CHAPTER 9

Measuring Species Diversity

Introduction

A great deal of research in community ecology documents patterns of species diversity and how these patterns differ in time and space (Rosenzweig 1995). These data are used to study how communities are assembled, and to address the accelerating loss of biodiversity in the face of the ever expanding human population. For example, at large spatial scales, more species of plants and animals can be found in the Tropics than in temperate latitudes (Hillebrand 2004). At small spatial scales, species richness and evenness often differ between locations with high versus low nutrient inputs (Huston 1980), and between locations with and without top predators (Sergio et al. 2005). Dozens of mechanisms have been proposed for these kinds of patterns (Rohde 1992).

Most ecology textbooks provide a catalog of species diversity patterns, a set of hypotheses to account for these patterns, and individual case studies that illustrate the patterns and test the mechanisms. This chapter will address a more basic question: How do we quantify species diversity? Before we can study something with the scientific method, we have to be able to quantify it, and that turns out to be surprisingly difficult to do in biodiversity studies.

A WALK THROUGH THE WOODS

Suppose we walk through a New England woodland and randomly sample 100 ants from the forest floor and low vegetation (Ellison et al. 2007). We preserve each specimen we collect in a small vial of ethanol, which includes a label listing the date of collection; the habitat; and the latitude, longitude, and elevation of the study site (which we would measure from a GPS unit or from a topographic map; see Agosti et al. 2000 for protocols). Back in the laboratory, we examine the specimens under the dissecting microscope and identify them each to species, using published guidebooks and taxonomic keys (Coovert 2006, Fisher and Cover 2007) as well as Web-based resources (antbase.org; Agosti and Johnston 2005). Because species identification is often a difficult task for an ecologist (Gotelli 2004), we might spend a few days working at a nearby natural history museum, where we could compare our specimens with a large reference collection of previously identified species and consult with taxonomic specialists who are experts in ant identification.

When we have finished our identifications, these data can be organized into a table in which each row is a species. The first column gives the species and genus name, and the second column gives the number of individuals collected (Table 9.1). These data can be plotted in a **rank abundance graph**. We first order the species from most abundant to least abundant, and put them in rank order on the x axis of the graph. We then draw a bar to represent each

Table 9.1 Hypothetical biodiversity data. This table represents the simplest kind of data collected in a biodiversity survey. Although this is a hypothetical data set, the species list and rank order were those found in a survey of ant species in a New York oak forest (see Ellison et al. 2007).

Ant species	*Number of individuals collected*
Aphaenogaster rudis	45
Formica neogagates	32
Myrmica punctiventris	12
Myrmica sculptilis	3
Formica subsericea	2
Stenamma impar	2
Temnothorax longispinosus	1
Lasius alienus	1
Lasius umbratus	1
Prenolepis imparis	1

species. The height of the bar is the abundance of each species, which is measured on the y axis (Figure 9.1).

This rank abundance graph contains all of the essential biodiversity information, and allows us to see the two components of biodiversity, species rich-

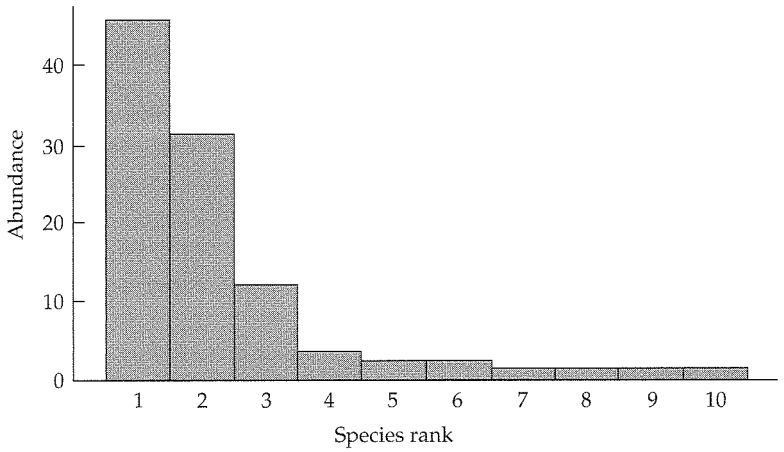

Figure 9.1 Rank abundance graph of the data in Table 9.1. Each bar represents a different species. The height of each bar is the abundance of the species in the sample.

ness and species evenness. **Species richness** refers to the number of species in the sample, which is simply the number of bars in the graph (10 in this example). **Species evenness** refers to the relative heights of the bars. In a maximally even sample, all species would be represented by the same abundance level. In this example, a maximally even distribution would be one in which each of the 10 species were represented by exactly 10 individuals each. A maximally uneven distribution would have the first species represented by 91 individuals and the remaining 9 species represented by 1 individual each. Both distributions contain 10 species and 100 individuals, but the first collection seems more diverse because the species evenness is higher.

Most real samples fall somewhere between these extremes of evenness. In this example, the two most abundant species are represented by 45 and 32 individuals, and the four most rare species are each represented by one individual. This is a typical pattern: most samples are dominated by a small number of species that comprise the bulk of the abundance or biomass of the assemblage. The other species in the assemblage are much rarer—some species may be represented by only one or two individuals each. The resulting rank abundance graph often has a long right-hand "tail" that reflects the abundances of the rarer species in the assemblage (McGill et al. 2007). As we will see later in this chapter, these rare species contain important information about "missing" species: species that are present in the assemblage but are so uncommon that they did not show up in our particular sample.

