PRACTICAL
DATA HANDLING

Visual Presentation of Data by Means of Box Plots

D.L. Massart,[a] J. Smeyers-Verbeke,[a] X. Capron[a] and Karin Schlesier[b]
[a]Vrije Universiteit Brussel, Belgium
[b]Bundesinstitut für Risikobewertung (Federal Institute for Risk Assessment), Berlin, Germany.

Data analysis should always start by (literally) looking at the data. An efficient way to do this is to use *box and whisker plots*, which, for short, are called box plots. All figures in this column are box plots and Figures 2 to 4 are *box plots* for real data sets. In this column we will explain how to construct them and how they can help you to learn more about your data.

Robust Statistics

The box plot is based on *robust statistics*. Robust statistics are called thus because they are more resistant (robust) to the presence of outliers than the classical statistics based on the normal distribution. Consider, for instance, the determination of the mean, one of the simplest statistical operations possible and suppose the following series of results was obtained: 2, 3, 2, 4, 3, 4, 3. The *mean* of these data is 3. It describes one of the aspects of this data set, the so called *central tendency*, in an adequate way. Now suppose an outlier is present, perhaps because an error was made or because one of the objects measured belongs to another population than the other six. One of the 4s is replaced by a 11: 2, 3, 2, 4, 3, 11, 3. The mean is now 4, which is no longer a representative measure for the central tendency of the data set since five of the seven values are smaller than 4 and only one is larger: the mean is not robust towards outlying observations. It can be shown that this is also true for the usual measure of dispersion or spread within the data, the *standard deviation*.

The Median

Robust statistics replaces the mean by other measures of central tendency and the most common of those is the *median*. When the number of data is odd, the median is the middle observation in a ranked series of data. When it is even, it is the mean of the two middle observations. In the first example of the preceding section the ranked data are: 2, 2, 3, 3, 3, 4, 4. There are seven data and, therefore, the fourth in the series is the median: it is equal to 3. The second example yields the ranked series 2, 2, 3, 3, 3, 4, 11. The median is still 3 and is, therefore, not affected by the outlier: it is robust towards the outlier. Even if the outlier were 100, the median would still be 3.

Interquartile Range

The *interquartile range* or IQR is a robust way of describing the dispersion of the data. It is the range within which the middle 50% of the ranked data are found. This is also the range between what is called the *lower quartile value* and the *upper quartile value*. Let us consider a somewhat larger data set to show how the IQR is determined (See Table 1).

First the median is obtained. Since the number of data is even, the median is equal to the mean of the two middle values, 13.1 and 13.3, (i.e., the median is 13.2). The data is then split into an upper and a lower half. When the number of data is odd, which is not the situation in the example, the median value is included in both halves. The median of the first half is the lower quartile value, (i.e., the value below which the 25% lowest values are found). Here it is the sixth value, (i.e., 12.0). Likewise the upper quartile value is the value above which the 25% highest values are found. It is equal to 14.5 and the IQR is the difference between the two quartile values, (i.e., IQR = 14.5 – 12.0 = 2.5).

The Box and The Whiskers

The median and the IQR are used to construct the box. It has a height equal to the IQR and is drawn so that it starts at the lower quartile value and stops at the upper quartile value. A horizontal bar is drawn at the height of the median. In our example (Figure 1(a)) the box encompasses the values from 12.0 to 14.5, and the horizontal bar for the median is drawn at 13.2. The whiskers indicate the range of the data and they are usually represented as vertical lines ending in a small horizontal line. In our situation, they are drawn at 11.0 and 16.0.

A provision is also made for the representation of extreme values. An upper

Table 1: Data set 1.

11.0	11.2	11.5	11.6	11.9	12.0	12.2	12.8	12.9	12.9	13.1
13.3	13.4	13.8	13.9	14.2	14.5	14.5	14.6	15.3	15.5	16.0

extreme value limit is computed as the upper quartile range + 1.5 × IQR and the lower extreme value limit as the lower quartile range − 1.5 × IQR. These limits are therefore 12.0 − 1.5 × 2.5 = 8.25 and 14.5 + 1.5 × 2.5 = 18.25. None of the data exceeds these values and, therefore, the box plot is left unchanged.

