CHAPTER 5

Measurement of species diversity

Brian A. Maurer and Brian J. McGill

5.1 Introduction

One of the most conspicuous aspects of biodiversity is the fact that individual organisms are organized into relatively discrete units referred to as species. Although there is much variability in what can be called a species, generally a species represents a distinct genetic lineage of organisms that interact with the environment in similar ways and are generally reproductively compatible. For both theoretical and practical reasons, it is often desirable to know how many species are found in a given region of space-time (Chapter 4) and to know something about how abundant each species is relative to others in the same community (Chapter 9). This relatively simple objective, however, becomes greatly complicated because there are so many different ecological circumstances in which species diversity is measured. Because of this there have been a large number of quantities suggested as appropriate measures of species diversity (Box 5.1) (Pielou 1975; Krebs 1989; Magurran 2004). This plethora of indices makes it difficult to evaluate which method is appropriate in what particular circumstances. The most commonly used indices are used primarily because they have been used before, and not necessarily because they provide useful information.

In this chapter we consider the problem of species diversity by focusing firstly on precise definitions of the term. We then describe the necessary statistical sampling theory that follows from the definition. There are two different types of sampling issues, both of which are important to developing an understanding of species diversity. First, there is the issue of how ecological circumstances act as a probabilistic 'filter' in determining which specific species can be found in a region. We will refer to this as the *ecological sample* in what follows. Second, when collecting data on species abundances within a specified ecosystem, it is often not appropriate to assume that every last individual of every species has been identified. Hence, much field data are subsamples of a larger, unknown community. We will call this the *empirical sample* below. Empirical samples are often used to estimate quantities assumed to represent the unmeasured parameters that describe the process thought to have given rise to the ecological sample (Green & Plotkin 2007).

No ecosystem remains unchanged across space and time. However, there may be ecological conditions that are sufficiently similar that they may give rise to similar levels of diversity. How is it possible to determine whether multiple communities arrayed across space and/or time give rise to an increase in diversity? In other words, given a set of communities that can be sensibly aggregated into a larger entity, is it possible to partition the overall diversity of the aggregate into a component due to within-community diversity and a component attributable to between-community diversity? The former has been termed 'α diversity' and the latter 'β diversity' (Whittaker 1975; Chapter 6). Estimating these diversity components in communities that result from some combination of deterministic and random processes provides a unique statistical challenge.

5.2 State of the art

First it is necessary to define the underlying structure of the statistical population from which the ecological sample of a community is derived. To do this, we start with a definition of a community for which a measure of species diversity is desired.

Box 5.1 Measures of species diversity and eveness

There are a number of descriptors that are based on some concept such as evenness, diversity, or dominance that are not derived from any probability distribution. Of course many of the parameters from probability distributions could fit these goals as well (Box 9.2). The notation used throughout this section is: N_i is the abundance of the ith species after sorting (so N_1 is the abundance of the most abundant species), S is the number of species observed, and N is the total abundance ($N = \sum_{i=1}^{S} N_i$) and p_i is the proportion of abundance for species i($p_i = N_i/N$). S_i is the number of species with an abundance i, so S_1 is the number of singletons. To provide some order, we have tried to group indices with similar goals and to use a consistent notation.

Number of individuals (N) – Total number of individuals. This is another easily calculated yet powerful descriptor. Note that in neutral theory N is often denoted by J, but we use the more traditional N here.

I. Richness metrics (S)

Richness (S) – Species richness: the total number of species identified in the sample. It is among the simplest descriptors of community structure.

Margalef diversity ($S_{Margalef}$) – Margalef (Clifford & Stephenson 1975) noted that species richness increases with N, and in particular increases non-linearly and roughly logarithmically with N. $S_{Margalef} = (S - 1)/\ln N$

Menhinick diversity ($S_{Menhinick}$) – In a similar vein (Clifford & Stephenson 1975), Menhinick proposed adjusting species richness by the similarly shaped square root of N. $S_{Menhinick} = S/\sqrt{N}$

Chao estimated diversity (S_{Chao}) – Another way to make species richness S comparable between sites with different sample sizes N is to extrapolate to the richness of an infinite sample. Chao (1987) proposed a simple robust estimator for this: $S_{Chao} = S + S_1^2/(2S_2)$.

Chao estimated variance – Although not an estimate of species richness, Chao provided an analytical formula for the variance in S_{Chao} that can be used to place error bars on $S_{Chao}(S_{Chao} \pm 1.96\sqrt{S_{ChaoVar}})$ and is given by $S_{ChaoVar} = S_2[(S_1/S_2)^4/4 + (S_1/S_2)^3 + (S_1/S_2)^2/2]$.

