APPENDIX I
Identification and Handling of Outliers

I-1. Purpose.

I-1.1. Outliers are measurements that are extremely large or small relative to the rest of the data and, therefore, are suspected of misrepresenting the population from which they were collected. Outliers influence statistics if used in calculations, and statistical tests based on parametric methods are generally more sensitive than nonparametric methods to outliers. Outliers may result from transcription errors, data-coding errors, or measurement system problems, such as instrument breakdown. However, outliers may also represent true extreme values of a distribution and may indicate more variability in the population or a different underlying distribution for the population than what was initially assumed. For example, a point that appears as an outlier under the assumption that the underlying distribution is normal will not necessarily appear as an outlier if it were initially assumed that the distribution is lognormal. Not removing true outliers or removing false outliers can lead to a distortion of estimates of population parameters.

I-1.2. Statistical outlier tests give the analyst probabilistic evidence that an extreme value (potential outlier) does not fit with the distribution of the remainder of the data and is a statistical outlier. These tests should only be used to *identify* data points that require further investigation. Tests alone cannot determine whether a statistical outlier should be discarded or corrected within a data set; this decision should be based on judgment and scientific reasoning. (See EPA 600/R-96/084, Gilbert, 1987, for further details on identifying and handling outliers.)

I-2. Methods. Five steps are involved in treating extreme values or outliers:

- Identify extreme values that may be potential outliers.
- Apply a statistical test.
- Scientifically review statistical outliers and decide on their disposition.
- Conduct data analyses with and without statistical outliers.
- Document the entire process.

Potential outliers can be identified through graphical representations. Graphs, such as the box-and-whisker plot, normal probability plot, and time plot, can be used to identify observations that are much larger or smaller than the rest of the data. (Appendix J presents these graphical tools.) If potential outliers are identified, the next step is to apply one of the statistical tests described below.

I-2.1. *Dixon's Test*. Dixon's extreme value test can be used to test for statistical outliers when the sample size is less than or equal to 25. This test considers extreme values that are much smaller or larger than the rest of the data. Because this test assumes that the data without the suspected outlier are normally distributed, it is necessary to test for normality in the data without the suspected outlier before applying Dixon's test. If the data are not normally distributed, a transformation that normalizes the data should be applied, or a different test should be used. Directions for the extreme value test are contained in Paragraph I-2.1.1 followed by an example in Paragraph I-2.1.2. Dixon's test should be used when only one outlier is suspected in the data. If more than one outlier is suspected, the extreme value test may lead to masking, in which two or more outliers close in value obscure one another. Therefore, if the analyst decides to use the extreme value test for multiple outliers, it should be applied to the least extreme value first; otherwise, Rosner's test should be used to test for multiple outliers. Rosner's test is discussed below.

I-2.1.1. *Directions for the Extreme Value Test (Dixon's Test)*. Let $x_{(1)}, x_{(2)},\ldots,x_{(n)}$ represent the data ordered from smallest to largest. Check that the data without the suspected outlier are normally distributed, using one of the methods in Appendix F.

I-2.1.1.1. If normality fails, transform the data or apply a different outlier test.

I-2.1.1.2. Case 1: $x_{(1)}$ is a potential outlier. Compute the test statistic C, where

$$C = \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}} \text{ for } 3 \le n \le 7, \quad C = \frac{x_{(3)} - x_{(1)}}{x_{(n-1)} - x_{(1)}} \text{ for } 11 \le n \le 13,$$

$$C = \frac{x_{(2)} - x_{(1)}}{x_{(n-1)} - x_{(1)}} \text{ for } 8 \le n \le 10, \quad C = \frac{x_{(3)} - x_{(1)}}{x_{(n-2)} - x_{(1)}} \text{ for } 14 \le n \le 25.$$

I-2.1.1.3. If C exceeds the critical value from Table B-5 of Appendix B for the specified significance level α, $x_{(1)}$ is an outlier and should be further investigated.

I-2.1.1.4. Case 2: $x_{(n)}$ is a potential outlier. Compute the test statistic C, where

$$C = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}} \text{ for } 3 \le n \le 7, \quad C = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(2)}} \text{ for } 11 \le n \le 13,$$

$$C = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}} \text{ for } 8 \le n \le 10, \quad C = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(3)}} \text{ for } 14 \le n \le 25.$$

I-2.1.1.5. If C exceeds the critical value from Table B-5 of Appendix B for the specified significance level α, $x_{(n)}$ is an outlier and should be further investigated.

