[illegible]

[illegible], University of Vermont, Burlington, VT, USA
[illegible], National Tsing Hua University, Hsin-Chu, Taiwan

[illegible]

Biotic similarity A measure of the degree to which two or more samples or assemblages are similar in species composition. Familiar biotic similarity indices include Sørensen's, Jaccard's, Horn's, and Morisita's indices.

Hill numbers A family of diversity measures developed by Mark Hill. Hill numbers quantify diversity in units of equivalent numbers of equally abundant species.

Individual-based (abundance) data A common form of data in biodiversity surveys. The data set consists of a vector of the abundances of different species. This data structure is used when an investigator randomly samples individual organisms in a biodiversity survey.

Nonparametric asymptotic estimators Estimators of total species richness (including Chao1, Chao2, abundance-based coverage estimator (ACE), incidence-based coverage estimator (ICE), and the jackknife) that do not assume a particular form of the species abundance distribution (such as a log-series or log-normal distribution). Instead, these methods use information on the frequency of rare species in a sample to estimate the number of undetected species in an assemblage.

Phylogenetic diversity Adjusted diversity measures that take into account the degree of relatedness among a set of species in an assemblage. Other things being equal, an assemblage of closely related species is less phylogenetically diverse than a set of distantly related species.

Rarefaction A statistical interpolation method of rarefying or thinning a reference sample by drawing random subsets of individuals (or samples) in order to standardize the comparison of biological diversity on the basis of a common number of individuals or samples.

Sample-based (incidence) data A common form of data in biodiversity surveys. The data set consists of a set of sampling units (such as plots, quadrats, traps, and transect lines). The incidence or presence of each species is recorded for each sampling unit.

Species accumulation curve A curve of rising biodiversity in which the -axis is the number of sampling units (individuals or samples) from an assemblage and the -axis is the observed species richness. The species accumulation curve rises monotonically to an asymptotic maximum number of species.

Species diversity A measure of diversity that incorporates both the number of species in an assemblage and some measure of their relative abundances. Many species diversity indices can be converted by an algebraic transformation to Hill numbers.

Species richness The total number of species in an assemblage or a sample. Species richness in an assemblage is difficult to estimate reliably from sample data because it is very sensitive to the number of individuals and the number of samples collected. Species richness is a diversity of order 0 (which means it is completely insensitive to species abundances).

[illegible]

Mea ing Bi l gical Di e i

The notion of biological diversity is pervasive at levels of organization ranging from the expression of heat-shock proteins in a single fruit fly to the production of ecosystem services by a terrestrial ecosystem that is threatened by climate change. How can one quantify diversity in meaningful units across such different levels of organization? This article describes a basic statistical framework for quantifying diversity and making meaningful inferences from samples of diversity data.

In very general terms, a collection of "elements" are considered, each of which can be uniquely assigned to one of several distinct "types" or categories. In community ecology, the elements typically represent the individual organisms, and the types represent the distinct species. These definitions are generic, and typically are modified for different kinds of diversity studies. For example, paleontologists often cannot identify fossils to the species level, so they might study diversity at higher taxonomic levels, such as genera or families. Population geneticists and molecular biologists might be interested in more fine-scale "omics" classifications of biological materials on the basis of unique DNA sequences (genomics), expressed mRNA molecules (transcriptomics), proteins (proteomics), or metabolic products (metabolomics). Ecosystem ecologists might be concerned not with individual molecules, genotypes, or species, but with broad functional groups (producers, predators, and decomposers) or specialized ecological or evolutionary life forms (understory forest herbs and filter-feeding molluscs). However, to keep things simple, this article will refer throughout to "species" as the distinct categories of biological classification.

Although the sampling unit is often thought of as the individual organism, many species, such as clonal plants or colonial invertebrates, do not occur as distinct individuals that can be counted. In other cases, the individual organisms, such as aquatic invertebrate larvae, marine phytoplankton, or soil microbes are so abundant that they cannot be practically counted. In these cases, the elements of biodiversity will

http://dx.doi.org/10.1016/B978-0-12-384719-5.00424-X

diversity (PD)) or indirectly, based on their function (referred to as functional diversity). These metrics relax the second assumption discussed in the section Species Richness and Traditional Species Diversity Metrics (all species are ‘‘equally different’’ from one another) by weighting each species by a measure of its taxonomic classification, phylogeny, or function.