When the highest value 16.0 is replaced by 19.0, which is quite high compared with the rest of the series (See Table 2), the highest value now exceeds 18.25 and is, therefore, considered extreme. The box plot emphasizes the presence of the extreme value by giving it a specific symbol; in Figure 1(b), a + sign. The box does not change except that the highest value within the extreme value limit is now 15.5, instead of 16.0 and the upper whisker is now drawn at this value.

Using Box Plots

Box plots can be used in many ways and, to illustrate this, we will make use of data from a European Union funded project, called Wine-D(ata)B(ase). The aim of the project is to be able to authenticate the origin of wines from certain countries outside the European Union. For five such countries each year 100 samples are taken according to a sampling plan that guarantees representativity as well as possible. Some of these samples are commercial wines, some are wines obtained from grapes collected by official oenological institutes at exactly known places and vinified according to known procedures (authentic wines). Initially close to 100 variables were measured for each sample, such as classical oenological parameters, macroelements, isotopic ratios, inorganic and organic minor and trace compounds such as trace elements, biogenic amines, alcohols and esters. This yields very large databases and advanced data analysis is needed. Many of the techniques used are, of course, multivariate, such as principal component analysis (PCA),[1] but in an initial phase, it is also necessary to analyse each variable individually (i.e., in a *univariate* way). In what follows we will give examples of box plots obtained for these data. Since the results have not been published yet, we will not identify the countries nor the variables.

The most evident application is to identify samples with extreme characteristics.

Extreme Values (Outliers)

The most evident application is to identify samples with extreme characteristics. Figure 2 gives examples of box plots obtained for four countries and three variables. Figure 2(a) tells us that this variable is very well behaved: no single sample is considered extreme. In Figure 2(b) there are a few extreme values. Country B has a low value for sample 093 and one low and one high value for country D (samples 455 and 465 respectively). In Figure 2(c) we see a very different picture. There are now extreme values for all countries, a rather large one (sample 112) for country C and clusters of extreme samples for countries B and D.

A question might be how conservative the box plot is: does it tend to find many or only few extreme values? For a normal distribution, the IQR is equal to about 5 standard deviations, which means that about 2% of all data coming from such a distribution would be considered extreme. This is somewhat more than for the usual outlier test where the recommended level would be 1%. Depending on the software, some box plots are more conservative in the sense that in addition to the extreme level as computed above, they use a second extreme level, more distant from the box than the first one. This is the situation in Figure 3, where a distinction is made between close extreme values and far away extremes (the * symbol).

Classical outlier tests would not be effective for the data set described here.

Table 2: Data set 2.

11.0	11.2	11.5	11.6	11.9	12.0	12.2	12.8	12.9	12.9	13.1
13.3	13.4	13.8	13.9	14.2	14.5	14.5	14.6	15.3	15.5	19.0

Figure 1: Box plots for 22 data points (see text). (a) Without and (b) with one extreme value.

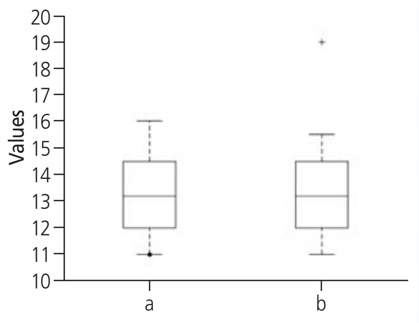
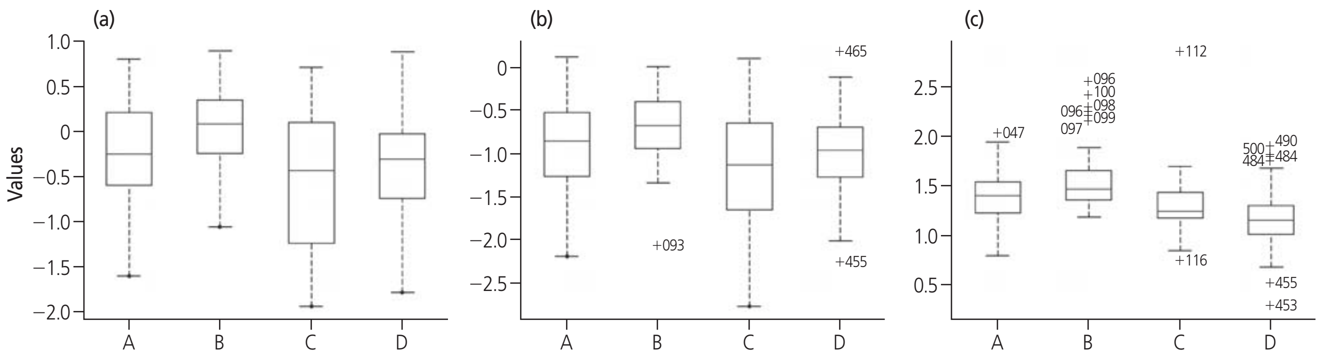


