OIKOS

Forum

A conceptual guide to measuring species diversity

Michael Roswell, Jonathan Dushoff and Rachael Winfree

M. Roswell (https://orcid.org/0000-0002-8479-9184) ✉ (mroswell.rutgers@gmail.com), Graduate Program in Ecology and Evolution, Rutgers Univ., New Brunswick, NJ, USA. – MR and R. Winfree (https://orcid.org/0000-0002-1271-2676), Dept of Ecology, Evolution and Natural Resources, Rutgers Univ., New Brunswick, NJ, USA. MR, Dept of Entomology, Univ. of Maryland College Park, College Park, MD, USA. – J. Dushoff (https://orcid.org/0000-0003-0506-4794), Dept of Biology, McMaster Univ., Hamilton, ON, Canada.

Oikos
130: 321–338, 2021
doi: 10.1111/oik.07202

Subject Editor: Jarrett Byrnes
Editor-in-Chief: Andrew Gonzalez
Accepted 22 December 2020

Three metrics of species diversity – species richness, the Shannon index and the Simpson index – are still widely used in ecology, despite decades of valid critiques leveled against them. Developing a robust diversity metric has been challenging because, unlike many variables ecologists measure, the diversity of a community often cannot be estimated in an unbiased way based on a random sample from that community. Over the past decade, ecologists have begun to incorporate two important tools for estimating diversity: coverage and Hill diversity. Coverage is a method for equalizing samples that is, on theoretical grounds, preferable to other commonly used methods such as equal-effort sampling, or rarefying datasets to equal sample size. Hill diversity comprises a spectrum of diversity metrics and is based on three key insights. First, species richness and variants of the Shannon and Simpson indices are all special cases of one general equation. Second, richness, Shannon and Simpson can be expressed on the same scale and in units of species. Third, there is no way to eliminate the effect of relative abundance from estimates of any of these diversity metrics, including species richness. Rather, a researcher must choose the relative sensitivity of the metric towards rare and common species, a concept which we describe as 'leverage.' In this paper we explain coverage and Hill diversity, provide guidelines for how to use them together to measure species diversity, and demonstrate their use with examples from our own data. We show why researchers will obtain more robust results when they estimate the Hill diversity of equal-coverage samples, rather than using other methods such as equal-effort sampling or traditional sample rarefaction.

Keywords: coverage, Hill numbers, rarefaction, rarity

Synthesis

Most species are rare, and therefore efforts to measure and compare biodiversity suffer from sampling issues related to undetected species. How can we best recognize and mitigate sampling limitations in biodiversity measurement? This guide recommends using two tools: coverage to measure and equalize sample completeness, and Hill diversity as a unifying concept to link different measures of diversity. The guide explores how biodiversity metrics work and tradeoffs among them. Using both conceptual and applied examples, the guide shows how to use coverage and Hill diversity together to grapple productively with sampling limitations, and make more meaningful biodiversity measurements and comparisons.

Introduction

Species diversity is one of the more frequently measured quantities in ecology, yet how to measure it is complex, and sometimes contentious. The past decade has seen great advances in comparing and unifying various diversity metrics, and also in developing ways to standardize samples prior to measuring diversity (Jost 2006, Ellison 2010, Chiarucci et al. 2011, Chao and Jost 2012, Colwell et al. 2012, Chao and Chiu 2016, Cox et al. 2017, Chao et al. 2019a, 2020). This latter step is necessary because – in contrast to many variables ecologists measure, for which a random sample from a community provides a reasonably unbiased estimate of the community itself – most species diversity values estimated from samples are a biased measure of the diversity of the larger community. This is mainly because the true relative abundance of rare species is poorly captured in samples, in which those species tend to appear only once or not at all. Here, we provide a conceptual guide to best practices for comparing the level of biodiversity of two or more communities, based on samples from those communities. We begin by reviewing methods for standardizing samples, which is an important but often overlooked step in measuring diversity. In this section we review 'coverage,' a conceptually elegant, but under-used, method for standardizing samples. We then provide a guide to using Hill diversity. We try to make this concept more intuitive to ecologists by showing how different Hill diversities all calculate the mean rarity of the species in the community, but using different types of means (arithmetic, geometric and harmonic). We also draw parallels between Hill diversity and a tool familiar to many ecologists: the link functions of generalized linear models. We use 'relative abundance' throughout to mean the proportion of individuals belonging to a given species, but in most cases other measures, like proportional biomass or percent cover, could be used instead.

We assume that ecologists wish to determine which communities are more and less diverse, and by how much; in other words, that they aim to measure an 'effect size' (Chao and Jost 2012, Chase and Knight 2013). Thus, we advocate for methods that will accurately reflect relative (but not necessarily absolute) differences in diversity. To demonstrate the preferred tools for standardizing samples and quantifying diversity – coverage and Hill diversity – we analyze a small data set on wild bees we collected from four meadows.

Equalizing samples

Diversity can only be meaningfully compared across communities that have been sampled equivalently in some way. Unfortunately, there are multiple ways to standardize samples, and the choice of sample standardization method can strongly influence results. In this section, we consider three main ways ecologists standardize their samples: by equalizing effort, equalizing sample size or equalizing coverage.

Conceptual problems with traditional methods of equalizing samples

Many ecologists build equal-effort sampling into their study designs. Effort can be measured as the amount of time spent sampling, the area sampled, the number of traps set out or the like. This seems like the right way to compare communities: sample the same way and the same amount in each, and any differences should reflect only the diversity of each community, and not how the communities were sampled. But this is not true. In reality, two factors determine how well the sample represents the **true diversity** of the community: how hard one looks, and also how many species there are and in what relative abundances. Equal-effort sampling only deals with the first factor. A key problem with equal-effort sampling is that sample size generally varies across communities given equal effort, and sample size partly determines how well the observed abundance distribution matches the true species abundance distribution of the community. For instance, a small sample is likely to contain only a few species, all of them common. As samples contain more individuals, the number of species rises and **sample diversity** grows (Preston 1948). In sum, diversity estimates (especially richness) based on equal-effort sampling underestimate community diversity from samples that contain fewer individuals, because these samples often include fewer species by chance alone, regardless of the community from which they are drawn (Gotelli and Colwell 2001).

A second way ecologists standardize samples is by sample size; for example, by removal of individuals from larger samples until all samples have the same number of individuals (**rarefaction**). However, rarefaction does not provide unbiased samples either, because it still does not account for the distribution of relative abundances in the whole, larger community (Willis 2019). Because more diverse communities usually have both more and also rarer species, they also require more effort to characterize. Furthermore, it is not always possible to predict, from smaller samples, which of two communities would appear more diverse with much larger samples. In sum, sample-size standardization leads to larger underestimates of diversity for more diverse communities (Chao and Jost 2012).

Coverage: a solution

Sample-size and effort-based standardization do not fairly represent community diversity because they do not account for the underlying species abundance distribution of the community being sampled (Brose et al. 2003, Cao et al. 2007, Beck and Schwanghart 2010, Willis 2019). In contrast, a newer method, **coverage** (Box 1), accounts for both the amount of sampling and, to a much greater extent than the other methods, the true diversity of the community. Coverage thereby recognizes that more diverse communities require more sampling in order to be equally well-characterized. Coverage was discovered in the 1940s by the founder of computer science, Alan Turing, but was only recently introduced as a tool for

Box 1. What is coverage?

Coverage is a measure of how completely a community has been sampled. Specifically, it estimates the total true relative abundance in the community of all the species represented in the sample. Coverage can be visualized as the complement of the slope of a species accumulation curve (Fig. B1). Coverage increases more slowly as sample size increases and more and more species are detected. In ecological communities, this slowing is often quite dramatic because, while most species in an ecological community are likely rare, most individuals in the community belong to common species.

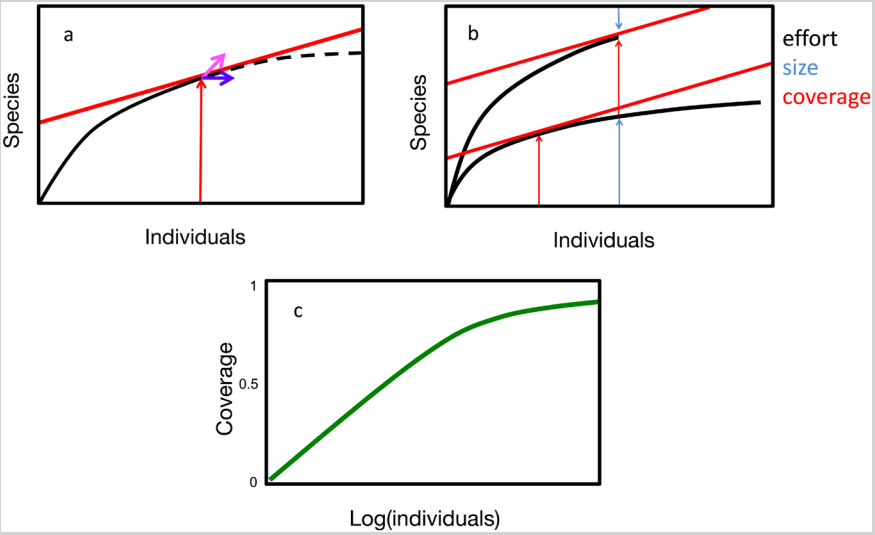

Figure B1. (a) As an ecologist collects individuals, there is some probability that the next individual would be from an already detected species (short horizontal purple arrow), or from a new one (short diagonal pink arrow). The chance that it is from a new species is the slope of the species accumulation curve (red line) at that point. The probability of not picking a new species (1-the slope) is the coverage, which approaches 1 as the curve flattens out. (b) Two species accumulation curves at equal effort (ends of black curves), equal size (light blue arrows) and equal coverage (red arrows). These three data standardization methods often result in different diversity estimates. (c) At higher values of coverage, to obtain even modest gains in coverage, sample sizes may need to increase by orders of magnitude.

To estimate coverage, only three parameters are needed (Chao and Jost 2012):

f_1, the number of singletons (species represented by only 1 individual) in the sample
f_2, the number of doubletons (species represented by only 2 individuals) in the sample
n, the total number of individuals in the sample

Chao and Jost (2012) provide the following equation for coverage (C):

$$C = 1 - \frac{f_1}{n}\left[\frac{(n-1)f_1}{(n-1)f_1 + 2f_2}\right] \qquad \text{(B1)}$$

R code that will estimate coverage from sample abundances is available with the functions 'iNEXT' and 'estimateD' in the R package *iNEXT* (Hsieh et al. 2016). If researchers do not sample individuals, occurrence counts in grids or other subunits may be used.