Bi ic Simila i

These concepts of species diversity apply to metrics that are used to quantify the diversity of single assemblages. However, the concept of diversity can also be applied to the comparison of multiple assemblages. Suppose again that a person visits two woodlands, both of which have 10 trees species, each species contributing 10% to the abundance of individual trees within the woodland. Thus, in terms of species richness and species diversity, the two woodlands are identical. However, the two woodlands may differ in their species composition. At one extreme, they may have no species in common, so they are biologically distinct, in spite of having equal species richness and species diversity. At the other extreme, if the list of tree species in the two woodlands is the same, they are identical in all aspects of diversity (including taxonomic, phylogenetic, and functional diversity). More typically, the two woodlands might have a certain number of species found in both woodlands and a certain number that are found in only one.

B quantifies the extent to which two or more sites are similar in their species composition and relative abundance distribution. The concept of biotic similarity is important at large spatial scales for the designation of biogeographic provinces that harbor distinctive species assemblages with both endemic and shared elements. Biotic similarity is also a key concept underlying the measurement of beta diversity, the turnover in species composition among a set of sites. In an applied context, biotic similarity indices can quantify the extent to which distinct biotas in different regions have become homogenized through losses of endemic species and the introduction and spread of nonnative species. Differences among species in evolutionary histories and functional trait values can also be incorporated in similarity measures.

Bia in he E ima i n f Di e i

The true species richness and relative abundances in an assemblage are unknown in most applications. Thus species richness, species diversity, and biotic similarity must be estimated from samples taken from the assemblage. If the sample relative abundances are used directly in the formulas for traditional diversity and similarity measures, the maximum likelihood estimator (MLE) of the true diversity or similarity measure is obtained. However, the MLEs of most species diversity measures are biased when sample sizes are small. When sample size is not sufficiently large to observe all species, the unobserved species are undersampled, and – as a consequence – the relative abundance of observed species, on average, is overestimated.

Because biotic diversity at all levels of organization is often high, and biodiversity sampling is usually labor intensive, these biases are usually substantial. Even the simplest comparison of species richness between two samples is complicated unless the number of individuals is identical in the two samples (which it never is) or the two samples represent the same degree of coverage (completeness) in sampling. Ignoring the sampling effects may obscure the influence of overall abundance or sampling intensity on species richness. Attempts to adjust for sampling differences by algebraic rescaling (such as dividing S by or by sampling effort) lead to serious distortions and gross overestimates of species richness for small samples. Thus, an important general objective in diversity analysis is to reduce the undersampling bias and to adjust for the effect of undersampled species on the estimation of diversity and similarity measures. Because sampling variation is an inevitable component of biodiversity studies, it is equally important to assess the variance (or standard error) of an estimator and provide a confidence interval that will reflect sampling uncertainty.

[illegible]

This article introduces a common set of notation for describing biodiversity data (Colwell ., 2012). Consider an assemblage consisting of $N^{□}$ total individuals, each belonging to one of S distinct species. Species has N individuals, so that $\sum^{S}_{¼\,1} N$ ¼ $N^{□}$. The relative frequency of species is $N/N^{□}$, so that $\sum^{S}_{¼\,1}$ ¼ 1. Note here that $N^{□}$, S, N, and represent the ‘‘true’’ underlying abundance, species richness, and relative frequencies of species. These quantities are unknowns, but can be estimated, and one can make statistical inferences by taking random samples of data from such an assemblage. This article distinguishes between two sampling structures.

Indi id al-Ba ed (Ab ndance) Da a

The ′ ′ ′ ′ ′ is a collection of individuals, drawn at random from the assemblage with N□ total individuals. In the reference sample, a total of S_{obs} species are observed, with X individuals observed for species , so that $\sum^{S}_{\frac{1}{4}\,1} X$

more individuals are sampled, but the slope becomes shallower because progressively more sampling is required to detect the rare species. As long as the sampling area is sufficiently homogeneous, all of the species will eventually be sampled and the curve will flatten out at an asymptote that represents the true species richness for the assemblage. For incidence data, a similar accumulation curve can be drawn in which the -axis represents the number of sampling units and the -axis is the number of species recorded.

