ATSDR Division of Community Health Investigations
Exposure Point Concentration Guidance for Discrete Sampling

Citation:
[ATSDR] Agency for Toxic Substances and Disease Registry. 2023. Exposure Point Concentration Guidance for Discrete Sampling. Atlanta, GA: U.S. Department of Health and Human Services, Public Health Service.

List of Exhibits

List of Abbreviations

ADS	Associate Director for Science
ATSDR	Agency for Toxic Substances and Disease Registry
DCHI	Division of Community Health Investigations
EDG	Exposure Dose Guidance
EPA	U.S. Environmental Protection Agency
EPC	exposure point concentration
GOF	goodness of fit
LSE	levels of significant exposure
MLE	maximum likelihood estimation
MRL	minimal risk level
PAH	polycyclic aromatic hydrocarbon
PHAGM	Public Health Assessment Guidance Manual
PHAST	Public Health Assessment Site Tool

ppb	parts per billion
PPCC	Probability Plot Correlation Coefficient
ROS	regression on order statistics
SAS	Statistical Analysis System
UCL	upper confidence limit
95UCL	95 percent upper confidence limit of the arithmetic mean

Note

This guidance mentions several software applications and programming languages by name, including ProUCL, R, Microsoft Excel, Minitab, SAS, and SPSS. Use of these trade names is for identification purposes only and does not constitute an endorsement of their use.

1.0 INTRODUCTION

Environmental samples provide insights on contamination levels at sites of interest. But those insights are never complete: health assessors will never have sampling data for every single fish in a lake, from every volume of air that people breathe, or from every drop of water that people drink. Health assessors instead must use available environmental sampling data to estimate human exposures to contamination levels that are not fully known. This document presents ATSDR's guidance for estimating exposure point concentrations (EPCs) for use in the public health assessment process.

ATSDR has developed an EPC Tool that calculates EPCs using the procedures identified in this guidance. The ATSDR EPC Tool allows users to upload data and calculate EPCs using a standardized process.

ATSDR has developed an EPC Tool that calculates EPCs using the procedures identified in this guidance. The ATSDR EPC Tool allows users to upload data and calculate EPCs. Once the EPCs are calculated, the *Results Summary* screen provides a table summarizing the EPCs and a button for exporting the EPCs into a Public Health Assessment Tool (PHAST) import template. The *Results Details* page provides supplementary statistics and figures for each calculated EPC, and the *Resources* page contains supporting documents and files. For questions or assistance using the ATSDR EPC Tool, please contact the ATSDR technical team at PHAST@cdc.gov and include the phrase "ATSDR EPC Tool" in the subject line.

The ATSDR EPC Tool was developed to make the process of calculating EPCs more consistent and reliable. However, health assessors may also choose to use the procedures outlined in this document along with other statistical techniques to develop EPCs using other tools such as ProUCL and R. Developing EPCs using non-ATSDR tools should be coordinated with the Associate Director of Science (ADS) in the health assessor's ATSDR office and an ATSDR subject matter expert in statistics and data analysis.

This guidance is based on robust statistical procedures that specifically consider uncertainties associated with environmental sampling data to generate reasonable, health-protective EPCs. By applying this guidance, health assessors should be confident that their estimated EPCs do not understate actual exposures, despite the statistical uncertainties associated with environmental sampling data.

This guidance allows health assessors:

- To determine the most appropriate EPC statistic to use when evaluating acute, intermediate, and chronic exposure durations (see Section 2.0)

- To calculate 95 percent upper confidence limits of the arithmetic mean (95UCL) when this statistic is warranted using either ProUCL or R (see Section 3.0)

In many cases, the 95UCL will be the appropriate statistic for the EPC. This value equals or exceeds an exposure unit's (EU) actual arithmetic mean of site concentrations 95 percent of the time and is therefore considered a health-protective estimate of the actual mean. Calculating 95UCLs is not a new concept, and many textbooks, articles, and other publications present equations and procedures for doing so. However, health assessors should rely on this guidance, which was developed to address

unique aspects of calculating 95UCLs specifically for environmental data sets and for use in ATSDR's public health assessment process.

This guidance prescribes specific 95UCL computational approaches that health assessors are expected to follow whenever calculating EPCs, and the approaches were developed to apply to an extremely broad range of site-specific scenarios (e.g., small and large data sets, data sets with and without non-detects, data sets with highly variable measurements). Computational approaches other than those prescribed in this guidance should not be used, unless first approved by an ADS group.

