional ordination by plotting a scatter plot in the three-dimensionsl space of S, N and E. If additional information is desired, a dominance (C) measure (such as relative or McNaughton dominance) and/or a rarity (R) measure such as log skewness (with more independence from S) or PctRare N/S (with good small sample behaviour) would also make sense. ZSM m works well but requires very large samples and very advanced computer skills.
- **Comparing unequal samples or samples from unrelated communities:** When comparing data from, say, the tropics and temperate zone or data with different sample sizes, direct comparison of S and N is not meaningful. Gotelli and Colwell (Chapter 4) present methods to make S more comparable. However, it may become desirable to use other measures from the S and N groups. Margalef's diversity or Fisher's α (the latter especially once sample sizes reach into the hundreds) are good, quickly converging representatives of the S group. Representing the N group, the log-series c or the Smith–Wilson evenness have good sampling behaviours but the latter can be confused with an evenness measure and should probably be avoided. The remaining recommendations would be the same as the previous section (Shannon evenness or m_{logit} to capture evenness, E, and then, if desired, a dominance measure, C, and a rarity measure, R).
- **Very small sample sizes ($<$ 100s):** When less than a few hundred individuals are sampled, the gross inaccuracy of most metrics at representing the larger community must dominate goals of independence. In this case, the log-series c, the Hurlbert/Simpson and Shennon diversity, relative or McNaughton dominance, and PctRare5% and PctRareN/S and possibly Shannon evenness (less rapidly convergent than the others, but the best of the E cluster) make the most sense.

9.4 Prospectus

There can be little doubt that sophisticated multivariate analyses are perhaps the most rigorous way of analyzing abundance data. However, the long-standing popularity of simpler (visual and metric-based) approaches to SADs suggests that not everybody can or will make the switch, therefore it is important to pursue a more full understanding of the simpler approaches as well. In this pursuit, a balance is needed between adding new measures that might be better and keeping the number of measures under control. Ecology has tended to see new methods added without old methods decisively removed. Science does not progress this

way (Platt 1964). I have attempted here to suggest some new approaches (e.g. m_{logit} and direct measures of rare species by metrics like PctRare5%) as well as to perform some pruning. The results are fairly messy. More efforts are needed in both directions.

In addition, the results have suggested some tantalizing patterns worth pursuing that might ultimately lead us to a more mechanistic analysis. The appearance that Smith–Wilson evenness (called E_{var} by them) does not align with other evenness measures is a surprise and requires more analysis. The fact that the Zipf–Mandelbrot distribution converges quickly and the two parameters are independent suggests that this probability model deserves more usage and understanding. The seeming links between richness (S) and rarity (R) and between abundance/sample size (N) and dominance (C) are provocative. These may prove to be mere artefacts of the limited number of ways one can pack individuals into species, but it is not immediately obvious that they are. Finally, although the results of principal component analysis were very weak, the singling out of the importance of the number of intermediate abundance species is suggestive. Gray noted that intermediate abundant species were the ones that responded most to environmental change (Gray 1979), and Kempton (1979) suggested that metrics focused on species of intermediate abundance had the greatest ability to discriminate between sites. As yet, few metrics have been targeted at this area.

9.5 Key points

1. Species abundance distributions have been widely used because they are easy to measure and have simple methods of analysis (visual and various metrics).
2. I suggest that using a rescaled ECDF (empirical cumulative distribution function) is the optimal way to plot SADs.
3. Most SAD metrics are very inefficient and usually biased. Very large sample sizes are needed. In general, field workers should target sampling 1000 individuals and if this goal is not met, only a few metrics are appropriate.
4. The metrics proposed to date seem to span three broad features, richness/diversity, evenness, and abundance. These three broad groups have some reasonably consistent subgroups, suggesting a total of seven different groupings of metrics: richness (S), diversity (D), rarity (R), evenness (E), with a subgroup focused on median abundance (E50), and abundance (N), with a subgroup focused on relative dominance of common species (C).
5. Three specific scenarios with regards to different types of data are described. All of these scenarios can be analysed visually using the ECDF. When using metrics, different sets of metrics are recommended for the three different scenarios.

Acknowledgements

I thank the working group on SADs held at National Center for Ecological Analysis and Synthesis for many discussions and ideas that clarified my thinking about SADs and to NCEAS for funding this working group. I dedicate this chapter to John Gray, who was one of the leaders of this working group but passed away between our second and third meetings. John was one of those rare individuals who readily spanned the empirical and the theoretical divide in ecology and thought strategically. His 1979 application of SADs to detect the effects of pollution on communities and his 1987 review of SADs were both way ahead of their time. John's personality was as rare as his mind—he was in science purely for the intellectual curiosity and got excited about ideas wherever they came from, even if it was from a lowly graduate student (as I was when I first began interacting with John) or if it contradicted his own ideas. His enthusiasm and kindness are sorely missed.