CHAPTER 4

Estimating species richness

Nicholas J. Gotelli and Robert K. Colwell

4.1 Introduction

Measuring species richness is an essential objective for many community ecologists and conservation biologists. The number of species in a local assemblage is an intuitive and natural index of community structure, and patterns of species richness have been measured at both small (e.g. Blake & Loiselle 2000) and large (e.g. Rahbek & Graves 2001) spatial scales. Many classic models in community ecology, such as the MacArthur–Wilson equilibrium model (MacArthur & Wilson 1967) and the intermediate disturbance hypothesis (Connell 1978), as well as more recent models of neutral theory (Hubbell 2001), metacommunity structure (Holyoak et al. 2005), and biogeography (Gotelli et al. 2009) generate quantitative predictions of the number of coexisting species. To make progress in modelling species richness, these predictions need to be compared with empirical data. In applied ecology and conservation biology, the number of species that remain in a community represents the ultimate 'scorecard' in the fight to preserve and restore perturbed communities (e.g. Brook et al. 2003).

Yet, in spite of our familiarity with species richness, it is a surprisingly difficult variable to measure. Almost without exception, species richness can be neither accurately measured nor directly estimated by observation because the observed number of species is a downward-biased estimator for the complete (total) species richness of a local assemblage. Hundreds of papers describe statistical methods for correcting this bias in the estimation of species richness (see also Chapter 3), and special protocols and methods have been developed for estimating species richness for particular taxa (e.g. Agosti et al. 2000). Nevertheless, many recent studies continue to ignore some of the fundamental sampling and measurement problems that can compromise the accurate estimation of species richness (Gotelli & Colwell 2001).

In this chapter we review the basic statistical issues involved with species richness estimation. Although a complete review of the subject is beyond the scope of this chapter, we highlight sampling models for species richness that account for undersampling bias by adjusting or controlling for differences in the number of individuals and the number of samples collected (rarefaction) as well as models that use abundance or incidence distributions to estimate the number of undetected species (estimators of asymptotic richness).

4.2 State of the field

4.2.1 Sampling models for biodiversity data

Although the methods of estimating species richness that we discuss can be applied to assemblages of organisms that have been identified by genotype (e.g. Hughes et al. 2000), to species, or to some higher taxonomic rank, such as genus or family (e.g. Bush & Bambach 2004), we will write 'species' to keep it simple. Because we are discussing estimation of species richness, we assume that one or more *samples* have been taken, by collection or observation, from one or more *assemblages* for some specified group or groups of organisms. We distinguish two kinds of data used in richness studies: (1) *incidence data*, in which each species detected in a sample from an assemblage is simply noted as being present, and (2) *abundance data*, in which the abundance of each species is tallied within each sample. Of course, abundance data can always be converted to incidence data, but not the reverse.

Box 4.1 Observed and estimated richness

S_{obs} is the total number of species observed in a sample, or in a set of samples.

S_{est} is the estimated number of species in the assemblage represented by the sample, or by the set of samples, where *est* is replaced by the name of an estimator.

Abundance data. Let f_k be the number of species each represented by exactly k individuals in a single sample. Thus, f_0 is the number of *undetected* species (species present in the assemblage but not included in the sample), f_1 is the number of *singleton* species, f_2 is the number of *doubleton* species, etc. The total number of individuals in the sample is $n = \sum_{k=1}^{S_{obs}} f_k$.

Replicated incidence data. Let q_k be the number of species present in exactly k samples in a set of replicate incidence samples. Thus, q_0 is the number of *undetected* species (species present in the assemblage but not included in the set of samples), q_1 is the number of *unique* species, q_2 is the number of *duplicate* species, etc. The total number of samples is $m = \sum_{k=1}^{S_{obs}} q_k$.

Chao 1 (for abundance data)

$S_{Chao1} = S_{obs} + \frac{f_1^2}{2f_2}$ is the classic form, but is not defined when $f_2 = 0$ (no doubletons).

$S_{Chao1} = S_{obs} + \frac{f_1(f_1-1)}{2(f_2+1)}$ is a bias-corrected form, always obtainable.

$\text{var}(S_{Chao1}) = f_2\left[\frac{1}{2}\left(\frac{f_1}{f_2}\right)^2 + \left(\frac{f_1}{f_2}\right)^3 + \frac{1}{4}\left(\frac{f_1}{f_2}\right)^4\right]$ for $f_1 > 0$ and $f_2 > 0$ (see Colwell 2009, Appendix B of *EstimateS User's Guide* for other cases and for asymmetrical confidence interval computation).

Chao 2 (for replicated incidence data)

$S_{Chao2} = S_{obs} + \frac{q_1^2}{2q_2}$ is the classic form, but is not defined when $q_2 = 0$ (no duplicates).

$S_{Chao2} = S_{obs} + \left(\frac{m-1}{m}\right)\frac{q_1(q_1-1)}{2(q_2+1)}$ is a bias-corrected form, always obtainable.

$\text{var}(S_{Chao2}) = q_2\left[\frac{1}{2}\left(\frac{q_1}{q_2}\right)^2 + \left(\frac{q_1}{q_2}\right)^3 + \frac{1}{4}\left(\frac{q_1}{q_2}\right)^4\right]$ for $q_1 > 0$ and $q_2 > 0$ (see Colwell 2009, Appendix B of *EstimateS User's Guide* for other cases and for asymmetrical confidence interval computation).

ACE (for abundance data)

$S_{rare} = \sum_{k=1}^{10} f_k$ is the number of *rare* species in a sample (each with 10 or fewer individuals).

$S_{abund} = \sum_{k=11}^{S_{obs}} f_k$ is the number of *abundant* species in a sample (each with more than 10 individuals).

$n_{rare} = \sum_{k=1}^{10} k f_k$ is the total number of individuals in the rare species.

The sample coverage estimate is $C_{ACE} = 1 - \frac{f_1}{n_{rare}}$, the proportion of all individuals in rare species that are not singletons. Then the ACE estimator of species richness is $S_{ACE} = S_{abund} + \frac{S_{rare}}{C_{ACE}} + \frac{f_1}{C_{ACE}}\gamma^2_{ACE}$, where γ^2_{ACE} is the coefficient of variation,

$$\gamma^2_{ACE} = \max\left[\frac{S_{rare}}{C_{ACE}}\frac{\sum_{k=1}^{10} k(k-1)f_k}{(n_{rare})(n_{rare}-1)} - 1, 0\right]$$

The formula for ACE is undefined when all rare species are singletons ($f_1 = n_{rare}$, yielding $C_{ACE} = 0$). In this case, compute the bias-corrected form of Chao1 instead.