At present, standardizing samples based on coverage faces two unsolved issues. First, the best available method for rarefying to equal coverage (Chao and Jost 2012, Chao et al. 2014a, Hsieh et al. 2016) does so indirectly. Both expected coverage and expected diversity increase with sample size: for any given sample, increasing sample size can only increase both coverage and diversity (Chao and Jost 2012, Chao et al. 2014a, Hsieh et al. 2016). Coverage-based rarefaction and extrapolation extends traditional rarefaction methods (Colwell et al. 2012) by transforming the x-axis from sample size to expected coverage through a theoretically derived equation that uses observed sample frequences and sample size as its parameters (Chao and Jost 2012, Chao et al. 2014a, Hsieh et al. 2016). While on average, if one knows how much bigger or smaller a sample is than one's reference, one knows how much more or less coverage that sample will have, in a particular case the actual coverage may be higher or lower than expected. This is another way of saying there is uncertainty in the x-axis in coverage-based rarefaction and extrapolations.

Second, as a result of the indirect rarefaction procedure, the confidence intervals (CI) provided for coverage-standardized samples are too narrow (i.e. anti-conservative). The uncertainty in the x-axis (coverage) is not propagated up to the y-axis (diversity), and as a result, the estimate of the uncertainty in the y-axis is smaller than it needs to be. Newer approximate CI are not formally conditioned on sample size (Chao et al. 2020). Although at smaller sample sizes these also exhibit anti-conservatism, they become exact asymptotically (Anne Chao, pers. comm.).

standardizing samples in ecology (Alroy 2010, 2017, Jost 2010, Chao and Jost 2012, Chao et al. 2014a). Coverage is a theoretically elegant way to standardize samples, and is increasingly used in the ecological literature.

Coverage describes how well a sample captures the true diversity of the whole community, including species that have not yet been detected. More precisely, coverage estimates the proportion of individuals in the (whole) community that belong to species present in the sample. As this proportion increases, the share of individuals in the community that belong to undetected species falls. For example, a coverage of 0.98 means that 2% of the individuals in the community being sampled belong to species the researcher has missed. For a sample to contain enough species to represent 98% of the individuals in a more diverse community, it usually must be larger than a sample with 98% coverage from a less diverse community. Thus, when ecologists standardize samples by coverage, they compare samples that have more individuals from some communities than others. This results in more balanced information from each community.

Sampling with equal coverage isn't quite what we might want: to sample each community until the same proportion of its diversity had been recorded. For example, using species richness as the metric, one could imagine sampling until 90% of the species in each community had been detected. In this case, the comparison would be fair. Unfortunately, this method is not possible (Chao et al. 2020), because it is not usually possible to know how many species are truly there, nor in what proportions – if it were, we wouldn't need to estimate diversity. Given that this ideal cannot be implemented, coverage is a practical approach to achieving more comparable samples, using information available to researchers.

The key insight behind coverage is that the proportion of individuals in the community belonging to undetected species can be estimated reliably, based only on the frequencies of species already in the sample (Good and Toulmin 1956, Chao and Jost 2012, Zhang 2016). This concept is best illustrated with a species accumulation curve (Box 1, Fig. B1(a)). Imagine being at the endpoint of the curve, about to sample one more individual. The pool that individual will be sampled from contains all the as-yet-unsampled individuals in the community, most of which belong to species already detected, but some of which do not. If the next individual obtained is a new species, the species accumulation curve goes up one step for a slope of 1. If it is not a new species, the curve moves horizontally one step for a slope of 0 (Fig. B1(a), arrows). Thus, the expected slope of the species accumulation curve represents the probability that the next individual sampled will belong to a new species. This slope is (1-coverage). As coverage approaches 1, the species accumulation curve approaches its asymptote.

While ecologists have long used the slope of the species accumulation curve to measure sampling completeness, the advantage of the more recent formalization is that even when sampling is incomplete, samples can be compared at equal coverage (Fig. B1(b)). This comparison is 'fair' in the sense that the same proportion of individuals from each community is represented by the species in each sample. In sum, while it cannot remove sample diversity's dependence on sample completeness (Willis 2019), coverage is the fairest available way to standardize samples because it standardizes what is known (the sample) relative to what is there (the true community).

What coverage clarifies about species richness

Ecologists may be attracted to species richness as a diversity metric because true richness depends only on the number of species, but not on their relative abundances. However, estimates of richness are, in fact, highly sensitive to the relative abundances of species in the community being sampled. The concept of coverage offers a nice demonstration of how this is so.

Although sample coverage increases with sampling, the rate of this increase slows as sampling proceeds (Fig. B1(c)). This is because after initial sampling, the vast majority of individuals in a community do belong to species represented in the sample, and it takes a lot of work to find those comparatively few individuals belonging to the new, rare species. This means that sample richness depends on how rare the rare species are. For example, imagine two communities, one in which all species have the same abundance, and the second in which a few species are very common, but most are very, very rare. In the first community, at low and medium sample sizes, finding a new species with additional sampling remains quite likely. In the second, once samples are large enough that the common species have been detected, the chance of detecting a new (and very rare) species with additional sampling is low. This means that even if both communities had the same richness, samples from the first community would usually contain more species. In sum, species richness estimates are not only sensitive to the size of the sample and the true number of species in the community, but also to species relative abundances in the community, just like other diversity indices.

Diversity metrics

In this section, we briefly review problems with species richness, and the traditional Shannon and Simpson indices, which are the ways ecologists most often measure the diversity of a community (Magurran and McGill 2011). We then explain **Hill diversity**, a general approach that includes, as special cases, species richness and modified versions of the traditional Shannon and Simpson indices. There is an increasing consensus that Hill diversity is the preferred way to measure not only the species diversity of a community, which is the subject of this paper, but also differentiation among communities (Jost 2007, Ellison 2010, Chao and Chiu 2016, Botta-Dukát 2018, Chao et al. 2019b, Ohlmann et al. 2019) functional and phylogenetic diversity (Chao et al. 2014b, Kang et al. 2016), genetic diversity (Jost 2008, Sherwin et al.

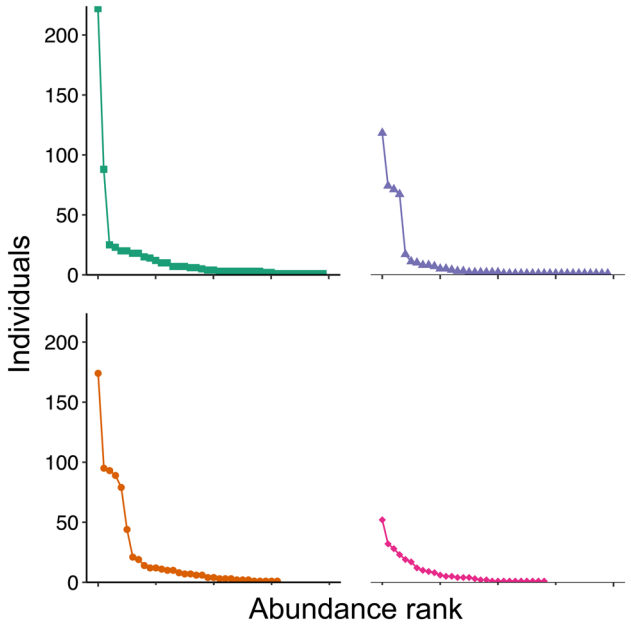

Figure 1. Observed rank–abundance distributions for the bee samples from our four meadows. The sample from the green square community has strong dominance by a small number of species with a long 'tail' of rare species. The sample from the purple triangle community also has strong dominance and a long a tail of rare species, although it has fewer species of intermediate rarity. The sample from the orange circle community has a much shorter tail of rare species. The pink diamond community sample has the least variation in rarity, and the fewest species. Diversity metrics summarize these distributions to enable quantitative comparisons.

2017, Alberdi and Gilbert 2019) and evenness (Chao et al. 2019a).

To illustrate our main points, we present some analyses of a small data set extracted from a larger study (Roswell et al. 2019a, b). This data set includes wild bees we collected with hand nets from four meadows, using equal effort: seven person-hours over three consecutive days in each meadow (Fig. 1). In the study from which these data are taken, we sampled the entire meadow, collecting bees that contacted the reproductive parts of any flower within a 1-m radius semicircle in front of the observer, in timed 30-min transects that crossed back and forth throughout the meadow. The meadows were all mowed annually, and some were also burned to maintain an early state of succession, and had been seeded within the last four years with a mix of 'pollinator plants,' though many of the plants from which we sampled colonized naturally.

In the first meadow (green squares in Fig. 1), we collected 578 individual bees that we identified to 40 bee species. In the second meadow (purple triangles), we collected 442 individuals of 40 species. In the third meadow (orange circles), we collected 745 individuals of 32 species. In the fourth (pink diamonds), we collected 225 individuals of 29 species. In each case, the pool we sampled included the entire bee fauna that foraged in each meadow during the three days of sampling, our operational definition of a 'bee community.' The question we seek to answer is, 'which bee communities are more and less diverse, and by how much?' While in fact, these data were not collected along an ecological gradient that would provide a hypothesis-driven reason to answer this question, the reader can imagine scenarios that might have caused differences in diversity in the meadows, such as landscape context, disturbance history or the composition of the plant community. Although our study design does not let us test these ideas, we use this dataset to show how the differences in diversity we would report, if the goal of the original study were to compare diversity, could vary with the amount of data collected, the method chosen to standardize samples, and the diversity metric used. In this section, we use our bee community data to illustrate points we make about Hill diversity. In the final section of the paper, we use these same dataset to conduct a demonstration of how researchers can use sample standardization together with Hill diversities to analyze their own data.

