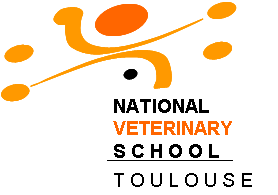
**Reference guide for**

**Reference Value Advisor v2.0**

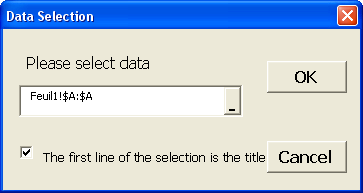
** November 2011**

ReferenceValue Advisor is a set of Excel macros that compute reference intervals from data contained in spreadsheet. It closely follows the CLSI guideline [1].

**Data**

The data should be contained in columns whose first line indicates the name of the analyte for which a reference interval is to be computed.

A click on the button RvA gives the following menu



Select the data with the mouse and click on the OK button.

**Results given by Reference Value Advisor**

Reference Value Advisor creates

* for each analyte a spreadsheets containing the report of the analysis,
* for all analytes a spreadsheets containing information about outliers,

As an example open the file Creat

This file contains two columns: Creat and Urea

Selecting the columns Creat and Urea produces three spreadsheets: Report for Creat, Report for Urea, Outliers Analysis.

The spreadsheet Outliers Analysis contains outliers' information for both Creat and Urea.

Let us see now in detail the information given in the report spreadsheet.

This spreadsheet contains 2 pages reproduced hereafter.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***Reference Value Advisor v2.0*** | ***Results for Creat*** | | | | |
|  |  |  | *Date* | *10/01/2010* |
|  |  |  | *Performed by* | *Didier* |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  | Untransformed data | | Box-Cox transformed data | |  |
| Method | Standard | Robust | Standard | Robust | Nonparametric |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| N | 1460 | 1460 | 1460 | 1460 | 1460 |
| Mean | 93.0 |  | 3.4 |  |  |
| Median | 88.0 | 89.4 | 3.4 | 3.4 |  |
| SD | 26.8 | 27.2 | 0.2 | 0.2 |  |
| Minimum | 35.0 | 35.0 | 2.7 | 2.7 |  |
| Maximum | 407.0 | 407.0 | 4.2 | 4.2 |  |
|  |  |  |  |  |  |
| *λ1* coefficient **Box-Cox** |  |  | -8.5 | -8.5 |  |
| *λ2* coefficient **Box-Cox** |  |  | -0.126 | -0.126 |  |
| p-value **Anderson-Darling** | 0.000 | 0.000 | 0.000 | 0.000 |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| Outliers **Dixon** |  |  |  |  |  |
| Outliers **Tukey** | 8 | 8 | 1 | 1 |  |
| Suspect data **Tukey** | 28 | 28 | 26 | 26 |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| Lower limit of reference interval | **40.4** | **36.0** | **54.4** | **54.6** | **53.0** |
| Upper limit of reference interval | **145.5** | **142.8** | **155.3** | **156.4** | **150.5** |
|  |  |  |  |  |  |
| 90% CI for lower limit | 36.0 | 31.4 | 53.4 | 53.6 | 52.0 |
|  | 44.3 | 40.3 | 55.5 | 55.7 | 53.0 |
| 90% CI for upper limit | 140.5 | 138.3 | 151.4 | 152.4 | 148.0 |
|  | 151.2 | 147.4 | 159.4 | 160.5 | 159.0 |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| **Comments** |  |  |  |  |  |
| Possible outliers detected according to Tukey or Dixon. IFCC-CLSI C28-A3 recommends that unless outliers are | | | | | |
| known to be aberrant observations, the emphasis should be on retaining rather than deleting them. | | | | | |
| The sample size is large enough to compute the nonparametric reference interval : [53 ; 150.5]. | | | | | |
| The 90% CI of one (or more) limit is larger than recommended in IFCC-CLSI C28-A3. | | | | | |
|  |  |  |  |  |  |
|  |  |  |  |  |  |



**The standard method**

The standard method assumes the data distribution is Gaussian and the 95% reference interval is computed as

 (eq. 1)

where

*N* is the sample size named as N in the output table,

 is the sample mean. Its value is given in the Mean row,

 is the sample standard deviation whose value is given in the SD row,

 is the 0.975 quantile of a Student distribution with *N-1* degrees of freedom.

The lower (resp. upper) limit of the reference interval is given in the row entitled Lower limit of reference interval (resp. Upper limit of reference interval) for the columns named Untransformed data.

**The robust method**

The robust method assumes that the data distribution is symmetric without being necessarily Gaussian. It is built using robust statistics (median and median absolute deviation) that are less sensitive to outliers. This reference interval is obtained by a process that iteratively gives a smaller weight to data far from the central location. The method is fully explained in [2] and the tuning parameters that need to be used for this computation are given in [1].

The 95% reference interval is computed with the robust method is given by

 (eq. 2)

where

*N* is the sample size named as N in the output table,

 is an estimate of the central location of the data distribution. Its value is given in the Median row,

*SD* is an estimate of the data dispersion whose value is given in the SD row

 is the 0.975 quantile of a Student distribution with *N-1* degrees of freedom.

The lower (resp. upper) limit of the reference interval is given in the row entitled Lower limit of reference interval (resp. Upper limit of reference interval).

