**ARCHIVE:** **Archived because new guidance from U.S. EPA renders this guidance unnecessary. Refer to VA30007.05.001 for the updated document.**

**TITLE:** **Special Considerations in the Analysis of Environmental Data Sets**

**DATE EFFECTIVE:** January 2003

**HISTORY:** Update of VA30007.00.005 - Revision was necessary to reflect changes in rule citations that became effective on October 21, 2002. However, the archived TDC document remains accurate under the 1996 VAP rules.

**KEYWORDS:** sampling; field data; environmental data sets; lognormal distributions; skewed distributions; exploratory data analysis; outliers; non-parametric; 95% UCL

**RULE/ AUTHORITY:** OAC 3745-300-07(D)(1)(e) and 3745-300-07(D)(6)(c)

**QUESTION:** When the 95% upper confidence limit of the arithmetic mean (95% UCL) is calculated for certain data sets, the result can greatly exceed the maximum value in the original data set. What are some valid approaches to analyze problematic data sets?

**ANSWER:** In order to cover the wide variety of sampling conditions that are encountered during a VAP Phase II property assessment, a set of regulatory requirements have been promulgated to guide sampling activity. The VAP requirements for sampling environmental media are found in Paragraph (D)(1)(e) of Rule 3745-300-07 of the OAC.

This rule requires that sampling data must be representative of actual site conditions, and accounts for any spatial or temporal variation in the concentrations of chemicals which may be expected to occur. Sampling of all appropriate environmental media must be sufficient to fulfill the requirements set forth in OAC 3745-300-07(D)(2) through (D)(10). In particular, OAC 3745-300-07(D)(6)(c) requires that data sets be “comprised of a sufficient number and quality of samples as to derive a normal, log-normal, or [an]other applicable frequency distribution.” Alternately, for sites where sampling activities can be reliably biased to the location of the highest concentration (see OAC

3745-300-07(D)(6)(c)(ii)), then the volunteer may use the maximum concentration from a data set of at least three points.

Under highly variable field conditions there is no overall prescription for how many samples to collect; the number of samples collected is almost entirely situational. However, as long as the extent of contamination is well characterized and the sampling is conducted in accordance with the VAP rules, the number of samples is not, of itself, a crucial factor. Yet, for these same reasons, sampling for environmental data often can yield small data sets.

One of the fundamental assumptions of risk assessment is that any receptor will tend to “average” its exposure to a contaminant source and this averaging is related to the type and extent of its habitat and activities, which also affect the frequency of exposure. Therefore, the arithmetic mean of a set of measures of contaminant concentrations is commonly taken as an estimate of the exposure point concentration. In order to reflect the variance in the contamination, risk assessments typically use the 95% upper confidence limit of the arithmetic mean (95% UCL). The 95% UCL is an estimate of the value that, 95% of the time, the true mean will not exceed.

When the number of samples is low, and variation in data is high, the determination of a 95% UCL may yield a value substantially higher than the maximum concentration. Most importantly, these estimates are critical factors in the decision-making process leading up to whether or not a remedial action is necessary. This fact alone underscores the importance that the conclusions drawn from environmental data sets are objective, accurate, and reproducible.

**Issues in the analysis of environmental data sets**

Environmental data sets are often small in size. Data sets which have 20 or fewer data points challenge the ability of parametric confidence limits to accurately reflect an upper confidence limit. The low number of samples may not have a distribution, and even if a distribution does exist, it tends to yield poor estimates of the true mean and variance. In addition, these data sets can have non-normal distributions, which may also exhibit a considerable degree of skewness. These seemingly aberrant characteristics may result from samples for which an analyte was not detected or quantifiable (i.e., a non-detect). Environmental data sets commonly have data points which are spatially autocorrelated (i.e., non-independence of data points due to their signficant spatial relationships) due to topographic relationships,

an inappropriate sampling scale, and extreme values that may or may not be outliers.