I-2.1.2. *Example for the Extreme Value Test (Dixon's Test)*. Consider the following subsurface background chromium data in order of magnitude from smallest to largest: 3.84, 4.26, 4.53, 4.60, 5.28, 5.29, 5.74, 5.86 (in mg/kg). Suppose there was an additional sample with a result of 10 mg/kg. As this additional sample is much larger than the other values, it is suspected that this point might be an outlier. The required level of significance for an outlier is 5%.

I-2.1.2.1. Testing the data for normality using the Shapiro-Wilk test (without the extreme value) indicated that the data were normal. Therefore, the extreme value test may be used to determine if the largest data value is an outlier.

$$C = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}} = \frac{10.0 - 5.86}{10.0 - 4.26} = 0.72 \ .$$

I-2.1.2.2. Because $C = 0.72 > 0.512$ (from Table B-5 of Appendix B with $n = 9$ and $\alpha = 0.05$), there is evidence that $x_{(n)}$ is an outlier at a 5% significance level and should be further investigated.

I-2.2. *Discordance Test*. The discordance test can be used to test if one extreme value is an outlier. This test considers two cases: i) where the extreme value (potential outlier) is the smallest value of the data set; and ii) where the extreme value (potential outlier) is the largest value of the data set. The discordance test assumes that the data are normally distributed; therefore, it is necessary to perform a test for normality before applying the discordance test. If the data are not normally distributed, transform the data, apply a different test, or consult a statistician. Note that the test assumes that the data without the outlier are normally distributed, so the test for normality should be performed without the suspected outlier. Directions and an example of the discordance test are contained in Paragraphs I-2.2.1 and I-2.2.2, respectively.

I-2.2.1. *Directions for the Discordance Test*. Let $x_{(1)}, x_{(2)},\ldots,x_{(n)}$ represent the data ordered from smallest to largest. Check that the data without the suspect outlier are normally distributed, using one of the methods of Appendix F, Paragraph F-11. If normality fails, transform the data or apply a different outlier test.

I-2.2.1.1. Compute the sample mean, $\bar{x}$, and the sample standard deviation, s, without the suspected outlier. If the minimum value $x_{(1)}$ is a suspected outlier, compute the test statistic

$$D = \frac{\bar{x} - x_{(1)}}{s} \ .$$

I-2.2.1.2. If D exceeds the critical value from Table B-4 of Appendix B, $x_{(1)}$ is an outlier and should be further investigated.

I-2.2.1.3. If the maximum value $x_{(n)}$ is a suspected outlier, compute the test statistic

$$D = \frac{x_n - \bar{x}}{s} \ .$$

I-2.2.1.4. If D exceeds the critical value from Table B-4 of Appendix B, $x_{(1)}$ is an outlier and should be further investigated.

I-2.2.2. *Example for the Discordance Test*. Consider the following subsurface background chromium data from smallest to largest: 3.84, 4.26, 4.53, 4.60, 5.28, 5.29, 5.74, 5.86 (in mg/kg). Suppose there was an additional sample with a result of 10 mg/kg. Because this additional sample is much larger than the other values, it is suspected that this point might be an outlier. The required level of significance for an outlier is 5%.

I-2.2.2.1. Testing the data for normality using the Shapiro-Wilk test (without the extreme value) indicated the data were normal. Therefore, the discordance test may be used to determine if the largest data value is an outlier.

$\bar{x} = 5.48$ mg/kg and $s = 1.82$ mg/kg without the suspected outlier.

I-2.2.2.2 Because the maximum value $x_{(n)}$ is a suspected outlier, do the following:

$$D = \frac{x_n - \bar{x}}{s} = \frac{10.0 - 5.48}{1.82} = 2.48 \, .$$

I-2.2.2.3. Because $D = 2.48 > 2.110$ (from Table B-4 of Appendix B with $n = 9$ and $\alpha = 0.05$), there is evidence that $x_{(1)}$ is an outlier at a 5% significance level and should be further investigated.

I-2.3. *Rosner's Test*. Rosner developed a parametric test that can be used to detect up to 10 outliers for sample sizes of 25 or more. This test assumes that the data are normally distributed; therefore, a test for normality should be performed before applying it. If the data are not normally distributed, transform the data, apply a different test, or consult a statistician. Note that the test assumes that the data without the outlier are normally distributed, so the test for normality may be done without the suspected outlier. Directions for Rosner's test are contained in Paragraph I-2.3.2 and an example is contained in Paragraph I-2.3.3.