II. Diversity metrics (D)

Diversity is traditionally taken to be a function of both richness and evenness, with less even communities being less diverse than their richness alone would indicate (should a species with only one individual count towards diversity the same as an abundant species?).

Shannon diversity (H' or $D_{Shannon}$) – Shannon's information theory can be used to calculate the information in a community as an estimate of diversity. There is a finite population size version known as Brillouin's index (equation 5.9, main text) which should probably be used but usually isn't. $D_{Shannon} = -\sum p_i \ln p_i$

Simpson diversity (1/D or $D_{Simpson}$) – Simpson noted that $D = \sum p_i^2$ gave the probability that two individuals drawn at random from an infinite community would belong to the same species. This has precedent in population genetics of the probability of getting two alleles the same. As such D is the inverse of diversity and some form of inverse is needed to create a diversity index. Although variations including $1 - D$ (related to the variance within species and to Hurlbert's PIE below) and $-ln(D)$ (related to the Hill measure H_2) have been used, the most common way (e.g. MacArthur 1972) of converting homogeneity into diversity is $D_{Simpson} = 1/D$.

Hurlbert diversity (1 - PIE or $D_{Hurlbert}$) – Hurlbert (1971) argued that a biologically meaningful measure of diversity is the odds that a given interaction between two species is interspecific (PIE = $1 - D$ = probability of interspecific encounter). With a correction for finite sample size we have $D_{Hurlbert} = 1 - \sum(n_i/N)[(n_i - 1)/(N - 1)]$.

Diversity numbers ($D_{Hill,\alpha}$) – Hill (1973) proposed using information-based criteria to obtain 'weighted' counts of species, based on the degree of dominance. Also see Chapter 6. The weighted counts, H_α, are obtained by choosing an appropriate value of α, with small α weighting rare species most and large α weighting common species most. Note that the Hill numbers are the exponential of the Renyi entropies given in equation 5.6: $H_\alpha = \exp(R_\alpha)$. This set of measures contains many common measures as special cases. $H_{-\infty} = 1/p_S$ (reciprocal of the proportional abundance of the rarest species), $H_0 = S$, $H_1 = \exp(H')$, which some people have argued should be used instead of H' (where H' is the Shannon diversity or $D_{Shannon}$ above), $H_2 = 1/D$(i.e. $D_{Simpson}$ above), and $H_\infty = 1/C_{Rel}$) (reciprocal of the Berger–Parker index). Kempton (1979) found that α between 0 and 0.5 provided the best discrimination in empirical data known to come from different communities.

$$H_\alpha = \left[\sum p_i^{\alpha}\right]^{1/(1-\alpha)}$$

III. Evenness metrics (E)

Evenness is a measure of how different the abundances of the species in a community are from each other (Smith & Wilson 1996). A community where every species had the same abundance would be perfectly even. All natural communities are highly uneven, so evenness is a relative statement. Most evenness indices are scaled to approximately run from 0 = maximally uneven to 1 = perfectly even.

Shannon evenness (J' or $E_{Shannon}$) – If diversity is a mixture of richness and evenness, then removing richness should produce evenness. This is the logic behind Shannon's evenness measure; the highest value of $D_{Shannon}$ when all species are equally abundant can readily be seen to be $\ln(S)$ so dividing by $\ln(S)$ will give an index from 0 to 1. $E_{Shannon} = D_{Shannon}/\ln(S)$

Simpson evenness ($1/D/S$ or $E_{Simpson}$) – The same logic applies to Simpson's diversity, giving $E_{Simpson} = D_{Simpson}/S$.

Camargo evenness (or $E_{Camargo}$) – The highest possible evenness is when $p_i = p_j = 1/S$, so Camargo et al. (1993) suggested a direct measurement of deviation from this ideal. $E_{Camargo} = 1 - \sum |p_i - p_j|/S$, where the sum is taken over $i = 1 \ldots S, j = i + 1 \ldots S$.

Smith Wilson evenness (E_{var} or $E_{SmithWilson}$) – Smith and Wilson (Smith & Wilson 1996) reviewed an array of evenness indices and assessed them on some core properties. Two core properties they identified are spanning the whole range 0–1 and being independent of unit of measure (evenness of biomass measured in grams should be equal to evenness of biomass measured in kilograms). They invented an index that performed well on these goals, which they called E_{var}. The formula is based on the variance of log abundances (centered on the mean of log abundances) then appropriately scaled to cover 0–1. $E_{SmithWilson} = 1 - \frac{2}{\pi} \arctan\left[\frac{1}{S}\sum(\ln(n_i) - \mu_{\ln})^2\right]$ where $\mu_{\ln} = \frac{1}{S}\sum\ln(n_i)$.