Once a sampling plan has been executed, and the samples analyzed, the investigator may attempt to fit the data to a standard probability distribution. While it is difficult to exactly fit a standard probability distribution to field data, these models can often provide reasonable fits, which are evaluated by goodness-of-fit tests. A probability distribution (e.g., normal, lognormal, exponential) can streamline the modeling of an environmental data set in that each probability distribution provides a unique set of formulae for the mean, variance, and percentiles of the data set. However, the minor deviation of the field data distribution from a standard probability distribution can produce spurious results in subsequent calculations.

For example, when distribution-based 95% UCL methods such as that of Land's H (Land, 1975), or Student's t are applied to skewed data sets which are only approximately lognormal or normal, respectively, a common result is a wider confidence interval than that which would be determined from a less-skewed, true lognormal, or normal data set. This is in part because the H or t parameter in these applications is taken from a generalized, theoretical distribution function. The high variation contained in a specific environmental data set can force the point estimate out of bounds with respect to the actual range of the data. Therefore, it is important to be fully aware of the utility and limitations of the data set at hand. These issues can be addressed *a priori* by considering data quality standards to make sure that questions about the site can be answered, based on a sample size that gives sufficient resolution.

## Exploratory Data Analysis

Due to the wide variety of settings and chemicals of concern in the environment, each environmental data set has a different characteristic degree of variation. This is why it is important to consider the overall structure of the data set, and particularly the presence and location of outliers. An outlier is a data point that does not conform to the pattern or variance structure of the overall data distribution. For example, the presence of one outlier may indicate equipment malfunction, a manual error in recording a data point, or some other random source of error. However, two or more outliers can indicate that the level of spatial or temporal variation in the environment was higher than anticipated, and therefore not well-characterized by the sampling. Sometimes outliers may represent the

so-called “hotspot”. A hotspot is a location with extremely high concentrations of the chemical of concern, relative to its immediate surroundings. There is also the chance that a single data set with several outliers is actually composed of multiple, statistically-independent subsets of data.

An environmental data set which exhibits skewness may be modeled with the lognormal distribution, which provides some relief for these complications due to extreme values. Since the lognormal distribution can easily incorporate and otherwise model extreme values, there is a chance that statistical outliers will be included in the model.

Calculation of univariate statistics such as the sample mean, median, range, variance, and coefficient of variation will give the investigator an idea of the spread and skewness of data. Many statistical and spreadsheet software packages are capable of calculating a full range of univariate test statistics. A box-and-whisker plot will display this information graphically and give some preliminary indication of the presence and location of data points which may also be candidates for outlier analysis. Normal probability plots for both raw and log-transformed data will help to determine whether or not straightforward parametric approaches will be valid. In the event that the data distribution is not normal, the analyst should try various transformations (logarithmic, arcsin-square root, etc.) in an attempt to normalize the data distribution. Goodness-of-fit tests facilitate fitting field data to a wide variety of distributions (e.g., Weibull, Beta, Gamma), and can be used as another check in judging the fit of data to normal or lognormal distributions. The presence or absence of spatial structure or topography can both affect a particular measurement. The consideration of physical setting in the data validation process is enhanced by a complete set of field notes and sampling maps.

Several approaches to estimation of the 95% UCL of the mean are outlined in this document; it is assumed that the investigator uses exploratory data analysis (EDA) techniques to guide initial assumptions about the structure of the data set. This EDA will allow the analyst to maximize objectivity in the treatment of the data. If outliers are identified, there may be some basis for removing these points. Indeed, it may be the case that without the outliers, a new distribution may be more appropriate. The analyst should approach outlier analysis with extreme caution and fully document any alterations made to the data set. Whether or not to remove outliers is outside of the scope of this TDC; a wide variety of statistical texts are available for consultation on this matter.

## Calculation of 95% UCLs

### i. Parametric approaches

If the underlying distribution of data is normal, then the calculation of the 95% UCL is straightforward. In the case that the distribution is lognormal with some degree of skew, then the investigator should consider using the Land method (Land, 1975) or consult the USEPA guidance document *The Lognormal Distribution in Environmental Applications* (Singh et al., 1997).

