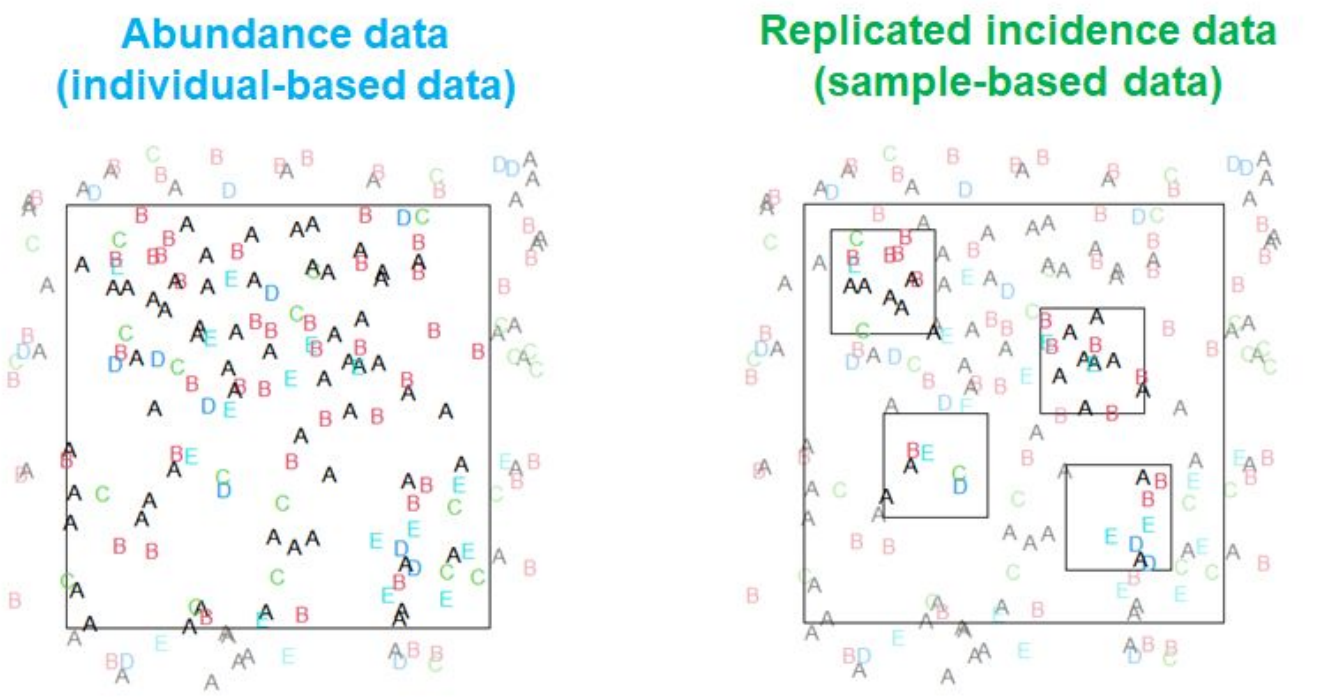
Comparing diversity between communities

All diversity measures suffer from the same problem – they depend on the sampling effort, i.e. energy and time spent by a researcher to discover all species present in a community. The higher the sampling effort, the more species may be discovered. The sampling effort may include the area (or volume) surveyed, the number of individuals, or the number of sampling units (e.g. vegetation plots). If the diversity comparison between communities should be fair, the sampling effort behind their data should be comparable. If the sampling effort is not the same, one solution is to focus on Shannon or Simpson diversity instead of species richness. Species richness is the measure most sensitive to rare species, since each species has equal weight when calculating richness. Rare species (i.e. those represented by few individuals, small biomass or low cover) are those most easily undetected and require disproportionately greater sampling effort than common species. Focusing on Shannon or Simpson diversity shifts the focus on common and dominant species that have higher probability that they will not go undetected.

An alternative solution is to standardize sampled communities to the same sampling effort. In some cases, this may need to be done already in the field (e.g. sampling vegetation by plots of the same site), but sometimes we can standardize data even during the analysis. Below, we will introduce the method of rarefaction and extrapolation curves, which can standardize data with uneven numbers of individuals (or sampling units) or different levels of completeness.