I-2.3.1. *Caveats*. Rosner's test is not as easy as the preceding tests to apply. To apply this test, first determine an upper limit r_0 for the number of outliers ($r_0 \leq 10$), then order the r_0 extreme values from most extreme to least extreme. Rosner's test statistic is then based on the sample mean and sample standard deviation computed without the $r = r_0$ extreme values. If this test statistic is greater than the critical value given in Table B-18 of Appendix B, there are r_0 outliers. Otherwise, the test is performed again with the $r = r_0 - 1$ extreme values. This process is repeated until either Rosner's test statistic is greater than the critical value or $r = 0$.

I-2.3.2. *Directions for Rosner's Test for Outliers*. Let $x_{(1)}, x_{(2)},...,x_{(n)}$ represent the ordered data points. By inspection, identify the maximum number of possible outliers, r_0. Check that the data are normally distributed, using one of the methods in Appendix F, Paragraph F-11.

I-2.3.2.1. Compute the sample mean, $\bar{x}$, and the sample standard deviation, s, for all of the data. Label these values $\bar{x}^{(0)}$ and $s^{(0)}$, respectively. Determine the observation farthest from $\bar{x}^{(0)}$ and label this observation $y^{(0)}$. Delete $y^{(0)}$ from the data and compute the sample mean, labeled $\bar{x}^{(1)}$, and the sample standard deviation, labeled $s^{(1)}$. Then determine the observation farthest from $\bar{x}^{(1)}$ and label this observation $y^{(1)}$. Delete $y^{(1)}$ and compute $\bar{x}^{(2)}$ and $s^{(2)}$. Continue this process until r_0 extreme values have been eliminated.

I-2.3.2.2 In summary, after the above process the analyst should have

$$\left[\bar{x}^{(0)}, s^{(0)}, y^{(0)}\right]; \left[\bar{x}^{(1)}, s^{(1)}, y^{(1)}\right]; \ldots, \left[\bar{x}^{(r_0-1)}, s^{(r_o-1)}, y^{(r_0-1)}\right]$$

where

$$\bar{x}^{(i)} = \frac{1}{n-i}\sum_{j=1}^{n-i} x_j, s^{(i)} = \left[\frac{1}{n-i}\sum_{j=1}^{n-1}\left(x_j - \bar{x}^{(i)}\right)^2\right]^{1/2}$$

and $y^{(i)}$ is the farthest value from $\bar{x}^{(i)}$. (Note the above formulas for $\bar{x}^{(i)}$ and $s^{(i)}$ assume that the data were renumbered after each observation was deleted.)

I-2.3.2.3. To test if there are r outliers in the data, compute

$$R_r = \frac{\left|y^{(r-1)} - \bar{x}^{(r-1)}\right|}{s^{(r-1)}} .$$

Compare R_r to λ_r in Table B-18 of Appendix B. If, $R_r \geq \lambda_r$ conclude that there are r outliers. First, test if there are r_0 outliers (compare R_{r_0} to λ_{r_0}). If not, test if there are $r_0 - 1$ outliers (compare R_{r_0-1} to λ_{r_0-1}). If not, test if there are $r_0 - 2$ outliers, and continue until it is determined there are a certain number of outliers or no outliers at all.

I-2.3.3. *Example for Rosner's Test for Outliers*. Consider the following subsurface site copper data in order from smallest to largest: 1.99, 2.19, 2.34, 2.42, 2.45, 2.64, 2.70, 2.79, 2.82, 2.85, 2.86, 2.93, 3.10, 3.19, 3.21, 3.23, 3.25, 3.26, 3.28, 3.43, 3.55, 3.66, 3.71, 3.76, 3.83, 3.91, 3.92, 3.97, 3.98, 4.48, 5.0, 11.1, 11.6, 12.3, 32.1, 44.2.

I-2.3.3.1. By inspection, five potential outliers are suspected. Testing the data for normality using the Shapiro-Wilk test (without the extreme values) indicated that the data were normal. So Rosner's test for outliers may be used to determine if there are five or fewer outliers.