The Organization of Biodiversity Data

In order to develop a quantitative framework for describing biodiversity, we need a conceptual model for how diversity is organized in nature. Our model will be that a community consists of discrete, individual organisms. Each organism can be recognized as belonging to a distinct "type," usually a species. Alternatively, individuals could be grouped into higher taxonomic units, such as genera or families. This is a common practice for fossil assemblages, where it is not always possible to identify specimens to the species level. We can also quantify biodiversity at taxonomic levels other than the individual and the species. For example, instead of counting individuals, we could count species and classify them into genera or other taxonomic units. Other classification schemes that are not based on evolutionary relationships can also be used, such as trophic status (producer, herbivore), functional status (grazer, filter feeder), or growth form (shrub, tree). Whatever the grouping, each individual is recognized and counted as a distinct entity and is assigned unambiguously to only one of the categories in the classification scheme.

To introduce some quantitative notation, suppose our sample consists of N individuals that represent S species. We rank these species in order of abundance from $i = 1$ (the most abundant) to S (the least abundant). Let n_i = the abundance of the ranked species i. Note that:

$$N = \sum_{i=1}^{S} n_i \qquad \text{Equation 9.1}$$

We also define the proportion p_i of the total assemblage that is represented by species i as:

$$p_i = \frac{n_i}{N} \qquad \text{Equation 9.2}$$

Because these p_i values are proportions of the total N, they sum to 1.0:

$$1.0 = \sum_{i=1}^{S} p_i \qquad \text{Expression 9.1}$$

This framework of recognizing, classifying, and counting individuals works well for most sexually reproducing animal species. However, there are many plant and invertebrate species, such as grasses and corals, that grow and reproduce asexually, as clones or colonies. For these kinds of organisms, we cannot count discrete "individuals," so species diversity is usually quantified by measuring the percent cover or the biomass of each species in a sample. Unfortunately, we cannot use percent cover or biomass data for the models in this chapter, which are based on the sampling properties of discrete entities (i.e., individuals). Near the end of this chapter, we will discuss a method for quantifying biodiversity that can be used with data that do not represent counts of individual organisms.

THE CANDY JAR OF DIVERSITY

A good analogy for quantifying biodiversity is to think of the environment or habitat that contains the community as a large candy jar full of multicolored jelly beans (Longino et al. 2002). Each jelly bean represents an individual organism, and the different colors represent the different species present in the community. The complete information about biodiversity in this community would then consist of the total number of jelly beans, the total number of jelly bean colors, and the number of jelly beans of each color in the jar. These quantities would correspond to the total number of individuals, the total number of species, and the relative abundance of each species in the community.

Unfortunately, nature's candy jar is so large we can never count everything that is inside it. Instead, we have to make inferences about what the entire jar

contains based on a sample of jelly beans taken from the jar. When we compare two communities, we are effectively drawing a handful of jelly beans from each of two jars and trying to describe how the contents of the two jars differ based on our analysis of the samples.

We face two problems in trying to sample biodiversity from the candy jar. The first problem is that the more jelly beans (individuals) we sample, the more colors (species) we will find; unless we are very careful about the counts of individuals that are used to estimate species number, we can be led astray. In particular, we expect to find more species in a large sample of individuals than in a small sample. Therefore, if two communities differ in the number of individuals that are sampled from each, we cannot necessarily conclude that the larger sample has more species in it.

The second problem is that the relative abundance of different species in an assemblage is rarely perfectly even—that is, we almost never find that all the resident species in a community have equal abundances. In most communities, there is a small number of species that are very common and a large number of species that are uncommon or rare. In other words, the candy jar may contain a few colors that are relatively common, but most colors may be represented by only a few jelly beans (or even a single jelly bean). If we sample only a small handful of jelly beans, we will mostly get the common colors. If we want to collect the rare colors, we will have to take a larger sample (or more samples).

CARABID BEETLES IN PINE PLANTATIONS

A case study from northern Europe will illustrate these issues more clearly. Niemelä et al. (1988) sampled carabid beetles with a set of pitfall traps placed in young (< 20 years old) pine plantations and in old (20–60 years old) pine plantations. The first three columns of Table 9.2 give the number of individuals of each beetle species collected in the two habitats.

There are several interesting patterns in these data. First, in both habitats there are a number of rare species, represented by only one or two individuals. As we will see, these **singletons** and **doubletons** are important numbers for estimating total species richness in the assemblage. At the same time, there are a handful of very common species. In, the young-plantations data, the four most common species (*Calathus micropterus, Trechus secalis, Pterostichus oblongopunctatus,* and *Pterostichus adstrictus*) accounted for 49% of the 243 individuals in the sample $\{100 \times [(48 + 30 + 23 + 23)/243]\}$. In the old-plantations data, the two most common species (*Calathus micropterus* and *Trechus secalis*) accounted for 70% of the 63 individuals sampled $\{100 \times [(29 + 15)/63]\}$.

A total of 31 species was collected in the young plantations, whereas only 9 species were collected in the old plantations; thus it appears that young

Table 9.2 Number of individuals of carabid beetle species captured in pitfall traps in young and old pine plantations in northern Europe. A blank space indicates that zero individuals of that species were collected. The third column gives the results for a random draw of 63 individuals from the young-plantations sample. (Primary data from Niemelä et al. 1988.)