In e la ing S ecie Richne i h Ra efac i n

A single empirical sample of individuals or a pooled set of sampling units represents one point on the species accumulation curve, but the investigator has no way of directly determining where on the curve this point lies. To compare the richness of two different samples, they must be standardized to a common number of individuals, for abundance samples (Sanders, 1968; Gotelli and Colwell, 2001, 2011). Rarefaction represents an interpolation of a biodiversity sample to a smaller number of individuals for purposes of comparison among samples. Typically, the abundance of the larger sample is rarefied to the total abundance of the smaller sample to determine if species richness (or any other biodiversity index) differs for a common number of individuals (Figure 4). For incidence data, rarefaction interpolates between the reference sample and a smaller number of sampling units.

Let S_{ind}( ) represent the expected number of species in a random sample of individuals from the reference sample of individuals ($<$). If the true probabilities ($_1$, $_2$, …,) of each of the S species in the assemblage were known, and species frequencies (X_1, X_2, …, X_S) follow a multinomial distribution for which the total of all frequencies is , and cell probabilities ($_1$, $_2$, … , $_S$), then

$$S$$

Estimation) and β ¼ $\binom{R\ □}{} \Big/ \binom{R}{}$ for $\leq R$

eqns [9] and [10] need modification; the modified variances are available in P4kP¼P¼1fnS

[illegible]

augmented area *A*

which measures the probability that two randomly chosen individuals (selected with replacement) belong to two different species. The measure 1 □ H_{GS} ¼ $\sum^{S}_{¼\,1}$ ^{2} is referred to as the Simpson index. With an adjustment for $N^{□}$, the total number of individuals in the assemblage, the Gini–Simpson index is closely related to the ecological index *PIE* (Hurlbert, 1971), the probability of an interspecific encounter:

PIE ¼ ½$N^{□}$/ð$N^{□}$ □ 1Þ□H_{GS} ½22□

which measures the probability that two randomly chosen individuals (selected replacement) belong to two different species. Both PIE and the Gini–Simpson index have a straightforward interpretation as a probability. When PIE is applied to species abundance data, it is equivalent to the slope of the individual-based rarefaction curve measured at its base.

However, the units of the Gini–Simpson index and *PIE* are probabilities that are bounded between 0 and 1, and the units of Shannon entropy are logarithmic units of information. These popular complexity measures do not behave in the same intuitive way as species richness (Jost, 2007).

The ecologist MacArthur (1965) was the first to show that Shannon entropy (when computed using natural logarithms)

For an integer , ≥ 2, a similar derivation leads to a nearly unbiased estimator for ' D.

$$\text{'}\hat{D}\ \text{¼}\ \left\{\sum^{S}_{\text{¼}\ 1}\text{½}X\ \text{ð}X\ \square\ 1\text{Þ}\cdots\text{ð}X\ \square\ ,\ \text{þ}\ 1\text{Þ}\square\right.$$

Species), can be regarded as a special case of an ultrametric tree. In contrast, if the branch lengths are proportional to the number of base-pair changes in a given gene, or some other measure of genetic or morphological change, some branch tips may be farther in absolute time from the basal node than other branch tips, and such trees are nonultrametric.

Pielou (1975) was the first to notice that the concept of diversity could be broadened to consider differences among species. The earliest taxonomic diversity measure is the

′ (*CD*), which is defined as the total number of taxa or nodes in a taxonomic tree that encompasses all of the species in the assemblage (Vane-Wright ′ ., 1991). Another pioneering work is Faith’s (1992) *PD*, which is defined as the sum of the branch lengths of a phylogeny connecting all species in the target assemblage. In both *CD* and *PD*, species abundances are not considered.