ICE (for incidence data)

$S_{infr} = \sum_{k=1}^{10} q_k$ is the number of *infrequent* species in a sample (each found in 10 or fewer samples).

$S_{freq} = \sum_{k=11}^{S_{obs}} q_k$ is the number of *frequent* species in a sample (each found in more than 10 samples).

$n_{infr} = \sum_{k=1}^{10} k q_k$ is the total number of incidences in the infrequent species.

The sample coverage estimate is $C_{ICE} = 1 - \frac{q_1}{n_{infr}}$, the proportion of all incidences of infrequent species that are not uniques. Then the ICE estimator of species richness is $C_{ICE} = S_{freq} + \frac{S_{infr}}{C_{ICE}} + \frac{q_1}{C_{ICE}}\gamma^2_{ICE}$, where γ^2_{ICE} is the coefficient of variation,

$$\gamma^2_{ICE} = \max\left[\frac{S_{infr}}{C_{ICE}}\frac{m_{infr}}{(m_{infr}-1)}\frac{\sum_{k=1}^{10} k(k-1)q_k}{(n_{infr})^2} - 1, 0\right]$$

The formula for ICE is undefined when all infrequent species are uniques ($q_1 = n_{infr}$, yielding $C_{ICE} = 0$). In this case, compute the bias-corrected form of Chao2 instead.

Jackknife estimators (for abundance data)

The first-order jackknife richness estimator is

$$S_{jackknife1} = S_{obs} + f_1$$

The second-order jackknife richness estimator is

$$S_{jackknife2} = S_{obs} + 2f_1 - f_2$$

Jackknife estimators (for incidence data)

The first-order jackknife richness estimator is

$$S_{jackknife1} = S_{obs} + q_1\left(\frac{m-1}{m}\right)$$

The second-order jackknife richness estimator is

$$S_{jackknife2} = S_{obs} + \left[\frac{q_1(2m-3)}{m} - \frac{q_2(m-2)^2}{m(m-1)}\right]$$

By their nature, sampling data document only the verified *presence* of species in samples. The absence of a particular species in a sample may represent either a true absence (the species is not present in the assemblage) or a false absence (the species is present, but was not detected in the sample; see Chapter 3). Although the term 'presence/absence data' is often used as a synonym for incidence data, the importance of distinguishing true absences from false ones (not only for richness estimation, but in modelling contexts, e.g. Elith et al. 2006) leads us to emphasize that incidence data are actually 'presence data'. Richness estimation methods for abundance data assume that organisms can be sampled and identified as distinct individuals. For clonal and colonial organisms, such as many species of grasses and corals, individuals cannot always be separated or counted, but methods designed for incidence data can nonetheless be used if species presence is recorded within standardized quadrats or samples (e.g. Butler & Chazdon 1998).

Snacking from a jar of mixed jellybeans provides a good analogy for biodiversity sampling (Longino et al. 2002). Each jellybean represents a single individual, and the different colours represent the different species in the jellybean 'assemblage'—in a typical sample, some colours are common, but most are rare. Collecting a sample of biodiversity data is equivalent to taking a small handful of jellybeans from the jar and examining them one by one. From this incomplete sample, we try to make inferences about the number of colours (species) in the entire jar. This process of statistical inference depends critically on the biological assumption that the community is 'closed,' with an unchanging total number of species and a steady species abundance distribution. Jellybeans may be added or removed from the jar, but the proportional representation of colours is assumed to remain the same. In an open metacommunity, in which the assemblage changes size and composition through time, it may not be possible to draw valid inferences about community structure from a snapshot sample at one point in time (Magurran 2007). Few, if any, real communities are completely 'closed', but many are sufficiently circumscribed that that richness estimators may be used, but with caution and caveats.

For all of the methods and metrics (Box 4.1) that we discuss in this chapter, we make the closely related statistical assumption that sampling is *with replacement*. In terms of collecting inventory data from nature, this assumption means either that individuals are recorded, but not removed, from the assemblage (e.g. censusing trees in a plot) or, if they are removed, the proportions remaining are unchanged by the sampling.

This framework of sampling, counting, and identifying individuals applies not only to richness estimation, but also to many other questions in the study of biodiversity, including the characterization of the species abundance distribution (see Chapter 9) and partitioning diversity into α and β components (see Chapters 6 and 7).

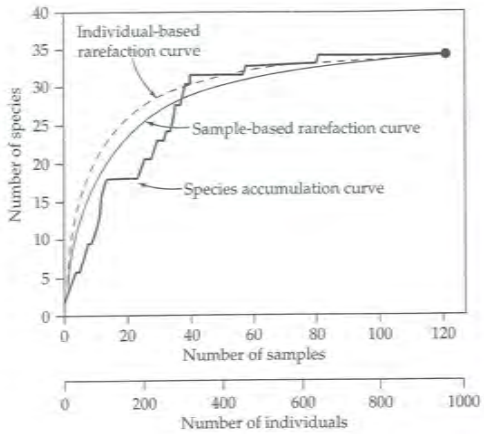

Figure 4.1 Species accumulation and rarefaction curves. The *jagged line* is the *species accumulation curve* for one of many possible orderings of 121 soil seedbank samples, yielding a total of 952 individual tree seedlings, from an intensive census of a plot of Costa Rican rainforest (Butler & Chazdon 1998). The cumulative number of tree species (*y*-axis) is plotted as a function of the cumulative number of samples (upper *x*-axis), pooled in random order. The smooth, *solid line* is the *sample-based rarefaction curve* for the same data set, showing the mean number of species for all possible combinations of 1, 2, ..., m^*, ..., 121 actual samples from the dataset—this curve plots the statistical expectation of the (sample-based) species accumulation curve. The *dashed line* is the *individual-based rarefaction curve* for the same data set—the expected number of species for (m^*)(952/121) individuals, randomly chosen from all 952 individuals (lower *x*-axis). The black dot indicates the total richness for all samples (or all individuals) pooled. The sample-based rarefaction curve lies below the individual-based rarefaction curve because of spatial aggregation within species. This is a very typical pattern for empirical comparisons of sample-based and individual-based rarefaction curves.

4.2.2 The species accumulation curve

Consider a graph in which the *x*-axis is the number of individuals sampled and the *y*-axis is the cumulative number of species recorded (Fig. 4.1, lower *x*-axis). Imagine taking one jellybean at a time from the jar, at random. As more individuals (jellybeans) are sampled, the total number of species (colours) recorded in the sample increases, and a *species accumulation curve* is generated. Of course, the first individual drawn will represent exactly one species new to the sample, so all species accumulation curves based on individual organisms originate at the point [1,1]. The next individual drawn will represent either the same species or a species new to the sample. The probability of drawing a new species will depend both on the complete number of species in the assemblage and their relative abundances. The more species in the assemblage and the more even the species abundance distribution (see Chapter 9), the more rapidly this curve will rise. In contrast, if the species abundance distribution is highly uneven (a few common species and many rare ones, for example), the curve will rise more slowly, even at the outset, because most of the individuals sampled will represent more common species that have already been added to the sample, rather than rarer ones that have yet to be detected.