Conceptual problems with traditional diversity metrics

The number of species in a sample (sample richness) is a very flawed measure of diversity. Richness is strongly associated with the number of individuals in the sample, especially at the earlier stages of sampling (Fig. 2). Furthermore, as sampling proceeds, the accumulation curves representing

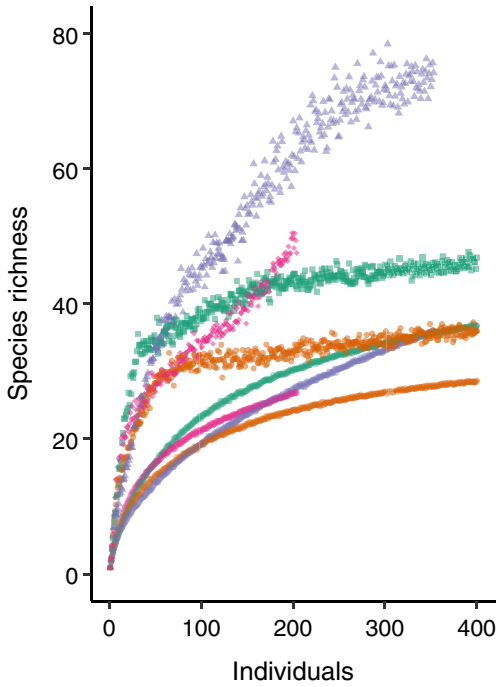

Figure 2. Species accumulation curves for number of species observed (y) versus the number of individuals sampled (x) for the bee communities in four meadows. The clouds of points represent the Chao1 estimates for the meadow of the same color. Chao1 predictions seem likely to continue increasing in most examples.

different communities often cross (Lande et al. 2000; colored lines in Fig. 2). This means that the relative richness of two communities, as measured at a smaller sample size, does not predict their relative richness at a larger sample size well (Cao et al. 2007, Coddington et al. 2009, Haegeman et al. 2013). This is often true even when estimators such as Chao1 are used to predict true diversity (colored clouds of points in Fig. 2).

Because richness is so sensitive to sampling effort and relative abundance, its estimation can hinge on how samples are standardized. Even the best asymptotic richness estimators, such as Chao1 (Gotelli and Colwell 2011), cannot reliably predict the true community diversity (Chao and Jost 2015, Fig. 2). The problem is that both sample richness and sample-based richness estimators are strongly influenced by the rarest species, which are precisely the species that we know least about. This is another way of saying that richness has high uncertainty. In fact, in the context of estimating and comparing community diversity from samples, this uncertainty is often insurmountable (Haegeman et al. 2013).

The traditional diversity indices that explicitly include relative abundance (Magurran and McGill 2011), such as the Shannon (Shannon and Weaver 1963) and Simpson (Simpson 1949) indices, are more robust than richness to the sampling problems outlined above. However, their use creates a new set of problems: these indices have different

Box 2. Problems with the traditional Shannon and Simpson indices

The first problem with traditional diversity indices is that they measure very different things (Tuomisto 2010). Species richness, of course, measures the number of species. The Shannon index measures uncertainty about the identity of species in the sample, and its units quantify information (bits; Hurlbert 1971), while the Gini–Simpson (1 − Simpson's original index) measures a probability, specifically, the probability that two individuals, drawn randomly from the sample, will be of different species (Simpson 1949, Hurlbert 1971). Because species richness, the Shannon index and the Gini–Simpson index do not measure the same quantities, justifying the choice of one of them to represent diversity is particularly difficult.

A second problem is that the Shannon and Gini–Simpson indices behave in ways that do not make sense for a metric of diversity. For example, if a diverse community (Fig. B2(a)) loses 1/3 of its species, the traditional Shannon and Gini–Simpson indices show only small proportional changes (Fig. B2(b)). Even a loss of 2/3 of species does not result in dramatic changes in index values (Fig. B2(c)). In contrast, all of the Hill diversity measures presented in this guide would give values of 30, 20 and 10 for the three communities. This property of Hill diversities is called the 'replication principle' (Hill 1973, Chao et al. 2014a). Note that although in the illustrations, individuals are lost along with their species, the values of all diversity metrics would be the same if total abundance were held constant even as species were lost. That is because all the diversity metrics discussed in this guide consider only relative, not absolute, abundance.

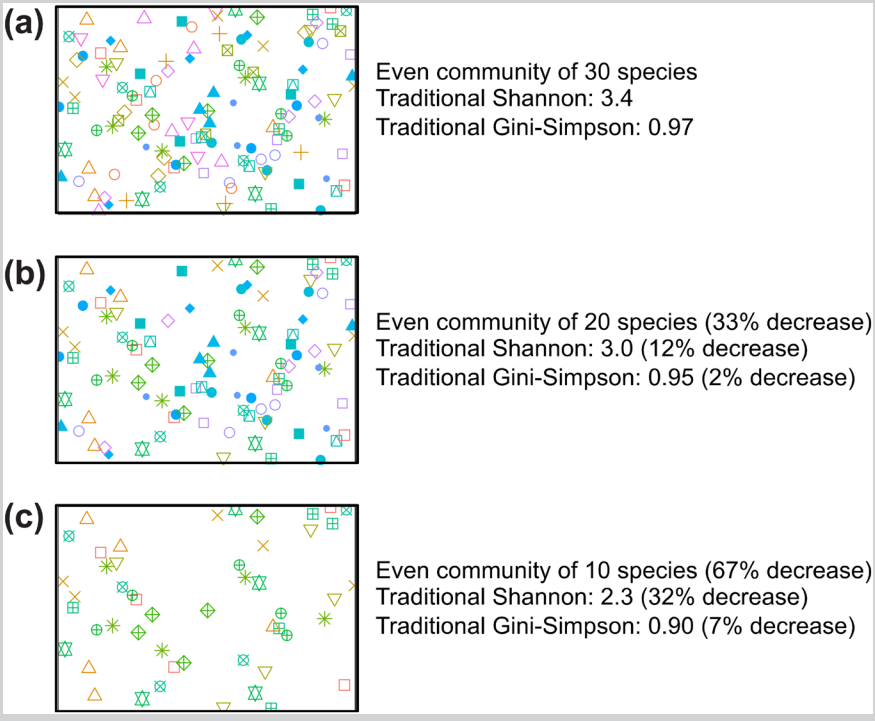

Figure B2. Values for the traditional Shannon and Gini–Simpson indices, calculated for communities that have decreasing numbers of species.

Box 3. Defining Hill diversity

We define Hill diversity by the equation

$$D = \left(\sum_{i=1}^{S} p_i \left(r_i \right)^{\ell} \right)^{1/\ell} \qquad \text{(B2)}$$

where D is diversity, S is the number of species, p_i is the proportion of all individuals that belong to species i, r_i is the rarity of species i, defined as $1/p_i$, and ℓ is the exponent that determines the rarity scale on which the mean is taken.

An elegant aspect of Hill diversities is that Eq. B2 is the equation for the generalized weighted mean, or Hölder mean (Bullen 2003). We intentionally use the exponent ℓ rather than 'q' (Jost 2006) to highlight this insight; it is easily shown that our equation is algebraically equivalent to Jost's, with $\ell = 1 - q$ (Supplementary information).

Hill diversities are a type of average. Specifically, they measure the mean rarity of the species in the sample, where the rarity of a species is the reciprocal of its relative abundance (Patil and Taillie 1982). When computing this average, the rarity of each species is first scaled by the exponent ℓ, and then weighted by the relative abundance of that species. This average is then back-transformed onto the diversity scale because of the outer exponent, the power of $1/\ell$. It may be helpful to think of the exponent ℓ as determining the leverage provided to rare species, and to recognize that for all values of the exponent, each species is weighted by its relative abundance. We discuss this idea further in Box 5.

Hill diversity formalizes a simple truism: a community consisting of species that are, on average, more rare has higher diversity (Patil and Taillie 1982, Tuomisto 2010, Botta-Dukát 2018, Kondratyeva et al. 2019).

units, and do not scale intuitively, or even similarly, with species gain and loss (Box 2; Jost 2009, Tuomisto 2010). These problems have led to the suggestion that diversity lacks any conceptual grounding (Hurlbert 1971).

Hill diversity: a solution

A unified method for measuring diversity was developed by Hill (1973), and re-introduced to ecologists by Jost (2006). This method takes as its starting point that both the number and the relative abundance of species are components of diversity, and that these components cannot be fully separated. The diversity metric developed by Hill (1973) consists of a single equation that, depending on the value taken by its sole parameter, the exponent that we call 'ℓ,' can vary from counting all species equally, even if they are vanishingly rare, to heavily emphasizing the species that are most common (Box 3).

Hill diversity has several important advantages. First, Hill diversities behave in ways that are logically reasonable for a measure of diversity (Hurlbert 1971, Jost 2009, Tuomisto 2010). For example, if some proportion of a community's species were randomly removed, all Hill diversities decrease by that proportion. Traditional diversity indices fail this and other common-sense expectations.

But how do Hill diversities do this? One interpretation is that Hill diversities express the diversity of a community in terms of an imaginary community with that same diversity, but in which all species are equally abundant (Jost 2006). For example, imagine comparing two communities using a given Hill diversity (i.e. with a given exponent in Eq. B2). Imagine that community A has a diversity of 5 and community B has a diversity of 25. This means that community A has the same diversity as a perfectly even community with five species, and community B has the same diversity as a perfectly even community with 25 species. Thus, there is a concrete sense in which community B is five times more diverse than community A. All Hill diversities can be interpreted this way.

A second advantage is that the calculation of Hill diversity is simple and already familiar to ecologists. Like the traditional diversity indices, Hill diversity summarizes relative (but not absolute) abundances, and the only data required to compute the sample Hill diversity are the relative abundances of species in a sample. The three forms of Hill diversity most commonly used by ecologists are species richness, and modifications of the traditional Shannon and Simpson indices. The key insight of Hill (1973) was that these three measures are special cases of the same general equation (Box 4). These three forms of Hill diversity – which we will refer to as **species richness**, **Hill–Shannon** diversity and **Hill–Simpson** diversity – differ only in how they scale rarity (Box 5). Richness uses an arithmetic rarity scale, which gives high **leverage** to, and therefore remains very sensitive to, rare species; Hill–Simpson diversity uses a reciprocal scale, which shifts leverage towards, and is thus dominated by common species; Hill–Shannon uses a logarithmic scale, and falls between the two.

A third, elegant aspect of Hill diversity is that each of its forms are a type of average. Specifically, here we develop the idea that rarity can be defined as the reciprocal of relative abundance, and that Hill diversities calculate the mean of the rarities of the species in the sample (Patil and Taillie 1982). If a community includes many species, all rare, that community has high mean rarity. In contrast, a community with only a few species, none of which is rare, has low diversity and low mean rarity. This way of understanding what Hill diversities 'really are' may be intuitive for many ecologists, who are accustomed to thinking about rarity in the context of diversity.

Box 4. Three particularly useful Hill diversity metrics

While Hill diversities are a continuous function of the exponent ℓ in Eq. B2, three particular integer values of ℓ produce versions of metrics that are already familiar to ecologists: species richness, Hill–Shannon and Hill–Simpson.