RAD beta or NHC evenness (E_{NHC}) – The slope of the rank abundance diagram (RAD) has long been interpreted as a measure of evenness (a perfectly horizontal line would represent perfect evenness). Nee, Harvey, and Cotgreave (Nee et al. 1992) proposed taking the slope of the regression line through the points in the RAD as a measure of evenness. This runs from $(-\infty,0)$. Some authors (e.g. Smith & Wilson 1996) rescale this to go from 0 to 1 –i.e. $-2/\arctan(\beta)$, but we prefer to keep the simple geometric interpretation β, where β is the OLS slope of log abundance vs. rescaled rank (divide by S so rank goes from $1/S$ to 1). $E_{NHC} = \beta$ as above.

Diversity number ratios (E_{Hill}) – Using diversity numbers H_α (Hill 1973), one can take ratios of different diversity numbers $E_{\alpha,\beta} = H_\alpha/H_\beta$ to express the degree of evenness among species within an ecological sample. The most natural case is $\beta = 0$, then $H_0 = S$, which is the maximal value (i.e. on a perfectly even community with $p_i = N/S$) of H_α for all α. For example, $\log(H_2)/\log(H_0)$ gives Shannon evenness.

IV. Dominance or common species metrics (C)

Dominance is a measure of how much one or a few species dominate the community numerically (McNaughton & Wolf 1970). In some ways it is the inverse of evenness, but it is specifically focused on the right side of the SAD (very common species).

Absolute dominance (C_{Abs}) – The simplest measure of dominance is simply N_1, the abundance of the most abundant species. Although it might seem that N_1 would be so heavily dependent on total abundance N as to be useless, in some systems N_1 can stay surprisingly constant even while N varies. $C_{Abs} = N_1$

Relative dominance (Berger–Parker) (C_{Rel}) – The easiest way to correct for the effects of N is to divide by N, producing p_1 the relative abundance of the most abundant species (Berger & Parker 1970). $C_{Rel} = p_1$

McNaughton dominance ($C_{McNaught}$) – McNaughton (1970) made a more robust measure that was less subject to the vagaries of a single species by looking at the proportional abundance of the two most abundant species (and rescaling to 0–100). A similar index based on the abundance of the three most abundant species was made by Misra and Misra (1981).
$C_{McNaught} = (p_1 + p_2) \times 100 = [(N_1 + N_2)/2N] \times 100$

V. High rarity metrics (R)

Rarity – the opposite of dominance metrics – focus on an assessment of rare species. Since abundance is bounded at 1, rarity metrics focus on the number of species with specific abundances in contrast to commonness metrics which focus on the abundance of specific species.

LogSkew ($R_{LogSkew}$) – Skew is the third moment of a probability distribution, measuring asymmetry. Right skew (positive numbers) indicates more probability on the right (abundant) side. Left skew (negative numbers) indicates more probability on the left side. All species abundance distributions are strongly right skewed on an arithmetic

Continued

Box 5.1 (*Continued*)

scale, so the more interesting measure is skew on the log scale. This measures asymmetry relative to the log-normal. A negative number indicates an excess of rare species (McGill 2003). $R_{\text{LogSkew}} = \left[\sum(\log(n_i) - \mu)^3/S\right] \Big/ \left[\sum(\log(n_i) - \mu)^2/S\right]^{3/2} S/(S-2)\sqrt{[(S-1)/S]}$, where μ is the mean of $\log(n_i)$.

% Singletons ($R_{\text{Singleton}}$) – A simple measure of rare species is to count the number of singletons. $R_{\text{Singleton}} = S_1$

PctRare1% ($R_{1\%}$) – Like the first two dominance measures, $R_{\text{Singleton}}$ focuses exclusively on one abundance class, potentially making it a noisy metric. By focusing on multiple abundance classes, this problem can be avoided. The challenge is to define which species count as rare. A simple one is to call any species with an abundance less than 1% of total abundance rare. A major shortcoming of this is that no species can have an abundance less than 1% of N if $N < 100$. This measure is only useful when N is at least several hundred. $R_{1\%} = (S_1 + S_2 + \ldots + S_T)/S$, where T is the largest integer less than $0.01 \times N$.

PctRare5% ($R_{5\%}$) – An alternative, more expansive, definition that can work with $N > 20$ is that a species is rare if its abundance is $< 5\%$ of N.
$R_{5\%} = (S_1 + S_2 + \ldots + S_T)/S$, where T is the largest integer less than $0.05 \times N$

PctRareN/S ($R_{N/S}$) – Both the 5% and 1% cut-offs are relative to N only, not S. A thousand individuals with 500 species is bound to have many rare species by these definitions but 1000 individuals with 10 species will be very different. A simple measure that attempts to correct for this is to count a species as rare if its abundance is $< N/S$. N/S gives the average abundance of a species. Because of the strong right skew, the average abundance will be greater than the median (50th percentile) abundance and thus quite large. $R_{N/S} = (S_1 + S_2 + \ldots + S_T)/S$, where T is the largest integer less than N/S

VI. Semi-parametric metrics

Several metrics are not parameters from the probability distributions listed in Box 9.1, but are related to those or other probability distributions.