Regardless of the species abundance distribution, this curve increases monotonically, with a decelerating slope. For a given sample, different stochastic realizations of the order in which the individuals in the sample are added to the graph will produce species accumulation curves that differ slightly from one another. The smoothed average of these individual curves represents the statistical expectation of the species accumulation curve for that particular sample, and the variability among the different orderings is reflected in the variance in the number of species recorded for any given number of individuals. However, this variance is specific, or *conditional*, on the particular sample that we have drawn because it is based only on re-orderings of that single sample. Suppose, instead, we plot the smoothed average of several species accumulation curves, each based on a different handful of jellybeans from the same jar, each handful having the same number of beans. Variation among these smoothed curves from the several independent, random samples represents another source of variation in richness, for a given number of individuals. The variance among these curves is called an *unconditional* variance because it estimates the true variance in richness of the assemblage. The unconditional variance in richness is necessarily larger than the variance conditional on any single sample.

4.2.3 Climbing the species accumulation curve

In theory, finding out how many species characterize an assemblage means sampling more and more individuals until no new species are found and the species accumulation curve reaches an asymptote. In practice, this approach is routinely impossible for two reasons. First, the number of individuals that must be sampled to reach an asymptote can often be prohibitively large (Chao et al. 2009). The problem is most severe in the tropics, where species diversity is high and most species are rare. For example, after nearly 30 consecutive years of sampling, an ongoing inventory of a tropical rainforest ant assemblage at La Selva, Costa Rica, has still not reached an asymptote in species richness. Each year, one or two new species are added to the local list. In some cases these species are already known from collections at other localities, but in other cases they are new to science (Longino et al. 2002). In other words, biodiversity samples, even very extensive ones, often fall short of revealing the complete species richness for an assemblage, representing some unspecified milestone along a slowly rising species accumulation curve with an unknown destination.

A second reason that the species accumulation curve cannot be used to directly determine species richness is that, in field sampling, ecologists almost never collect random individuals in sequence. Instead, individual plants or mobile animals are often recorded from transects or points counts, or individual organisms are collected in pitfall and bait traps, sweep samples, nets, plankton tows, water, soil, and leaf litter samples, and other taxon-specific sampling units that capture multiple individuals (Southwood & Henderson 2000). Although these samples can, under appropriate circumstances, be treated as independent of one another, the individuals accumulated within a single sample do not represent independent observations. Although individuals contain the biodiversity 'information' (species identity), it is the samples that represent the statistically independent replicates for analysis. When spatial and temporal autocorrelation is taken into account, the samples themselves may be only partially independent. Nevertheless, the inevitable non-independence of individuals within samples can be overcome by plotting a second kind of species accumulation curve, called *a sample-based* species accumulation curve, in which the *x*-axis is the number of samples and the *y*-axis is the accumulated number of species (Fig. 4.1, upper *x*-axis). Because only the identity but not the number of individuals of each species represented within a sample is needed to construct a sample-based species accumulation curve, these curves plot incidence data. This approach is therefore also suitable for clonal and colonial species that cannot be counted as discrete individuals.

4.2.4 Species richness versus species density

The observed number of species recorded in a sample (or a set of samples) is very sensitive to the number of individuals or samples observed or collected, which in turn is influenced by the effective area that is sampled and, in replicated designs, by the spatial arrangement of the replicates. Thus, many measures reported as 'species richness' are effectively measures of species density: the number of species collected in a *particular* total area. For quadrat samples or other methods that sample a fixed area, *species density* is expressed in units of species per *specified* area. Even for traps that collect individuals at a single point (such as a pitfall trap), there is probably an effective sampling area that is encompassed by data collection at a single point.

Whenever sampling is involved, species density is a slippery concept that is often misused and misunderstood. The problem arises from the non-linearity of the species accumulation curve. Consider the species accumulation curve for rainforest seedlings (Butler & Chazdon 1998) in Fig. 4.2, which plots the species of seedlings grown from dormant seed in 121 soil samples, each covering a soil surface area of 17.35 cm^2 and a depth of 10 cm. The *x*-axis plots the cumulative surface area of soil sampled. The *slopes* of lines A, B, and C represent species density: *number of species observed* (*y*), divided by *area-sampled* (*x*). You can see that species density

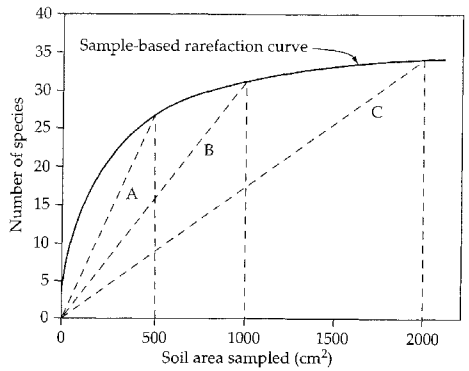

Figure 4.2 Species richness and species density are not the same thing. The *solid line* is the *sample-based rarefaction curve* for the same data set as in Fig 4.1, showing the expected species richness of rainforest tree seedlings for 1, 2, ..., m^*, ..., 121 soil samples, each covering a soil surface area of 17.35 cm^2 and a depth of 10 cm. *Species richness* (*y*-axis) is plotted as a function of the total *soil surface area* sampled (*x*-axis). Because *species density* is the ratio of richness (*y*-coordinate) to area (*x*-coordinate) for any point in the graph, the *slopes* of lines A, B, and C quantify species density for 500, 1000, and 2000 cm^2, respectively. Clearly, species density estimates depend on the *particular* amount of area sampled. All of the species density slopes over-estimate species number when extrapolated to larger areas, and species density estimates based on differing areas are not comparable.

depends critically not just on area, but on the *specific* amount of area sampled. For this reason, it *never* works to 'standardize' the species richness of samples from two or more assemblages by simply dividing observed richness by area sampled (or by any other measure of effort, including number of individuals or number of samples). Estimating species density by calculating the ratio of species richness to area sampled will always grossly overestimate species density when this index is extrapolated to larger areas, and the size of that bias will depend on the area sampled.