The only data required to calculate the Hill diversity of a sample are the number of individuals of each species found in each sample. The equations below have only two types of parameters:

S = number of species in the sample
p_i = (number of individuals of species i)/(total number of individuals in the sample)

Species richness emphasizes (provides higher leverage to) rare species, and can be simply calculated as:

$$S$$

This is equivalent to Eq. B2 when $\ell = 1$.

Hill–Shannon diversity emphasizes neither rare nor common species. It is defined as the limit of Eq. B2 as ℓ approaches 0, and is calculated with the base of the natural logarithm, e, raised to the power of the traditional Shannon entropy index:

$$e^{-\sum_{i=1}^{S} p_i \ln(p_i)} \quad \text{(B3)}$$

Hill–Simpson diversity emphasizes (provides higher leverage to) the common species. It is equivalent to Eq. B2 when $\ell = -1$. It has been described some authors as ENSPIE (Chase and Knight 2013), and it is equivalent to the inverse of the traditional Simpson index:

$$\frac{1}{\sum_{i=1}^{S} (p_i)^2} \quad \text{(B4)}$$

Sample Hill diversities can be computed using the function 'renyi' in the R package *vegan* (Oksanen 2016) and the function 'rarity' in the R package *MeanRarity* (Roswell and Dushoff 2020), and Hill diversities of equal-sized or equal-coverage samples can be approximately compared using the functions 'iNEXT' and 'estimateD' in the R package *iNEXT* (Hsieh et al. 2016). Estimates for asymptotic values of Hill diversity are available in *SpadeR* (Chao and Jost 2015, Chao et al. 2015).

The difference between richness, Hill–Shannon diversity and Hill–Simpson diversity is that they calculate mean rarity using different types of means: the arithmetic, geometric and harmonic means, respectively. An important point, which has generally been overlooked in the literature, is that these means differ not in how they weight the values they average (because in all cases, each value is weighted by its frequency) but instead by how they scale these values. Each type of mean locates a balance point among a set of items. But the different means spread these items apart and squish them together differently. Thus, they provide greater leverage to either higher or lower values, i.e. to either common or rare species. Many ecologists are already familiar with this scaling process as it is directly analogous to the use of link functions in generalized linear models. We explore this new way of visualizing Hill diversities, and the different forms of means generally, in Box 5.

Which Hill diversity to use?

Which variant of Hill diversity to use, then? There is no one answer to this question. As Southwood quipped, about diversity indices in general, 'There can be no universal 'best-buy,' although there are rich opportunities for inappropriate usage' (Southwood 1978). Hill diversity diminishes these opportunities, because Hill diversities require researchers to consciously choose how much leverage they want to afford to rare species. This decision is reflected in the value of the exponent ℓ. We discuss some advantages and disadvantages of using different values of ℓ below.

Species richness ($\ell = 1$) is not recommended by any of the authors who have systematically tested diversity metrics (Hurlbert 1971, Kempton 1979, Magurran and McGill 2011, Chase and Knight 2013, Haegeman et al. 2013), because it is difficult to estimate accurately outside of an experimental setting. Sample richness varies drastically with sample size and sample equalization method. This is because it is very sensitive to the rarest species. The same problem affects asymptotic richness estimators (Melo 2004, Chao and Jost 2015). Species richness is best reserved for special cases, such as when the community is completely known, or possibly, when there is enough information to parameterize an occupancy model (Guillera-Arroita et al. 2019).

Hill–Simpson diversity ($\ell = -1$) may be a good choice for a research question that mainly concerns the patterns in the relative abundances of common species, requires confidence that the expected diversity would not change substantially

with additional sampling, or relates to the probability that two randomly selected individuals are the same species (Simpson 1949, Hurlbert 1971). The reciprocal scale used to calculate Hill–Simpson diversity spreads low rarity values apart and squishes high ones together (Box 5). Therefore, Hill–Simpson diversity is most sensitive to the differences in low rarity values (i.e. the relative abundance of common species). The expected value of sample Hill–Simpson diversity tends to be robust to sample standardization and to change little as sample sizes increases. Furthermore, true Hill–Simpson diversity

Box 5. A new way to visualize mean rarity

When ecologists calculate Hill diversity, they effectively calculate the arithmetic, geometric or harmonic mean species rarity. The exponent ℓ in Eq. B2, which scales the rarities and determines what type of mean is calculated, could also be thought of as a link function. Every ecologist has used a link function to transform values onto a scale at which they will be additive, calculated the mean, and then back-transformed the mean onto the original scale. In fact, we do this just to calculate the standard deviation of a sample.

To calculate the standard deviation, we raise each difference from the mean to the power 2, add these new, squared values together, divide by the sample size, and then back-transform to the original scale of the data by raising the computed mean to the power of 1/2. In other words, we use the quadratic link function. The root mean square error of a model is computed the same way.

A generalized linear model with an identity link estimates the arithmetic mean of the data, and could be thought of as raising each value to the power of 1, taking the mean, and back transforming the mean by raising to the power of 1. Of course, this is the same as not transforming at all. When a log link is used in a generalized linear model, the data are transformed by taking the logarithm, and then typically, the mean is back transformed to the original scale by exponentiating. Thus, the mean that is calculated with the log link is the geometric mean (which is the limit of the generalized mean when the scaling exponent ℓ approaches 0). The harmonic mean uses the reciprocal function as the link (to transform, raise to the power of -1; to back transform, raise to the power of -1). A similar link function is used with gamma error structures in generalized linear models.

But what do these transformed scales, these link functions, look like (Fig. B3)?

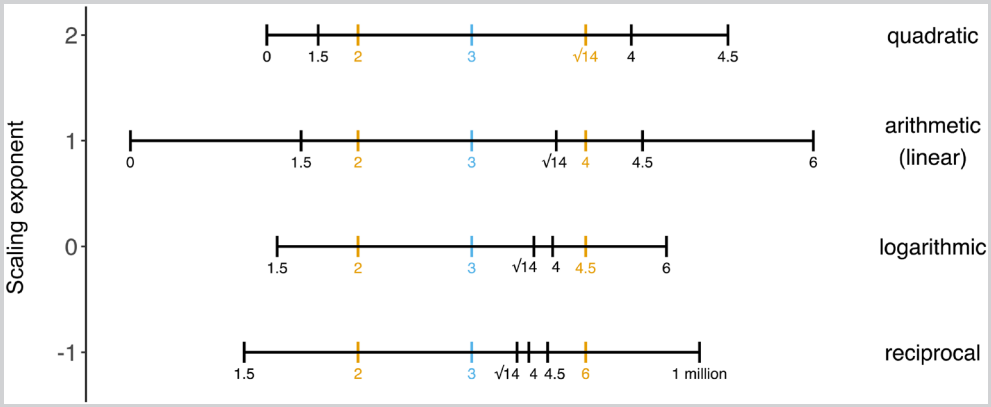

Figure B3. A link function can be visualized in terms of how it scales values. On each scale, a power transformation with the scaling exponent shown on the left, the two yellow points are equidistant from the blue one. It is not possible to align more than 2 points on two different scales; here, each scale is aligned to show that 2 and the higher yellow value are equidistant from 3. On the quadratic scale, the distance between two values is the difference between their squares. Thus, the distance between 2 and 3 is equal to the distance between 3 and $\sqrt{14}$ (~3.74) because 2^2 (4) and $(\sqrt{14})^2$ (14) both differ from 3^2 (9) by 5. On the arithmetic scale, distances between pairs of values are their arithmetic differences. Thus, the distance between 3 and 2 is equal to the distance between 3 and 4; both differ from 3 by 1. On the logarithmic scale, the distance between two values is the factor (proportion) by which the two values differ. Thus, the distance between 3 and 2 is equal to the distance between 3 and 4.5, because both differ from 3 by a factor of 1.5. On the reciprocal scale, the distance between two values is equal to the difference in their reciprocals. Thus, the distance between 3 and 2 is equal to the distance between 3 and 6.

The three 'link functions' used in computing the arithmetic, geometric and harmonic means correspond to scaling exponents of 1, 0 and -1 in Eq. B2, respectively. The mean of a set of values, when put on the appropriate scale, is the balance point between them. This could be visualized as the fulcrum on a balanced lever. The scales differ in which values are spaced farthest apart (Fig. B3), and thus which extreme values will be most displaced from the center, or given the highest leverage. As the scaling exponent decreases, the leverage afforded to high values shrinks, and the leverage afforded the lowest values grows. For example, relative to the arithmetic scale (exponent = 1), a log-transformation (exponent = 0) spreads the small values out but compresses the largest values together.

When thinking about the different Hill diversities, it may be useful to consider this leverage metaphor. Historically, the differences between Hill diversities with different exponents (for example, the difference between species richness, Hill–Shannon and Hill–Simpson) have been discussed in terms of how heavily the exponents 'weight' rare or abundant species (Jost 2006, Magurran and McGill 2011). From Eq. B2, it is clear that this is not the simplest interpretation. Regardless of the exponent, each species is always weighted by its relative abundance, and every individual 'counts' towards the average by the same amount. What changes with the exponent is the scaling of the species' rarities, or how far apart rarity values fall (Fig. B4).

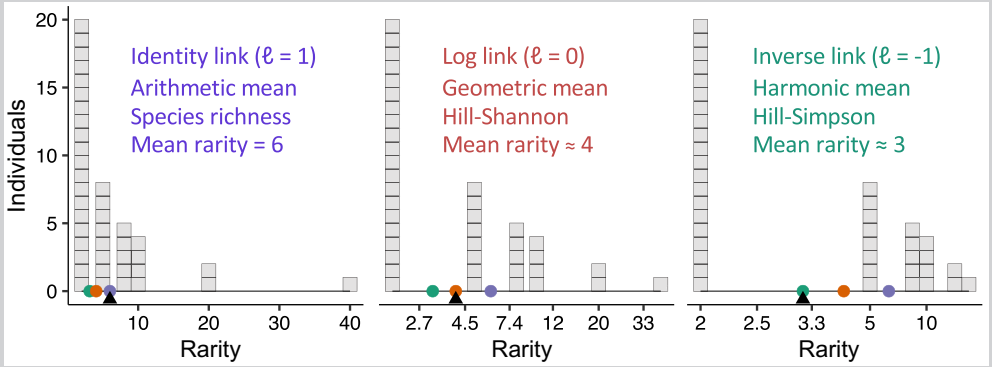

Figure B4. Diversity, or mean rarity, is the balance point for the community along the rarity scale. Each panel shows the same ecological community of 40 individuals and 6 species with abundances 20, 8, 5, 4, 2 and 1, and rarities 2, 5, 8, 10, 20 and 40, respectively. Each block represents an individual; in this metaphor, the 'mass' of each 'block' is the same. Each individual's x-axis value is its species' rarity, which is the reciprocal of its relative abundance. All three panels display the same community: the same individuals, the same species, and the same rarities; only the scaling of the rarities changes between panels. The community's balance point along the rarity scale, pictured in this figure as the fulcrum in each panel, is the mean rarity, or diversity, of the community. To ease comparison across scales, in each panel, we marked the arithmetic mean with a rose dot, the geometric mean with a blue dot and the harmonic mean with a green dot.