I-2.3.3.2. First the sample mean and sample standard deviation were computed for the entire data set, $\bar{x}^{(0)}$ and $s^{(0)}$. Subtraction showed that 44.20 was the farthest data point from $\bar{x}^{(0)}$, so $y^{(0)} = 44.20$. Then 44.20 was deleted from the data and the sample mean, $\bar{x}^{(1)}$, and the sample standard deviation, $s^{(1)}$, were computed. Subtraction showed that 32.10 was the farthest value from $\bar{x}^{(1)}$. This value was then dropped from the data and the process was repeated again on 12.30 and 11.60 to yield the values below.

i	$\bar{x}^{(i)}$	$s^{(i)}$	$y^{(i)}$
0	5.88	8.43	44.20
1	4.79	5.36	32.10
2	3.99	2.51	12.30
3	3.74	2.07	11.60
4	3.49	1.54	11.10

I-2.3.3.3. To apply Rosner's test, it is first necessary to test if there are five outliers ($r = 5$) by computing

$$R_5 = \frac{\left|y^{(4)} - \bar{x}^{(4)}\right|}{s^{(4)}} = \frac{|11.10 - 3.49|}{1.54} = \frac{7.61}{1.54} = 4.94$$

and comparing R_5 to λ_5 in Table B-18 of Appendix B with $n = 36$ and $\alpha = 0.05$. Because $R_5 = 4.94 > \lambda_5 = 2.94$, there are five outliers in the data set.

I-2.3.3.4. Suppose $R_5 > \lambda_5 = 2.94$.

I-2.4. *Walsh's Test*. Walsh developed a nonparametric test to detect multiple outliers in a data set. This test requires a large sample size: $n > 220$ for a significance level of $\alpha = 0.05$, and $n > 60$ for a significance level of $\alpha = 0.10$. However, as the test is nonparametric, it may be used whenever the data are not normally distributed. Directions for the Walsh test for large sample sizes are provided in Paragraph I-2.4.1, followed by an example in Paragraph I-2.4.2.

I-2.4.1. *Directions for Walsh's Test for Large Sample Sizes*. Let $x_{(1)}, x_{(2)},...,x_{(n)}$ represent the data ordered from smallest to largest. If $n \le 60$, do not apply this test. If $60 < n \le 220$, then $\alpha = 0.10$. If $n > 220$, then $\alpha = 0.05$.

I-2.4.1.1. Identify the number of possible outliers, *r*. Note that *r* can equal 1.

I-2.4.1.2. Compute

$$c = \left[\sqrt{2n}\right],\ k = r + c,\ b^2 = 1/\alpha$$

and

$$a = \frac{1 + b\sqrt{(c - b^2)/(c-1)}}{c - b^2 - 1}$$

where [ ] indicates rounding the value up to the next largest integer (i.e., 3.24 becomes 4).

I-2.4.1.3. The *r* smallest points are outliers (with an α % level of significance) if

$$x_{(r)} - (1+a)\,x_{(r+1)} + a\,x_{(k)} < 0.$$

I-2.4.1.4. The *r* largest points are outliers (with an α % level of significance) if

$$x_{(n+1-r)} - (1+a)\,x_{(n-r)} + a\,x_{(n+1-k)} > 0.$$

I-2.4.1.5. If both of the inequalities are true, small and large outliers are indicated.

I-2.4.2. *Example for Walsh's Test for Large Sample Sizes*. Consider that the following surface soil lead data from Site 2 in order from smallest to largest: 11.7, 13.9, 14.4, 15.1, 17.2, 19.1, 19.3, 19.5, 19.6, 19.9, 20.8, 21.2, 21.8, 23.4, 24.2, 24.3, 25.8, 26.4, 27.4, 28.1, 29.1, 34.3, 35.3, 36, 37.9, 39.8, 43.8, 45.4, 51.4, 65.4, 74.4, 78.5, 87, 93.3, 105, 108, 120, 134, 135, 136, 143, 150, 178, 186, 194, 203, 214, 216, 232, 251, 263, 268, 277, 283, 300, 421, 446, 510, 811, 1260, 5320.