Abundance data vs replicated incidence data

Aim of diversity comparison is usually to compare the true diversity of the communities. However, it is usually impossible (or impractical) to survey the community in its entirety, and we need to rely on less complete data, where limited number of individuals or sampling plots were recorded. For the purpose of further analyses, two alternative types of community data can be recognized: abundance data, containing the number of individuals for each species, and replicated incidence data, containing a set of samples in each of which species' presences and absences were reported. In reality, the third type of data, with multiple samples each containing counts of individuals for species can also be available; such data can be transformed to either abundance data (by pooling individuals from all samples together), or to replicated incidence data (by converting abundances of species in each sample into presences only). Fig. 1 shows the difference between abundance and replicated incidence data sampled in the same community.



	sp A	sp B	sp C	sp D	sp E
indiv.	58	35	15	8	15

	sp A	sp B	sp C	sp D	sp E
samp. 1	1	1	0	1	1
samp. 2	1	1	1	0	1
samp. 3	1	1	1	1	1
samp. 4	1	1	0	0	1
freq.	4	4	2	2	4

Figure 1: The true community has five species (A, B, C, D, and E), and we survey it within a limited area (black rectangle). In the case of surveying abundance data (left), we record the number of individuals and assign each to a species; as a result, we get a vector containing the abundance of individuals for each species. In the case of replicated incidence data (right), we record a set of samples (e.g., vegetation plots), and for each sample, we record species presence or absence; as a result, we obtain a set of samples, which can be pooled into a vector containing the frequency of each species.

Accumulation and rarefaction curves

For abundance data, if we record the first individual, we have one species, and with each new individual, the number of species either increases or remains the same (since the new individual is of a species we already recorded). This can be visualized by an accumulation curve (Fig. 2, from Gotelli & Colwell 2001). The jagged curve represents one scenario of randomly choosing individuals one by one from our sampled pool. The smooth rarefaction curve on abundance data is constructed in a different way: it can be read from right to left, and shows an estimated number of species for a lower number of individuals. Similar accumulation and rarefaction curves can also be constructed for replicated incidence data (instead of individuals, the x-axis represents the number of samples, each containing one or more species). The curves for incidence (sample) data on Fig. 2 are lower than curves for data based on individuals, since multiple individuals/species are clustered in spatially restricted samples. Both curves, however, merge at the end, since they come from the same dataset.

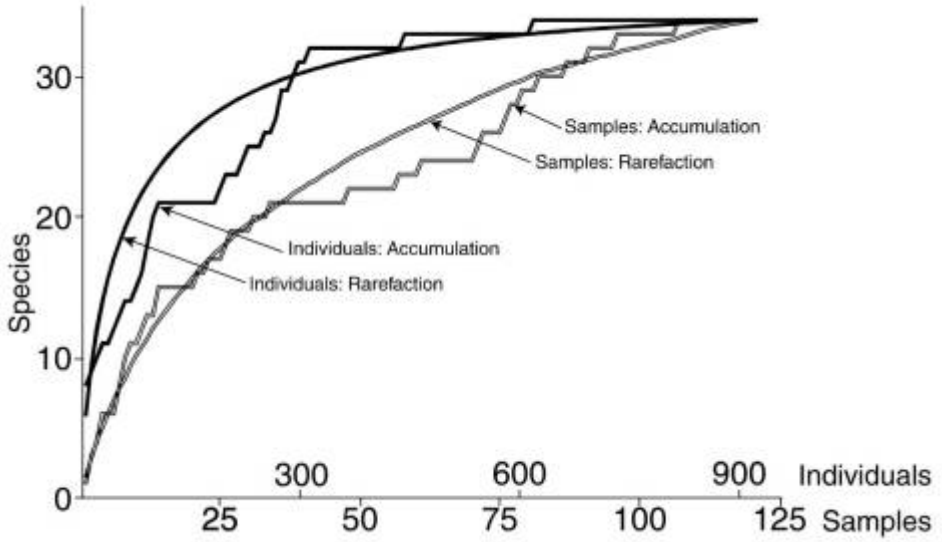


Figure 2: Accumulation and rarefaction curves (Gotelli & Colwell 2001)).

