Chapter 3

Data pre-processing: irregular data

Contents

Objectives

✓ To deal with the most common irregular data in CoDa: missing data, values below detection limit and zeros.

✓ To distinguish the type of zeros and accordingly decide the procedure for dealing with them.

✓ To know how to detect potential outliers in CoDa.

In this chapter we present the techniques to deal with no-common data, that is, data that are unusual and have special features. We call these data *irregular data*. We consider several types of irregular data: non-available data (*missing data*), values below detection limit (*bdl*), zeros and outliers. The analysis and treatment of irregular data must be done before the statistical analysis (cluster, regression, MANOVA, etc) of the data. This step is known as *data pre-processing* or simply *pre-processing*. Because each type of irregular data requires its own specific pre-processing we introduce the particular techniques for doing so in following subsections. In any case, for this section and for the remainder of the chapter, we assume that groups in the data set do not exist. In other words, if there are groups in the data set then the techniques presented in this chapter should be applied separately to each group. In the next chapter we will introduce some basic multivariate techniques to deal with the groups of a data set.

3.1. Missing data

The first type of irregular data are the non-available data, usually labelled as "NA" as an entry in the incomplete data matrix. It is important to observe that there are different types of missing data because each type will need its own particular treatment. Let $\mathbf{x}$ be an incomplete multivariate sample which can be split into two parts, *observed* and *missing*, that is, $\mathbf{x}$=(*observed* part,*missing* part)=($\mathbf{x}_{obs}, \mathbf{x}_{mis}$).

According to [**LR87**], there are three different types of missing data or missingness mechanism:

- Missing Completely At Random (**MCAR**): $\mathbf{x}_{mis}$ are a simple random sample of all data values. Missingness does not depend on the data values.
  Example: in a questionnaire, the accidental omission of an answer.
- Missing At Random (**MAR**): the probability that one value is missing depends on the $\mathbf{x}_{obs}$ part but not on $\mathbf{x}_{mis}$.
  Example: in a questionnaire, the probability of omitting an answer depends on the answer to other questions.
- Not Missing At Random (**NMAR**): the probability that one value is missing depends also on the $\mathbf{x}_{mis}$ part.
  Example: a question on a questionnaire has been deliberately skipped by the participant.

In the particular case of CoDa we can distinguish other simple cases (Table 3.1). In the next section we justify why the most usual NMAR value is the rounded zero value (e.g. Obs. 1). Furthermore, observe that the case when only one value is randomly missing in a closed composition where the sum of the observed part is less than κ (constant constraint sum) has an easy solution: impute the residual part to get the constraint sum value. Consequently, we deal with more complicated situations where the observed part $\mathbf{x}_{obs}$ holds the constraint sum (e.g. Obs. 2) or we have more than one missing part (e.g. Obs. 3). Note that cases such as the Obs. 2 are equivalent to the cases where the samples are not closed. That is, when we have a non-closed sample with missing values and we apply the closure operation using the sum of $\mathbf{x}_{obs}$ as the denominator, the result will be a vector similar to Obs. 2.