The robust method is sensitive to asymmetry of the distribution. For this reason, the data distribution should be first carefully inspected.

**Confidence intervals of the limits**

Confidence intervals (CI) of the reference limits are computed for all methods.

For the standard method the 90% CI is obtained using parametric bootstrap when N≤20.

In all other cases a non parametric bootstrap is used.

Reference Interval Advisor spends most of the computation time to compute these confidence intervals.

The 90% confidence interval of the lower (resp. upper) reference limit is given in the rows 90% CI for lower limit (resp. 90% CI for upper limit).

The width of the confidence intervals depends on the sample size N and on the data dispersion. When N is low, the confidence intervals are wide while large N give narrow confidence intervals. The confidence interval of the reference limits should be given whatever the sample size used for the computation. They give useful information on the precision of the computed reference interval.

**The generalized Box-Cox transformation**

The generalized Box-Cox transformation columns contains information about the data after the following Box-Cox transformation



The parameters ** and **that enter in the Box-Cox transformation are chosen so that the transformed data distribution is as Gaussian as possible. The ** values given in the *λ* coefficient **Box-Cox** rows optimize the likelihood. An interval is first build using equation (eq. 1) on the Box-Cox transformed data. This interval is then back-transformed using  to give the reference interval given in the rows Lower/Upper limit of reference interval for the columns Box-Cox transformed data. The function  is given by

.

**Normality and symmetry tests**

Normality is required by the standard method for an accurate computation of the reference interval. The row p-value **Anderson-Darling**/**symmetry test for Robust** gives the result of the normality test for both the untransformed and the Box-Cox transformed data. When it is little than 0.05, normality should be questioned. Non normality can be due to a non normal distribution or presence of outliers. There exist other tests for normality such as Kolmogorov-Smirnov test, Chi-square test… We deliberately chose to retain only the Anderson-Darling test as it mainly looks for deviation from normality in the tails of the distribution which are of primary importance for the reference interval determination.

The robust method required the data distribution to be symmetrical about its median. For the robust method, the result of the symmetry test is contained in the row p-value **Anderson-Darling**/**symmetry test for Robust**. When this P-value is less than 0.05, symmetry should be questioned and the robust method should not be used. This test is completely described in [4].

**Outliers**

An outlier is a value that has a small probability to have been observed in the reference population. Outliers can distort a reference interval determination. There exist three kinds of outliers:

1. data that are the results of a mistake. These data should be removed from the analysis.
2. data that have been collected on a patient that does not belong to the reference population. If the reference population contains values for healthy patients, other results such as analytes concentrations/ clinical examination can help to detect these patients. When detected, these data should be removed from the analysis.
3. data that are far from the other. This situation is touchier than the two others as it can be explained by 1°), 2°) or by a wrong choice of distribution to describe the data. It can be identified by a large number of values detected as suspect or as outliers. When the chosen distribution does not fit the data, it is reasonable to change the distribution not the data.

For this reason, we suggest

1. to use the robust method [2] described hereafter that is less sensitive to the shape of the chosen distribution.
2. to compute the reference interval after removing these data and to compare the reference intervals obtained with and without the potential outliers. If the reference interval does not change too much then keeps this reference interval.
3. to get more data (at least 120) so that the nonparametric method can be used. As we will see, the nonparametric method does not assume any specific shape for the data distribution.
4. to change the distribution. As the current version of this code does not propose other distributions, a solution is to consult your preferred statistician.

The three lines Outliers Dixon, Outliers Tukey and Suspect data Tukey should help you to detect such values.

The row Outliers **Dixon** gives the results of the tests for the minimum and the maximum values of the untransformed/Box-Cox transformed data. The corresponding cells table can contain Min, Max or Min/Max respectively meaning that the minimum, the maximum, both the minimum and the maximum values are outliers. When these cells are empty, the Dixon test did not detect that these values as outliers. This test is known to be rather insensitive.

The rows Outliers **Tukey** and Suspect data **Tukey** respectively contain the number of outliers and suspect data identified using the Tukey method [3].

Tukey differentiated "mild" and "extreme" outliers that we here respectively named suspect and outliers data.

If Q1, Q2, Q3 respectively represent the first, second and third quartiles of the distribution and IQ=Q3-Q1 is the inter-quartile range, outliers are data smaller than Q1-3 IQ or greater than Q3+3\*IQ and suspect data are data contained in the interval [Q1-3 IQ; Q1-1.5 IQ] or in [Q3+1.5\*IQ ; Q3+3\*IQ].

The suspect data and the outliers are respectively written in orange and red cells in the spreadsheet named "Outliers analysis" represented hereafter.

References

1. CLSI. Defining, establishing, and verifying reference intervals in the clinical laboratory; approved guideline. Third ed. Wayne, PA: CLSI, 2008.
2. Horn PS. A biweight prediction interval for random samples. J Am Stat Assoc 1988;83:249-256.
3. John W. Tukey. "*Exploratory Data Analysis*". [Addison-Wesley](http://en.wikipedia.org/wiki/Addison-Wesley), Reading, MA. 1977.
4. McWilliams TP. A Distribution-Free Test for Symmetry Based on a Runs Statistic. J Am Stat Assoc 1990;85:1130-1133.