Sometimes, however, ecologists or conservation biologists are interested in species density, for some particular amount of area, in its own right. For example, if only one of two areas, equal in size and cost per hectare, can be purchased to establish a reserve, species density at the scale of the reserve is clearly a variable of interest. Because *species density* is so sensitive to area (and, ultimately, to the number of individuals observed or collected), it is useful to decompose it into the product of two quantities: *species richness* (number of species represented by some *particular* number, *N*, of individuals) and *total individual density* (number of individuals *N*, disregarding species, in some *particular* amount of area *A*):

$$\left(\frac{species}{area\ A}\right) = \left(\frac{species}{N\ individuals}\right) \times \left(\frac{N\ individuals}{area\ A}\right)$$

(James & Wamer 1982). This decomposition demonstrates that the number of species per sampling unit reflects both the underlying species richness and the total number of individuals sampled. If two samples differ in species density, is it because of differences in underlying species richness, differences in abundance, or some combination of both? In other words, how do we meaningfully compare the species richness of collections that probably differ in both the number of individuals and the number of samples collected? Until recently, many ecologists have not recognized this problem. The distinction between species density and species richness has not always been appreciated, and many papers have compared species density using standard parametric statistics, but without accounting for differences in abundance or sampling effort.

One statistical solution is to treat abundance, number of samples, or sample area as a covariate that can be entered into a multiple regression analysis or an analysis of covariance. If the original data (counts and identities of individuals) are not available, this may be the best that we can do. For example, Dunn et al. (2009) assembled a global database of ant species richness from a number of published studies. To control for sampling effects, they used the area, number of samples, and total number of individuals from each sample location as statistical covariates in regression analyses. However, they

did not make the mistake of trying to 'standardize' the richness of different samples by dividing the species counts by the area, the number of individuals sampled, or any other measure of effort. As we have repeatedly emphasized, this rescaling produces serious distortions: extrapolations from small sample ratios of species density inevitably lead to gross over-estimates of the number of species expected in larger sample areas (Fig. 4.2 and Figure 4–6 in Gotelli & Colwell 2001).

4.2.5 Individual-based rarefaction

The species accumulation curve itself suggests an intuitive way to compare the richness of two samples (for the same kind of organism) that differ in the number of individuals collected. Suppose one of the two samples has N individuals and S species, and the other has n individuals and s species. The samples differ in the number of individuals present ($N > n$) and will usually differ in the number of species present (typically $S > s$). In the procedure called *rarefaction*, we randomly draw n^* individuals, subsampling *without* replacement from the larger of the two original samples, where $n^* = n$, the size of the smaller original sample. (This re-sampling, without replacement, of individuals from within the sample does not violate the assumption that the process of taking the sample itself did not change the relative abundance of species). Computing the mean number of species, $\bar{s}^*$, among repeated subsamples of n^* individuals estimates $E(s^*|n^*)$, the expected number of species in a random subsample of n^* individuals from the larger original sample (Fig. 4.1, lower *x*-axis). The variance of (s^*), among random re-orderings of individuals, can also be estimated this way along with a parametric 95% confidence interval, or the confidence interval can be estimated from the bootstrapped values (Manly 1991).

A simple test can now be conducted to ask whether s, the observed species richness of the complete smaller sample, falls within the 95% confidence interval of s^*, the expected species richness based on random subsamples of size n from the larger sample (Simberloff 1978). If the observed value falls within the confidence interval, then the hypothesis that the richness of the smaller sample, based on all n individuals, does not differ from the richness of a subsample of size n^* from the larger sample cannot be rejected at $P \leq 0.05$. If this null hypothesis is not rejected, and the original, unrarefied samples differed in species density, then this difference in species density must be driven by differing numbers of individuals between the two samples. Alternatively, if s is not contained within the confidence interval of s^*, the two samples differ in species richness in ways that cannot be accounted for entirely by differences in abundance and/or sampling effort (at $P \leq 0.05$).

Rarefaction can be used not only to calculate a point estimate of s^*, but also to construct an entire *rarefaction curve* in which the number of individuals randomly subsampled ranges from 1 to N. Rarefaction can be thought of as a method of *interpolating* $E(s^*|n^*)$ the expected number of species, given n^* individuals ($1 \leq n^* \leq N$), between the point [1, 1] and the point [S, N] (Colwell et al. 2004). With progressively smaller subsamples from $N - 1$ to 1, the resulting *individual-based rarefaction curve*, in a sense, is the reverse of the corresponding species accumulation curve, which progressively builds larger and larger samples.

Because this individual-based rarefaction curve is conditional on one particular sample, the variance in s^*, among random re-orderings of individuals, is 0 at both extremes of the curve: with the minimum of only one individual there will always be only one species represented, and with the maximum of N individuals, there will always be exactly S species represented. Hurlbert (1971) and Heck et al. (1975) give analytical solutions for the expectation and the *conditional* variance of s^*, which are derived from the hypergeometric distribution. In contrast, treating the sample (one handful of jellybeans) as representative of a larger assemblage (the jar of jellybeans) requires an estimate of the *unconditional* variance (the variance in $s^*|n^*$ among replicate handfuls of jellybeans from the same jar). The unconditional variance in richness, S, for the full sample of N individuals, must be greater than zero to account for the heterogeneity that would be expected with additional random samples of the same size taken from the entire assemblage. Although Smith & Grassle (1977) derived an estimator for the unconditional variance of $E(s^*|n^*)$,

it is computationally complex and has been little used. R.K. Colwell and C.X. Mao (in preparation) have recently derived an unconditional variance estimator for individual-based rarefaction that is analogous to the unconditional variance estimator for sample-based rarefaction described in Colwell et al. (2004), and discussed below.

Regardless of how the variance is estimated, the statistical significance of the difference in rarefied species richness between two samples will depend, in part, on n, the number of individuals being compared. This sample-size dependence arises because all rarefaction curves based on individuals converge at the point [1,1]. Therefore, no matter how different two assemblages are, rarefaction curves based on samples of individuals drawn at random will not appear to differ statistically if n is too small. In some cases, rarefaction curves may cross at higher values of n, making the results of statistical tests even more dependent on n (e.g. Raup 1975).

To compare multiple samples, each can be rarefied down to a common abundance, which will typically be the total abundance for the smallest of the samples. At that point, the set of s^* values, one for each sample, can be used as a response variable in any kind of statistical analysis, such as ANOVA or regression. This method assumes that the rarefaction curves do not cross (which may be assessed visually), so that their rank order remains the same regardless of the abundance level used. Alternatively, multiple samples from the same assemblage can be used in a *sample-based rarefaction*, which we describe below.

Rarefaction has a long history in ecology and evolution (Sanders 1968; Hurlbert 1971; Raup 1975; Tipper 1979; Järvinen 1982; Chiarucci et al. 2008).The method was proposed in the 1960s and 1970s to compare species number when samples differed in abundance (Tipper 1979), but the same statistical problem had been solved many decades earlier by biogeographers who wanted to estimate species/genus ratios and other taxonomic diversity indices (Järvinen 1982).