Figure 2: Box plots for wine samples collected in four countries (A–D) for three variables (a–c).



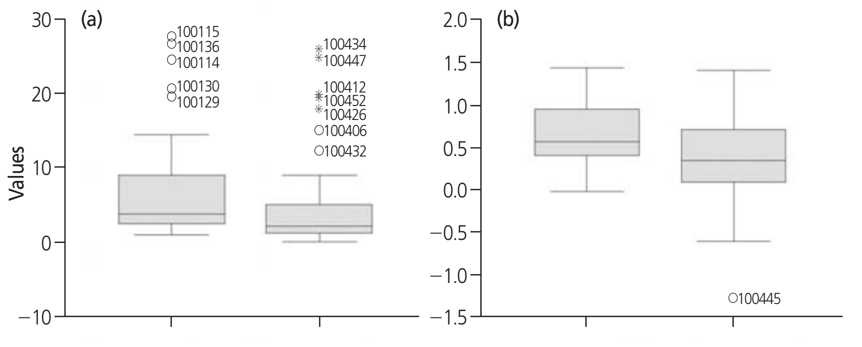
This should not be understood as a statement that outlier tests are never useful. For instance, when a laboratory repeatedly measures the same sample and one of the results appears to be far away from the rest, an outlier test can be useful to make a decision. The outlier test which is most accepted in the analytical community is Grubb's test. It is the test recommended by ISO to investigate replicate determinations. There is a single Grubb's test, used to decide on a statistical basis if one single result is an outlier and there is a double Grubb's test, which is used when there are two suspect values. While this test would probably perform correctly for the data of Figures 2(a) and (b), it is, however, clear that it cannot be used for the variable of Figure 2(c). For many sets of data, such as the one here, the visual analysis by box plots is to be preferred.

Finding extreme values, and sometimes many extreme values, is the rule rather than the exception in data like this. Therefore, samples are not eliminated because one or two variables show outlying values, except when it is found that this is because of technical reasons, for instance malfunction of an instrument. A sample with uncharacteristic values may truly represent part of the population of samples and should not be deleted. This is also the reason why we prefer in this context the term "extreme value" to "outlier". Such samples should be marked as "suspect" or "extreme," but still be processed.

Skewed Distributions

Contrary to the implicit belief of many analytical chemists, data are often not normally distributed. Data with low values may for instance be more frequent than data with a high value, leading to skewed distributions with a long tail towards the high end of the distribution. Such distributions are often found for data sets consisting of environmental, food and other natural samples and can easily be detected with a box plot. Because there are more data with low values, the median is shifted towards the low end of the data range and is found towards the bottom of the box. Usually there are also several samples with extreme high results. They might (wrongly) be considered to be outliers and deleted from the data set. If they are kept in, they have too large an influence in the data analysis. Moreover, many statistical tests, such as the *t-test*, assume normal distribution of the data. When a skewed distribution is encountered it is, therefore, always better to try transforming it into a more symmetric and preferably a normal distribution.

Figure 3: Box plots of two series of wine samples. (a) Original data, (b) after log transformation.



An example of a skewed distribution is shown in Figure 3(a). The box plots concern the same trace compound in two sets of about 50 wine samples. In Figure 3(b) the same data as in Figure 3(a) are shown, but instead of the raw data, their logarithms are plotted. This is called a *log transform*. There are now no longer extreme values and the median has moved towards the centre of the box. This is indicative of what is called a *log normal distribution*. Statistical tests like the *Kolmogorov–Smirnov* test can be applied to confirm that the data are indeed log normally distributed. For the wine data set it was shown that nearly all, close to 100, variables were log normally distributed and the log values were systematically substituted for the original data in the subsequent data analysis.