OAC 3745-300-07(D)(6)(c)(i) directs the investigator to use peer-reviewed statistical methodology for normal or lognormal distributions, or as stated earlier in this paper, default to the maximum concentration value form a data set comprised of at least three points (see OAC 3745-300-07(D)(6)(c)(ii)). However, if the data set is distribution-free, then the investigator may want to explore nonparametric methods of analysis. The nonparametric methods are not presently widely used, yet address the problems typically encountered when using parametric methods to obtain 95% UCLs. These nonparametric approaches would be considered by VAP risk assessors to be an alternative to a 95% UCL, and acceptable for review as long as full documentation is provided.

### ii. A nonparametric method to calculate the estimated 95$^{th}$ percentile

Data sets which display neither normal nor lognormal distributions should be analyzed with nonparametric (also referred to as “distribution-free”) techniques. In general, nonparametric techniques rely on a ranking of the data set to determine percentiles. A non-parametric methodology adapted from Gilbert (1987) has been used in the calculation of a 95% UCL, and is appropriate for data sets with >20 points (L. Sirinek, personal communication, 2000). This data analysis procedure is best carried out on a computer spreadsheet program. A step-by-step procedure for the calculation of a nonparametric 95% UCL follows a data set (Table 1) which represents the concentration of a hypothetical chemical of concern in soil:

Table 1

| Rank | Conc. (mg/kg) | Percentile |
|---|---|---|
| 1 | 70 | 0.64 |
| 2 | 70 | 0.64 |
| 3 | 70 | 0.64 |
| 4 | 70 | 0.64 |
| 5 | 70 | 0.64 |
| 6 | 70 | 0.64 |
| 7 | 70 | 0.64 |
| 8 | 70 | 0.64 |
| 9 | 70 | 0.64 |
| 10 | 70 | 0.64 |
| 11 | 70 | 0.64 |
| 12 | 70 | 0.64 |
| 13 | 70 | 0.64 |
| 14 | 70 | 0.64 |
| 15 | 70 | 0.64 |
| 16 | 70 | 0.64 |
| 17 | 1400 | 0.68 |
| 18 | 1425 | 0.72 |
| 19 | 1510 | 0.76 |
| 20 | 1600 | 0.80 |
| 21 | 4500 | 0.84 |
| 22 | 4880 | 0.88 |
| 23 | 4985 | 0.92 |
| 24 | 7350 | 0.96 |
| 25 | 7500 | 1.00 |

1) Data are ranked in ascending order. For a data set with 25 points, the first rank is the lowest concentration (1), and the 25th rank is the highest concentration. Calculate the percentile for each data point by dividing its rank by 25, which is the total number of data points. In this procedure, it is customary to count non-detect data points. The "ties" are counted as separate observations and constitute a single percentile, as opposed to an averaged percentile. Non-detect data points are considered to be in the range of zero to the analytical lower-detection limit. The single point value estimate of each non-detect is defined as one-half of the analytical lower detection limit.

2) Calculate the arithmetic mean from the data set. This would be the sum of all twenty-five numbers, divided by 25. In this example, the arithmetic mean is 1450.

3) Determine the rank for the arithmetic mean by interpolation. This is done by locating the mean between the data points which lie above and below the calculated arithmetic mean. These data points "bracket" the arithmetic mean. In this example, the percentiles of ranks 18 (1425) and 19 (1510) are 0.72 and 0.76, respectively. Since 1450 is between 1425 and 1410, linear interpolation is used to determine the percentile of the arithmetic mean. A proportion can be set up to determine the percentile of the arithmetic mean:

$$1425(0.72)^{-1} = (1450)(x)^{-1}$$

Solve for x to get 0.73

4) The data point (u) whose rank represents the 95% confidence level on the percentile of the arithmetic mean is given by this equation:

$$u_{95\%\ UCL} = p(n+1) + Z_{1-\alpha}\,[np(1-p)]^{1/2}$$

Where p is the percentile of the arithmetic mean calculated in step 3, $Z_{1-\alpha}$ is the one-sided Z variate at an alpha level of 0.95, which is 1.645, and n is the number of samples in the data set. For this example, this calculation yields the rank 22.675.