Fisher's α (α or S_{Fisher}) – The log-series distribution parameter is called c, but as discussed in Box 9.2 for the log-series, α can be calculated as a function of N and c by $\alpha = N(1 - c)/c$. Although typlcally smaller (sometimes by a factor of 2–10) than S, it is strongly correlated with S. It has been recommended as a sample-size-independent estimator of richness (Rosenzweig 1995).

Lognormal CV – As discussed in the section on the lognormal distribution (Box 9.2) (also see Limpert et al. 2001), the coefficient of variation $CV = \mu/\sigma$ is not a parameter but is perhaps the best single descriptor of shape for the log-normal (making it analogous to the gamma and Weibull shape parameters). A high CV indicates many rare species (high unevenness).

Prop LN μ^*, Prop LN σ^*, Prop LN CV – In the log-normal distribution, both μ^* and σ^* are in units of abundance and scale with increasing sample size N (i.e. $\mu = \log(\mu^*)$ and $\sigma = \log(\sigma^*)$ scale as $\log(N)$). One way to adjust for this is to calculate the log-normal parameters (mean, standard deviation, CV) on the log of the relative abundances, $\log(p_i)$. This removes the heavy dependence on N.

Gambin α – The Poisson-gamma distribution (which leads to the logseries distribution) is discussed in Box 9.2. An alternative sampling distribution to the Poisson is the binomial (Green & Plotkin 2007), where the gamma distribution gives the probability p of the species appearing, which is then passed through binomial sampling. This gives the binomial gamma, which might be a good model for SADs (Ugland et al. 2007). Since the gamma distribution runs to ∞, it is necessary to truncate the right tail (say at the 99th percentile). By scaling without loss of generality so that the maximum value is 1, the binomial gamma or GamBin (Ugland et al. 2007) is well defined with only one parameter, the gamma shape parameter. It has been shown that the GamBin fits many datasets well and that the parameter α may be a good proxy for the habitat complexity from which the community is sampled (Ugland et al. 2007).

m_{logit} and i_{logit} – As discussed in Chapter 9, fitting a sigmoidal logistic function to the empirical cumulative distribution function (ECDF) on a log proportional abundance scale is tantamount to hypothesizing a log–logit probability distribution (Evans et al. 1993; Williamson & Gaston 2005). The logistic function (and log–logit probability distribution) has two parameters: i is a scale parameter and gives the location of the inflection point on the x-axis. In this context, since the inflection occurs at 50% of species accumulated, it gives the median relative abundance on a log scale. Since there are many more rare species than common, this is tantamount to a form of measurement of how many rare species there are. Similarly, m represents the slope of the function at the intercept, and as such is a proxy for evenness. When $m = 0$ every species has different abundance and when $m = \infty$ the logit function becomes a step function and all species are

equally abundant. $ECDF(p) = 1/\{1 + \exp[-m \times (p - i)]\}$, which can be fitted to the ECDF by least squares.

m_{genlog}, i_{genlog}, and a_{genlog} – The logit function assumes that the ECDF is symmetric about the inflection point, which it may not be. A generalized logit with three parameters can allow asymmetry. By the appropriate choice of a three-parameter sigmoidal logit-like function, the parameters m and i can retain their meanings and a single parameter a can describe the degree of asymmetry. $ECDF(p) = 1/\{1 + a \times \exp[-m \times (p - i)]^{1/a}\}$, which can be fitted to the ECDF by least squares.

Assume that a community can be described by a set of J individual 'sites' occupied by a single individual from one of S species. For some species, an individual may not be a discrete unit, but may exist as a distributed network of 'nodes', such as a tree species that generates above-ground stems from a network of roots. In such situations, it is important to define precisely what part of the network is being counted (e.g. stems). To model this situation, let the random variable X_{ij} be defined as follows: $X_{ij} = 1$ if site j is occupied by species i and $X_{ij} = 0$ otherwise. Since each site can only be occupied by a single individual, it is convenient to combine the random variables X_{ij} across each species at site j as a single random vector $\mathbf{X}_j$. The limiting distribution of $\mathbf{X}_j$ can be generally considered to be a multinomial distribution with a single observation. Given this general structure, we can calculate the expected value of the random vector $\mathbf{X}_j$ over species as a vector of probabilities $\mathbf{q}_j = [q_{ij}]$, where q_{ij} is the probability that species i is found on site j. Collecting the probability vectors from all sites into a single matrix gives the $S \times J$ matrix $\mathbf{Q}$, which summarizes the probabilities of species distributed across all sites. The expected abundance (N_i) of each species is obtained as