Comparing Series of Results

Comparing two or more series of results is one of the most often performed data analysis tasks. Classical statistical methods are the *t-test* for comparing means and the F-test for comparing variances of two series of data, and *analysis of variance* (ANOVA) for more than two series. These methods are vulnerable to the presence of outliers and are based on assumptions such as normal distributions and (depending on the test) equal variance.

The juxtaposition of box plots is an excellent way to investigate if there are differences between the data sets and can be applied without any statistical assumptions. In our wine example, the analyst would like to know whether certain variables allow a distinction to be made between wines of different countries. In Figure 4 the discrimination potential of three variables is investigated. It is immediately clear that variable (a) is useless: the box plots show that the four countries yield similar results. Variable (b) shows some potential since it discriminates most of the B samples from those from country D; also C is to some extent separated from both B and D. Alone, this variable is not able to perform a complete discrimination between any pair of countries, but together with other variables, (i.e., using a multivariate method), it might still prove useful. Variable (c) achieves excellent separation of country D from the others.

Conclusion

Always look at your data!

Acknowledgement

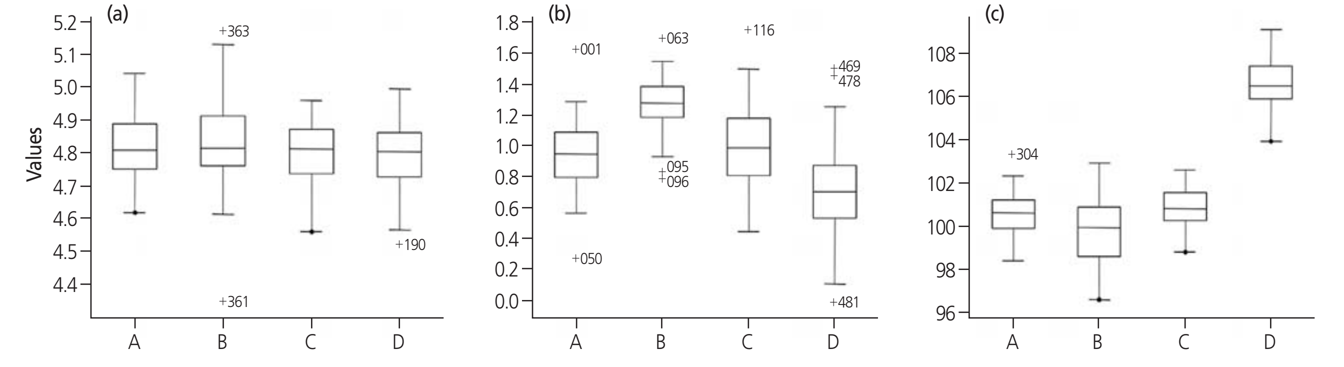
This publication contains results from the EU project "Establishing of a wine data bank for analytical parameters for wines from third countries (wine DB)", a shared-cost RTD project funded under the Fifth Framework Programme of the European Community, within the Competitive and Sustainable Growth Programme (G6RD-CT-2001-00646). The authors are solely responsible for the content of the publication. It does not represent the opinion of the Community nor is the Community responsible for any use that might be made of the data herein.

References

1. D.L. Massart and Y. Vander Heyden, *LC•GC Eur.*, **17**(11) 586–591, (2004).

Column editor, **Desire Luc Massart** is a part-time professor at the Vrije Universiteit Brussel, Belgium. He performs research on chemometrics in process analysis and its use in the detection of counterfeiting products or illegal manufacturing

Figure 4: Box plots comparing the results obtained for wine samples from four countries (A–D) for three variables (a–c).



processes.

Johanna (An) Smeyers-Verbeke is a professor at the Vrije Universiteit Brussel and is head of the department of analytical chemistry.

Xavier Capron is a PhD student at the same department.

Karin Schlesier, PhD in natural sciences, employed at the Federal Institute for Risk Assessment, Berlin, Germany, is responsible for the coordination of the multinational WINE DB project.