Beetle species	*Young plantations*	*Old plantations*	*Random draw from young-plantations sample*
Calathus micropterus	48	29	11
Pterostichus oblongopunctatus	23	9	6
Notiophilus biguttatus	3	1	
Carabus hortensis	2		1
Carabus glabratus	15	1	6
Cychrus caraboides	6	1	2
Amara brunnea	2		1
Trechus secalis	30	15	13
Leistus terminatus	3	1	1
Amara familiaris	1		
Amara lunicollis	7		
Bembidion gilvipes	2		1
Bradycellus caucasicus	1		
Calathus melanocephalus	3		1
Carabus nitens	1		
Carabus violaceus	1		
Cicindela sylvatica	10		2
Cymindus vaporariorum	3		1
Harpalus quadripunctatus	7		
Harpalus sp.	1		1
Leistus ferrugineus	1		
Miscodera arctica	13		5
Notiophilus aestuans	2		2
Notiophilus germinyi	9		1
Notiophilus palustris	9		1
Pterostichus adstrictus	23		5
Pterostichus cupreus	1		1
Pterostichus diligens	1	2	
Pterostichus niger	7		
Pterostichus strenuus	4	4	1
Synuchus vivalis	4		
Total individuals (abundance)	243	63	63
Total species (species richness)	31	9	20

plantations support more beetle species. But is this conclusion valid? The 31 young-plantation species came from a sample of 243 individuals, where as the 9 old-plantation species came from a sample of only 63 individuals. So is it really surprising that we should find more species in the young plantations, from which we collected almost four times as many individuals?

Notice also that the species composition of the old plantations is a *nested subset* of the young plantation sample: every species represented in the old plantations is also present in the young plantations. However, there are many species found in the young plantations that were not collected from the old plantations.

These patterns suggest a simple null hypothesis: perhaps total species richness does not differ between young and old pine plantations. Instead, the difference in observed species richness between the two samples might simply reflect the different number of individuals collected in the two habitats. In other words, the old-plantations data might represent "a smaller handful of jelly beans drawn from the same jar." How can we test this null hypothesis?

Rarefaction

The candy-jar sampling analogy suggests a straightforward way to compare diversity in the two plantations. Suppose we had sampled only 63 individuals each from both the old and young pine plantations. Then the comparison of species richness would be valid because the sampling effort would be identical for both assemblages; any observed differences in species richness would not be confounded by differences in abundance. We cannot go back to the pine plantations and sample again, but we can randomly draw 63 individuals from the young-plantations data and see how many species are present. Note that we are effectively taking a random *subsample* of data that are themselves a small sample of the community.

We could fill a real candy jar with 243 colored jelly beans to match the data in Table 9.2, and then randomly draw a handful of 63 jelly beans and see how many colors (species) we get. Or, we can write a short computer program to do the same thing. The program conducts a random draw from the list of individuals, **sampling without replacement** (meaning that once a particular individual is drawn from the list, that same individual cannot be drawn a second time). Column four of Table 9.2 illustrates what we get from a single random subsample of 63 individuals from the young-plantations data. This subsample contains 20 species—certainly less than the 31 species observed in the original sample, but still quite a few more than the 9 species observed for the sample of 63 individuals from the old plantations. This method of drawing a random subsample and generating from that subsample the

expected species richness (or any other diversity statistic) is called **rarefaction** (from the word *rarefy*, which means "to thin" or "to make less dense").

Of course, a different random subsample would contain a different mix of individuals and therefore may contain a different number of species. If we repeat the sampling process 1000 times, we generate a histogram of species-richness values, shown in Figure 9.2. Notice that, although we are sampling each individual without replacement, all those individuals are returned to the jar when we begin the next random subsample (**sampling with replacement**). The species richness values in Figure 9.2 range from 13 to 28. In other words, out of 1000 simulations of a random draw of 63 individuals, one of the simulations contained only 13 species and another contained 28 species. Thus, the single simulation illustrated in the fourth column of Table 9.2 was fairly typical, with 20 species represented. The average of the 1000 simulations was 19.92 species, and 95% of the random draws generated between 16 and 24 species.

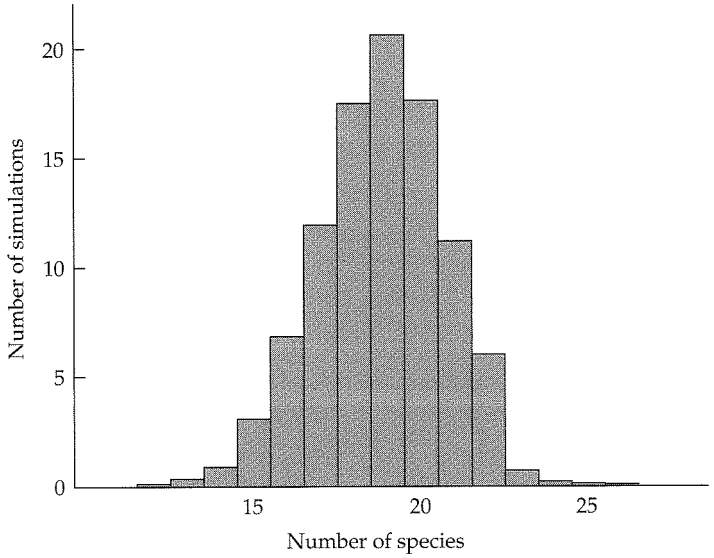

Figure 9.2 Histogram of 1000 random subsamples of species richness. A total of 63 individuals were randomly drawn from the list of beetle species for the young-plantations data in column 2 of Table 9.2. Column 4 of Table 9.2 illustrates the results of one such random draw. The histogram shows the distribution of species richness values from these simulations. The range was from 13 to 28 species. The average was 19.92 species, with a 95% confidence interval of 16 to 24 species. Notice that all these values are considerably greater than the 9 species represented by the 63 individuals sampled from the old plantations (column 3 of Table 9.2).