Rarefaction can be accompanied by extrapolation (Chao et al. 2014). While rarefaction predicts the number of species for fewer individuals (abundance data) or fewer samples (incidence data), extrapolation does the same for more individuals/samples. When merged together, they create a rarefaction-extrapolation curve (Fig. 3). Both rarefaction and extrapolation can be calculated using dedicated formulas. It is also possible to construct a confidence interval for the curves, using other formulas, or simply using bootstrap resampling. For $q = 0$, the extrapolation part of the curve can be constructed only to twice the number of individuals/samples in the original dataset; beyond this limit, the estimation is imprecise. For $q = 1$ or 2, the extrapolation can be to infinity.

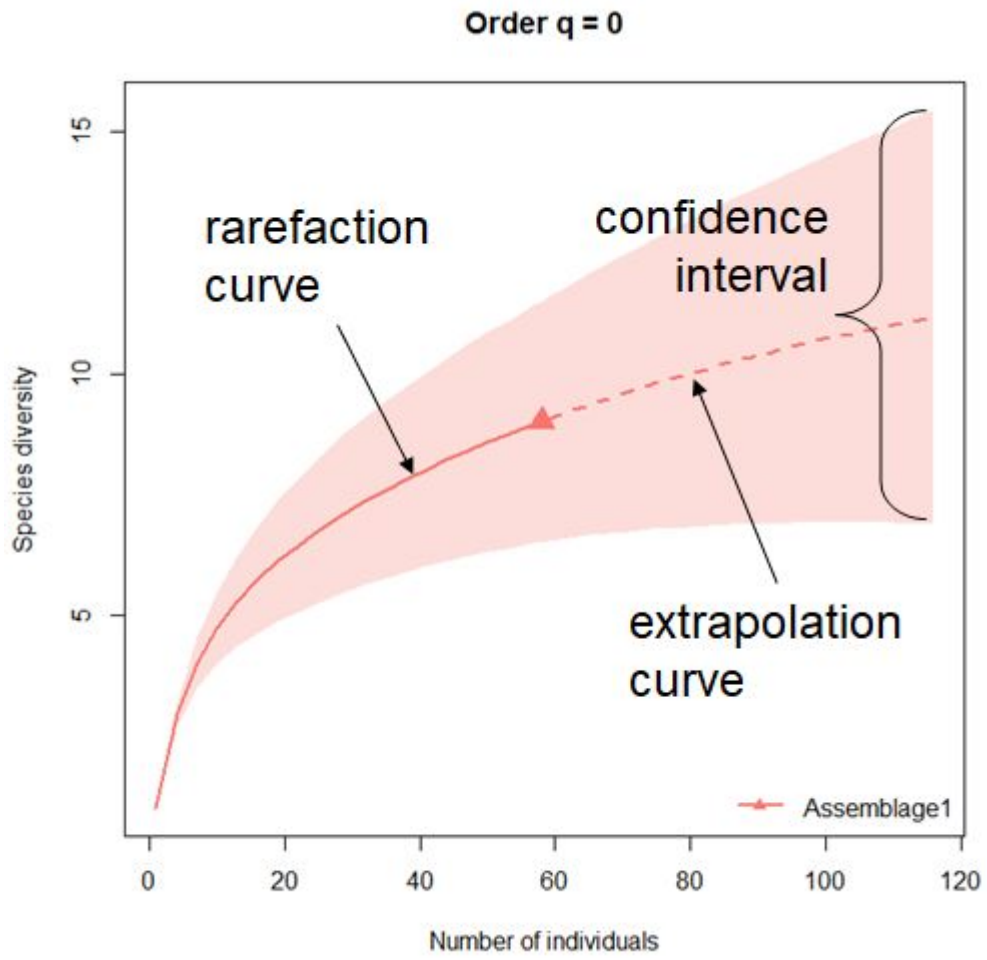


Figure 3: Rarefaction and extrapolation curve. The relationship is based on abundance data.

From:
https://anadat-r.davidzeleny.net/ - **Analysis of community ecology data in R**

Permanent link:
https://anadat-r.davidzeleny.net/doku.php/en:rarefaction

Last update: **2026/05/24 21:26**