THE MEASUREMENT OF SIZE DIVERSITY

This webpage provides an R script for measuring the size diversity and the size evenness of a size distribution. The R script "SizeDiversity_2018" was programmed by J.J. Egozcue and O. Martínez-Abella following the method presented in Quintana et al. (2008) *Limnology and Oceanography: Methods*, 6:75-86 and updated by Quintana et al. (2016) *Limnology and Oceanography: Methods*, 14:408-413. It is based on a non-parametric kernel estimation, after data standardization by division of sample geometric mean. This method is applicable to any type of size distribution, even if it doesn't fit a parametric function, and allows its immediate comparison with the size diversity of other distributions independently of the method used for size estimation. See the above references for more details on the method.

1. THEORETICAL FRAMEWORK

The Shannon-Wiener index is used to measure size diversity. Since size is not a discrete, but a continuous variable, the goal is to estimate a Shannon size diversity index $\mu_2(X)$ corresponding to the probability density function $p_X(x)$ of the size of the individuals, which takes the following integral form:

$$\mu_2(X) = -\int_0^{+\infty} p_X(x) \log_2 p_X(x)\, dx$$

Size diversity is computed by means of a non-parametric kernel estimation. The expression used for the computation of the size diversity is:

$$\hat{\mu}_{\text{kerMC}}(X) = \bar{y} - \frac{1}{n}\sum_{k=1}^{n} \log_2 \left[\frac{1}{n\sqrt{2\pi}\,\sigma} \sum_{j=1}^{n} \exp\left(-\frac{1}{2} \frac{(y_k - y_j)^2}{\sigma^2} \right) \right]$$

This equation coincides with equation (9) in Quintana *et al.* (2008), except that the base of the logarithm is binary instead of a natural logarithm (see reference above for further details on notation).

Data are standardized by means of the division by the geometric mean. According to Quintana *et al.* (2008), this standardization has the advantages that 1) the same size diversity value is obtained when using original size or log-transformed data and 2) size measurements with different dimensionality (longitudes, areas, volumes or biomasses) may be immediately compared with the simple addition of ln k, where k is the dimensionality (1, 2, or 3, respectively). See below for more details (section 4 "Dimensionality").

Size diversity values may be negative (as in sample 3 of example 2, see below). The model is based on a continuous probability density function, which may take local probability values > 1. Thus, 0 is not the minimum value of size diversity. Negative values may be found when there is a high accumulation of data in a determinate size, meaning extremely low size diversity values.

Size e-evenness is computed by means of the division of the size diversity by the maximum diversity possible for a size distribution with the same standard deviation, according to Quintana *et al*. (2016):

$$J_e(X) = \frac{\exp(\mu(X))}{\exp(\mu(\mathrm{LN}))} = \frac{2^{\mu_2(X)}}{2^{\mu_2(\mathrm{LN})}}, \quad 0 \leq J_e(X) \leq 1.$$

where the denominator is the size diversity of a log-normal distribution with the same standard deviation. However, J_e values > 1 may eventually appear, as is justified in the mentioned reference (see pg 410):

When the comparison is done using a kernel estimate of the pdf, … parameters should be those of the kernel estimated pdf. When using a standardization of data dividing by the geometric mean of the data $m_{ln} \approx 0$, but not exactly equal to 0 due to kernel estimation. Similarly, σ^2_{ln} must be that corresponding to the kernel estimation of the pdf, which is not equal to the logarithmic variance estimated from the data but only an approximation. Then, e-evenness function using the log-variance of the standardized data can be larger than that of the corresponding Log-Normal distribution. However, this seldom occurs.

2. PROGRAM PROCEDURE

1) Open the R script SizeDiversity_2018 and run it from line 70 until the end

2) Prepare an excel file (or directly an R data frame) with three columns:

the sample identifier, de body size and the abundance.

Variable names must be, respectively: "SampleID", "x" and "abund".

Eventually you may have no abundances, that is, every size measure belongs to a single individual; in this case, prepare a file with only two columns: "SampleID" and "x". Sort your data file by "SampleID".

3) Enter these data in an R data frame called "div.data" (see details in the script)

4) Introduce in line 56 the dimensionality of your sample. DO NOT FORGET THIS STEP!

"measuredim=1" when your body size data are lengths (L),

"measuredim=2 when your body size data are recovering (L^2) or

"measuredim=3" if your body size data are biomass units (L^3) (e.g. dry weight, carbon, biovolume...).

5) Run lines 51 to 61 and check the results.

IMPORTANT: If you find two (or more) individuals with the same size, DO NOT AGGREGATE them to a single size measurement with double (or more) abundance. The sample size is the real number of individuals measured, not the total abundance of individuals which may vary a lot depending on the units used.

3. OUTPUTS

The script generates an R matrix, with the name "res1", including the following variables for each SampleID:

"**diversity2**": The size diversity using base2 logarithms in the Shannon expression

"**eevenness**": The size e-evenness. In practice it is calculated by dividing $2^{diversity2}/2^{\text{logNdiversity2}}$

"**logNdiversity2**": The size diversity, using base2 logarithms, of a log normal distribution with the same standard deviation of the sample. It represents the maximum size diversity for this given standard deviation.