Notice that this **95% confidence interval** (16 to 24) does not include the number 9, which was the number of species observed in the sample of 63 individuals from the old plantations. In other words, a random draw of individuals from the young plantations is very unlikely to have generated a sample with only 9 species. Therefore, when the old and new plantations are compared using samples of 63 individuals for both, there are significantly more species in the young plantations. The null hypothesis (that species richness does not differ between the two plantation types) can be rejected with these data.

EXTENDING THE RAREFACTION CURVE

In the beetle example, we rarefied the original young-plantations sample of 243 individuals down to a subsample of 63 so that we could make a valid comparison with the old-plantations data. However, a complete **rarefaction curve** can be extended down across the entire range of abundances. In this case, the rarefaction curve is drawn as a graph with abundance (number of individuals) on the x axis and S (number of species) on the y axis.

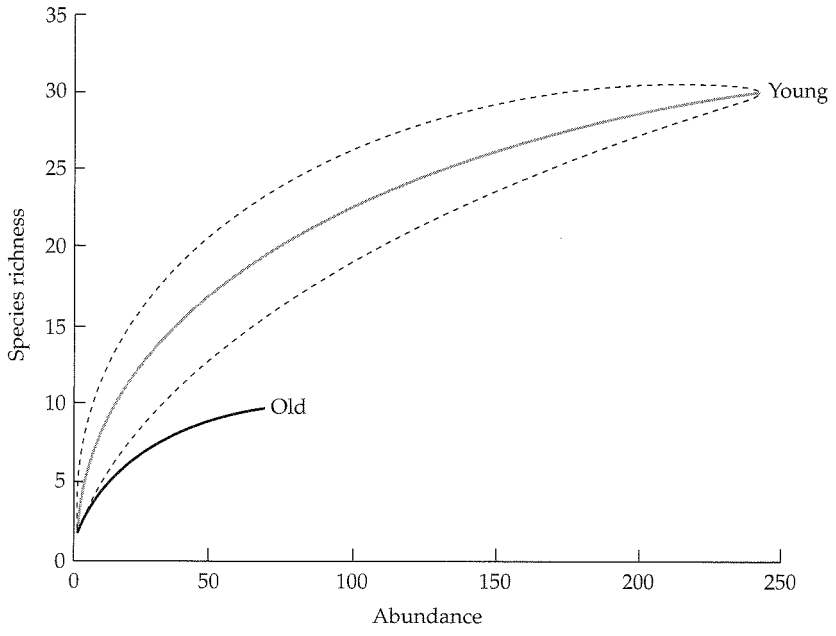

Figure 9.3 Rarefaction curves for the carabid beetle data in Table 9.2. The upper curve is the rarefaction of the young-plantations data; the lower curve is the rarefaction of the old-plantations data. For the young plantations data, the 95% confidence interval is shown as a dashed line. (Adapted from Gotelli and Graves 1996.)

Even without a computer simulation, we already know the location of two points in this curve. The first point is the observed sample itself, which contains S species and N individuals and is plotted somewhere in the upper right hand corner of the graph. The second point is close to the axis. Regardless of the biodiversity data or the classification scheme, if we draw only a single individual, we will count only one species. Therefore, the rarefaction curve must end at the point [1,1] (representing 1 individual on the x axis and 1 species on the y axis) in the lower left-hand corner of the graph. Between these two extremes, the rarefaction curve is constructed by drawing subsamples from the original sample and extending the curve from the right to the left by interpolation.

Figure 9.3 shows the rarefaction curves for both the young and the old plantations. This figure was constructed by computer simulation of randomly drawing a specified number of individuals from the two original samples in Table 9.2. For the young-plantations data, the curve bracketing the 95% confidence interval is also illustrated.

ASSUMPTIONS OF RAREFACTION

The rarefaction model entails a number of assumptions if it is going to be used to compare samples from different communities.

- ✔ ***The underlying community is constant, and its membership is closed.*** In other words, we assume that "the candy jar is closed," so that its contents are not greatly influenced by individuals migrating in from other habitats. As we will see, however, there are some assemblages where the candy jar appears to "leak" and the total species richness in the jar does not stay fixed from one time period to the next.
- ✔ ***Sampling has been sufficient to allow for a useful comparison of two or more communities.*** Because all rarefaction curves converge at small sample sizes to the point [1,1], we may not be able to distinguish different curves easily unless we have sampled a sufficient number of individuals. In practice, it is difficult to determine whether sampling has been sufficient because we do not know how close our original sample is to the asymptote of the rarefaction curve. Asymptotic species richness estimators (described later in this chapter) can be useful for this purpose.
- ✔ ***Individuals in different communities have been sampled with identical methods.*** All sampling methods have their biases. Therefore we cannot, for example, validly compare insect diversity samples from pitfall traps and sticky traps. However, we could use rarefaction to compare the sampling efficiency of pitfall traps and sticky traps used at the same site

(e.g., Ellison et al. 2007). We could also compare diversity in two sites by pooling data from pitfall traps and sticky traps, but only if both sampling methods were used in the same standardized way in both sites.

