

CL Statistics Calculator FAQs

Q. So is it exactly the same as the Local Authority model?

A. Yes. They're exactly the same at the moment, but we may add additional features as we get feedback from users.

Q. The calculator doesn't identify low outliers. Is there any reason for this?

A. The reason that it's not in is that it isn't part of the new guidance to spot low outliers. But we'll note this as something to perhaps add to the next version of the software. It's a useful feature for understanding and zoning the site.

Q. Can you use the calculator for the assessment of lead – given that the EA guidance indicates that you should do the US95 test on the log of the data?

Q. Since page 6 of the user manual gives lead as an example, I assume that we are now to use the arithmetic mean instead of the geometric mean for lead.

A. As far as ESI Ltd understand the CL:AIRE/CIEH guidance on which the Statistics Calculator is based does not replace any guidance issued by the Environment Agency for the assessment of lead.

The user may enter the log of the soil contamination concentrations in the *Data* worksheet if they wish to compare it against the log of the Critical concentration which is entered into the *Summary* worksheet. In this instance either of the cells to enter the dataset name or units on the *Data* worksheet could be used to record that the log of the data are being used.

Q. It would be useful to have more columns on the data page (e.g. lithology, depths or just user-defined) & a separate tab so that I could summarise the SGVs etc - so then it be possible to use the stats package as my main data source – any chance?

Q. Would it be possible to add the outlier values to the individual summary sheets?

A. Whilst these useful features could be added to the Statistics Calculator, there are no plans to develop the tool any further. Should this situation change ESI Ltd will be considering adding the additional functionality suggested. Please continue to check www.esinternational.com for further information.

Q. When I choose to go to a contaminant individual summary the outlier test changes from log-normal to normal – can this be stopped?

A. There is no way that this can be stopped in this version of the Statistics Calculator.

Q. I have some data where the Shapiro-Wilk test tells me that it's not-normally distributed and the histogram looks a better fit with a log-normal curve than a normal curve. However the "outlier test" page has "normal" highlighted in the drop down box and with that that there is an outlier. I assumed that this is because another dataset was normally distributed and so affected this one ("It is noted that the choice in this drop down box will be applied to all datasets in the spreadsheet. The user must decide and choose from the drop down box the distribution which should be used to

test for outliers for each individual dataset” – page 11 of the user manual). So, I changed this drop-down to log-normal and there is now no longer an outlier.

My confusion comes from the premise that the Grubbs test relies on normally-distributed data (other than the outlier). If there is an outlier, it could skew the data so giving an erroneous result for the normality test. It feels that I am then caught in a cycle (the main dataset, minus the outlier, needs to be normal for the grubbs test, but to find out if it is normal you need to know if there is an outlier that needed to be removed). I'm not sure I'm making a lot of sense – as I said, I'm confused.

A. The automatic assessment of whether a dataset is normal or non-normal shown on the *Summary* worksheet, is based on all corresponding data in the *Data* worksheet. If outliers are permanently excluded or deleted from a dataset the normality test is updated automatically. The choice of assuming either a *normal* or *log-normal* distribution for the outlier test is independent of the automatic normality test.

For specific advice on the use of Grubb's Test the user should refer to Appendix A of the CL:AIRE/CIEH guidance.

Q. I have a sample population that is non-normally distributed and I apply a log-normal outlier test identifying NO outliers. I then complete the test and because I have a significantly elevated concentration within the sample population my UCL value is very high, greatly exceeding the critical concentration resulting in acceptance of the null hypothesis under the Planning Regime.

It is obvious from visual assessment that I have an outlier in the population and when I apply the outlier test normally as opposed to lognormally the outlier is identified. This is then removed from the sample population and the UCL generated falls below the CC allowing rejection of the null hypothesis.

It is clear there is an area of the site where outliers are being generated that needs to be considered separately however if I blindly follow guidance and apply a lognormal test for outliers for all non-normally distributed data sets these 'significant concentrations' are not recognised as outliers and the generated UCL is not truly representative of the area being considered.

Should I be logtransforming for all outlier tests where I have a non-normally distributed data set? (My interpretation of Appendix B of CLAIRE Guidance)

A. When dealing with outliers, the CL:AIRE/CIEH guidance states that various 'outlier' tests are available and Grubb's Test is given in Appendix B as 'one such test'. The guidance explains that Grubb's Test assumes that the data set excluding the suspect value (the maximum) is normally distributed. The guidance suggests that if the dataset proves not to be normally distributed, consider transforming the data by taking the natural logarithm and checking the normality of the transformed dataset. If the dataset excluding the maximum (be it logged or not) is not normally distributed then the assumptions needed for Grubb's Test are not met. In other words, you should consider using another more suitable test (the guidance offers a good reference for these).

If the dataset (excluding the maximum) is normally distributed select the *normal* option in the Outlier Test. If the dataset (excluding the maximum) when log-transformed is normal select the *log-normal* option for the test.

The Statistics Calculator only implements Grubb's Test. If you can't show that the dataset excluding the maximum is normally distributed (and hence satisfies the assumptions needed for Grubb's Test) then you need to find another more appropriate test. It is important to fully understand the validity of any statistical test. To be clear, the Statistics Calculator, does not include an Outlier Test for looking at non-normal data.

To check the suitability of Grubb's Test you can automatically assess the normality of a dataset excluding the maximum using the Statistics Calculator by temporarily copying the relevant data to a new column in the Data worksheet and looking at the Normality Test result. Logged data could also be checked in this manner if you wanted. There is currently no automatic function in the Statistics Calculator to log the data and assess the normality of the transformed data. Please note that the final paragraph in Section 5 of the guidance says that it does not recommend log transformation as a means of normalising soil contamination data when calculating confidence limits.

Irrespective of this, you appear to have reached the conclusion that there is an area of site that should be considered separately and hence you might wish to consider Section 5.3.2 of the guidance "*In general, however, outliers should be excluded from a dataset ONLY where they clearly indicate that more than one soil population exists within the dataset and this can be justified by (or informs the further development of) the conceptual model – in which case the different population expressed by the outlier(s) should be explored in more detail, either by reviewing and refining zoning decisions and treating outlier values as a separate population or even individually or, if necessary, by undertaking further site sampling to verify conditions in the vicinity of outlier values.*".

Q. When I attempt Step 2ii of Example 2 the value for T_n given by the software is 1.249 not 1.96 as stated in the user manual. Why is this?

A. The user manual (May 2008) states that log-normal distribution should be used to test for outliers in the example when in fact normal distribution should be used because the dataset without the maximum value is approximately normal. The user should confirm that *normal* distribution is selected in the drop down box instead. Note that this will not affect the outcome of the outlier test in the example.