PBDB Intensive Summer Course 2007
Paleoecology Section – Thomas Olszewski

Measurement of Diversity

Diversity

Diversity = The variety of organisms in a community; a fundamental property of ecological systems.

Diversity is of interest because it is thought to reflect ecological and evolutionary processes (e.g., immigration, emigration, speciation, extinction, competition, predation, productivity, etc.).

Caveat. There are many different aspects of "variety" (genetic, functional, taxonomic) and there are many different definitions of "community"; often the exact meanings of these terms must be gleaned from context. Much confusion about the measurement and interpretation of diversity has resulted from unclear or inconsistent use of the term *diversity*.

Nature of the Data

Community = a group of populations that occur together; may be spatially delimited, may be trophically delimited, may be taxonomically delimited, may be functionally delimited

Fundamental sampling unit for community studies = a list of species from a specified area/stratigraphic interval (species can be grouped into higher taxonomic categories or functional guilds); abundances – measured as number or percentage of individual specimens, biomass, biovolume, areal coverage, etc. – are commonly also included ("300 rule" – recommended sample size based on binomial distribution)

Defintion of terms:

S = number of species in sample
i = rank of species (highest abundance = 1; lowest abundance= S)
n_i = number of individuals in species i
$N = \sum n_i$ = total number of individuals in sample
p_i = proportion of species i in sample*
s_n = number of species with n individuals

(*- p_i is estimated using n_i/N, but because abundance counts are discrete values (i.e., integers), p_i can never be smaller than $1/N$; this causes many metrics described below to be biased, particularly at small sample sizes)

Graphical depiction

1) Rank Abundance Distributions (RAD)
   Directly depict species abundances in rank order
   x-axis – taxa in rank order from most abundant to least abundant
   y-axis – logarithm of abundance or proportion for each taxon

2) Abundance Frequency Distrbutions (AFD)

Depict number of species in abundance bins

x-axis – abundance bins, typically in $\log_2$ (i.e., Preston's "octaves")

y-axis – number of species in abundance bin (note – following May (1975), species on a bin boundary are handled by counting ½ to each bounding bin)

AFDs are easy to generate from RADs, but the opposite is not true because binning obscures the actual abundances of species; in addition, when fitting curves to data, AFDs typically have fewer degrees of freedom because the curve is based on the number of bins rather than the total number of species.

Theoretical Abundance Distributions

Implicit Assumption – A species' abundance reflects the amount of resource it controls or its "importance" to the community in some way

1) Uniform

Function

RAD generated by:

$$n_i = N/S \qquad (1)$$

Interpretation

individuals of each species are sampled from an equiprobable, underlying distribution; species use resources independently of one another and to equal degrees

2) Geometric Series

Function

AFD generated by:

$$n_i = NC_k k(1-k)^{i-1} \qquad (2)$$

k = constant; proportion of remaining niche space occupied by each successively colonizing species

$C_k = [1-(1-k)^S]^{-1}$ = scaling constant that ensures $\Sigma n_i = N$

Interpretation

each species arrives at regular time intervals and preempts a constant fraction of remaining resources (compare with log series interpretation – log series is effectively a geometric series with stochastic noise incorporated into the resources allocated to each species)

3) Broken Stick

Function

RAD generated by:

$$n_i = \frac{N}{S}\sum_{r=i}^{S}\frac{1}{r} \qquad (3)$$

Interpretation

A one dimensional resource axis is simultaneously and randomly partitioned by S species

4) Log Series

Function

AFD generated by:

$$s_n = \alpha x^n/n \qquad (4)$$

x is estimated iteratively using: $\frac{S}{N} = -\ln(1-x)\left(\frac{1-x}{x}\right)$ (4a)

and $\alpha = N\left(\frac{1-x}{x}\right)$ (4b)

Interpretation

each species arrives at random time intervals and preempts a constant fraction of remaining resources; expected result of an infinitely large neutral community (see zero-sum multinomial distribution); α has been widely used as a diversity metric, whether the AFD of the sample conforms to a log series or not

5) Log Normal

Function

AFD generated by:

$$S_R = S_o e^{-R^2/2\sigma^2} \qquad (5)$$

S_R = number of species in octave R

So = number of species in modal octave

σ = standard deviation of AFD

Interpretation

Species populations grow exponentially and respond independently to different factors; see Bulmer (1974) for the poisson log normal function, which is a discrete function equivalent to the continuous log normal