Brewer & Williamson (1994) and Colwell & Coddington (1994) pointed out that a very close approximation for the rarefaction curve is the Coleman 'passive sampling' curve,

$$E(s^*) = \sum_{i=1}^{S} \left[1 - (1 - n^*/N)^{n_i}\right], \qquad (4.1)$$

in which i indexes species from 1 to S, and n_i is the abundance of species i in the full sample. As a null model for the species–area relationship (see Chapter 20), the Coleman curve assumes that islands of different area randomly intercept individuals and accumulate different numbers of species (Coleman et al. 1982). The individual-based rarefaction curve is very closely analogous to the Coleman curve (and, although mathematically distinct, differs only slightly from it) because relative island area is a proxy for the proportion n^*/N of individuals subsampled from the pooled distribution of all individuals in the original sample (Gotelli 2008).

4.2.6 Sample-based rarefaction

Individual-based rarefaction computes the expected number of species, s^*, in a subsample of n^* *individuals* drawn at random from a *single* representative sample from an assemblage. In contrast, *sample-based rarefaction* computes the expected number of species s^* when m^* *samples* ($1 \leq m^* \leq M$) are drawn at random (without replacement) from a *set of samples* that are, collectively, representative of an assemblage (Fig. 4.1, upper x-axis) (Gotelli & Colwell 2001; Colwell et al. 2004). (This re-sampling, without replacement, of samples from within the sample set does not violate the assumption that the process of taking the sample itself did not change the relative abundance of species.) The fundamental difference is that sample-based rarefaction, by design, preserves the spatial structure of the data, which may reflect processes such as spatial aggregation or segregation (see Chapter 12) both within and between species. In contrast, individual-based rarefaction does not preserve the spatial structure of the data and assumes complete random mixing among individuals of all species. Thus, for sample-based rarefaction, $E(s^*|m^*)$ is the expected number of species for m^* pooled samples that express the same patterns of aggregation, association, or segregation as the observed set of samples. For this reason, sample-based rarefaction is a more realistic treatment of the independent sampling units used in most biodiversity studies. Because sample-based rarefaction requires only incidence data, it can also be used for clonal organisms or for species in which individuals in a sample cannot be easily distinguished or counted.

Operationally, sample-based rarefaction can be carried out by repeatedly selecting and pooling m^* samples at random from the set of samples, and computing the mean and conditional (on the particular set of samples) variance and 95% confidence interval for s^*. On the other hand, $E(s^*|m^*)$ is more easily and accurately computed from combinatorial equations based on the distribution of *counts*, the number of species found in exactly 1, 2, ..., m^* samples in the set (Ugland et al. 2003; Colwell et al. 2004; see Chiarucci et al. 2008 for a history of this approach). Colwell et al. 2004 also introduced a sample-based version of the Coleman rarefaction model, the results of which closely approximate the true sample-based rarefaction curve.

Ugland et al. (2003) provide an expression for the conditional variance in richness estimates from sample-based rarefaction. Colwell et al. (2004) derived an unconditional variance estimator for sample-based rarefaction that treats the observed set of samples, in turn, as a sample from some larger assemblage, so that the variance in S for all M samples, pooled (the full set of samples), takes some non-zero value. This unconditional variance (and its associated confidence interval (CI)) accounts for the variability expected among replicate sets of samples. Based on unconditional variances for two sample-based rarefaction curves, richness can be compared for any common number of samples (or individuals, as explained below). Using eigenvalue decomposition, Mao & Li (2009) developed a computationally complex method for comparing two sample-based rarefaction curves in their entirety. A much simpler, but approximate, method is to assess, for a desired value of m^*, whether or not the two (appropriately computed) confidence intervals overlap. If the two CIs (calculated from the unconditional variance) are approximately equal, for a type I error rate of $P < 0.05$, the appropriate CI is about 84% (Payton et al. 2003; the z value for 84% CI is 0.994 standard deviations). Basing the test on the overlap of traditional 95% CIs is overly conservative: richness values that would differ significantly with the 84% interval would often be declared statistically indistinguishable because the 95% intervals for the same pair of samples would overlap (Payton et al. 2003).

An important pitfall to avoid in using sample-based rarefaction to compare richness between sample sets is that the method does not directly control for differences in overall abundance between sets of samples. Suppose two sets of samples are recorded from the same assemblage, but they differ in mean number of individuals per sample (systematically or by chance). When plotted as a function of number of samples (on the x-axis) the sample-based rarefaction curve for the sample set with a higher mean abundance per sample will lie above the curve for the sample set with lower mean abundance because more individuals reveal more species. The solution suggested by Gotelli & Colwell (2001) is to first calculate sample-based rarefaction curves and their variances (or CIs) for each set of samples in the analysis. Next, the curves are re-plotted against an x-axis of individual abundance, rather than number of samples. This re-plotting effectively shifts the points of each individual-based rarefaction curve to the left or the right, depending on the average number of individuals that were collected in each sample. Ellison et al. (2007) used this method to compare the efficacy of ant sampling methods that differed greatly in the average number of individuals per sample (e.g. 2 ants per pitfall trap, versus > 89 ants per plot for standardized hand sampling). Note that if sample-based rarefaction is based on species occurrences rather than abundances, then the rescaled x-axis is the number of species occurrences, not the number of individuals.

4.2.7 Assumptions of rarefaction

To use rarefaction to compare species richness of two (or more) samples or assemblages rigorously, the following assumptions should be met:

1. *Sufficient sampling.* As with any other statistical procedure, the power to detect a difference, if there is one, depends on having

large enough individuals or samples, especially since rarefactions curves necessarily converge towards the origin. Although it is difficult to give specific recommendations, our experience has been that rarefaction curves should be based on at least 20 individuals (individual-based rarefaction) or 20 samples (sample-based rarefaction), and preferably many more.