$$N_i = \Sigma_j q_{ij} \tag{5.1}$$

The vector $\mathbf{N} = [N_1\ N_2 \ldots N_S]$ is sometimes called the species abundance distribution (see Chapter 9) and has been a major focus of research in ecology for the past several decades. Because of the underlying probabilistic definition of $\mathbf{N}$, it is convenient to consider it as a random vector. An active area of current investigation is the examination of alternative stochastic processes that can model the evolution of $\mathbf{N}$ in space-time (Alonso & McKane 2004; Etienne 2005). Consideration of these models is beyond the scope of this chapter; suffice it to say that the distribution of $\mathbf{N}$ that is most useful is the equilibrium distribution of the underlying stochastic process generating the ecological sample. Here we will use the multinomial distribution to represent this limiting distribution. Finally, it is also useful to consider the *relative* species abundance distribution to be represented by the random vector

$$\mathbf{p} = \mathbf{N}/N \tag{5.2}$$

where $N = \Sigma_i N_i$.

5.2.1 Species diversity as variance

We are now in a position to consider definitions of species diversity. Diversity is often intended to represent two different aspects of the species abundance distribution. The first is termed 'species richness' or, simply, the number of species in the ecological sample. For reasons that will become evident shortly, the number of species in the ecological sample may not be equal to S (the number of species that have nonzero probabilities of being found in at least one site in the community). In other words, it is possible that n_i might equal zero for some species in a particular ecological sample, even though the species could possibly be found in the community. Let S_e be the number of species with non-zero abundances in a single ecological sample, then obviously $S_e \leq S$ is the species richness of the sample. The other component of species diversity is the degree to which the relative abundances are similar among species. This has been called 'evenness' in the ecological literature, but in actuality the underlying concept of interest is the covariance in relative abundances among species.

There are two sources of variation in relative abundances among species. The first is the variation within a species and the second is the

variability among species. The variance within a species is

$$\text{var}(p_i) = p_i(1 - p_i) \tag{5.3}$$

where p_i is the ith element of the vector of relative abundances (**p**). Since relative abundances of species are necessarily correlated, a measure of between species variability is the covariance of the relative frequencies of species i and k, that is

$$\text{cov}(p_i, p_k) = -p_i p_k \tag{5.4}$$

The total variance across all species is obtained by summing equations (5.3) and (5.4) across all species and pairs of species, which gives

$$V = 1 - \Sigma_i p_i^2 - 2\Sigma_{i<k} p_i p_k \tag{5.5}$$

In the two extreme cases when all species are equally abundant or when all species have zero abundance except for one, V reaches its minimum value of zero (no variability among species). Thus, this variance itself is probably not useful as a measure of species diversity because of this. However, equation (5.5) is useful heuristically because it illustrates how the two separate aspects of species diversity, richness and evenness, might be related.

The two terms in equation (5.5) represent different aspects of species diversity. The quantity $D = \Sigma_i p_i^2$ is the familiar metric of species diversity first suggested by Simpson (1949). It has often been called a measure of 'dominance'. In equation (5.5), $1 - D(= 1 - \Sigma p^2)$ represents the total variance attributable to within-species variability. D is well known to be correlated with species richness, and in the context used here it could be considered to be a probabilistic measure of richness. As a measure of species diversity, however, it is incomplete because it does not include information about the variability in relative abundances among species given by the last term in equation 5.5. The last term in equation 5.5 is particularly interesting because it is a measure of the degree to which abundances covary among species. Intuitively, the summed covariances in relative abundances among species capture the essence of the 'evenness' component of species diversity.

To summarize, if species diversity is defined as the number and relative abundances of species within a community, one way to represent it is to partition the total variation in species abundances among species into a within-species component and a between-species component. The within-species component represents the richness aspect of species diversity and the between-species component represents the evenness aspect.

5.2.2 Species diversity as information

In the previous section, species diversity was defined by partitioning the total variance of abundances across all species in the ecological sample. An alternative approach developed in the early 1960s was based on measuring the information content of a long string of symbols developed by information theorists (Pielou 1975 and references therein). The basic idea of the analogy is to view an ecological sample of species as a 'message' with individual organisms as pieces of 'information'. The relevant information is the taxon to which each organism belongs, and the measurement of this 'taxonomic information' is obtained from the relative abundances of species. A general measure of information content per symbol in an infinitely large set of symbols for which some 'code' exists is given by