I-2.4.2.1. The possible outliers are 811, 1260, 5320. So *r* = 3.

$$c = \left[\sqrt{2n}\right] = \left[\sqrt{2 \times 63}\right] = [11.22] = 12$$

$$k = r + c = 3 + 12 = 15$$

$$b^2 = 1/\alpha = \frac{1}{0.10} = 10$$

$$a = \frac{1 + b\sqrt{(c - b^2)/(c-1)}}{c - b^2 - 1} = \frac{1 + 3.16\sqrt{(12-10)/(12-1)}}{12 - 10 - 1} = 2.347$$

$$x_{(n+1-r)} - (1+a)\,x_{(n-r)} + a\,x_{(n+1-k)} > 0$$

$$x_{(63+1-3)} - (1+2.347)\,x_{(63-3)} + 2.347\,x_{(63+1-15)} > 0$$

$$x_{(61)} - (1+2.347)x_{(60)} + 2.347x_{(49)} > 0$$

$$811 - (1+2.347)\,510 + 2.347\,(214) > 0$$

$$-393.712 \not> 0\,.$$

I-2.4.2.2. Therefore the largest points, 811, 1260, 5320, are not outliers at $\alpha = 0.10$.

I-2.5. *Fourth-Spread Outlier Test*. A graphical qualitative method for identifying outliers entails creating box-and-whisker plots. Paragraph J-3 of Appendix J describes how to create such a plot. The process of identifying outliers by generating box-and-whisker plots is the same as identifying outliers using the "fourth-spread" outlier test (Hoaglin et al. 1983). The fourth-spread outlier test can identify one or more outliers from either end of the range of sample results.

I-2.5.1. A box-and-whisker plot identifies mild and extreme outliers. A mild outlier is a statistical outlier that is any result less than the difference of the 25th percentile and 1.5 times the inter-quartile range (IQR), or any result greater than the sum of the 75th percentile and 1.5 × IQR. An extreme outlier is a statistical outlier that is any result less than the difference of the 25th percentile and 3 × IQR, or any result greater than the sum of the 75th percentile and 3 × IQR. Extreme outliers are more severe than mild outliers and should be considered more influential.

I-2.5.2. The advantages of this test are that it does not have any sample size requirements and can identify one or more outliers. A disadvantage of the test is that no level of significance is placed on the decision to declare a result an outlier. However, it should be noted that, for a normally distributed variable X with a standard deviation of σ, 1.5 × IQR is approximately 2σ and there is slightly less than a 1% chance that points will be greater than $X_{0.75}$ + 1.5 × IQR or less than $X_{0.25}$ − 1.5 × IQR. Otherwise, the choice of 1.5 times the inter-quartile range is "somewhat arbitrary, but experience with many data sets indicates that this definition serves well in identifying values that may require special attention" (Hoaglin et al., 1983).

I-2.6. *Multivariate Outliers*. Multivariate analysis, such as factor analysis and principal components analysis, involves the statistical analysis of several variables simultaneously. Outliers in multivariate analysis are values that are extreme in relationship to one or more variables. As the number of variables increases, identifying potential outliers using graphical representations becomes more difficult. Special procedures are required to test for multivariate outliers. Details of these procedures are beyond the scope of this document, but are contained in statistical textbooks on multivariate analysis (see Gnanadesikan, 1997).

I-3. Retaining or Discarding Outliers. Once outliers are identified, the project team should review outliers and determine, case-by-case, if there is an explanation for each outlier. Furthermore, any suspect data point, whether identified as a statistical outlier or not, should be reviewed. Unexpected values, especially those identified as statistical outliers, should not be removed from any data evaluations unless a specific reason for the unexpected measurements can be determined.

I-3.1. If a data point is found to be an outlier, the analyst may: i) correct the data point; ii) discard the data point from analysis; or iii) use the data point in all analyses. Removing outliers should be based on scientific reasoning *in addition to* the results of the statistical test. An outlier should *never* be discarded based solely on a statistical test. Instead, the decision should be based on some scientific or quality assurance basis. Discarding an outlier from a data set should be done with extreme caution, particularly for environmental data sets, which often contain legitimate extreme values.

I-3.2. According to EPA 530-SW-89-026, a value may be corrected or dropped only if one can determine that an error has occurred. If an error can be identified, the correction should be made and the correct value used. Data points containing transcription errors should be corrected whether they are outliers or not. A value that is identified as incorrect may be deleted from the data set. Valid reasons for removing outliers or unexpected values include, for example, evidence they are the result of contaminated sampling equipment, laboratory errors, malfunctioning instrumentation, transcription errors, sampling of differing geological strata, or a non-typical sampling location taken for background. If a plausible reason cannot be found for removing an unexpected value or a statistical outlier, the result should be treated as a true but extreme value and retained in the data.