- ✔ ***The spatial distribution of individuals is random.*** If the individuals of each species are clumped or aggregated in their occurrence, then individual-based rarefaction will overestimate the number of species that would be found in a random sample taken from a small area.
- ✔ ***Individuals are sampled independently and at random.*** This assumption is probably the one that is violated most often in sampling studies. The unit of sampling is rarely a single individual. Instead, we use pitfall traps, plot surveys, fish seines, timed counts, soil cores, or other sampling units that capture multiple individuals but are themselves independent of one another. As we will see, the rarefaction model can be extended to use with sample-based data, which can also help to overcome the effects of clumping of individual species occurrences at small spatial scales (assumption #4).
- ✔ ***The samples to be compared are taxonomically "similar."*** If two samples have identical rarefaction curves but none of the species in the two samples are the same, we would not necessarily want to conclude that they were drawn from the same assemblage. Rarefaction estimates species richness and takes into account the relative evenness of species, but it does not directly quantify or compare the composition of two communities. Similarity indices, which also have sampling distributions, can be used to directly assess the composition of two communities via the expected and observed number of shared species (Chao et al. 2005).

INDIVIDUAL- AND SAMPLE-BASED RAREFACTION

The examples in this chapter describe the biodiversity sampling as a random selection of individual organisms, and the statistics described here are based on this model of random sampling of individuals. In reality, biodiversity data are rarely sampled this way. The unit of study is usually a pitfall trap, quadrat, point count, trawl sample, food bait, or some other sampling unit that attracts or captures multiple individuals.

The problem is that the sample represents the independent statistical unit of replication, but it is the individuals within the sample that contain the information on biodiversity (Gotelli and Colwell 2001). The examples in this chapter describe **individual-based rarefaction**, in which individuals are randomly chosen from a single large sample. However, data more typically consist of a collection of many samples, each of which contains multiple indi-

viduals and species. For such data, it is possible to conduct **sample-based rarefaction**, in which a random subset of entire samples is drawn from the data, and from these randomly selected samples we calculate the number of species observed.

In sample-based rarefaction, we often do not count the individual organisms in a sample, but simply record whether or not a given species is present in that sample. Sample-base rarefaction also results in a rarefaction curve, but the x axis of the curve is the number of samples, not the number of individuals. Also, the origin of the sample-based rarefaction curve may no longer occur at the point [1,1] because a single sample will usually contain more than a single species.

Sample-based rarefaction uses the sample as the proper unit of replication, but there is still a potential problem of differences in underlying abundance. Suppose we compare two habitats with sample-based rarefaction curves and estimate the number of species expected for, say, 100 samples. If the number of individuals per sample is much greater in one of the habitats than the other, the sample-based rarefaction still does not allow for a proper comparison of diversity in the two regions because they differ in their abundance levels. The solution is to rescale the x axis of the sample-based rarefaction curve from samples back to individual abundances (Gotelli and Colwell 2001). Thus, the rarefaction curves will potentially be shifted to the left or to the right, depending on the average number of individuals that would be expected for a given number of samples.

Incidentally, when using sample-based rarefaction, we can relax the assumption that the spatial distribution of individuals is random (assumption #4). The samples preserve any inherent heterogeneity in spatial distributions, so that diversity per sample is properly estimated. In the remainder of this chapter, we will focus on individual-based rarefaction. See Colwell et al. (2004) for more details on sample-based rarefaction.

CALCULATING THE RAREFACTION CURVE AND ITS VARIANCE

When we calculate a rarefaction curve, we want to know the average or expected species richness for a given number of individuals. As we saw, we expect some variation among the random samples, so there is also a variance and a confidence interval that can be associated with the average.

The rarefaction curve and its variance can be calculated (or estimated) in three different ways. First, we can use a computer to repeatedly simulate random draws of individuals from the original collection, as we described earlier for the beetle data.

A second method for calculating the rarefaction curve is to use statistical sampling theory, and directly calculate the expected number of species and

its variance, without resorting to a simulation (Heck et al. 1975). This calculation is based on statistical equations that specify the number of permutations of the data that are possible with a sample of a given size.

Finally, a very good approximation to the rarefaction curve can be calculated from the following equation (Brewer and Williamson 1994):

$$E(S_m) \approx S - \sum_{i=1}^{S} (1 - p_i)^N \qquad \text{Equation 9.3}$$

where S is the number of species, N is the total number of individuals, and p_i is the proportion of individuals of species i in the original sample. $E(S_m)$ is the expected number of species to be found in a subsample of m individuals drawn randomly from the original collection of N individuals.

Carefully compare Equation 9.3 to Equation 7.6 in the Island Biogeography chapter. In that chapter, we developed a *passive sampling model* (Coleman et al. 1982) to predict the number of species on an island of a certain size. In the passive sampling model (see p. 170), individuals were seen as analogous to darts, species were analogous to the colors of the darts' feathers, and islands were analogous to targets that were randomly hit by the darts. Each island randomly "sampled" a fraction of the total number of individuals in the archipelago. Equation 9.3 represents a parallel derivation based on the proportion of the total number of individuals contained in the subsample. In fact, we could say that the passive sampling model for islands is simply a rarefaction curve constructed from a pooled list of all the individuals and species that are present in the archipelago. The number of individuals "sampled" by the island is proportional to the relative area of the island.

It is beyond the scope of this textbook to derive or present equations for the variance of the rarefaction curve. However, we would note that in individual-based rarefaction, the variance (and the confidence interval) converges to 0.0 at the maximum and at the minimum abundance. At the minimum abundance of 1 individual, we will always count exactly one species, so there is no variance. At the maximum abundance of the original number of individuals S, we will always count exactly S species, so the variance will also be zero at that point (see Figure 9.3).