$$R_\alpha = \text{In}\left[\Sigma_i p_i^\alpha\right]/(1 - \alpha) \tag{5.6}$$

where α is an arbitrary integer (Hill 1973; Pielou 1975). These are called the Renyi entropies of order α. Different values for α produce different weightings of the information content inherent in the relative abundances of species with low values of α (< 0) weighting in favour of rare species (in the limit $R_{-\infty}$ is a function only of the p_i of the rarest species) and high values of α emphasizing weighting in favour of common species (again R_∞ being a function only of the p_i for the most common species). Three values of α are of particular interest to ecologists. The first, R_0, is simply the logarithm of the number of species in the ecological sample. When $\alpha = 2$, equation (5.6) yields $R_2 = -\log D$, that is, the negative logarithm of Simpson's diversity measure. The final value of interest is the limit of equation (5.6) when α approaches 1, which yields

$$H' = R_1 = -\Sigma_i p_i \log p_i \tag{5.7}$$

This is the well-known Shannon measure of species diversity (Pielou 1975), which is widely used. Hill numbers, which are simply the exponent of Renyi entropy ($H_\alpha = \exp(R_\alpha)$), are also commonly used in ecology (Box 5.1 and Hill, 1973) and Chapter 6, where Hill number H_a is denoted ${}^q D$.

The appeal of equation (5.6) as a measure of diversity is that it is convenient to define evenness quantitatively. Intuitively, evenness should be greatest if all species are equally common. If this is the case, equation (5.6) yields a value of log S_e regardless of what particular value α takes on. Hence, it is sometimes useful to rescale R_a by dividing it by its theoretical maximum, yielding

$$R_\alpha^* = R_a / \log S_e \qquad (5.8)$$

R_1^* (also denoted J') has been widely used as measure of evenness in the ecological literature.

Pielou (1975) points out that if there is a finite number of individuals in an ecological sample, then the information content per species for that particular sample is

$$H_B = (1/N)\, log\, [(N!)/\Pi_i\, N_i!] \qquad (5.9)$$

This form of information diversity is related to $H' = R_1$ because as the values of the abundances of species become very large, H_B converges on the Shannon diversity, $H' = R_1$ (Pielou 1975). However, for small collections of individuals (i.e. small, fully censused communities), equation (5.9) is the appropriate measure of information. It has become known as the Brillouin index and takes on values slightly less than Shannon values ($H_B < H = R_1$).

5.2.3 Traditional measures of various types of diversity

Given that species diversity has at least two different general formulations, that is, it can represent a partitioning of abundance in a community into between- and within-species variance components or it can represent shared or mutual information among species, a large number of measurements have been suggested to represent these different aspects of species diversity. Here we survey these measurements and indicate how they relate to our distinction between species diversity as variance and species diversity as information.

We have classified species diversity measurements into six categories based on how researchers have proposed to use them (Box 5.1). The first set of metrics contains those that attempt to express some basic aspect of 'richness'. The idea behind these metrics is to express some aspect of the number of species in the ecological sample. Some, such as Chao's estimators (Chao 1987), attempt to use information from an empirical sample to infer the species richness of the underlying ecological sample.

A second set of metrics used widely in the literature we refer to as 'diversity' metrics and include the most widely used metrics. In this context, diversity is used to mean a combination of both richness and evenness. Simpson's diversity metric is the within-species component of variance discussed above, but also is related to the information concept of diversity. Shannon's diversity metric (H' or R_1) is probably the most commonly used expression of species diversity.

Evenness metrics all attempt to examine how abundance is apportioned among species within a community. The basic concept underlying all of these measurements is that evenness is highest when a community is not dominated by a few species of very high abundance or equivalently that all species have an equal abundance. Low evenness implies that most species in the community are very rare, and consequently may contribute very little to the underlying ecological role the community plays within the ecosystem that contains it.

Dominance metrics are in many ways the converse of evenness. If the scientific objectives of a study focus on the most common species in a community, then dominance measures may be the most appropriate descriptors of species diversity. Likewise, in some studies it may be more important to focus on the rarest species. This might be particularly true in conservation studies where rare species may be of particular interest in determining the value of locations for the conservation of biological diversity. For such studies there are a variety of metrics that focus on the number of rare species found in a community.

Finally, a variety of metrics are based on various parametric or non-parametric descriptions of the probability distribution underlying the

apportionment of diversity among species. Some of these metrics are related to probability distributions that arise from assuming a certain type of mechanism underlying the dynamics of abundances among species over time (see Box 9.2 and Chapter 9). Generally, these metrics should be used to fit parameters in specific models that might underlie the structure of a community.

