Procedure for Dealing With Outliers

An outlier is defined as an observation or "data point" which does not appear to fall within the expected distribution for a particular data set. Outliers may be rejected outright if they are caused by a known or demonstrated physical reason, such as sample spillage, contamination, mechanical failure, or improper calibration. Data points which appear to deviate from the expected sample distribution for no known physical reason must be verified as outliers using statistical criteria.

Outliers can significantly alter the outcome of a method detection limit calculation. Including outliers in an MDL calculation leads to increased variability (larger standard deviation). An MDL calculated using outliers will be inaccurate and higher than the true detection limit. For this reason, it is important to recognize outliers, and to reject them from the calculation. Since the procedure requires at least seven replicates, rejecting one of only seven sample results will result in too few data points to calculate an MDL.

Table of Critical Values
(1% significance value)

# Observations	Critical Value
7	2.10
8	2.22
9	2.32
10	2.41
11	2.48
12	2.55
13	2.61
14	2.66

For the MDL procedure, all data sets will only be samples of the true population, and both the population mean (μ) and the population standard deviation (σ) will be unknown. The expected distribution for MDL observations is most closely represented by a log-normal distribution, and only one-sided outliers should be expected. Due to the nature of the MDL procedure (low-level precision), most outliers will be high-sided, and the only test necessary will be a single-sided outlier test. A low-sided outlier could occur, but the data would be unusable because it would most often appear as a "no detect".

One method for determining single sided outliers when both the population mean (μ) and the population standard deviation (σ) are unknown was described by Grubbs (F.E. Grubbs 1979) and is included in *Standard Methods*.

$$T_n = X_n - X_{ave}/s \quad \text{(high sided outliers)}$$
$$T_1 = X_{ave} - X_1/s \quad \text{(low sided outliers)}$$

Where X_n (X_1) is the data point in question, X_{ave} is the sample mean, and s is the sample standard deviation. The value T_n is then compared against a table of critical values. If T_n is greater than the critical value for the appropriate number of replicates at the 1% significance level, the questionable data point is an outlier, and it may be rejected. The critical values for various numbers of replicates at the 1% significance level are given in the sidebar.

Example 1: The following results were obtained for an MDL study: [10.2, 9.5, 10.1, 10.3, 9.8, 9.9, 11.9, 10.0] with $X_{ave}= 10.2$ and s= 0.726.   The analyst suspects 11.9 to be an outlier. Using the high-sided test:

$$T_n = 11.9 - 10.2 / 0.726 = 2.34$$

The calculated T_n value is now checked against the table. Since 2.34>2.22, 11.9 is indeed an outlier.

Example 2: The following results were obtained : [0.523, 0.562, 0.601, 0.498, 0.547, 0.525, 0.578, 0.503] with X_{ave}= 0.542 and s= 0.036. Is 0.601 an outlier?

$$T_n = 0.601 - 0.542 / 0.036 = 1.64$$

Checking the table shows that 1.64<2.22 and 0.601 is not an outlier and could be included in the MDL calculation.

References

Grubbs, F.E. 1979. Procedures for detecting outlying observations. In *Army Statistics Manual* DARCOM-P706-103, Chapter 3. U.S. Army Research and Development Center, Aberdeen Proving Ground, MD 21005.

American Public Health Association, *Standard Methods for the Examination of Water and Wastewater*, 17th, 18th or 19th Editions, (1989, 1992 or 1996).

This document was prepared by the DNR's Office of Technical Services, Laboratory Certification Program.