However, the observed collection is itself a random sample from a much larger community. If we were to take a second sample of N individuals from the same community, we would not necessarily obtain S species again. Therefore, the variance at the largest sample should not converge to zero. Colwell et al. (2004) derived variance estimators for sample-based rarefaction that take into account the estimated variability of sampling from a much larg-

er statistical universe. Software packages available for free on the internet can be used to calculate individual- and sample-based rarefaction curves and their variances (Colwell 2004, Gotelli and Entsminger 2007).

Species Richness and Species Density

When ecologists discuss "species richness," in fact they are usually referring to **species density**—the number of species collected per a standardized sampling effort (Simpson 1964). This standardized sampling effort might be defined by time (hours spent searching), space (area of a quadrat censused), or some other arbitrary but consistent sampling unit (e.g., number of species per trap/time).

We prefer to think of species richness as the expected number of species for a given number of individuals. In other words, we are referring to species richness as a standardization based on the number of individuals sampled. The distinction is important because species density actually has two components: species richness, and the *total density* or *abundance* of organisms in the sampling area (James and Wamer 1982). The relationship can be seen by noticing that the units of these terms cancel out as follows:

$$\begin{array}{lcccc} \text{DIVERSITY MEASURES} & \textit{Species density} & = & \textit{Species richness} & \times \quad \textit{Total density} \\ \text{UNITS} & \dfrac{\textit{Species}}{\textit{Area}} & = & \dfrac{\textit{Species}}{\textit{Individuals}} & \times \dfrac{\textit{Individuals}}{\textit{Area}} \end{array} \qquad \text{Expression 9.2}$$

Thus, species density is affected both by species richness (number of species per standardized number of individuals) and by the total density of organisms in the study area (individuals per unit area or other sampling unit). Although most ecologists are interested in species density, the total density of organisms determines the number of individuals sampled and potentially can have a big effect on measured species density.

The challenge is that the total density of organisms is affected by the sampling design the investigator uses. Great care must be taken when comparing the diversity between areas that were not sampled with equal effort. Even when sampling is directly comparable, total density can be affected by local sampling conditions such as such as air temperature or cloud cover, which influence the movement of animals and therefore the sampling efficiency of traps. Finally, total density can also be affected by important biological factors. For example, habitats with high nutrient inputs may have greater biomass and abundance than habitats with low nutrient inputs. Rarefaction is an important analytical method for biodiversity studies because it separates species density into the components of species richness and total density.

Asymptotic Species Richness Estimators

Rarefaction is a method standardizing species richness data through **interpolation**: we rarefy the observed data to determine how many species would have been found in a smaller sample. However, an equally important question is, How many total species are present in the study area (Colwell and Coddington 1994)? If we could sample enough individuals, we would eventually reach an asymptote in species richness, and further sampling would find no more new species. To estimate this asymptote, we can use **extrapolation** to make inferences beyond the limits of our sample.

Asymptotic estimators use information in a biodiversity sample to estimate the number of undetected species that are present in the community but missing from our sample. In other words, we use the information on the numbers and colors of jelly beans in a small handful to estimate the number of different colors in the entire jar.

The *Chao1* index (Chao 1984) is a particularly powerful and simple asymptotic estimator that uses information on the frequency of rare species in the sample to estimate the minimum number of missing species. Suppose we have a collection of N individuals with an observed number of species S_{obs}. Let f_1 = the number of singletons (species that are represented by exactly one individual in the sample), and let f_2 = the number of doubletons (species represented by exactly two individuals). Intuitively, it would seem that the more of these "rare" species that are present in the sample, the more undiscovered species there are in the community. The *Chao1* estimate of the total number of species is:

$$S_{est} = S_{obs} + \left(\frac{(f_1)^2}{2f_2} \right) \qquad \text{Equation 9.4}$$

where S_{est} is the estimated number of species, S_{obs} is the observed number of species, and the ratio $(f1)^2/(2f_2)$ is the estimated number of undiscovered species. For example, for the carabid beetle data in the old pine plantations (Table 9.2, column 3), 9 species were observed, 4 of which were represented by only 1 individual ($f_1 = 4$) and 1 of which was represented by 2 individuals ($f_2 = 1$). Thus, the estimated total species richness is 17, with 8 undiscovered species. For a number of reasons, the *Chao1* index is a conservative estimate, so it should be interpreted as estimating the minimum number of undetected species. A modified form of this index is available for sample-based data, and variance and confidence intervals can also be calculated; see Colwell (2004) for details.

How much sampling effort is needed to reach S_{est}? Equations are now available to answer this question (Gotelli et al., in press), but Equation 9.4 immediately yields a heuristic "stopping rule": the asymptote will be reached when all species are represented by at least 2 individuals in the sample. From Equation 9.4, when the number of singletons (f_1) equals 0, $S_{est} = S_{obs}$ and there are no further undiscovered species. Unfortunately, to reach this goal may require a lot of extra sampling. When enough individuals are accumulated to eliminate the singletons in the initial sample, previously undetected species will start appearing. These new species will initially be represented by singletons, and it would take much additional sampling to capture more individuals of these species so that they become doubletons.