2. *Comparable sampling methods.* Because all sampling methods have inherent and usually unknown sampling biases that favour detection of some species but not others (see Chapter 3), rarefaction cannot be used to compare data from two different assemblages that were collected with two different methods (e.g. bait samples vs pitfall traps, mist-netting vs point-sampling for birds). However, rarefaction can be used meaningfully to compare the efficacy of different sampling methods that are used in the same area (Longino et al. 2002; Ellison et al. 2007). Also, data from different sampling methods may be pooled in order to maximize the kinds of species that may be sampled with different sampling methods (e.g. ants in Colwell et al. (2008)). However, identical sampling and pooling procedures must to be employed to compare two composite collections.
3. *Taxonomic similarity.* The assemblages represented by the two samples should be taxonomically 'similar'. In other words, if two samples that differ in abundance but have rarefaction curves with identical shapes do not share any taxa, we would not want to conclude that the smaller collection is a random subsample of the larger (Tipper 1979). Rarefaction seems most useful when the species composition of the smaller sample appears to be a nested or partially nested subset of the larger collection. Much more powerful methods are now available to test directly for differences in species composition (Chao et al. 2005).
4. *Closed communities of discrete individuals.* The assemblages being sampled should be well circumscribed, with consistent membership. Discrete individuals in a single sample must be countable (individual-based rarefaction) or species presence in multiple samples must be detectable (sample-based rarefaction).
5. *Random placement.* Individual-based rarefaction assumes that the spatial distribution of individuals sampled is random. If individuals within species are spatially aggregated, individual-based rarefaction will over-estimate species richness because it assumes that the rare and common species are perfectly intermixed. Some authors have modified the basic rarefaction equations to include explicit terms for spatial clumping (Kobayashi & Kimura 1994). However, this approach is rarely successful because the model parameters (such as the constants in the negative binomial distribution) cannot be easily and independently estimated for all of the species in the sample. One way to deal with aggregation is to increase the distance or timing between randomly sampled individuals so that patterns of spatial or temporal aggregation are not so prominent. An even better approach is to use sample-based rarefaction, again employing sampling areas that are large enough to overcome small-scale aggregation.
6. *Independent, random sampling.* Individuals or samples should be collected randomly and independently. Both the individual-based and sample-based methods described in this chapter assume that sampling, from nature, does not affect the relative abundance of species (statistically, sampling with replacement). However, if the sample is relatively small compared to the size of the underlying assemblage (which is often the case), the results should be similar for samples collected with or without replacement. More work is needed to derive estimators that can be used for sampling without replacement, which will be important for cases in which the sample represents a large fraction of the total assemblage. Unfortunately, as we have noted earlier, biodiversity data rarely consist of collections of individuals that were sampled randomly. Instead, the data often consist of a series of random and approximately independent samples that contain multiple individuals.

4.2.8 Estimating asymptotic species richness

Consider the species richness of a single biodiversity sample (or the pooled richness of a set of samples) as the starting point in a graph of richness versus abundance or sample number (the dot at the right-hand end of the curves in Fig. 4.1). Rarefaction amounts to *interpolating* 'backward' from the endpoint of a species accumulation curve, yielding estimates of species richness expected for smaller numbers of individuals or samples. In contrast, using this starting point to estimate the complete richness of the assemblage, including species that were not detected by the sample, can be visualized as *extrapolating* 'forward' along a hypothetical projection the accumulation curve (Colwell et al. 2004, their Figure 4). Two objectives of extrapolation can be distinguished: (1) estimating the richness of a larger sample and (2) estimating the complete richness of the assemblage, visualized as the asymptote of the accumulation curve. Once this asymptote is reached, the species accumulation curve is flat and additional sampling will not yield any additional species.

Why should the species accumulation curve have an asymptote? On large geographical scales, it does not: larger areas accumulate species at a constant or even an increasing rate because expanded sampling incorporates diverse habitat types that support distinctive species assemblages (see Chapter 20). As a consequence, the species accumulation curve continues to increase, and will not reach a final asymptote until it approaches the total area of the biosphere. The subject of *species turnover* is covered by Jost et al. and Magurran (Chapters 6 and 7) and *species–area relationships* are the subject of Chapter 20. In this chapter, we focus on the estimation of species richness at smaller spatial scales—scales at which an asymptote is a reasonable supposition and sampling issues are substantially more important than spatial turnover on habitat mosaics or gradients (Cam et al. 2002). In statistical terms, we assume that samples were drawn independently and at random from the local assemblage, so that the ordering of the samples in time or space is not important. In fact, *un*importance of sample order is diagnostic of the kinds of sample sets appropriately used by ecologists to assess local species richness (Colwell et al. 2004).

The most direct approach to estimating the species richness asymptote is to fit an asymptotic mathematical function (such as the Michaelis–Menten function; Keating & Quinn (1998)) to a rarefaction or species accumulation curve. This approach dates back at least to Holdridge et al. (1971), who fitted a negative binomial function to smoothed species accumulation curves to compare the richness of Costa Rica trees at different localities. Many other asymptotic functions have since been explored (reviewed by Colwell & Coddington (1994), Flather (1996), Chao (2005), and Rosenzweig et al. (2003)). Unfortunately, this strictly phenomenological method, despite the advantage that it makes no assumptions about sampling schemes or species abundance distributions, does not seem to work well in practice. Two or more functions may fit a dataset equally well, but yield drastically different estimates of asymptotic richness (Soberón & Llorente 1993; Chao 2005), and variance estimates for the asymptote are necessarily large. Residual analysis often reveals that the popular functions do not correctly fit the shape of empirical species accumulation curves (O'Hara 2005), and this curve-fitting method consistently performs worse than other approaches (Walther & Moore 2005; Walther & Morand 2008). For these reasons, we do not recommend fitting asymptotic mathematical functions as a means of estimating complete species richness of local assemblages.

Mixture models, in which species abundance or occurrence distributions are modelled as a weighted mixture of statistical distributions, offer a completely different, non-parametric approach to extrapolating an empirical rarefaction curve to a larger sample sizes (or a larger set of samples) (reviewed by Mao et al. (2005), Mao & Colwell (2005), and Chao (2005)). Colwell et al. (2004), for example, modelled the sample-based rarefaction curve as a binomial mixture model. However, these models are effective only for a doubling or tripling of the observed sample size. Beyond this point, the variance of the richness estimate increases rapidly. Unless the initial sample size is very large, projecting the curve to an asymptotic value usually requires much more than a doubling or tripling of the initial sample size (Chao et al. 2009), so this method is not always feasible, especially for hyperdiverse taxa (Mao & Colwell 2005).

