1 Compositional data

Compositional or closed data are multivariate data with positive values that sum up to a constant, in our case 100 being the data expressed in percentages. In mathematical notation a composition from a laboratory is given by $x = (x_1, \ldots, x_D)'$ where D is the number of the components. So we assume that each component is a positive number, $x_i > 0$, and $\sum_{i=1}^{D} x_i = 100$.

Due to this last constraint, standard statistical methods can lead to questionable results if they are directly applied to the original, closed data.

Instead it is well established practice to transform the data using transformations that preserve the specific geometry of compositional data on the simplex (also called Aitchison geometry). In particular we will use the centered logratio (clr) transformation for the composition x:

$$y = (y_1, \ldots, y_D)' = \left(\log \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \ldots, \log \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right).$$

Note that this transformation results in collinear data because $\sum_{i=1}^{D} y_i = 0$.

2 Dealing with zeros

It is important to note that the statistical analysis of compositional data is based on logratios of parts, then if is not suitable when zeros are present in a data set.

We can distinguish two kinds of zeros: *essential zeros*-or absolute absence of the part in the observation and *rounded zeros*- or presence of a component, but below detection limit. In our case we have faced to this second kind of zeros and it seems reasonable to replace them by a suitable value.

We use a multiplicative replacement strategy (mrs) where the original composition x is replaced by $r = (r_1, \ldots, r_D)$. Each replaced part r_i is calculated accorging to the following formula

$$r_i = \begin{cases} \Delta_i, & \text{if } x_i = 0 \\ \left(1 - \frac{\sum_{k|x_k=0} \Delta_k}{100}\right) x_i, & \text{if } x_i > 0. \end{cases}$$

where Δ_i is the imputed value on the part x_i. In our case we have set the imputed value Δ_i equal to the detection limit 0.005.

3 Statistical estimate of the reference value and z-score

In intercalibration study a widespread used estimate of the unknown reference value is given by the mean of the observed values for all laboratories. In presence of outlier values, instead of discharging the values, we use a robust estimate given by the median of the data,

$$\text{reference value} = \text{median}_i \, x_i \, .$$

For an assessment of laboratory performance, a z-score is usually calculated according to the formula:

$$z_k = \frac{x_k - \text{reference value}}{s} \, ,$$

where k denotes the laboratory and x_k is the value of the laboratory k and s the standard deviation of the data.

The z-score expresses the difference between the value of the laboratory and the reference value estimated by the median of the values observed in all laboratories, normalized by the statistical variability. Performance of the laboratory is considered to be acceptable when the absolute value of this difference is less than or equal to two. The measurement is regarded to be questionable when the absolute value is greater than two and less or equal to three, unsatisfactory when this is greater than 3.

These limits works well if the data show a symmetric distribution as the normal distribution without outlier values. Compositional data are intrinsically skewed because bounded between 0 and 100, so we have adopted this procedure for evaluating the performance of the laboratories.

First of all we checked if rounded zeros are present. If any, the corresponding compositions were modified using msr. Then we transformed the data x using the clr transformation and getting the part of the composition y. Finally, the z-score has been evaluated by

$$z_k = \frac{y_k - \text{median}_i \, y_i}{1.486 \times \text{MAD}}$$

where y_k is the transformed compositional value of the laboratory k. As a measure of variability we use the robust measure given by the Mean Absolute Deviation (MAD)

$$\text{MAD} = \times \, \text{median}_i \, (\, |y_i - \text{median}_j(y_j)| \,) \, ,$$

corrected by the scale factor 1.486 to be consistent with the normal distribution.