C.R. Rao’s , ′ was the first diversity measure that accounted for both phylogeny and species abundances (Rao, 1982). It is a generalization of the Gini–Simpson index:

$$Q_{\mathrm{Rao}} \text{ ¼ } \sum_{,} \qquad \text{½26a}$$

where denotes the phylogenetic distance between species and , and and denote the relative abundance of species and . This index measures the average phylogenetic distance between any two individuals randomly selected from the assemblage. For the special case of no phylogenetic structure (all

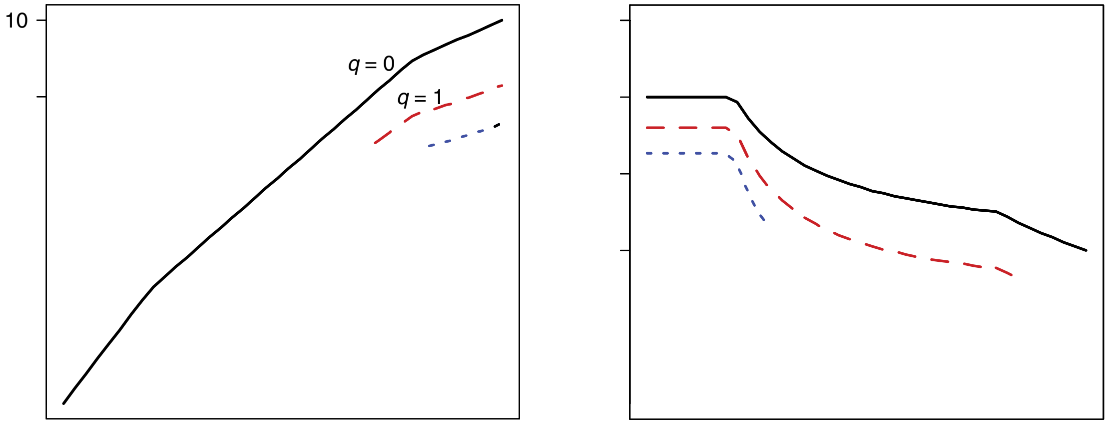
This measure ʻ$\overline{D}$ðTÞ gives the mean effective number of maximally distinct lineages (or species) *T* time steps in the past. The diversity of a tree with ʻ$\overline{D}$ðTÞ¼ in the time period [□ *T*, 0] is the same as the diversity of an assemblage consisting of equally abundant and maximally distinct species with all branch lengths *T*.

The branch diversity or phylogenetic diversity ʻPD(*T*) of order , through *T* time steps before present is defined as the product of ʻ$\overline{D}$ðTÞ and *T* . The measure ʻPD(*T*) given below quantifies "the total effective number of lineage-lengths or lineage-time steps" (Chao ʻ ., 2010)

$$ʻ\mathrm{PD}ðTÞ¼T□ʻ\overline{D}ðTÞ¼\left\{\sum_{∈B_T}L\left(\frac{}{T}\right)ʻ\right\}^{1/ð1□,Þ} \qquad ½28□$$

If , ¼0 and *T*¼*T*□ (tree height), then ^{0}PD(*T*) reduces to Faith's PD. It also reduces to CD in a taxonomic tree if the branching of each Linnaean taxonomic category is assigned a time step of unit length. A PD profile can be constructed by plotting both ʻPD(*T*) and ʻ$\overline{D}$ðTÞ as a function of *T* for , ¼0, 1, and 2. It is also informative to construct another diversity profile by plotting ʻPD(*T*) and ʻ$\overline{D}$ðTÞ as a function of order , for some selected values of temporal perspective *T*. See Figure 8 for a numerical example. In most applications, ecologists are interested in the case *T*¼*T*□ (tree height) or the divergence time between the species group of interest and its nearest outgroup. The divergence time of the most recent common ancestor of all extant taxa is another useful comparison.

For nonultrametric trees, the time parameter *T* is generalized to $\overline{T}$, where $\overline{T}$¼$\sum_{∈B_{\overline{T}}}L$ represents the abundance-weighted mean base change per species and $\boldsymbol{B}_{\overline{T}}$ denote the set of branches connecting all focal species. The diversity of a nonultrametric tree with mean evolutionary change $\overline{T}$ is the same as that of an ultrametric tree with a time step $\overline{T}$.



Therefore, the diversity formula for a nonultrametric tree is obtained by replacing T in the '$\overline{D}$ðTÞ and '$\mathrm{PD}(T)$ with $\overline{T}$. Equation [27] can also describe taxonomic diversity, if the phylogenetic tree is a Linnaean tree with L levels (ranks), and each branch is assigned unit length. It also describes functional diversity, if a dendrogram can be constructed from a trait-based distance matrix using a clustering scheme (Petchey and Gaston, 2002). Thus, Hill numbers can be effectively generalized to incorporate taxonomy, phylogeny, and function and provide a unified framework for measuring biodiversity (Chao and Jost, in press).