6) Zipf-Mandelbrot

Function

AFD generated by:

$$s_n = Sn^{-k} \qquad (6)$$

k = a scaling constant

Interpretation

Later colonists have more specific requirements and are therefore rarer than initial colonists

7) Zero-Sum Multinomial

Function

AFD generated by:

$$s_n = \theta \frac{J!}{n!(J-n)!} \frac{(\gamma-1)!}{(J+\gamma-1)!} \int_0^{\gamma} \frac{(n+y-1)!}{y!} \frac{(J-n+\gamma-y-1)}{(\gamma-y-1)} e^{(-y\theta/\gamma)} dy \qquad (7)$$

$$\gamma = \frac{m(J-1)}{1-m} \qquad (7a)$$

J = size of local community (a fixed value)

m = immigration rate (chance of replacement from outside the community)

θ = log series α = number of speciation events per individual per unit time

See papers by Alonso, D. & McKane, A. J. Ecol. Lett. 7, 901–910 (2004); Etienne, R. S. Ecol. Lett. 8, 253–260 (2005); and Etienne, R. S. & Alonso, D. Ecol. Lett. 8, 1147–1156 (2005) for a solution to the RAD of the zero-sum multinomial.

Interpretation

Expected species abundance distribution in a local community of size J in which all individuals, regardless of species identity, are equivalent, so their abundances reflect demographic random walks (i.e., demographic stochasticity); equivalent to log series as $J \rightarrow \infty$

8) Dynamical/Niche-based

Models governed by rules of niche division/pre-emption; see Tokeshi (1990, 1993) and He (2005) for examples of species abundance distributions based on specific models of species interaction, migration, and speciation/extinction

Veil "line" – Samples are never complete pictures of the real world – the rarest species may not be sampled and moderately common species may be represented by few individuals and therefore appear rare. Preston (1948) depicted this by simply truncating a log normal distribution. However, Dewdney (1998) showed that sampling theory does not support the idea that distributions are simply truncated; rather, their shape depends on the intensity of sampling. In addition, at small sampling sizes, it can be very difficult to distinguish different species abundance distributions, so extrapolation based on any particular model can be dangerously unsupported.

Measures of Diversity

Basic Concepts

Summarizing abundance distributions into one or a few fundamental parameters makes comparison of communities explicit and straightforward. However, expressing a distribution with a single number also reduces the total amount of information about the community.

Diversity Index = a single number summarizing the number of species and their relative abundances

Richness = number of species

Evenness = uniformity of species abundances (complement of "dominance" – i.e., dominance and evenness reflect the same information)

Berger-Parker Index

$$d = \max(p_i) \qquad (8)$$

Interpretation. If the Berger-Parker index is high, this means that the community is dominated by the most common species – i.e., it is not even.

Comments. This metric is a useful back-of-the-envelope estimator, but it is highly biased by sample size (N) and richness (S). In addition, it does not make use of all the information available from the sample. Rarely used in paleontology; increasingly avoided by ecologists.

Shannon Information Index or "Entropy"

$$H = -\sum_{i=1}^{S} p_i \ln(p_i) \qquad (9)$$

Interpretation. This measure, which is based on information theory, summarizes the "entropy" of a community. It addresses the following question: if rare species carry more "information" than common species and their information value is proportional to the logarithm of their p_i, what is the average amount of "information" in the community? That average amount of "information" is the entropy.

H incorporates both richness and evenness – i.e., it increases as both a function of the number of species and the uniformity of their abundance values. Note that *H* is the absolute value of the mean of $\ln(p_i)$.

Comments. This metric is the single most common measure of diversity. It is biased by sample size (*N*), but the error is on the order of *(S-1)/2N* (i.e., if this value is much smaller than *H* calculated using equation (2), then the calculated value is an acceptable estimate of the true value; note that this error estimate is only correct if *H* is calculated using *natural logarithms*).

This metric is the basis for SHE analysis, a means of characterizing a community's diversity using richness (*S*), entropy (*H*), and a measure of evenness (*E*):

$$E = e^{H/\ln(S)} \qquad (10)$$

(e is the base of the natural logarithm).

$\ln(S)$ is the maximum value *H* can have for a sample of given richness (*S*), so $H/\ln(S)$ is a means of removing the influence of richness on the entropy. See Hayek and Buzas (1997) for a complete description of this approach.