1.1 When to Use This Guidance

During the public health assessment process, health assessors perform many activities, including developing a site conceptual model, evaluating exposure pathways, identifying exposure units, compiling and reviewing environmental data, and screening those data against health-based comparison values. ATSDR has developed other guidance to assist health assessors with these and many other steps in the public health assessment process. For example, the *Public Health Assessment Guidance Manual (PHAGM)* presents guidance on developing conceptual site models and evaluating exposure pathways (ATSDR 2022a); and ATSDR's *Exposure Unit Guidance* outlines key considerations for defining exposure units for public health assessment purposes (ATSDR 2020). It is important that health assessors carefully define exposure units before applying this guidance.

In cases where contaminants in completed or potential exposure pathways are selected for health evaluations (e.g., when maximum concentrations in environmental samples exceed comparison values), the health assessor must perform exposure dose calculations for the environmental media and exposure pathways requiring further evaluation; and the health assessor needs to know what EPCs to use in these calculations. The calculated exposure doses then form the basis for the health evaluations, and different exposure doses should be calculated for the various exposure units that the health assessor has identified for a site. Calculated exposure doses are then compared to established toxicity values for non-cancer (e.g., ATSDR Minimal Risk Levels, EPA Reference Doses) and cancer (e.g., oral slope factors) health effects to determine whether harmful health effects might be possible. This guidance walks health assessors through the process of determining what statistics to use for EPCs and how to calculate them. This guidance supersedes all information in *PHAGM* relevant to determining what EPC to use in exposure dose calculations for health effects evaluations.

This guidance applies to determining EPCs for environmental data generated only by "discrete sampling." For purposes of this document, a discrete sample refers to an individual environmental sample from a given point and time that is independent of other samples. For

What is a "sample"?

Though a seemingly straightforward term, the word "sample" has different connotations among scientific disciplines:

- For statisticians, a "data sample" typically refers to all observations from a data set. For instance, if 100 out of 125 people submit a survey, the data sample in this case would be the 100 completed surveys.
- For environmental scientists, an "environmental sample" typically refers to a physical quantity of an environmental medium—soil, water, air, and food items—that is collected for measurement. For instance, six ounces of tap water collected in a vial for laboratory analysis is one environmental sample. Unless otherwise noted, all references to "sample" in this guidance describe environmental samples.

instance, at a site with contaminated soils, investigators might collect dozens of surface samples from grid locations throughout an area of interest and have a laboratory analyze each sample individually. Additionally, at a lake with suspected contamination in fish tissue, investigators might catch one fish from several different areas and have a laboratory analyze each fish individually. Both are examples of discrete sampling programs for which this guidance should be used to determine EPCs.

The computational methods to estimate EPCs prescribed in this guidance apply only to "discrete" samples collected in a random manner. However, in some cases, sampling will be done in a non-random manner such that a greater density of samples is collected from more highly contaminated areas or during periods when contamination is expected to be highest. For example, an air quality study might collect samples every 12 days during seasons when air pollution levels tend to be lowest and every 3 days during seasons that have historically found elevated air pollution levels. In these cases, the universe of sampling results available to the health assessor presents a biased account of the contamination levels within an exposure unit. As another example, a field investigation might collect soil samples only in suspected "hot spots" or areas known to have previous spills, and not throughout an entire exposure unit. In these cases, the statistical methods outlined in this guidance for random sampling are not appropriate. Health assessors can either (1) redefine the exposure unit to better align with a random sampling method and then apply the methods of calculating EPCs prescribed in this guidance or (2) consider the alternative statistical approaches briefly described for non-random or judgmental samples in Section 3.8 of this guidance. Health assessors should consult with their ADS group on how to proceed with such data if they have any questions or concerns.

The following section identifies other types of sampling program designs for which this guidance should not be used.

1.2 Topics Not Covered by This Guidance

While this guidance is expected to apply to a wide range of site-specific scenarios, it does not apply to all topics. As shown in the top half of Figure 1 and further described here, this guidance does not apply to:

- *Asbestos or lead.* This guidance should not be used for determining EPCs of asbestos or lead. In the case of asbestos, health assessors should consult with their ADS group or with an ATSDR subject matter expert about preferred EPC approaches. In the case of lead, the arithmetic mean should be used as the EPC in lead models (e.g., EPA's Integrated Exposure Uptake Biokinetic Model for Lead in Children model and EPA's Adult Lead Model) [EPA, 2016].

- *Radiological contaminants.* This guidance should generally not be used for radiological contaminants, except those with a minimal risk level (MRL).