5.2.4 Addressing the difference between the empirical and ecological samples: estimating species diversity components using empirical samples

A fundamental difficulty that has not been addressed up to this point is the nature of the 'object' being measured when ecologists collect information on the abundances of species at a specific location. This 'object' is typically called a 'community', and is defined as all the organisms belonging to a set of species found at a given point in space and time. The existence of such an object in any real sense might be questioned on many grounds (Maurer 1999; Ricklefs 2008), yet there is enough accumulated evidence to suggest that counting organisms of different kinds of species in local regions of space-time is of enormous practical value (Chapters 17,18, and 20). Here we focus on how to analyse and interpret the data obtained from such counts given that not all individuals can be counted and some species that have appreciable populations in the region may in fact not show up in these empirically derived counts. In practice, nearly all data obtained by ecologists form an empirical sample rather than an ecological sample. It is therefore technically incorrect to calculate a diversity measure (previous section and Box 5.1) on an empirical sample and claim it is the correct value for the ecological sample.

Estimating the number of species, S_e, in an ecological sample from empirical samples assumes that in a local community conditions remain constant enough over the sample period to assume that there are no changes in the relative abundances and incidence of species. If this is the case, then one can assume a 'collector's curve' exists, so that as the total number of individuals sampled increases (the size of the empirical sample approaches the size of the ecological sample), the total number of species identified begins to asymptote, reaching a theoretical maximum of S_e. S_e is commonly denoted just S, but it is expected in practice to recognize the distinction between the empirical sample and the ecological sample, and use one of the techniques below to estimate this value despite the notational imprecision.

There are two basic approaches to estimating S_e. First, it is possible to assume that some parameterized distribution function can be used as a model for a given species abundance distribution. If this is true, then at least for some statistical distributions, S_e is a parameter (or function of parameters) of the distribution that can be estimated using a sufficiently large empirical sample (Pielou 1975; Magurran 2004). The second general approach is to observe and extrapolate the empirical pattern of accumulation of species as the number of individuals in the empirical sample accumulates. The problem with this approach is that there is no logical way of choosing the sequences with which individuals in the empirical sample are accumulated. If there are n individuals in an empirical sample, there are $n!$ possible ways of accumulating individuals. Estimates of accumulation curves can be constructed either by random sampling (without replacement) samples of various sizes from the empirical sample or by examining the average rate of accumulation using rarefaction (Simberloff 1972). More details on estimating species richness are discussed in Chapters 4 and 20.

Estimating the diversity and richness of ecological samples is limited by the amount of information available on the ecological sample being studied. Generally, both the total number of organisms, N, and the number of species, S_e, in the ecological sample are unknown. If both of these quantities are large, then any single empirical sample may be inadequate to fully characterize the entire ecological sample (Pielou 1975; Peet 1974; Magurran 2004). Empirical samples that are much smaller than the ecological sample are unlikely to contain all species found in the ecological sample. In particular, rare species may show up in relatively few empirical samples.

The solution to the dilemma posed in the previous paragraph is to examine the behaviour of species diversity measurements among many empirical samples taken of the same ecological sample (also see Chapter 9). By taking multiple empirical samples from a single empirical sample, we gain information about the variance involved in the empirical sampling process. If a certain number of sampling units are drawn from the larger ecological sample, then it is possible to calculate measures of species diversity as a function of the number of sample units aggregated (Pielou 1975; Magurran 2004). This assumes, of course, that each sampling unit is drawn from the same ecological sample. On the other hand, if conditions change across space and time, then aggregating sampling units may not be appropriate. This assumption may be problematic if most ecological communities are open systems (Maurer 1999). In the next section we consider how to evaluate whether several empirical samples are drawn from the same empirical sample. It is still common in practice to ignore the distinction between the empirical sample and the ecological sample, and simply calculate diversity statistics (other than richness discussed above) simply on the empirical sample and report them as if they were the true value for the ecological sample. Despite being common practice, this is incorrect (especially for small samples) and the approach just outlined is superior.

5.2.5 Testing for heterogeneity among ecological samples

In the simplest case, suppose two empirical samples have been obtained from a specified location. The question to be answered is whether these two samples can be considered to be samples of a single larger community or whether they are different. The species abundance distribution for each location can written as a vector where $j = 1, 2$ indexes locations. The abundance for species i at location j is given by n_{ij}. The relative abundances are then

$$p_{ij} = n_{ij} / \sum_i n_{ij} \tag{5.10}$$

The index i goes from 1 to S_c, which is the number of species found in at least one of the samples. Note that some of the relative abundances may be zero. Writing the relative abundances as a vector gives two relative abundance vectors $\mathbf{p}_1$ and $\mathbf{p}_2$. Finally, we can calculate the relative abundance vector for the combined two samples as

$$\mathbf{p}_c = \left[\left(\sum_i n_{i1}\right)\mathbf{p}_1 + \left(\sum_i n_{i2}\right)\mathbf{p}_2\right] \Big/ \left(\sum_i n_{i1} + \sum_i n_{i2}\right) \tag{5.11}$$

Here we assume that each of the abundance distributions can be approximated by a multinomial distribution. If this is the case, then we have two candidate models to describe the data. The first model assumes the two samples come from the same community, hence there is a single multinomial distribution that describes both samples and there are S_c parameters (the actual relative abundances in the single ecological sample). Note that we here assume that $S_c = S_e$, that is, all the species in the ecological sample are found in at least one of the two empirical samples. The second model assumes that each empirical sample comes from a different ecological sample, which means that we would have two different multinomial distributions, each with S_c parameters (the relative frequencies of each species in each of the two different communities). The second model requires twice as many parameters to describe the data as the first model.