Simpson's Evenness

$$\text{For continuous data: } \Delta = 1 - \sum_{i=1}^{S}\left(\frac{n_i}{N}\right)^2 = 1 - \sum_{i=1}^{S} p_i^2 \qquad (11)$$

$$\text{For discrete data: } PIE = 1 - \sum_{i=1}^{S}\left(\frac{n_i}{N}\right)\left(\frac{n_i - 1}{N - 1}\right) = \frac{N}{N-1}\left(1 - \sum_{i=1}^{S} p_i^2\right) \qquad (12)$$

Interpretation. If you pick two specimens from a sample randomly, the probability that they will be two different species is given by *PIE*, which stands for the *Probability of Interspecific Encounter* (Hurlbert, 1971). In theory, this metric ranges from 0 (perfectly uneven) to 1 (perfectly even). It is a true probability value.

Comments. PIE is independent of sample size and is closely mathematically related to rarefaction estimates of richness. Although *PIE* can range from 0 to 1 in theory, in reality, its range is clipped because p_i is estimated using discrete values (see earlier note under definition of terms).

PIE is rather insensitive to significant differences among even communities (*PIE* > 0.9), but this issue can be easily addressed by using 1/(1-*PIE*) as the basis of comparison. This transformation gives the richness of a community (i.e., the number of species) that would have the same value of *PIE*, but in which all species had equal abundance (i.e., a perfectly even sample).

Estimating Richness Using Rarefaction

Richness is highly dependent on the size of a sample: *(Big samples have more species!)*

Rarefaction = a means of estimating the expected richness if a sample were smaller than it really is; provides a means of comparing richness among samples of different size.

Rarefaction curves show sample size on the x-axis and sample richness on the y-axis

Counting Rules

1) Multiplicative Rule. If you are drawing one specimen from each of *S* sets of specimens, with sizes $n_1, n_2, n_3,\ldots, n_S$, the number of different possible samples is:

$$n_1 n_2 n_3 \bullet \bullet \bullet n_S \qquad (13)$$

2) Permutations Rule. If you are drawing *m* specimens from a set of *N* and arranging the *m* specimens in a distinct order, the number of different possible samples is:

$$P_m^N = \frac{N!}{(N-n)!} \qquad (14)$$

$x! = x(x-1)(x-2)(x-3)\ldots 1;\ 1! = 1;\ 0! = 1$

3) Partitions Rule. If you are partitioning the specimens in a set of *N* into *S* groups, each consisting of $n_1, n_2, n_3,\ldots, n_S$ specimens ($\sum n_i = N$), then the number of different possible samples is:

$$\frac{N!}{\prod_{i=1}^{S} n_i!} \qquad (15)$$

$(\prod n_i! = n_1! n_2! n_3! \bullet \bullet \bullet n_S!)$

4) Combinations Rule. If you are drawing m specimens from a set of *N* specimens *without regard to the order* of *m* elements, the number of different possible samples is:

$$\binom{N}{m} = \frac{N!}{m!(N-m)!} \qquad (16)$$

Rarefaction Equation

n_i = number of individuals in species *i*
$N = \sum n_i$ = total number of individuals in sample
S = number of species in sample – i.e., richness
m = size of rarefied sample
s_m = estimated number of species in sample of size *m*

$$s_m = S - \frac{\sum_{i=1}^{S}\binom{N-n_i}{m}}{\binom{N}{m}} = \sum_{i=1}^{S}\left[1 - \frac{\binom{N-n_i}{m}}{\binom{N}{m}}\right] = S - \sum_{i=1}^{S}\prod_{j=0}^{n_i-1}\left(1 - \frac{m}{N-j}\right) = \sum_{i=1}^{S}\left[1 - \prod_{j=0}^{n_i-1}1 - \left(\frac{m}{N-j}\right)\right] \quad (17)$$

How to interpret this equation

$\binom{N}{m}$ = # of ways of choosing *m* specimens from *N*

$\binom{N-n_i}{m}$ = # of ways of choosing *m* specimens that do NOT include species *i*

$1 - \frac{\binom{N-n_i}{m}}{\binom{N}{m}}$ = proportion of subsamples of size *m* that DOES include species *i*

(i.e., the expected probability of species *i* being present in the subsample)

In the master sample, the probability of seeing at least one specimen of every species is 1.0 – the sum of these probabilities equals the richness (*S*). In a subsample, the probability of seeing any given species may be <1.0; taking the sum of the probability of each species being present in the subsample gives an estimate of the total number of species expected to be seen in the subsample.