Even for well-sampled faunas, a doubling or more of the initial sample size may be necessary to capture all the rare species. And in some cases, an asymptote may never be reached. In the tropical rain forests of La Selva, Costa Rica, an intensive collecting program continues to yield new, undescribed species of ants and other arthropods after 10 years of collecting (Longino et al. 2002). The problem lies not with the particular diversity estimator, but with the sampling framework. Many of the rare species collected at La Selva are probably migrants or strays that are not breeding locally. In other words, the biodiversity candy jar is "leaky" or porous and the assumption of a fixed sampling pool in a closed environment (assumption #1) is not warranted. If the species pool is changing through time, the asymptotic estimators may also be inaccurate because species represented by singletons collected in early samples may never recur in later samples (Magurran 2007).

Species Evenness

Now we will consider the problem of how to estimate species evenness. Dozens of species diversity indices and graphical methods have been proposed (Magurran 2004), most of which make use of the p_i values from the rank abundance curve. Perhaps the most famous is the Shannon-Wiener index:

$$H' = -\sum_{i=1}^{S} p_i \ln(p_i) \qquad \text{Expression 9.3}$$

For a given number of species S and a given number of individuals N, the higher the index, the more even the relative abundance.

However, there are some challenges with using metrics such as the Shannon-Wiener index to quantify species evenness:

1. Most diversity indices are calculated using the observed p_i values. However, these p_i values are sensitive to the number of individuals censused. As an extreme example, regardless of the shape of the rank abundance distribution, if only one individual is sampled, $p_1 = 1.0$ and $p_i = 0.0$ for all other species. As more individuals are sampled, the estimated values for p_i change, because rare species begin to appear in large samples. Consequently, estimates for most diversity indices (including species richness) may be affected by the number of individuals sampled. Chao and Shen (2003) provide a correction for the Shannon-Wiener index that accounts for the undersampling of rare species.
2. Most of the indices do not have definable "units" of diversity that can be easily understood. Jost (2007) provides simple formulae for converting common diversity indices into units of "effective species."
3. Most of the indices cannot be related to a statistical sampling model.

We will use a simple diversity index, Hurlbert's (1971) **probability of an interspecific encounter** (***PIE***) that largely overcomes these problems. As we will see, this index also has an important theoretical relationship to the rarefaction curve.

The *PIE* index answers a simple question: what is the probability that two individuals randomly drawn from a sample represent two different species? At one extreme, if all the individuals in the sample are of the same species, then the answer must be $PIE = 0.0$, because there is no chance that we will ever draw an individual representing a second species. At the other extreme, if the relative abundances of all the species are maximally even and there are an infinitely large number of species, $PIE = 1.0$, because the next individual drawn will always represent a new species. Although 1.0 is the theoretical maximum value for *PIE*, real assemblages, having less than an infinite number of species, will always have a value of *PIE* that is less than 1.0.

We can use some simple laws of probability to derive a mathematical formula to estimate *PIE*. First, recall that the p_i values represent the fraction of the total number of individuals that consists of species i. We can also interpret p_i as an estimator of the probability that a single individual drawn from the sample represents species i. Therefore, the probability that two randomly drawn individuals, sampled with replacement, both represent species i is:

$$P(\text{two individuals are both species } i) = (p_i)(p_i) = (p_i)^2 \qquad \text{Expression 9.4}$$

However, this equation gives an answer for just one of the species. To determine the probability that both individuals are the same for any species in the assemblage, we must add these probabilities across all species:

$$P(\text{two individuals are the same species}) = \sum_{i=1}^{S} (p_i)^2 \qquad \text{Expression 9.5}$$

The probability that the two individuals represent two different species is:

$$P(\text{two individuals are not the same species}) = PIE = 1.0 - \sum_{i=1}^{S} (p_i)^2 \qquad \text{Expression 9.6}$$

Finally, we will multiply by a small bias-correction factor based on N, the total number of individuals in the sample:

$$PIE = \left(\frac{N}{N-1}\right)\left(1.0 - \sum_{i=1}^{S} (p_i)^2\right) \qquad \text{Equation 9.5}$$

Once the sample size is reasonably large ($N > 20$), the bias-correction factor has almost no effect on the calculated value of *PIE*.

With Equation 9.5, the calculated value of *PIE* for the young-plantations data in Table 9.2 is 0.92 versus 0.72 for the old plantations.

There are several reasons for preferring *PIE* to other diversity indices. First, *PIE* is measured in units of probability, so that differences or changes in diversity can be meaningfully interpreted. Second, *PIE*, unlike many other diversity indices, is insensitive to sample size. Figure 9.4 illustrates a rarefaction-style calculation, in which random subsamples of individuals were repeatedly drawn from the old- and young-plantations data, and the average *PIE* value calculated. You can see that the average does not change, although the estimates become more variable at smaller sample sizes.