To compare such models, we suggest using the information theoretic approach described by Burnham (2002). The approach is based on estimating the log likelihood of each model given the data and substituting the empirical estimates of the parameters into the likelihood function. Interestingly, for the multinomial distribution, the negative log likelihood function for a given data set is simply the Shannon information measure times the number of individuals in the sample. Letting H_{1c} be the Shannon diversity for the combined data, the negative log likelihood for the model assuming only a single ecological sample is

$$L_c = \left(\sum_i n_{i1} + \sum_i n_{i2}\right) H_{1c} \tag{5.12}$$

For the second model, which assumes two different ecological samples, the negative log likelihood (L_2) is obtained as

$$L_2 = \left(\sum_i n_{i1}\right) H_{11} + \left(\sum_i n_{i2}\right) H_{12} \qquad (5.13)$$

where H_{11} and H_{12} are the Shannon diversities for each of the separate empirical samples. The log likelihoods for the two models are then compared by calculating the respective Akaike Information Criterion (AIC) for each model:

$$AIC_c = 2(L_c + S_c) \qquad (5.14)$$

$$AIC_2 = 2(L_2 + 2S_c) \qquad (5.15)$$

The best model is the one which has the lowest AIC. Generally, a difference between AICs of 2.0 or more indicates that the model with the lowest AIC has 'significantly' more support from the data (Burnham & Anderson 1998). Furthermore, it is possible to calculate model weights using AIC differences. The interested reader is referred to Burnham (1998) for further details.

This procedure can be generalized to evaluate whether several different empirical samples come from the same ecological sample. As the number of empirical samples being compared increases, the number of possible models increases rapidly, making it impractical to compute all possible comparisons. In such cases it may be best to use some independent criterion (such as distance between samples) to group empirical samples into a small number of aggregate samples that can be examined using AICs.

The model selection procedure provides a basis for asking questions about so-called 'β diversity'. In the example described above, the first model, which assumes that the two empirical samples are drawn from the same ecological sample, there is no β diversity. The two samples are describing the same relative abundance distribution. The second model assumes that the two empirical samples are drawn from different ecological samples, which implies that there is β diversity (i.e. a turnover in abundances among communities). The degree to which the second model is supported by the data is related to how much β diversity exists between the two samples (Chapter 6). With several sites sampled, it is also possible to partition diversity into between-site (β) and within-site (α) diversity components (Chapter 6) (Whittaker 1975; Lande 1996; Crist et al. 2003; Crist & Veech 2006).

5.3 Prospectus

While easy to conceptualize, diversity (and evenness) are hard to measure. We present a basic framework here. Several developments are needed to provide a truly firm foundation to the measurement of diversity. First, more attention to and development of methods to account for the fact that collected data are sampled data are needed (i.e. the distinction between the empirical sample and the ecological sample). Second, rather than developing new measures of diversity by ad hoc processes we hope to see a further focus on fundamental ideas like variance and information.

5.4 Key points

1. To properly measure diversity requires recognition that the data usually collected represent just samples (empirical samples) from the actual community (ecological sample), which is in turn a probabilistic, imperfect representation of the potential community.
2. There are a great many approaches proposed in the literature to measure aspects of diversity, including richness, evenness, and the combination (diversity). Most of these approaches have been fairly ad hoc. There are probably at least two or three times as many measures proposed as the ones we cover in Box 5.1. We have tried to highlight the most commonly used and successful measures.
3. We present a uniform framework for building diversity measures. Two of the oldest measures of diversity, Simpson and Shannon, and their corresponding evenness measures, turn out to be directly related to two fairly deep concepts of diversity: variance and information.
4. The best way to use empirical samples to get at the ecological sample is to take multiple empirical samples of the same ecological sample. This provides information about the variability

induced by the empirical sampling process and allows for the development of an asymptotic approach that can then be extrapolated to the properties of the ecological sample. One common method is to plot the measure of interest vs the number of empirical samples and extrapolate.

5. We present a method for testing whether two (or any combination of more than two) empirical samples are drawn from a single ecological sample using the multinomial distribution and likelihood/AIC methods. It turns out that the likelihood is directly related to Shannon's diversity.