Panels differ in which exponent is used to transform the rarity scale. The arithmetic scale provides high leverage to very rare species; although they carry little weight (few individuals), these species influence the mean a great deal because they sit far to the right of the rarity scale. The arithmetic mean rarity of the community is the Hill diversity when $\ell = 1$, and is equal to species richness (value = 6). The logarithmic scale provides less leverage to very rare species. Thus, the geometric mean rarity of the community is lower (value ≈ 4). The geometric mean rarity is also known as the Hill–Shannon diversity, or the Hill diversity when $\ell = 0$. The reciprocal scale accords more leverage to low rarity values. Thus, the harmonic mean rarity, also known as the Hill–Simpson diversity, or Hill diversity when $\ell = -1$, is much lower still (value ≈ 3). An interactive online application that enables users to specify species abundances and the scaling parameter is available at <https://mean-rarity.shinyapps.io/rshiny_app1/>, and code for this and all main text figures is in the R package *MeanRarity* (Roswell and Dushoff 2020).

may be estimated with little bias (Simpson 1949, Chao and Jost 2015, Grabchak et al. 2017), although the uncertainty in these estimates shrinks slowly with additional sampling, and precise estimates remain difficult in more diverse, more even communities.

Hill–Shannon diversity ($\ell = 0$) lies between richness and Hill–Simpson diversity, and may be the 'just right' measure in many applications (Kempton 1979). The geometric mean affords leverage to extreme values according to their proportional, not absolute, difference from the mean. Thus, it can respond strongly to both very high and to very low rarity values. Another argument in favor of Hill–Shannon is that many species abundance distributions are approximately log-normal (Williamson and Gaston 2005, McGill et al. 2007), and thus their central tendency might be well described by the geometric mean. Observed Hill–Shannon diversity begins to stabilize at achievable sample sizes, and asymptotic estimators for Hill–Shannon diversity perform reasonably well (Beck and Schwanghart 2010). The Hill–Shannon diversity retains some of the sensitivity of Hill diversities with higher exponents (such as richness), and also the robustness to sampling and sample standardization of Hill diversities with lower exponents (such as Hill–Simpson diversity). As a result, Hill–Shannon may be a good choice for characterizing gradients in biodiversity in an ecologically meaningful way.

For research questions about diversity in a more general sense, researchers should consider using all three metrics, as well as intermediate values for the exponent ℓ (Fig. 3). Although Hill diversities with different scaling exponents tend to be highly correlated within communities (Magurran and McGill 2011), they emphasize different aspects of the community, and are not fully exchangeable (Hurlbert 1971, Patil and Taillie 1982). Using more than one diversity metric portrays the diversities of the communities most fully because, for example, one community can be the most diverse when its many rare species are given great leverage (when ℓ is large), but a different community most diverse when its more even distribution of more common species is emphasized (when ℓ is small) (Patil and Taillie 1982). Furthermore, Hill diversities with different exponents can be compared to describe evenness and dominance, and to fully describe the shape of a species abundance distribution (Hill 1973, Chao and Jost 2015, Chao and Ricotta 2019). Diversity profiles are complex and information rich, and therefore simple statistical methods to compare them are unavailable.

A diversity profile is constructed by estimating Hill diversity over the range of ℓ values. Researchers can do this in R with the functions 'Diversity' in the package *SpadeR* (for asymptotic and sample diversities with estimated uncertainty; plotting features built in), 'iNEXT' in the package *iNEXT* (for asymptotic and coverage-rarefied diversity estimates,

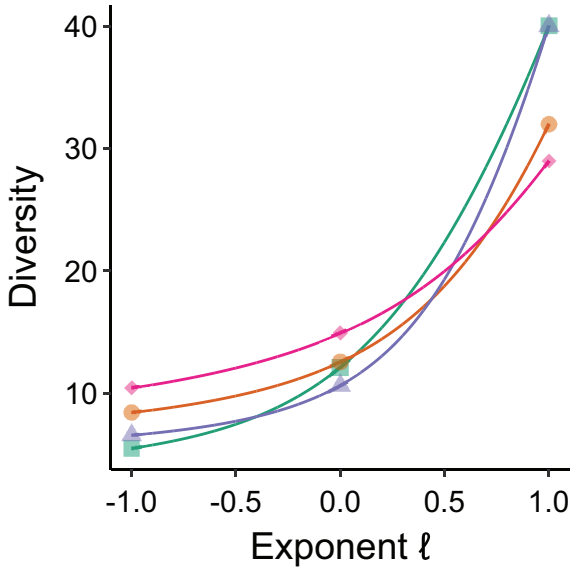

Figure 3. It can be useful to visualize a 'diversity profile' across values of the exponent ℓ. Here we show the sample diversities of our four bee communities plotted as a function of the exponent ℓ in Eq. B2; as ℓ increases, so does the leverage afforded to rare species. The y-axis is the value of the diversity metric, as calculated from the raw sample. The lines can cross because a sample can have, for example, a large number of rare species (high richness, rightmost points) but a small number of common species (low Hill–Simpson, leftmost points), as compared with another sample (middle points are Hill–Shannon).

with estimated uncertainty), and 'renyi' in the package *vegan* or 'rarity' in the package *MeanRarity* (raw sample diversity) (Chao et al. 2015, Hsieh et al. 2016, Oksanen 2016, Roswell and Dushoff 2020). Each of these packages parameterizes Hill diversity with the exponent $q = 1 - \ell$.

Ecologists rarely use Hill numbers with $\ell > 1$; these metrics are too sensitive to rare species to convey meaningful information about communities (Chao et al. 2014a). The Hill diversity as ℓ approaches negative infinity equals the relative abundance of only the most common species. This Hill diversity has been used as a dominance index (Berger and Parker 1970). However, Chao and colleagues suggest that little information about a community's diversity is revealed between $\ell = -2$ and $-\infty$ (Chao et al. 2014a), and suggest calculating Hill diversity profiles from $\ell = -2$ to 1 (Chao and Jost 2015).

Extrapolation and asymptotic estimators

A note on extrapolation per se

To this point, we have discussed standardizing via rarefaction, but not **extrapolation** – that is, extending the pattern of species detection to a greater sample size, effort or coverage than has actually been obtained. Inferring what one might have seen with additional sampling is obviously appealing, and compatible, at least in principle, with sample standardization. The past decade saw the introduction of unified methods for rarefaction and extrapolation for diversity estimation, based not only on sample size or effort, but also on sample coverage (Chao and Jost 2012, Colwell et al. 2012, Chao et al. 2014a).

Standardizing to a level that involves extrapolation for at least some samples could be preferable to, for example, rarefying to the size of the smallest sample and analyzing very incomplete samples for every community. The caveat is that the farther from the sample an extrapolation extends, the more sensitive it is to the extrapolation method's assumptions. A second issue is that neither empirical nor theoretical work yet guides what level of sample completeness enables robust comparison – though Chao and Jost (2012) suggest that extrapolating to double the observed sample size entails little risk even for species richness, and may allow researchers to take better advantage of data from well-sampled communities. Because of these complexities and unresolved issues, guidance on extrapolation per se is beyond the scope of this guide. However, below we discuss asymptotic diversity estimators, a popular technique that could be considered an extreme form of extrapolation.

Are asymptotic estimators the solution?

Standardizing samples and then calculating sample diversity is an imperfect approach to comparing true diversities; sample diversity is not expected to equal true diversity (Hurlbert 1971, Dauby and Hardy 2012, Chao et al. 2014a, Willis 2019). An alternative method is using asymptotic estimators to predict what diversity would be, if each community were censused completely (and therefore the diversity accumulation curve would reach its asymptote; Chao and Jost 2015). Asymptotic estimators that do this are quite popular. However, we believe that asymptotic estimators have two important limitations that ecologists sometimes overlook.

First, at feasible levels of sampling, asymptotic Hill diversity estimators frequently do not reach their own asymptote (Melo 2004, Beck and Schwanghart 2010, Chiu and Chao 2016). In other words, the estimate of the asymptotic, 'true' diversity value can change systematically with sampling (Fig. 2). This means that the diversity estimates given by asymptotic estimators depend on sample completeness, which hinders comparisons between communities and between studies.

Second, the uncertainty associated with asymptotic estimators can be large and difficult to quantify, particularly for richness (Haegeman et al. 2013). When sample coverage is low, the approximated confidence intervals (CI) around asymptotic diversity estimates for all Hill diversities are wide. Even so, they are not reliably wide enough: strictly speaking, a valid confidence interval contains the target parameter at least as often as the stated confidence level (Casella and Berger 2002). However, the CI for asymptotic Hill diversity estimators frequently do not overlap the true community diversity at their stated level (e.g. 95%;

Mao et al. 2017). For example, for a simulated community with a richness of 200, a Hill–Simpson diversity of 50, and a log-normal distribution of species relative abundances, the '95% CI' around Chao1 asymptotic estimates of richness include the true richness value less than 50% of the time for a random sample of a few hundred individuals (Supplementary information). While the Chao1 richness estimator (and closely related Chao2 for incidence data) are theoretically only 'lower bounds' for true species richness (Chao 1984, 1987), anticonservatism in the proposed intervals is evident for asymptotic Hill–Shannon and Hill–Simpson diversity estimates as well (for the sample sizes and log-normal abundance distribution tested here; Supplementary information).

Confidence intervals for Hill diversity estimates remain under development. The CI for sample Hill diversity – under traditional rarefaction – include the expected sample diversity at a rate closer to their stated confidence level than we observed for asymptotic estimators (Chao et al. 2014a, Chao and Jost 2015; Supplementary information). Of course, CI for sample diversities and CI for asymptotic estimators are trying to do two different things. CI for sample diversities aim simply to contain the expected diversity of a sample, conditioned on size or effort (Smith and Grassle 1977). The CI for asymptotic estimators, by contrast, aspire to contain the true diversity of the whole community – but often they do not. CI for expected diversity after standardizing by coverage are also anti-conservative (Box 1, Supplementary information).