Another classical approach to estimating asymptotic richness is to fit a species abundance

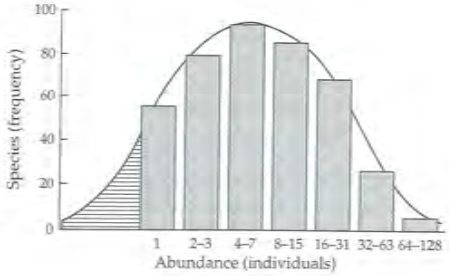

Figure 4.3 Estimation of asymptotic species richness by fitting a log-normal distribution to a species abundance distribution. The graph shows the number of species of ants in each of seven logarithmically-scaled abundance categories (a total of 435 species collected) in a long-term rainforest inventory in Costa Rica (Longino et al. 2002). The number of undetected species (21 additional species) is estimated by the area marked with horizontal hatching, yielding a predicted complete richness of 456 species.

distribution (see Chapter 9), based on a single sample, to a truncated parametric distribution, then estimate the 'missing' portion of the distribution, which corresponds to the undetected species in an assemblage. Fisher et al. (1943) pioneered this approach by fitting a geometric series to a large sample of moths captured at light traps. Relative incidence distributions from replicated sets of samples can be treated in the same way (Longino et al. 2002). The most widely used species abundance distribution for this approach is the log-normal (Fig. 4.3) and its variants (from Preston (1948) to Hubbell (2001)), but other distributions (geometric series, negative binomial, γ, exponential, inverse Guassian) have also been used. The challenges of fitting the log-normal have been widely discussed (e.g. Colwell & Coddington 1994; Chao 2004; Dornelas et al. 2006; Connolly et al. 2009). One of the limitations of this approach is shared with the extrapolation of fitted parametric functions: two or more species abundance distributions may fit the data equally well, but predict quite different assemblage richness. In addition, the species abundance distribution that fits best may be one that cannot be used to estimate undetected species, such as the widely used log-series distribution (Chao 2004).

The limitations of parametric methods inspired the development of non-parametric richness estimators, which require no assumptions about an underlying species abundance distribution and do not require the fitting of either *a priori* or *ad hoc* models (Chao 2004). These estimators have experienced a meteoric increase in usage in the past two decades, as species richness has become a focus of biodiversity surveys and conservation issues, and a subject of basic research on the causes and consequences of species richness in natural ecosystems. In Box 4.1, we have listed six of the most widely used and best-performing indices. All the estimators in Box 4.1 depend on a fundamental principle discovered during World War II by Alan Turing and I.J. Good (as reported by Good (1953, 2000)), while cracking the military codes of the German Wehrmacht Enigma coding machine: the abundances of the very rarest species or their frequencies in a sample or set of samples can be used to estimate the frequencies of undetected species. All of the estimators in Box 4.1 correct the observed richness S_{obs} by adding a term based on the number of species represented in a single abundance sample by only one individual (*singletons*), by two (*doubletons*), or by a few individuals. For incidence data, the added term is based on the frequencies of species represented in only one (*uniques*) sample, in two (*duplicates*), or in a few replicate incidence samples.

Fig. 4.4 shows how well one of these estimators, Chao2, estimates the asymptotic richness of the seedbank dataset of Figure 4.1, based on sets of m^* samples chosen at random. The estimator stabilizes after about 30 samples have been pooled. When all 121 samples have been pooled, the estimator suggests that 1–2 additional species still remain undetected.

Only four of the estimators in Box 4.1 (Chao1, ACE, and the two individual-based jackknife estimators) are appropriate for abundance data; the rest require replicated incidence data. Most of the incidence-based estimators were first developed, in

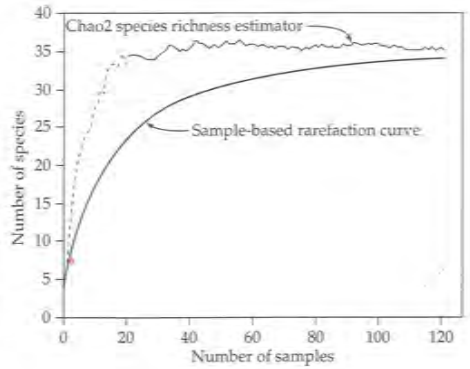

Figure 4.4 Asymptotic species richness estimated by the Chao2 non-parametric richness estimator for the seedbank dataset of Fig. 4.1. Plotted values for Chao2 are means of 100 randomizations of sample order. The estimator stabilizes after only about 30 samples have been pooled. When all 121 samples have been pooled (34 species detected), the estimator suggests that one or two additional species still remain undetected.

biological applications, for capture–recapture methods of population size estimation. The number of samples that include Species X in a set of biodiversity samples corresponds to the number of recaptures of marked Individual X in a capture–recapture study. In species richness estimation, the full assemblage of species, including those species not detected in the set of samples (but susceptible to detection), corresponds, in population size estimation, to the total population size, including those individuals never captured (but susceptible to capture) (Boulinier et al. 1998; Chao 2001, 2004).

Behind the disarming simplicity of Chao1 and Chao2 lies a rigorous body of statistical theory demonstrating that both are robust estimators of *minimum* richness (Shen et al. 2003). ACE and ICE are based on estimating *sample coverage*—the proportion of assemblage richness represented by the species in a single abundance sample (ACE) or in a set of replicated incidence samples (ICE). The estimators are adjusted to the 'spread' of the empirical species abundance (or incidence) distribution by a coefficient of variation term (Chao 2004). The Chao1 and Chao2 estimators also provide a heuristic, intuitive 'stopping rule' for biodiversity sampling: no additional species are expected to be found when all species in the sample are represented by at least two individuals (or samples). Extending this approach, Chao et al. (2009) provide equations and simple spreadsheet software for calculating how many additional individuals would be needed to sample 100% (or any other percentage) of the asymptotic species richness of a region based on the samples already in hand. Pan et al. (2009) have recently extended the Chao1 and Chao2 indices to provide an estimate of the number of shared species in multiple assemblages.

The jackknife is a general statistical technique for reducing the bias of an estimator by removing subsets of the data and recalculating the estimator with the reduced sample. In this application of the technique, the observed number of species is a biased (under-) estimator of the complete assemblage richness (Burnham & Overton 1979; Heltshe & Forrester 1983; Chao 2004). For a set of m replicate incidence samples, the kth order jackknife reduces the bias by estimating richness from all sets of m–k samples. The first-order jackknife (Jackknife1) thus depends only on the uniques (species found in only one sample) because the richness estimate is changed only when a sample that contains one of these species is deleted from a subset of samples. Likewise, the second-order jackknife (Jackknife2) depends only on the uniques and the duplicates (species found in exactly two samples). Similar expressions for abundance-based jackknife estimators are based on the number of singletons (species represented by exactly one individual) and doubletons (species represented by exactly two individuals; Burnham & Overton (1979)). These estimators can be derived by

letting the number of samples m tend to infinity in the equations for the incidence-based estimators.