I-3.3. The spatial context of outliers or potential outliers should be considered. If outliers occur at different locations for different analytes and tend to be located close to low concentrations, then sporadic high concentrations are simply a feature of the area; there is no reason to treat the data differently as a result of their presence. If outliers tend to occur in the same location for different analytes and are found close to other locations with elevated concentrations, it may be appropriate to consider the elevated locations separately.

I-3.4. If an outlier is discarded from the data set, all statistical analysis of the data should be applied to both the full and truncated data set so that the effect of discarding observations may be assessed. If scientific reasoning does not explain the outlier, it should not be discarded from the data set.

I-3.5. If any data points are found to be statistical outliers, this information should be documented along with the analysis of the data set, regardless of whether any data points are discarded. If no data points are discarded, the analyst should document that a process was implemented to identify any statistical outliers but none were found. If any data points are discarded, the analyst should document each data point, the statistical test performed, the scientific reason for discarding each data point, and the effect on the analysis of deleting the data points. Such information is critical for effective peer review.

I-4. Applications. This Paragraph provides a case study regarding outliers and how conclusions are affected by including or excluding outliers. This case study focuses on identifying outliers in background data.

I-4.1. A background metals study was conducted to determine background concentrations that may be compared to site concentrations. Regulators were concerned with identifying outliers in the background data and removing them from the background data set, based upon the *erroneous* assumption that unusually high concentrations cannot represent background conditions and necessarily represent site-related contamination. All background data (by metal), were evaluated for outliers using two outlier tests—the discordance test and fourth-spread test. For this investigation, the regulator required that any result identified as a statistical outlier be removed from the background data set, which *biased* the background sample towards smaller values. This case study focuses on the evaluation of antimony in surface soil.

I-4.2. Table I-1 presents the 20 samples associated with antimony concentrations from the background surface soil. Generally, the concentrations were quite small, ranging from 0.182 to 0.398 mg/kg. Outlier tests were performed on the highest concentration, 0.398 at sample BACK-005-005, to see if this concentration could be considered a statistical outlier.

I-4.3. First, a box-and-whisker box plot was generated to visualize the data and to perform the fourth-spread test. As Figure I-1 presents with the box plot, the highest concentration is a mild outlier.

Table I-1.
Background Surface Soil Data for Antimony

Sample ID	Result (mg/kg)	Sample ID	Result (mg/kg)
BACK-001-0005	0.235	BACK-0011-0005	0.202
BACK-002-0005	0.285	BACK-0012-0005	0.27
BACK-003-0005	0.202	BACK-0013-0005	0.298
BACK-004-0005	0.22	BACK-0014-0005	0.209
BACK-005-0005	0.398	BACK-0015-0005	0.182
BACK-006-0005	0.279	BACK-0016-0005	0.233
BACK-007-0005	0.215	BACK-0017-0005	0.186
BACK-008-0005	0.25	BACK-0018-0005	0.267
BACK-009-0005	0.279	BACK-0019-0005	0.273
BACK-0010-0005	0.23	BACK-0020-0005	0.28

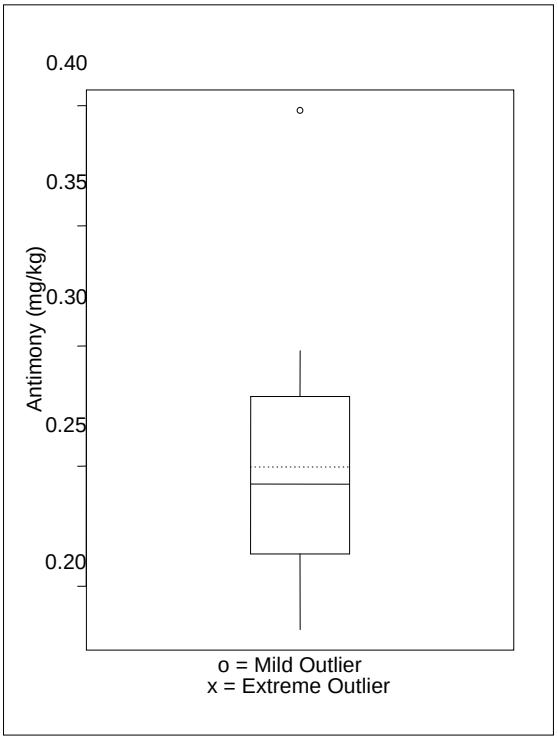

Figure I-1. Box-and-Whisker Plot for Antimony.

