

A nonparametric method for the measurement of size diversity with emphasis on data standardization

Xavier D. Quintana^{1*}, Sandra Bruce^{1,2}, Dani Boix¹, Rocío López-Flores¹, Stéphanie Gascón¹, Anna Badosa¹, Jordi Sala¹, Ramon Moreno-Amich¹, and Juan J. Egozcue³

¹Institute of Aquatic Ecology and Department of Environmental Sciences, Universitat de Girona, Girona, Spain

²Department of Freshwater Ecology, National Environmental Research Institute, Vejlsvøvej 25, DK-8600 Silkeborg, Denmark

³Department of Applied Mathematics III, Universitat Politècnica de Catalunya, Barcelona, Spain

Web Appendix B

Diversity of transformed variables—Standardization of samples for diversity index estimation are based on transformation of random variables. Relevant transformations and the corresponding diversity indexes are summarized. The original random variable is denoted X and the transformed one Y . In some of these results the support of the original variable can be the whole real line; then the integral defining the diversity index is extended from $-\infty$ to $+\infty$ accordingly.

Shift—The shift by a real constant a is defined as $Y = X - a$. The corresponding probability density is $p_Y(y) = p_X(y + a)$. The diversity index for the new variable is then

$$\mu(Y) = -\int_{-\infty}^{+\infty} p_X(y + a) \ln p_X(y + a) dy = \mu(X) \quad (13)$$

i.e., $\mu(X)$ is invariant under shifting of the random variable.

Scaling—Scaling by a positive constant c is defined as $Y = cX$. The pdf of Y is $p_Y(y) = c^{-1} p_X(c^{-1}y)$ and the diversity index for Y is

$$\begin{aligned} \mu(Y) &= -\int_0^{+\infty} c^{-1} p_X(c^{-1}y) (\ln p_X(c^{-1}y) - \ln c) dy \\ &= \mu(X) + \ln c \end{aligned} \quad (14)$$

Therefore, scaling by a constant is equivalent to add the logarithm of the absolute value of the constant to the diversity index.

Truncation of density—Sometimes a given density with support in $[a, +\infty)$ is conditioned to be less than a given value

$s, s > a$. This is normally called truncation of the variable to the support $[a, s]$. The density of X conditional to $X < s$ is

$$p_X(x|X < s) = \frac{1}{1 - F_X(s)} p_X(x) \mathbb{I}\{a < x < s\}$$

where F_X is the cumulative probability distribution of X . The size diversity of $X|X < s$ is then

$$\mu(X|X < s) = -\frac{1}{1 - F_X(s)} \int_a^s p_X(x) \ln p_X(x) dx + \frac{F_X(s)}{1 - F_X(s)} \ln(1 - F_X(s))$$

The integral $-\int_a^s p_X(x) \ln p_X(x) dx$ itself does not correspond to any size diversity but just a lower bound of the size diversity of X . The suppression of the normalizing term $1 - F_X(s)$ implies that $p_X(x|X < s)$ is not a probability density.

Logarithmic transformation—Define $Y = a \ln X$, being a , a positive real number. The pdf of Y is $p_Y(y) = a^{-1} p_X(\exp(y/a)) \exp(y/a)$. The diversity index of Y is

$$\mu(Y) = -\int_{-\infty}^{+\infty} a^{-1} p_X(\exp(y/a)) \exp(y/a) \ln(a^{-1} p_X(\exp(y/a)) \exp(y/a)) dy$$

and therefore,

$$\mu(Y) = \mu(X) - E[Y] + \ln a \quad (15)$$

Power transformation—A power transformation of X is $Y = aX^k$. Y for positive constants a and k . Consider $Z = \ln Y = \ln a + k \ln X$. Using Eqs. 13 and 15, $\mu(Z) = \mu(X) + \ln k - E[\ln X]$. Using again Eq. 15 to relate $\mu(Z)$ with $\mu(Y)$,

$$\mu(Y) = \mu(X) + \ln(ak) + (k - 1)E[\ln X]. \quad (16)$$