Is lacking valid confidence intervals a fatal flaw for a method to estimate diversity? We believe it depends on the application. Ecologists studying biodiversity will likely estimate biodiversity across many communities, and then use a statistical model to understand how biodiversity responds to predictors, such as forest cover or temperature. In the model, the uncertainty in the diversity estimates gets conflated with unmodeled but true variation between communities, and both contribute to the regression's error term. This problem can be remediated by increasing sample sizes (i.e. diversity estimates from more communities). For example, imagine sampling logged and unlogged forests to determine how logging affects species diversity (Chao and Jost 2012). Using a method such as standardizing by coverage or computing asymptotic diversity estimates may fail to provide a reliable estimate for any given site, but if enough sites are sampled, could reliably identify a group-level pattern. These methods would be preferable to a method that gave misleading estimates with better-known sampling uncertainty, such as sample diversity estimates under traditional rarefaction (Chao and Jost 2012).

Whether to use coverage or asymptotic estimators

Standardizing by coverage, like using asymptotic estimators, should preserve relative differences in diversity better than traditional rarefaction (Chao and Jost 2012, 2015, Chao et al. 2014a). Yet both of these approaches lack valid CI, and both are sensitive to sample completeness. Overall, we identify a subtle advantage to using sample diversity after standardizing by coverage, rather than using asymptotic diversity estimators.

Coverage is explicit about sampling completeness, whereas asymptotic estimators attempt to estimate the true diversity of the full community, a quantity that is not conditioned on sampling… but the resulting estimate often is. Because neither method is robust to sampling completeness, we advocate using the one (i.e. coverage) that both accounts for sample completeness and describes it in ecologically meaningful terms. Conditioning comparisons on sample completeness can help ecologists guard against interpreting patterns that reflect researcher decisions, rather than ecological processes.

Why not use coverage and asymptotic estimators together?

In practice, ecologists typically choose a method of sample standardization (effort, sample size, or increasingly, coverage), often involving rarefaction, or use asymptotic diversity estimators to extrapolate unstandardized samples. However, it is tempting to combine the two methods, because the asymptotic estimators themselves tend to be sensitive to sampling completeness (Close et al. 2018).

While this sounds promising, in our view there are important issues to resolve before coverage-based sample standardization should be combined with asymptotic estimators. How to combine the two tools has not yet been systematically developed or tested; there is not even evidence that any combination provides an advantage over using one or the other alone. Future theoretical and simulation-based work could build the case for a combined approach.

Standardizing samples, then calculating Hill diversity: a worked example with our bee data

In this section we provide a demonstration analysis, using some of our own data, to show how the researcher's choice of how to standardize samples and calculate diversity affects interpretation of diversity patterns. We use three data standardization methods (effort, size and coverage), as well as all three Hill diversities (richness, Hill–Shannon and Hill–Simpson). We also compare asymptotic Hill diversity estimates to the standardized sample diversities. Our purpose is not to determine the accuracy of these methods, which we cannot do: we do not know the true diversities of our bee communities. Rather, our goal is to show how our choice of standardization method and diversity metric, as well as our level of sampling, can determine the results. Because ecologists would likely use available uncertainty estimates when analyzing their own data, we have included these in our example.

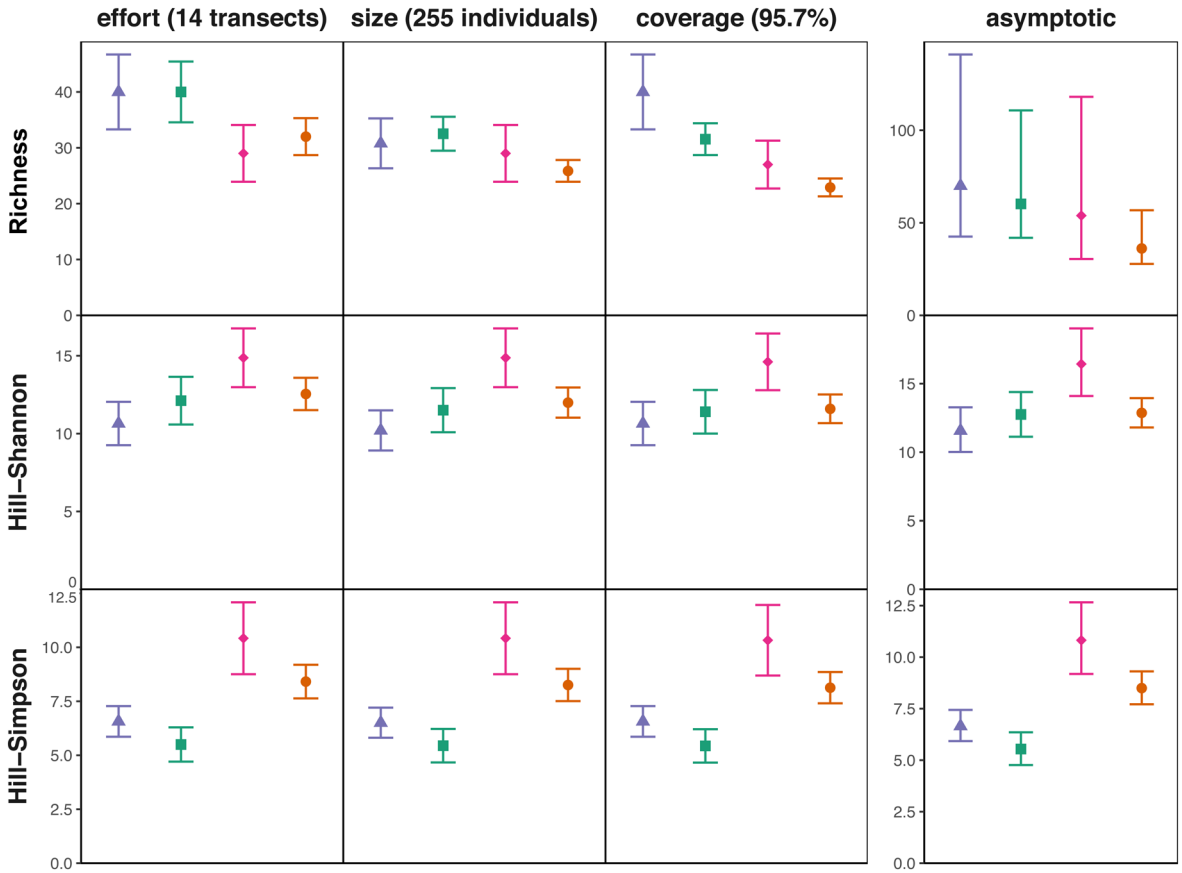

Figure 4. The answer to, 'which communities are more and less diverse, and by how much?' depends on both how the samples are standardized (columns), and which diversity metric is used (rows). Standardization method matters most when Hill diversity is strongly driven by the rarity of the rarest species (species richness, top row) and matters least when rare species have little leverage (Hill–Simpson, bottom row). Error bars are '95% CI' that assume uncertainty arises from the process of randomly sampling a fixed number of individuals (i.e. the number of individuals in the sample after standardization) from each community; raw (i.e. equal effort) samples used for asymptotic estimates. We plotted the asymptotic Hill diversity estimators with separate y-axes, to facilitate comparing relative differences in estimated diversity.

The data presented were not collected to compare local diversity between the sites shown (Roswell et al. 2019a, b), so we leave it to the reader to imagine conditions that could motivate such a comparison. We could imagine, for example, that only one site can be preserved, and the goal of measuring diversity is to identify the highest-value (highest diversity) site to protect, or that different plant restoration methods were deployed at different sites, and the goal of measuring diversity is to assess restoration success.

In our example dataset, we note first that which community we would judge most diverse, and by how much, depends on how we standardize our data. This can be seen by focusing on one row at a time in Fig. 4. When we standardize by size, we could conclude that species richness is fairly similar across the four bee communities, but when we standardize by effort or coverage, strong differences among communities emerge. These findings reinforce our argument that sample standardization is an important choice that researchers need to make carefully when measuring diversity.

Second, the choice of Hill diversity (richness, Hill–Shannon or Hill–Simpson) drives our understanding of the relative diversity of these four communities. We can see this by focusing on one column at a time in Fig. 4. For example, consider the column for which the data are standardized by coverage (Fig. 4, third column from left). Using richness as our metric indicates that there are large differences in diversity between the four communities, and that the purple community is the most diverse. Using Hill–Shannon or Hill–Simpson, however, leads us to the conclusion that the pink community is most diverse. Furthermore, when Hill–Shannon diversity is used, the pink diamond community appears around 25% more diverse than the purple triangle community, but when Hill–Simpson diversity is used, this difference increases to about 80%. We expected this result, because communities with many rare species need not also be 'most diverse' when using metrics that emphasize rare species less. This underscores the importance of researchers explicitly stating what aspects of diversity matter most for a particular question, and then choosing the appropriate Hill diversity to reflect those aspects.