5) By interpolation, the estimated 95% UCL concentration for the data set in this example is 4950 mg/kg.

This example illustrates how a non-parametric procedure can yield an upper-bound limit on the basis of the percentile and rank of the arithmetic mean.

**iii. Nonparametric method for estimating the confidence interval on the upper quartile of a distribution**

For particularly intractable data sets with less than 20, but more than 7 data points, another procedure is available for calculating a value that approximates the 90 to 95 percent confidence interval about the upper quartile of a continuous distribution (Conover 1980). The following guidance on how to conduct these calculations assumes some prior knowledge of probability distributions, and the operation of Crystal Ball® (Decisioneering Incorporated), or Excel® (Microsoft Corporation) software. This is accomplished by restricting the upper bound probability level of the confidence interval about the upper quartile:

$$\Pr(X^{(r)} \leq x_p \leq X^{(s)}) = 1 - \alpha \quad (1)$$

where p = 0.75 (i.e., the upper quartile) and the term (1 - α) determines the upper-bound (one-tail) confidence interval. The distribution of equation 1 is approximately binomial; the cumulative binomial distribution that is used in this determination can be either read off of Table A3 in Conover (1980), calculated with Crystal Ball®, or Excel® software. The parameters for Crystal Ball® users are as follows:

(1) Select a Binomial Distribution for an assumption cell.
(2) Enter 0.75 for p, which is the probability of the successful outcome for a given trial
(3) Click on Preferences in the Binomial Distribution window, then select Cumulative distribution.
(4) Place the mouse on the density function (i.e., the coloured bar as in a bar graph) above the integer of interest; the cumulative probability for Bin(n,p) will appear above the mouse pointer. Adjust significant digits to not less than 3 places.
(5) Write down the cumulative probability for each integer value up to and including the highest rank, which is n, the number of samples.

The same calculation can be made in Excel® with the function BINOMDIST (Number successes, Number trials (n), Probability of a successful outcome (p), Cumulative). The cumulative distribution is calculated for each trial, up to and including the total number of trials, n. The probability, p, is equal to 0.75, the upper quartile. The cumulative option is set to “TRUE”. The Excel® user needs to then calculate the cumulative probability for each trial (denoted $X_i$), $X_0$ to $X_n$.

To illustrate this procedure, an example was adapted from Conover (1980). Assume for the purposes of this example that we have a distribution of soil arsenic concentrations with n=15. The samples are independent and there is no spatial correlation among the datum. The ranked data is listed below in units of mg arsenic per kg dry soil:

Table 2

| Rank | Arsenic (mg/kg) |
|---|---|
| 1 | 16.4 |
| 2 | 32 |
| 3 | 53.5 |
| 4 | 56.8 |
| 5 | 71.6 |
| 6 | 98.6 |
| 7 | 103.4 |
| 8 | 148.7 |
| 9 | 185.8 |
| 10 | 274.1 |
| 11 | 274.8 |
| 12 | 286.5 |
| 13 | 305.6 |
| 14 | 329.4 |
| 15 | 859.0 |

The cumulative binomial probability distribution is calculated for each binomial “trial” up to n=15, with p=0.75; or, Bin(15, 0.75).