Comments on Rarefaction

1) Rarefaction curves CANNOT be extrapolated to larger samples
2) Rarefaction curves use the abundances of individual species in a sample, so they represent an alternative means of showing the same information as a species abundance distribution (RAD or AFD)
3) Rarefaction curves can be effectively approximated with logarithmic or power functions, which in turn can be used for extrapolation (i.e., see McGuiness, 1984 on the functions used to fit species-area curves)
4) Just as rarefaction curves depict how richness increases with number of specimens, *collector's curves* depict how diversity increases as collections are successively added:

$$s_a = \sum_{i=1}^{S_{total}}\left[1 - \frac{\binom{A-o_i}{a}}{\binom{A}{a}}\right] = S_{total} - \frac{\sum_{i=1}^{S_{total}}\binom{A-o_i}{a}}{\binom{A}{a}} \quad (18)$$

o_i = number of samples in which species *i* occurs

A = total number of samples; *a* = number of samples of interest (<*A*)

S_{total} = total number of species in all samples; s_a = estimated number of species in *a* samples

Hierarchical Theory of Diversity

Diversity is dependent on the scale of sampling – this is implicit in MacArthur & Wilson's theory of island biogeography (see Rosenzweig 1994) and was also formally acknowledged by Whittaker (1972) when he distinguished alpha, beta, and gamma diversity

Alpha diversity – local diversity, within-habitat diversity, within-patch diversity

Beta diversity – between-sampling unit diversity (NOT LIMITED to change along gradients)

Gamma diversity – total diversity of a region or landscape

Patch – an area over which there is no significant change in the taxonomic composition of a community

Habitat – a distinct set of environmental conditions; can include multiple community states

Landscape – an area including one or more patches and one or more habitats

Multiplicative Relationship Between Alpha, Beta, and Gamma:

$$\beta = \gamma/\alpha - 1 \qquad (19) \quad \text{(Whittaker, 1972)}$$

Additive Diversity Partitioning (ADP):

$$\gamma = \alpha + \beta_1 + \beta_2 + \beta_3 + \ldots \qquad (20) \quad \text{(Lande, 1996)}$$

ADP provides a means of assessing the contributions of various components of diversity in a complex, hierarchical system

Concavity – Diversity at any level in the hierarchy must be equal to or greater than the average diversity of the contributing diversity components

---

References for Measurement of Diversity

Bunge, J. and Fitzpatrick, M., 1993, Estimating the number of species: a review. Journal of the American Statistical Association, v. 88, p. 364-373. (Dealing with richness and its measurement from the formal view of a statistician.)

Colwell, R.K. and Coddington, J.A., 1994, Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society of London B, v. 345, p. 101-118. (The title says it all.)

Colwell, R.K., Mao, C.X., and Chang, J., 2004, Interpolating, extrapolating, and comparing incidence-based species accumulation curves. Ecology, v. 85, p. 2717-2727.

Dewdney, A.K., 1998, A general theory of the sampling process with applications to the "veil line". Theoretical Population Biology, v. 54, p. 294-302.

Gotelli, N.J. and Colwell, R.K., 2001, Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters, v. 4, p. 379-391. (A very up-to-date review of rarefaction in its many forms.)

Gotelli, N.J. and Graves, G.R., 1996, Null Models in Ecology. Smithsonian Institution Press, Washington, 368 p.

He, F., 2005, Deriving a neutral model of species abundances from fundamental mechanisms of population dynamics. Functional Ecology, v. 19, p. 187-193.

Heck, K.L., Jr., van Belle, G., and Simberloff, D., 1975, Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. Ecology, v. 56, p. 1459-1461. (Calculating rarefaction curves and their error bars.)

Hurlbert, S.H., 1971, The nonconcept of species diversity: a critique and alternative parameters. Ecology, v. 52, p. 577-586. (On explicit calculation of rarefaction richness and derivation of Hurlbert's PIE measure of evenness.)

Lande, R., 1996, Statistics and partitioning of species diversity and similarity among multiple communities. Oikos, v. 76, p. 5-13.

Magurran, A.E., 2004, Measuring Biological Diversity. Blackwell Publishing.