4.2.9 Comparing estimators of asymptotic species richness

Given the diversity of asymptotic estimators that have been proposed, which one(s) should ecologists use with their data? The ideal estimator would be *unbiased* (it neither over- or under-estimates asymptotic species richness), *precise* (replicates samples from the same assemblage produce similar estimates), and *efficient* (a relatively small number of individuals or samples is needed). Although there are many ways to estimate bias, precision, and efficiency (Walther & Moore 2005), none of the available estimators meet all these criteria for all datasets. Most estimators are biased because they chronically under-estimate true diversity (O'Hara 2005). The Chao1 estimator was formally derived as a *minimum* asymptotic estimator (Chao 1984), but all of the estimators should be treated as estimating the lower bound on species richness. Estimators of asymptotic species richness are often imprecise because they typically have large variances and confidence intervals, especially for small data sets. This imprecision is inevitable because, by necessity, these estimators represent an extrapolation beyond the limits of the data. In contrast, rarefaction estimators usually have smaller variances because they are interpolated within the range of the observed data. However, as noted earlier, the unconditional variance of richness as estimated by rarefaction is always larger than the variance that is conditional on a single sample (or set of samples). Finally, most estimators are not efficient and often exhibit 'sampling creep': the estimated asymptote itself increases with sample size, suggesting that the sample size is not large enough for the estimate to stabilize (e.g. Longino et al. (2002)).

Two strategies are possible to compare the performance of different estimators. The first strategy is to use data from a small area that has been exhaustively sampled (or nearly so), and to define that assemblage as the sampling universe. As in rarefaction, a random subsample of these data can then be used to calculate asymptotic estimators and compare them to the known richness in the plot (a method first suggested by Pielou (1975), but popularized by Colwell & Coddington (1994)). For example, Butler & Chazdon (1998) collected seeds from 121 soils samples from a 1 ha plot, on a 10 × 10 m grid in tropical rainforest in Costa Rica, yielding 952 individual seedlings representing a total of 34 tree species (Figure 4.1). Colwell & Coddington (1994) randomly rarefied these data, by repeatedly pooling m^* samples ($1 \leq m^* \leq M$), and found that the Chao2 index (illustrated in Fig. 4.4) and the second-order jackknife estimators were least biased for small m^*, followed by the first-order jackknife and the Michaelis–Menten estimator. Walther & Morand (1998) used a similar approach with nine parasite data sets and found that Chao2 and the first-order jackknife performed best. Walther & Moore (2005), using different quantitative measures of bias, precision, and accuracy, compiled the results of 14 studies that compared estimator performance, and concluded that, for most data sets, non-parametric estimators (mostly the Chao and jackknife estimators) performed better than extrapolated asymptotic functions or other parametric estimators.

In a second strategy for comparing diversity estimators, the investigator specifies the true species richness, the pattern of relative abundance, and the spatial pattern of individuals in a computer-simulated landscape. The program then randomly samples individuals or plots, just as an ecologist would do in a field survey. The estimators are then calculated and compared on the basis of their ability to estimate the 'true' species richness of the region. This kind of simulation can also be used to explore the effects of spatial aggregation and segregation, sampling efficiency, and the size and placement of sampling plots. Brose et al. (2003) carried out the most extensive analysis of this kind to date. In their analyses, which estimator performed best depended on the relative evenness of the rank abundance distribution, the sampling intensity, and the true species richness. As in the empirical surveys (Walther & Moore 2005), non-parametric estimators performed better in these model assemblages than extrapolated asymptotic curves (parametric estimators based on truncated distributions were not considered). One encouraging result was that environmental gradients and spatial autocorrelation (which characterize all biodiversity data at some spatial scales) did not have a serious effect on the performance of the estimators. These results are consistent with the findings of Hortal et al. (2006), who aggregated empirical data sets at different spatial grains and found that non-parametric estimators were not greatly affected by the spatial scale of the sampling.

O'Hara (2005) took a hybrid approach that used both empirical data and simulated assemblages. He first fit negative binomial and Poisson log-normal distributions to two very extensive (but incomplete) sets of survey data for moths. He used these fitted models to generate sample data for comparing non-parametric estimators, parametric estimators, and extrapolated asymptotic curves. As in other studies, true species richness was greater than predicted by the estimators. In each comparison, only one of the parametric estimators had a 95% confidence interval that encompassed the true richness. The catch is that this method worked well only when the 'correct' species abundance distribution was used. In other words, the investigator would need to know ahead of time that the negative binomial, Poisson log-normal, or some other distribution was the correct one to use (which rather defeats the value of using non-parametric estimators). Unfortunately, in spite of decades of research on this topic, there is still no agreement on a general underlying form of the species abundance distribution, and there are difficult issues in the fitting and estimation of these distributions from species abundance data (see Chapter 10). We hope that future work may lead to better species richness estimators. At this time, the non-parametric estimators still give the best performance in empirical comparisons, and they are also simple, intuitive, and relatively easy to use.

4.2.10 Software for estimating species richness from sample data

Free software packages with tools for estimating species richness from sample data include:

- *EstimateS* (Colwell 2009): http://purl.oclc.org/estimates
- *EcoSim* (Gotelli & Entsminger 2009): http://garyentsminger.com/ecosim/index.htm
- *SPADE*: http://chao.stat.nthu.edu.tw/software CE.html
- *VEGAN* (for *R*): http://cc.oulu.fi/~jarioksa/softhelp/vegan.html.

4.3 Prospectus

Estimates of species richness require special statistical procedures to account for differences in sampling effort and abundance. For comparing species richness among different assemblages, we recommend sample-based rarefaction using unconditional variances, with adjustments for the number of individuals sampled. Rarefaction methods for data that represent sampling from nature without replacement are still needed, for small assemblages, as are additional estimators for the number of shared species in multiple samples (A. Chao, personal communication). For many datasets, all existing methods for estimating undetected species seem to substantially under-estimate the number of species present, but the best methods nonetheless reduce the inherent undersampling bias in observed species counts. Non-parametric estimators (e.g. Chao1, Chao2) perform best in empirical comparisons and benchmark surveys, and have a more rigorous framework of sampling theory than parametric estimators or curve extrapolations.

4.4 Key points

1. Biodiversity sampling is a labour-intensive activity, and sampling is often not sufficient to detect all or even most of the species present in an assemblage.
2. Species richness counts are highly sensitive to the number of individuals sampled, and to the number, size, and spatial arrangement of samples.
3. Sensitivity to sampling effort cannot be accounted for by scaling species richness as a ratio of species counts to individuals, samples, or any other measure of effort.
4. Sample-based and individual-based rarefaction methods allow for the meaningful comparison of diversity samples based on equivalent numbers of individuals and samples.

5. Non-parametric estimators of species richness, which use information on the rare species in an assemblage to adjust for the number species present but not detected, are the most promising avenue for estimating the minimum number of species in the assemblage.

Acknowledgements

N.J.G. acknowledges US National Science Foundation grants DEB-0107403 and DEB 05-41936 for support of modelling and null model research. R.K.C. was supported by NSF DEB-0072702.

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.