Table 3

| n | trial ($X_i$) | ∑Bin(15,0.75) |
|---|---|---|
| 15 | 0 | 0 |
| | 1 | 0 |
| | 2 | 0 |
| | 3 | 0 |
| | 4 | 0.0001 |
| | 5 | 0.0008 |
| | 6 | 0.0042 |
| | 7 | 0.0173 |
| | 8 | 0.0566 |
| | 9 | 0.1484 |
| | 10 | 0.3135 |
| | 11 | 0.5387 |
| | 12 | 0.7639 |
| | 13 | 0.9198 |
| | 14 | 0.9866 |
| | 15 | 1.0000 |

We will use the ranked data in Table 2 and the cumulative probabilities in Table 3 to balance the two sides of equation 1. Reading down the cumulative probability column in Table 3, the probability 0.0566 is selected as being the closest to 0.05, and is termed α. The "trial" associated with this lower probability limit is 8, which is designated "r" (as in $X^{(r)}$, the data point with rank r ). The probability closest to the upper probability limit of 0.95 is 0.9198, which corresponds to the 13th "trial", which is designated as "s" (as in $X^{(s)}$, the data point with rank s) . In order to account for rank rounding errors, both r and s are increased by 1. This adjustment yields the ranks 9 and 14, which are the lower and upper bounds of the (1 - α) confidence interval, respectively. The data corresponding to these ranks are looked up in Table 2 of this example and are (185.8, 329.4). Therefore, we can describe this confidence interval as:

"the interval from 185.8 to 329.4 mg/kg arsenic, inclusive, is a 94.34% confidence interval for the upper quartile."

Yet, the arithmetic mean of a distribution is commonly used to represent the exposure point concentration. Note that the arithmetic mean (i.e., 206.4 mg/kg) of the data used in the foregoing example is within the bounds of the confidence interval, yet is closer to the lower bound. Although the mean may not always fall on the upper quartile, numerous simulations with highly-skewed data sets of similar size (i.e., n=15), show that the arithmetic mean tends to fall within the $67^{th}$ to $80^{th}$ percentile range.

At the user's discretion, Equation 1 can also be used to describe the confidence interval for any $p^{th}$ quantile ($\therefore p \in [0,1]$) of interest. The previous example does this and uses interpolation to determine the percentile and rank of the arithmetic mean for data sets with more than 20 samples. The present example can be used to guide the use of percentiles other than the $75^{th}$ for the calculation of the confidence intervals for data sets with at least seven, and fewer than 20 data points. The user will then have to recalculate a new cumulative probability distribution, based on the percentile and rank of the unique arithmetic mean.

In any case, it is nevertheless important to consider the impact of hotspots on small data sets. The presence of hotspots can make even normally robust nonparametric procedures yield spurious results, too. Apparently, these situations would merit more sampling to better characterize the nature of the contamination.

**REFERENCES:**

Conover, W.J., 1980. Practical Nonparametric statistics, 2nd ed., Wiley, New York.

Gilbert, R.O, 1987. Statistical methods for environmental pollution monitoring. Van Nostrand Publishing, Detroit.

Land, C.E., 1975. Tables of confidence limits for linear function of the normal mean and variance, in Selected Tables in Mathematical Statistics, Vol. III, American Mathematical Society, Providence, RI, pp. 385-419.

Singh, A.K., Singh, A., Englehardt, M., 1997. USEPA Technology Support Center Issue: The lognormal distribution in environmental applications. USEPA, ORD-OSWER, EPA/600/R-97/006.

Sirinek, L., 2000. Personal communication as Risk Assessment Coordinator, Remedial Response Program, Ohio EPA Division of Emergency and Remedial Response, Central Office, Columbus, OH.

U.S.EPA, 1992. Supplemental guidance to risk assessment guidance for Superfund (RAGS): calculating the concentration term. USEPA-OSWER. PB92-963373. OERR Hazardous Site Evaluation Division. Intermittent Bulletin Vol. 1, No.1.

**SUMMARY:** Under VAP rules, the Volunteer is required to adequately sample all environmental media which may contain chemicals of concern. What constitutes adequate sampling may differ with each situation, the nature of the contamination, and the physical setting of the property to be assessed. Yet, adequate sampling may in some cases yield small data sets with less than 20 points. An evaluation that includes exploratory data analysis and proper application of parametric or non-parametric techniques is recommended to determine the true tendency of the data set, in terms of confidence intervals and their limits.

**OHIO EPA CONTACT:** For any questions concerning this issue, please contact the VAP central office at (614) 644-2924.