I-4.4. The discordance test was done to determine if the maximum result might be considered a statistical outlier. Results of the discordance test show the maximum result is an outlier, as seen below.

I-4.4.1. *Normality Assumption*. The Shapiro-Wilk test was performed on the raw data, without the maximum result. The test statistic for this test was 0.9319 and the *p* value associated with this test statistic was 0.1878. Based on 95% level of confidence, because 0.1878 > 0.05, there is evidence to suggest the data without the maximum result were normal. Therefore, doing the discordance test on the raw data was appropriate.

I-4.4.2. *Test Statistic*. $D = \frac{X_n - \bar{x}}{s} = \frac{0.398 - 0.2418}{0.0366} = 4.268$.

I-4.4.3. *Critical Value*. 2.557 (based on $\alpha = 0.05$).

I-4.4.4. *Conclusion*. Because 4.268 > 2.557, there is evidence that the maximum result is an outlier.

I-4.5. As both outlier tests showed the maximum result is a statistical outlier, the maximum antimony result for surface soil was removed from the background data set at the request of the regulator even though the outlier appeared to be a valid result (i.e., it was not entered incorrectly or demonstrated to be the result of a non-complaint sampling or analytical procedure).

I-4.6. From a statistical perspective, it was probably inappropriate to remove the maximum detected concentration as an outlier for the antimony data set. To illustrate this conjecture, separate lists of summary statistics are presented in Table I-2 for all 20 antimony results and for the 19 antimony results without the maximum concentration.

Table I-2.
Summary Statistics for Antimony Background Surface Soil Data

	n	Minimum (mg/kg)	Maximum (mg/kg)	Median (mg/kg)	Mean (mg/kg)	Standard Deviation (mg/kg)	95% UCL (mg/kg)	Distribution	*p* value for Shapiro-Wilk test for original data	*p* value for Shapiro-Wilk test for log-transformed data
All Samples	20	0.182	0.398	0.2425	0.25	0.04988	0.270	Log-normal	0.0369	0.3309
All but Max.	19	0.182	0.298	0.235	0.242	0.0366	0.256	Normal	0.1878	0.1667

I-4.7. The most striking difference between the two data sets is their distribution. When all samples were evaluated, there was evidence that the data followed a lognormal distribution, but when all samples except the maximum were evaluated, there was evidence that the data followed

both a normal and lognormal distribution. (A data point from a lognormal distribution can appear as an outlier when it is erroneously assumed that the data set is normally distributed.) However, for this particular data set, the removal of the outlier (0.398 mg/kg) did not significantly affect decision-making because all of the antimony concentrations were less than the state-specified risk-based decision level of 2.7 mg/kg. Furthermore, fortuitously, similar statistical results were obtained with and without the outlier. Although the maximum detected concentration was eliminated, the sample median and mean were not seriously affected, and the difference between maximum concentrations was less than an order of magnitude. However, under different circumstances (e.g., had the risk-based decision limit or the difference between the two highest values been larger), the comparisons between the site and background data sets could have been adversely affected (e.g., a "false positive" could have resulted). *Data points should never be removed from any data set (background or otherwise) solely on the basis of an outlier test unless an independent weight of evidence indicates that the data points are not representative of the underlying population of interest.*

I-5. Recommendations. If the data are normally distributed, Rosner's test is recommended when the sample size is greater than 25 and the extreme value test is recommended when the sample size is less than 25. If only one outlier is suspected, the discordance test may be substituted for either of these tests. If the data are not normally distributed, or if the data cannot be transformed so that the transformed data are normally distributed, the analyst should apply a nonparametric test, such as the fourth-spread test, or Walsh's test. A summary of this information is contained in Table I-3. Recommendations on selecting a statistical test for outliers are listed.

Table I-3.
Recommendations for Selecting a Statistical Test for Outliers

Sample Size	Test	Assumes Normality	Multiple Outliers
$n \leq 25$	Extreme Value Test	Yes	No/Yes
$n \leq 50$	Discordance Test	Yes	No
$n \geq 25$	Rosner's Test	Yes	Yes
$n \geq 50$	Walsh's Test	No	Yes
Any sample size	Fourth-Spread Test	No	Yes