Estimation of phylogenetic and functional diversity from small samples has not been well studied. As with the estimation of simple Hill numbers, phylogenetic diversity '$\overline{D}$ðTÞ and '$\mathrm{PD}(T$

representation (and better statistical estimation) of the similarity of assemblages.

Ab ndance-Ba ed Simila i Indice

Assume that in the combined assemblages, there are *S* species. Denote the relative abundance vector for the *S* species in the th assemblage by ($_1$, $_2$,…,), some of them may be 0. Thus, for *N* assemblages, there are *N* sets of abundances {($_1$, $_2$,…,); ¼1,2,…,*N*}. A sample of equally weighted assemblages, beta diversity ${}^qD_\beta$ ranges between a minimum of 1 (when all assemblages are identical) and a maximum of N, the number of assemblages in each region (when all assemblages are completely distinct; i.e., there are no shared species). For example, a set of completely distinct sites in a region of three sites attains the maximum value of 3, whereas another set of completely distinct sites in a region of 10 sites attains the maximum value of 10. Because the maximum depends on the number of assemblages in the region, beta diversities usually cannot be compared directly among multiple regions. Instead, beta diversity should be compared with sample-based rarefaction to a common number of samples or to a common degree of completeness of samples in each region. However, beta diversity can be transformed to the C_{qN} measure in the range [0, 1] by the following nonlinear transform for N equally weighted assemblages:

$$C_{qN} = [(1/{}^qD_\beta)^{q-1} - (1/N)^{q-1}]/[1 - (1/N)^{q-1}] \qquad (36)$$

The transformed measure C_{qN} is unity (when all assemblages are identical) and 0 (when all assemblages are completely distinct).

This nonlinear transformation ensures that C_{qN} preserves an essential property of an overlap index: The transformed index C_{qN} gives the true q A/S for naCn2St5u(;3/nsformed)468(0 0510 1 Tfq78 0 Td()Tj9T13 1 Tif332 0 Td(for)Tj031 1 Tf10.718 0 Td(N)

One advantage of these measures is that the undersampling bias due to unseen, shared species can be evaluated and corrected. Chao ′ . (2005) used the frequencies of observed rare, shared species to obtain an appropriate adjustment term for U and V to account for the effect of ′′ shared species and thus remove most undersampling bias. Then the bias-corrected U and V estimators are substituted into the formulas to obtain Chao–Jaccard and Chao–Sørensen estimators. These measures are designed to be sensitive to rare shared species while still taking abundance into account, so they may increase sharply as more shared species are discovered. Because these measures match the total relative abundances of species shared between two assemblages, they are useful if the focus is to construct abundance-based ′ ′ (dissimilarity or distance) measures by subtracting each measure from one. This class of measures can also be extended to replicated incidence data; see Chao ′ . (2005) for details.

Ph l gene ic Simila i Indice

The classic Jaccard, Sørensen, and Morisita-Horn similarity measures all have their own phylogenetic generalizations. Most of the pioneering work was developed by microbial ecologists (Lozupone and Knight, 2005; Faith ′ ., 2009). The phylogenetic Jaccard and Sørensen measures are based on Faith's total branch lengths and have formulas similar to their classic versions. The phylogenetic Sørensen index can be expressed as 2L_{12}/(L_1 þ L_2), where L_1 and L_2 denote the total branch lengths in Assemblages 1 and 2, respectively, and L_{12} denotes the total length of the shared branches in the same time interval of interest (Lozupone and Knight, 2005). The phylogenetic Jaccard index takes the form of L_{12}/(L_1 þ L_2 □ L_{12}). When species relatedness is based on a simple Linnean taxonomic classification tree, L_1 and L_2 become the number of taxa in Trees 1 and 2, respectively, and L_{12} becomes the number of sh8535 30.4610()-3ene 1 932(nu)14(mb)18(er)-319(of)-4055812.435(e)-2336s27r9(of)5(es)-329

[illegible] A, [illegible] (2012) [illegible] 5: 3 21.

[illegible] A (1994) [illegible] B 345: 101 118.

[illegible] (2004) [illegible]