"**bandker**": the kernel bandwidth

"**devxlog**": the standard deviation of the body size data (log transformed)

"**gmeanx**": the geometric mean of the body size data (not log transformed)

"**meanx**": the arithmetic mean of the body size data (not log transformed)

Eventually you can convert the matrix to an R data frame:

res2<-data.frame(res1)

4. DIMENSIONALITY

One of the main advantages of the standardization by the geometric mean is that size distribution data measured with different dimensionalities are comparable, when an allometric relationship between them is assumed. For instance, let X be a random length, e.g., a radius of a sphere (or the equivalent spherical diameter), and let V be a scaled power of X, defined by $V = aX^k$, e.g., the volume of a sphere $a = 4\pi/3$, $k = 3$. Assume that both variables, V and X, are divided by the respective geometric means, the relationship of the corresponding diversities is:

$$\mu(V) = \log_2 k + \mu(X)$$

Where k is the dimensionality ($k = 3$ in this example). In practice, it means that data sets which differ in dimensionality may be easily compared by the simple addition (or subtraction) of $\log_2 k$ (see Quintana *et al*. (2008) for details).

By default, results of size diversity which appear in the R output file are normalised to dimensionality 3, that is, when dimensionality chosen in the input file is 1, the program adds $\log_2 3$ to the final result of size diversity; if dimensionality chosen is 3, it doesn't add anything. See in example 1 how dimensionality works.

4. EXAMPLES

EXAMPLE 1: Phytoplankton samples, counted by flow cytometry, where the size of each individual is measured. The size of a cell is measured using its equivalent spherical diameter (ESD, μm). The input file has this form:

SampleID	x
Phyto01	3
Phyto01	3
Phyto01	3
Phyto01	3
Phyto01	3
Phyto01	4
Phyto01	4
Phyto01	4
Phyto01	5
Phyto01	14
Phyto01	16
Phyto01	6
Phyto01	8
Phyto05	6
Phyto05	6
Phyto05	6
Phyto05	6
Phyto05	6
Phyto05	9
Phyto05	11
Phyto27	4
Phyto27	3
Phyto27	3
Phyto27	3
Phyto27	4
Phyto27	3
Phyto27	7
Phyto27	15

The first column of the data list is the sample identifier (sampleID). The second column is the size of a single individual. In this example, there are three samples (Phyto01, Phyto05 and Phyto27); in Phyto01 there are 13 individuals, 5 of them with an equivalent spherical diameter of 3 μm, etc.

No abundance values are included because each measurement belongs to one single individual. The variable "abund" is not necessary.

Size measurements are lengths (L), not surfaces or volumes. Thus "measuredim=1".

The resulting output file is:

```
> res1
        diversity2 eevenness logNdiversity2   bandker   devxlog   gmeanx    meanx
Phyto01   2.603439 0.8717503       2.801452 0.3568453 0.5622931 4.856536 5.846154
Phyto05   1.278161 0.8334514       1.540991 0.1685824 0.2347070 6.932870 7.142857
Phyto27   2.502942 0.8481647       2.740525 0.3769726 0.5390411 4.382524 5.250000
>
```

Using this example 1, we can now examine how dimensionality works. For each ESD value, the biovolume may be obtained ($V=4\pi r^3/3$):

SampleID	x(ESD)	x(biovolume)
Phyto01	3	14.13717
Phyto01	3	14.13717
Phyto01	3	14.13717
Phyto01	3	14.13717
Phyto01	3	14.13717
Phyto01	4	33.51032
Phyto01	4	33.51032
Phyto01	4	33.51032
Phyto01	5	65.44985
Phyto01	14	1436.75504
Phyto01	16	2144.66058
Phyto01	6	113.09734
Phyto01	8	268.08257
Phyto05	6	113.09734
Phyto05	6	113.09734
Phyto05	6	113.09734
Phyto05	6	113.09734
Phyto05	6	113.09734
Phyto05	9	381.70351
Phyto05	11	696.90997
Phyto27	4	33.51032
Phyto27	3	14.13717
Phyto27	3	14.13717
Phyto27	3	14.13717
Phyto27	4	33.51032
Phyto27	3	14.13717
Phyto27	7	179.59438
Phyto27	15	1767.14587

Because data are volumes, now we have to change the dimensionality: "measuredim=3". The resulting output file is:

```
> res1
        diversity2 eevenness logNdiversity2   bandker  devxlog    gmeanx    meanx
Phyto01   2.603439 0.8717503       2.801452 1.0705358 1.686879  59.97612 323.0202
Phyto05   1.278161 0.8334514       1.540991 0.5057473 0.704121 174.47685 234.8715
Phyto27   2.502942 0.8481647       2.740525 1.1309177 1.617123  44.07289 258.7887
>
```

Note that results differ in mean and standard deviation values, but not in diversity and e-evenness.

EXAMPLE 2: A sample of benthic invertebrates where body size is measured in dry weight (units in µgC), by measuring the length of the individuals and converting length in dry weight using allometric relationships found in the bibliography. Only a selected number of individuals are measured and the abundance of each size is estimated. The input file has this form:

SampleID	x	abund
1	1.12E+01	2
1	1.09E+01	8
1	3.00E+00	12
1	1.12E+01	3
1	8.86E+00	30
1	5.72E+00	17
1	9.28E+00	74
2	1.85E+01	7
2	4.35E+01	3
2	5.85E+00	25
2	6.27E+00	32
3	1.12E+01	28
3	3.00E+00	1

Now there is a new column in the data list indicating the abundance of each size. In sample 1, there are 2 individuals with an individual biomass of 1.11610E+01 µgC, 8 individuals with an individual biomass of 1.09070E+01 µgC, etc.

In this case data are biomass units (L^3), thus dimensionality must be: "measuredim=3".

The resulting output file is:

```
> res1
  diversity2 eevenness logNdiversity2   bandker   devxlog   gmeanx     meanx
1  0.1230053 0.7784707    0.484290694 0.2431281 0.3384923  8.04011  8.417534
2  0.7367664 0.7879828    1.080530463 0.4110823 0.5117230  7.46093  9.058060
3 -0.6305087 0.6502713   -0.009622377 0.2218023 0.2403622 10.70263 10.917241
>
```