Third, there are interactions between the choice of sample standardization method and the choice of diversity metric. As expected, although relative Hill–Shannon

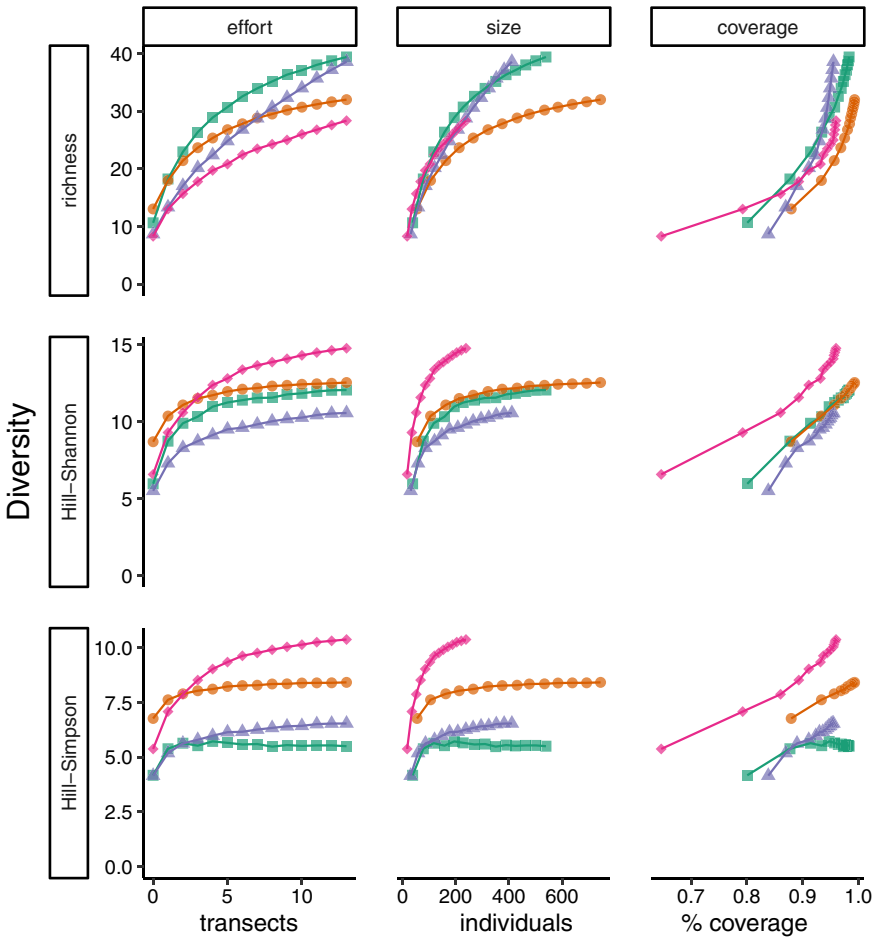

Figure 5. In addition to standardization method (columns), Hill diversities (rows) are sensitive to the amount of sampling (x-axis). To generate this figure, one to fourteen 30-min data collection events per community were resampled without replacement 9999 times. For each group of 9999 random subsamples, average effort (number of 30-min transects resampled), average size (number of individuals), or average coverage is plotted on the x-axis, and average sample diversity is plotted on the y (with tick marks places at log scaled intervals, but actual diversity values shown). The logarithmic y-axis reveals a constant relative difference in diversity as a constant distance between lines. Uncertainty estimates omitted for clarity. Diversity often increases rapidly as coverage gets very close to one, because in our communities (and in natural communities in general) there are many rare species, each of which makes up a small share of the total abundance.

diversities still depend on standardization, Hill–Shannon is far more robust than richness to standardization method. Hill–Simpson diversity is even more robust than Hill–Shannon diversity. While the absolute values of asymptotic diversity estimates are higher than the sample diversities, we see similar relative diversity patterns using asymptotic estimators and coverage-based rarefaction. The robustness of Hill–Shannon and Hill–Simpson to sample standardization method is a strong argument for using these Hill diversities, rather than richness.

Finally, we note that the relative diversities for our four samples are sensitive to sampling completeness, even after standardization (Fig. 5). If any Hill diversities were robust to sampling completeness, we would observe a constant distance between the lines for each community within a panel of Fig. 5 (i.e. the colored lines would increase in parallel). Clearly this is not the case for any of the Hill diversities shown, regardless of how we standardize samples. It is also not the case for asymptotic Hill diversity estimators, which exhibit different – but not less – sensitivity to sampling completeness up to the point we sampled each community (Supplementary information). This should be concerning to field ecologists, who rarely have the luxury of comparing complete samples. Even the sample diversity rankings, not to mention the relative differences in diversity, vary with sample completeness for all Hill diversities. In sum, sampling completeness almost always affects diversity estimates.

Overall, we find that if we were testing a hypothesis about biodiversity responses, or making management decisions based on diversity rankings, our findings could change dramatically with the choice of Hill diversity, the method of data standardization and the amount of sampling. The results

from this pedagogically selected dataset do not indicate what happens in all datasets, yet they illustrate that those choices can matter.

Premises and promises of Hill diversity and coverage

Two premises underlying the tools reviewed in this guide are: 1) when ecologists measure the biodiversity of a community, they have a notion of the spatial, temporal and taxonomic scale at which the community exists, and 2) that relative, but not absolute, abundance determines biodiversity. Furthermore, many of the tools assume that the pool of individuals sampled from a community is static, well-mixed, and in some cases infinite, whereas in reality species abundances and spatial distributions fluctuate through time, communities are rarely truly 'closed,' and ecologists often sample destructively, removing individuals from the pool as they sample.

Measuring diversity at only one or a few spatio-temporal scales may be insufficient to describe biodiversity gradients in nature (May et al. 2018). Hill diversity does enable describing the scale-dependent nature of biodiversity by partitioning diversity into alpha (local), beta (dissimilarity/turnover) and gamma (larger-scale) components that can be normalized to compare patterns across different regions and timescales (Chao and Chiu 2016). Hill diversities with different scaling exponents are not expected to respond identically as grain size, study extent or sampling intensity increase; contrasting these responses may reveal processes of ecological interest (Chase and Knight 2013, Chao and Chiu 2016, Chase et al. 2018). To date, research on the scale-dependent nature of biodiversity measurement focuses heavily on fitting curves to species abundance distributions (Williamson and Gaston 2005, Matthews et al. 2019), or richness- and occasionally Hill–Simpson diversity-based measures (Chase and Knight 2013, Chase et al. 2018, 2019, Antão et al. 2019, Ricotta et al. 2019).

Using a range of Hill diversities, including richness, may help ecologists refine hypotheses and models of biodiversity response to scale and global change, which local richness alone can, famously, fail to illustrate clearly (Chase and Knight 2013, Cardinale et al. 2018). When coverage and Hill diversity are used together, they should enable ecologists to separate patterns in relative abundances from patterns in total abundance, address artefacts of sampling completeness at different scales (Kraft et al. 2011, Engel et al. 2020), and develop richer hypotheses about patterns in species abundances over space and time (McGill et al. 2007).

Conclusions

The unavoidable truth is that when ecologists compare local diversity, they must choose how sensitive their comparison will be to the rarest species, which are always inadequately represented in samples. There is no robust way to simply 'count' the species in most natural communities; richness estimated from samples depends on species' relative abundances and sampling completeness.

Whereas ecologists usually cannot compare true species richness, we have shown how ecologists can compare communities using sample richness, Hill–Shannon and Hill–Simpson diversity, after rarefying samples to equal coverage. Using Hill diversities requires only minor modifications to the diversity metrics that ecologists already use. These small modifications make a big difference, as Hill diversities scale intuitively, are always expressed as rarities, and require that ecologists explicitly choose how sensitive their diversity metric is to rare species.

Standardizing samples by coverage improves upon simply acknowledging the fact that both sample size and the true distribution of species abundances drive diversity estimates. To capture the diversity of more diverse communities, larger samples are needed. Coverage measures how representative samples are of the communities from which they are drawn, so equalizing coverage before measuring diversity can reduce bias in biodiversity comparisons. When it is possible to estimate sample coverage while data collection is ongoing, this could help researchers allocate effort more efficiently, guiding which communities require more or less sampling (Rasmussen and Starr 1979).

Even though the tools contained in this guide are the best available at present, they are still under development. Ecologists still lack heuristics for identifying sufficient sample coverage levels, for choosing the appropriate Hill diversity scaling exponent(s) for a given question or dataset, and for robustly accounting for uncertainty in diversity estimates. Nonetheless, when researchers have a strong argument for comparing the species diversity of communities, the tools in this guide should facilitate doing so in a principled manner. Using coverage with Hill diversity, ecologists can assess relative differences in diversity between communities based on sample data, while clearly expressing the sample completeness upon which their inferences depend.

Data availability statement

Data and code available in the R package *MeanRarity* (Roswell and Dushoff 2020).

Acknowledgements – Thanks to Tina Harrison, Colleen Smith, Bob Holt, Sam Scheiner, Jimmy Peniston, Myla Aronson, Brooke Maslo, Jean Marie Hartman, Jarrett Byrnes, Andrew Gonzalez and Anne Chao for valuable suggestions on earlier drafts of this manuscript.
Funding– This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. In addition to NSF grant number NSFDGE1433187, MR was supported by an Excellence Fellowship from Rutgers University.

Glossary

Sample diversity	The (Hill) diversity of a sample. This quantity can be calculated directly, as the number of species and their relative abundances are known.
True diversity	The true (Hill) diversity of an entire community.
Asymptotic diversity	An estimate of the true (Hill) diversity of the community. This is known as the 'asymptotic' diversity because as the sample size increases, sample diversity and other diversity estimates converge on their true values, which are seldom known a priori.
Coverage	The proportion of individuals in the community belonging to species represented in a sample.
Hill diversity	Also called Hill numbers; the generalized mean species rarity.
Hill–Richness	The Hill diversity when $\ell = 1$, the arithmetic mean rarity, or the total number of species. Referred to simply as 'richness' throughout.
Hill–Shannon	The Hill diversity when $\ell = 0$, the geometric mean rarity, or the exponential of Shannon's entropy.
Hill–Simpson	The Hill diversity when $\ell = -1$, the harmonic mean rarity, or the inverse of Simpson's concentration index.
Leverage	The influence of a value on the mean depends on the frequency of that value ('weight'), but also its displacement from other values in the set ('leverage'). The farther a given value from the others, the more leverage that value has. Rescaling shifts leverage from low to high values, or vice-versa.
Rarefaction	A process of randomly subsampling by removing individuals or subsamples.
Extrapolation	An approach to estimating the diversity of an augmented sample that may resemble rarefaction 'in reverse.'
Rarity	1/relative abundance (Patil and Taillie 1982).

Author contributions

Michael Roswell: Conceptualization (lead); Data curation (lead); Formal analysis (lead); Funding acquisition (lead); Software (equal); Visualization (lead); Writing – original draft (equal); Writing – review and editing (equal). **Jonathan Dushoff**: Conceptualization (supporting); Formal analysis (supporting); Methodology (supporting); Software (equal); Supervision (supporting); Visualization (supporting); Writing – original draft (supporting); Writing – review and editing (equal). **Rachael Winfree**: Conceptualization (supporting); Data curation (supporting); Formal analysis (supporting); Funding acquisition (supporting); Resources (supporting); Supervision (lead); Writing – original draft (equal); Writing – review and editing (equal).

References

Alberdi, A. and Gilbert, M. T. P. 2019. A guide to the application of Hill numbers to DNA-based diversity analyses. – Mol. Ecol. Resour. 19: 804–817.

Alroy, J. 2010. The shifting balance of diversity among major marine animal groups. – Science 329: 1191–1194.

Alroy, J. 2017. Effects of habitat disturbance on tropical forest biodiversity. – Proc. Natl Acad. Sci. USA 114: 6056–6061.