May, R.M., 1975, Chapter 4. Patterns of species abundance and diversity. *In* (Cody, M.L. and Diamond, J.M., eds.) Ecology and Evolution of Communities, Belknap Press, Cambridge, Massachusetts, p. 81-120. (An excellent review of species abundance distributions up to that time including the original references and full derivations.)

McGuiness, K.A., 1984, Equations and explanations in the study of species-area curves. Biological Reviews, v. 59, p. 423-440. (A very useful review of species-area functions and a list of primary references. There have been some new developments, but this is a useful place to start.)

Olszewski, T.D., 2004, A unified mathematical framework for the measurement of richness and evenness within and among multiple communities. Oikos, v. 104, p. 377-387.

Sanders, H.L., 1968, Marine benthic diversity: a comparative study. American Naturalist, v. 102, p. 243-282. (A classic paper in ecology. Not only did Sanders originate an early form of rarefaction, he used it to pose the time-stability hypothesis.)

Shen, T.-J., Chao, A., and Lin, C.-F., 2003, Predicting the number of new species in further taxonomic sampling. Ecology, v. 84, p. 798-804.

Simpson, E.H., 1949, Measurement of diversity. Nature, v. 163, p. 688. (Measurement and error bars for dominance, which is the complement of evenness.)

Smith, B. and Wilson, J.B., 1996, A consumer's guide to evenness indices. Oikos, v. 76, p. 70-82.

Tipper, J.C., 1979, Rarefaction and rarefiction – the use and abuse of a method in paleoecology. Paleobiology, v. 5, p. 423-434. (A review of rarefaction in paleontology.)

Tokeshi, M., 1993, Species abundance patterns and community structure. Advances in Ecological Research, v. 24, p. 111-186.

Ugland, K.I., Gray, J.S., and Ellingsen, K.E., 2003, The species-accumulation curve and estimation of species richness. Journal of Animal Ecology, v. 72, p. 888-897.

Whittaker, R.H., 1972, Evolution and measurement of species diversity. Taxon, v. 21, p. 213-251.

Whittaker, R.J., Willis, K.J., and Field, R., 2001, Scale and species richness: towards a general, hierarchical theory of species diversity. Journal of Biogeography, v. 28, p. 453-470.

References for Confidence and Sampling

Austin, H.W., 1983, Sample size: How much is enough? Quality and Quantity, v. 17, p. 239-245.

Bennington, J.B. and Rutherford, S.D., 1999, Precision and reliability in paleocommunity comparisons based on cluster-confidence intervals: How to get more statistical bang for your sampling buck. Palaios, v. 14, p. 506-515.

Buzas, M.A., 1990, Another look at confidence limits for species proportions. Journal of Paleontology, p. 842-843.

Dennison, J.M. and Hay, W.W., 1967, Estimating the needed sampling area for subaquatic ecologic studies. Journal of Paleontology, p. 706-708.

Fatela, F. and Taborda, R., 2002, Confidence limits of species proportions in microfossil assemblages. Marine Micropaleontology, v. 45, p. 245-248.

Hayek, L.C. and Buzas, M.A., 1997, Surveying Natural Populations. Columbia University Press, New York, 563 p.

Patterson, R.T. and Fishbein, E., 1984, Re-examination of the statistical methods used to determine the number of point counts needed for micropaleontological quantitative research. Journal of Paleontology, v. 63, p. 245-248.

Raup, D.M., 1991, The future of analytical paleontology. *In* Gilinsky, N.L. and Signor, P.W. (eds.), Analytical Paleobiology, Short Courses in Paleontology Number 4, The Paleontological Society, p. 207-216.

Appendix: Websites for Diversity Analysis Software

http://viceroy.eeb.uconn.edu/estimates – EstimateS 6 computes randomized species accumulation curves, statistical estimators of true species richness (S), and a statistical estimator of the true number of species shared between pairs of samples, based on species-by-sample (or sample-by-species) incidence or abundance matrices. For comparative purposes, EstimateS also computes Fisher's alpha and the Shannon and Simpson diversity indexes for each sample, as well as the Jaccard, Morisita-Horn, and Sorensen (both incidence-based and abundance-based) indexes of biotic similarity between samples.

http://www.uga.edu/~strata/software/index.html – This site has two programs for doing rarefaction calculations. The first is based on the analytic solution and runs very quickly, even on large data sets. The second is based on a resampling routine and can be used to cross-check the analytic calculations.