Finally, *PIE* has an important conceptual link to the rarefaction curve: *PIE* measures the slope of the rarefaction curve at its origin (Olszewski 2004). Remember that the origin of the rarefaction curve is the point [1,1] (one individual and one species). The slope of this curve b is measured as:

$$b = \frac{\Delta Y}{\Delta X} = \frac{\Delta \text{ Species number}}{\Delta \text{ Number of individuals}} \qquad \text{Expression 9.7}$$

At the base of the rarefaction curve, the slope would represent the expected number of new species added as we add another individual. If the assemblage consisted of all individuals of one species, the rarefaction curve would be flat, and its slope would equal 0.0. As we described earlier, *PIE* also equals 0.0 when all the individuals are the same species because there is no chance that we will add a second species when we draw a second individual. At the other extreme, the steepest slope theoretically possible for the rarefaction curve is 1.0, which would happen if the second individual added always rep-

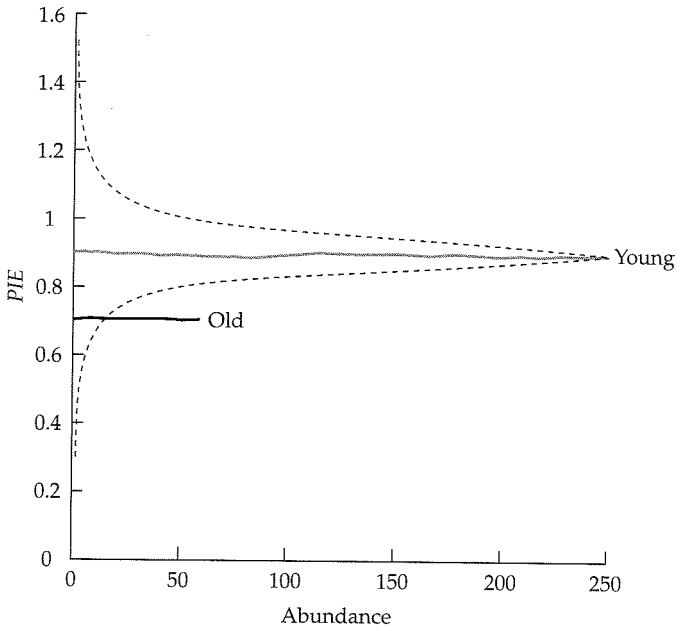

Figure 9.4 Rarefaction of the diversity index PIE, the probability of an interspecific encounter. This index (see Equation 9.5) measures the chances that two randomly drawn individuals represent two different species. Each line is the average of 1000 simulations. Unlike species richness, this index is not sample-size dependent, and the curves for the young- and old-plantations data do not change systematically as the number of individuals sampled is reduced. The dashed line indicates the theoretical 95% confidence interval for the young-plantations data. Although this theoretical confidence interval is drawn as a symmetric distribution, calculated values of PIE cannot exceed 1.0. (Adapted from Gotelli and Graves 1996.)

resented a new species. Similarly, for a maximally diverse community, *PIE* equals 1.0 because the probability that the next individual added is a new species is 1.0.

In the same way that the rarefaction curve and *PIE* are closely related, there is a relationship between species richness and all measures of species evenness. Species evenness influences the shape of the rarefaction curve, but the number of individuals and species sampled influences the relative evenness of the distribution. Although species richness and species evenness represent different components of biodiversity, they are not completely independent of one another.

There is a final bonus that comes from using *PIE* as an index of species evenness: it can be calculated for data that are expressed in units of percent

cover or biomass. Suppose that we survey the grasses in a 0.25 m^2 quadrat of grassland. We cannot count individual grass plants, but we can set up a grid of 100 points and record the species identity (or presence of bare soil) beneath each point. From these data, we could calculate:

$$p_i = \frac{\text{Number of points that touched species } i}{\sum_{i=1}^{S} \text{Number of points that touched species } i} \qquad \text{Expression 9.8}$$

Notice that if none of the points touch bare soil (i.e., there is a species present under every point), the denominator of Expression 9.8 equals 100. Using these p_i values, we can calculate the *PIE* index in the same way we did earlier.

Alternatively, suppose you have measured the biomass of each species rather than the abundance or percent cover. In this case:

$$p_i = \frac{\text{Biomass of species } i}{\sum_{i=1}^{S} \text{Biomass of species } i} \qquad \text{Expression 9.9}$$

However, because we are no longer randomly sampling individuals, the interpretations of *PIE* for percent cover or biomass data are subtly different: *PIE* now represents the probability that two randomly chosen points in space land on two different species. And *PIE* for the biomass example represents the probability that two randomly chosen tiny samples of animal (or plant) tissue represent two different species.

It is important to emphasize that the rarefaction model cannot be used unless we have sampled discrete individuals. However, even though we cannot calculate a rarefaction curve directly for percent cover and biomass data, we can calculate *PIE*, which is a diversity index that measures the slope of the rarefaction curve.

Summary

Quantifying and comparing patterns of species diversity is a basic activity of community ecologists. However, estimating species richness is challenging because many species have low abundances and are often undetected in biodiversity surveys. Moreover, the number and relative evenness of species is very sensitive to the number of individuals and samples collected. This chapter described three statistical tools that can help overcome these challenges.

1. Rarefaction, which standardizes comparisons of species richness to a common number of individuals.

2. The asymptotic estimator *Chao1*, which provides a minimum estimate of the number of species that were undetected in biological survey data.

3. *PIE*, the probability of an interspecific encounter, which measures the chances that two randomly drawn individuals from an assemblage represent two different species.

The statistical analysis of biodiversity data is a challenging and rapidly evolving field (Mao and Colwell 2005). Current and recent topics include the comparison of communities on the basis of similarity and species composition (Chao et al. 2005), estimation of sampling efforts needed to achieve asymptotic species richness (Gotelli et al., in press), and the partitioning of biodiversity into local and regional components (Jost 2007). Free software applications for the analysis of biodiversity data include EstimateS (Colwell 2004), SPADE (Chao and Shen 2006), and EcoSim (Gotelli and Entsminger 2007).

Problems

9.1 Using the ant data in Table 9.1, calculate *Chao1*, *PIE*, and the expected number of species for a random sample of 50 individuals.