Antão, L. H. et al. 2019. B-diversity scaling patterns are consistent across metrics and taxa. – Ecography 42: 1012–1023.

Beck, J. and Schwanghart, W. 2010. Comparing measures of species diversity from incomplete inventories: an update. – Methods Ecol. Evol. 1: 38–44.

Berger, W. H. and Parker, F. L. 1970. Diversity of planktonic foraminifera in deep-sea sediments. – Science 168: 1345–1347.

Botta-Dukát, Z. 2018. The generalized replication principle and the partitioning of functional diversity into independent alpha and beta components. – Ecography 41: 40–50.

Brose, U. et al. 2003. Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns. – Ecology 84: 2364–2377.

Bullen, P. S. 2003. Handbook of means and their inequalities. – Kluwer Academic Publishers.

Cao, Y. et al. 2007. Effects of sample standardization on mean species detectabilities and estimates of relative differences in species richness among assemblages. – Am. Nat. 170: 381–395.

Cardinale, B. J. et al. 2018. Is local biodiversity declining or not? A summary of the debate over analysis of species richness time trends. – Biol. Conserv. 219: 175–183.

Casella, G. and Berger, R. L. 2002. Statistical inference. – Wadsworth.

Chao, A. 1984. Nonparametric estimation of the number of classes in a population. – Scand. J. Stat. 11: 265–270.

Chao, A. 1987. Estimating the population size for capture–recapture data with unequal catchability. – Biometrics 43: 783–791.

Chao, A. and Chiu, C. H. 2016. Bridging the variance and diversity decomposition approaches to beta diversity via similarity and differentiation measures. – Methods Ecol. Evol. 7: 919–928.

Chao, A. and Jost, L. 2012. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. – Ecology 93: 2533–2547.

Chao, A. and Jost, L. 2015. Estimating diversity and entropy profiles via discovery rates of new species. – Methods Ecol. Evol. 6: 873–882.

Chao, A. and Ricotta, C. 2019. Quantifying evenness and linking it to diversity, beta diversity and similarity. – Ecology 100: 0–3.

Chao, A. et al. 2014a. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. – Ecol. Monogr. 84: 45–67.

Chao, A. et al. 2014b. Unifying species diversity, phylogenetic diversity, functional diversity and related similarity and differentiation measures through Hill numbers. – Annu. Rev. Ecol. Evol. Syst. 45: 297–324.
Chao, A. et al. 2015. Online Program SpadeR (Species-richness Prediction And Diversity Estimation in R). – <http://chao.stat.nthu.edu.tw/wordpress/wp-content/uploads/software/SpadeR_Introduction.pdf>.
Chao, A. et al. 2019a. Quantifying evenness and linking it to diversity, beta diversity, and similarity. Ecology 100: e02852.
Chao, A. et al. 2019b. Proportional mixture of two rarefaction/extrapolation curves to forecast biodiversity changes under landscape transformation. – Ecol. Lett. 22: 1913–1922.
Chao, A. et al. 2020. Quantifying sample completeness and comparing diversities among assemblages. – Ecol. Res. 35: 292–314.
Chase, J. M. and Knight, T. M. 2013. Scale-dependent effect sizes of ecological drivers on biodiversity: why standardised sampling is not enough. – Ecol. Lett. 16: 17–26.
Chase, J. M. et al. 2018. Embracing scale-dependence to achieve a deeper understanding of biodiversity and its change across communities. – Ecol. Lett. 21: 1737–1751.
Chase, J. M. et al. 2019. Species richness change across spatial scales. – Oikos 128: 1079–1091.
Chiarucci, A. et al. 2011. Old and new challenges in using species diversity for assessing biodiversity. – Phil. Trans. R. Soc. B 366: 2426–2437.
Chiu, C.-H. and Chao, A. 2016. Estimating and comparing microbial diversity in the presence of sequencing errors. – PeerJ 4: e1634.
Close, R. A. et al. 2018. How should we estimate diversity in the fossil record? Testing richness estimators using sampling-standardised discovery curves. – Methods Ecol. Evol. 9: 1386–1400.
Coddington, J. A. et al. 2009. Undersampling bias: the null hypothesis for singleton species in tropical arthropod surveys. – J. Anim. Ecol. 78: 573–584.
Colwell, R. K. et al. 2012. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. – J. Plant Ecol. 5: 3–21.
Cox, K. D. et al. 2017. Community assessment techniques and the implications for rarefaction and extrapolation with Hill numbers. – Ecol. Evol. 7: 11213–111226.
Dauby, G. and Hardy, O. J. 2012. Sampled-based estimation of diversity sensu stricto by transforming Hurlbert diversities into effective number of species. – Ecography 35: 661–672.
Ellison, A. M. 2010. Partitioning diversity. – Ecology 91: 1962–1963.
Engel, T. et al. 2020. Resolving the species pool dependence of beta-diversity using coverage-based rarefaction. – BioRxiv <doi: 10.1101/2020.04.14.040402>.
Good, I. J. and Toulmin, G. H. 1956. The number of new species and the increase in population coverage when a sample is increased. – Biometrika 77: 45–63.
Gotelli, N. J. and Colwell, R. K. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. – Ecol. Lett. 4: 379–391.
Gotelli, N. J. and Colwell, R. K. 2011. Estimating species richness. – In: Biological diversity. Frontiers in measurement and assessment, pp. 39–54.
Grabchak, M. et al. 2017. The generalized Simpson's entropy is a measure of biodiversity. – PLoS One 12: 1–11.
Guillera-Arroita, G. et al. 2019. Inferring species richness using multispecies occupancy modeling: estimation performance and interpretation. – Ecol. Evol. 9: 780–792.
Haegeman, B. et al. 2013. Robust estimation of microbial diversity in theory and in practice. – ISME J. 7: 1092–1101.
Hill, M. O. 1973. Diversity and evenness: a unifying notation and its consequences. – Ecology 54: 427–432.
Hsieh, T. C. et al. 2016. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). – Methods Ecol. Evol. 7: 1451–1456.
Hurlbert, S. H. 1971. The nonconcept of species diversity: a critique and alternative parameters. – Ecology 52: 577–586.
Jost, L. 2006. Entropy and diversity. – Oikos 113: 363–375.
Jost, L. 2007. Partitioning diversity into independent alpha and beta components. – Ecology 88: 2427–2439.
Jost, L. 2008. GST and its relatives do not measure differentiation. – Mol. Ecol. 17: 4015–4026.
Jost, L. 2009. Mismeasuring biological diversity: response to Hoffmann and Hoffmann (2008). – Ecol. Econ. 68: 925–928.
Jost, L. 2010. The relation between evenness and diversity. – Diversity 2: 207–232.
Kang, S. et al. 2016. Hill number as a bacterial diversity measure framework with high-throughput sequence data. – Sci. Rep. 6: 38263.
Kempton, R. A. 1979. The structure of species abundance and measurement of diversity. – Biometrics 35: 307–321.
Kondratyeva, A. et al. 2019. Reconciling the concepts and measures of diversity, rarity and originality in ecology and evolution. – Biol. Rev. 94: 1317–1337.
Kraft, N. J. B. et al. 2011. Disentangling the drivers of β diversity along latitudinal and elevational gradients. – Science 333: 1755–8.
Lande, R. et al. 2000. When species accumulation curves intersect: implications for ranking diversity using small samples. – Oikos 89: 601–605.
Magurran, A. E. and McGill, B. J. 2011. Biological diversity: frontiers in measurement and assessment. – Oxford Univ. Press.
Mao, C. X. et al. 2017. On the asymptotic variance of the Chao estimator for species richness estimation. – Stat. Sin. 27: 1193–1203.
Matthews, T. J. et al. 2019. Systematic variation in North American tree species abundance distributions along macroecological climatic gradients. – Global Ecol. Biogeogr. 28: 601–611.
May, F. et al. 2018. mobsim: an R package for the simulation and measurement of biodiversity across spatial scales. – Methods Ecol. Evol. 9: 1401–1408.
McGill, B. J. et al. 2007. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. – Ecol. Lett. 10: 995–1015.
Melo, A. S. 2004. A critique of the use of jackknife and related non-parametric techniques to estimate species richness. – Community Ecol. 5: 149–157.
Ohlmann, M. et al. 2019. Diversity indices for ecological networks: a unifying framework using Hill numbers. – Ecol. Lett. 22: 737–747.
Oksanen, J. 2016. Vegan: ecological diversity. – <https://cran.r-project.org/web/packages/vegan/vignettes/diversity-vegan.pdf>.
Patil, G. P. and Taillie, C. 1982. Diversity as a concept and its measurement. – J. Am. Stat. Assoc. 77: 548–561.
Preston, F. W. 1948. The commonness, and rarity, of species. – Ecology 29: 254–283.
Rasmussen, S. and Starr, N. 1979. Optimal and adaptive stopping in the search for new species. – J. Am. Stat. 74: 661–667.
Ricotta, C. et al. 2019. Rarefaction of beta diversity. – Ecol. Indic. 107: 105606.

Roswell, M. and Dushoff, J. 2020. MeanRarity: Hill diversity estimation and visualisation. – <https://github.com/mikeroswell/MeanRarity/>.
Roswell, M. et al. 2019a. Data from: Male and female bees show large differences in floral preference. – PLoS One 14: e0214909.
Roswell, M. et al. 2019b. Male and female bees show large differences in floral preference. – PLoS One 14: e0214909.
Shannon, C. E. and Weaver, W. 1963. The mathematical theory of communication. – Univ. Illinois Press 5: 1–131.
Sherwin, W. B. et al. 2017. Information theory broadens the spectrum of molecular ecology and evolution. – Trends Ecol. Evol. 32: 948–963.
Simpson, E. H. 1949. Measurement of diversity. – Nature 163: 688–688.
Smith, W. and Grassle, J. F. 1977. Sampling properties of a family of diversity measures. – Biometrics 33: 283–292.
Southwood, R. 1978. Ecological methods: with particular reference to the study of insect populations. – Chapman and Hall.
Tuomisto, H. 2010. A diversity of beta diversities: straightening up a concept gone awry. Part 2. Quantifying beta diversity and related phenomena. – Ecography 33: 23–45.
Williamson, M. and Gaston, K. J. 2005. The lognormal distribution is not an appropriate null hypothesis for the species-abundance distribution. – J. Anim. Ecol. 74: 409–422.
Willis, A. D. 2019. Rarefaction, alpha diversity and statistics. – Front. Microbiol. 10: 2407.
Zhang, Z. 2016. Statistical implications of Turing's formula. – Wiley.