- *Dioxins or polycyclic aromatic hydrocarbons (PAHs).* ATSDR has developed separate Exposure Dose Guidance (EDG) that applies specifically to dioxins (ATSDR 2022b) and PAHs (ATSDR 2022c*)*. These documents explain how health assessors should process environmental sampling data for those chemicals before applying this EPC guidance. Health assessors who need to address PAHs before the chemical specific EDG is available should consult with their ADS group about preferred approaches.

- *Data from "non-discrete sampling" programs.* Many different sampling strategies are used in environmental data collection. This guidance should not be used on all types of sampling data and the following list identifies two examples for which this guidance does not apply. ATSDR has developed separate EDG that applies to data collected using these two methods (ATSDR 2022d).

    - Composite sampling. Composite sampling refers to combining multiple discrete environmental samples into composite samples that are then sent to the laboratory for analysis. The approaches presented in this guidance generally do not apply when working with composite samples.

    - Incremental sampling. Incremental sampling methodology—for which MULTI INCREMENT® is a specific example[1]—is a technique designed to characterize the overall contamination level within an area of interest, thus minimizing the variability typically found in discrete or composite sampling efforts. This technique relies on careful statistical design for both sampling and laboratory analysis and is most commonly applied to soil sampling efforts. Health assessors should not apply this guidance to any data that are reported as being collected by "multi-incremental" or "incremental" sampling strategies.

- *Dependent data.* Some sampling programs are designed or implemented in a manner that leads to dependent data: instances where the results of one sample at least partially predicts the results of another sample. In these cases, data are described as having some degree of auto-correlation, whether in space or time; and this concept may best be described by examples. For instance, for an exposure unit defined as a stretch of river with known surface water contamination, surface water monitoring locations in very close proximity would be expected to have highly correlated data that do not provide independent measures of contamination. Similarly, for an exposure unit defined as a playing field with soil contamination of a persistent pollutant, repeated sampling of the same locations over time (i.e., time series) would also be expected to have highly correlated—and not independent—measures of contamination. In cases such as these, health assessors should evaluate whether observations were collected close enough in time or space to be serially correlated. If so, this guidance should not be applied since the 95UCL computational approaches (see Section 3.0) assume that the environmental sampling data are independent. Other statistical approaches (e.g., weighting samples collected through time so that they jointly represent a specific unit of time) may be more appropriate than what is prescribed in this guidance. Health assessors should consult with their ADS group for guidance when evaluating these types of dependent data sets.

1.3 Resources for Further Information

This guidance was developed to make EPC determinations—and 95UCL calculations—a straightforward process. Some health assessors may want to access additional resources for further information on the advanced and highly technical statistical topics. In those cases, health assessors should:

[1] *MULTI INCREMENT®* is a registered trademark of EnviroStat, Inc.

- develop training modules on calculating exposure point concentrations – we anticipate these will be available to health assessors after this guidance is finalized. Additionally, until training is available, health assessors should seek help from courses on environmental statistics which have been developed by several private vendors.

- Subscribe to U.S. EPA's TechDirect Clean-Up Information (CLU-IN) email list (https://clu-in.org/techdirect/). CLU-IN distributes monthly messages to subscribers summarizing new publications and training webinars related to site remediation and site assessment. Some training may be of interest to health assessors, and you may subscribe to the email list here: https://clu-in.org/newsletters/#s.

- Consult with their ADS group for other resources. ATSDR recommends this option to ensure that all health assessors consistently rely on a common set of documents when determining EPCs, rather than having health assessors individually seeking input from different (and perhaps inappropriate) sources.

1.4 How to Use This Guidance

Health assessors will find all EPC guidance in this document's text and flow charts (i.e., Figure 1 to Figure 6), Appendix A includes a glossary of key terms, and Appendixes B through F contain sample calculations. Further information is provided in text boxes, as follows:

Key Point
Blue text boxes concisely summarize major elements of this EPC guidance.

Additional Information
Yellow text boxes provide scientific background information on issues related to EPC calculation.

2.0 GUIDANCE RECOMMENDATION 1: SELECTING THE APPROPRIATE EPC STATISTIC

Key Point: What statistic to use for EPCs?
Health assessors should use the flow chart in Figure 1 to determine the most appropriate statistic to use for EPCs for acute, intermediate, and chronic exposure durations. They should also read this section for additional considerations on the recommended statistics.

For every contaminant requiring a health effects evaluation, health assessors must use environmental data, when available,[2] to select or calculate EPCs for each exposure unit. In many cases, multiple EPCs are needed for a given contaminant and exposure unit, because contamination levels may vary over an exposure unit's history (e.g., before and after remediation) and might differ across acute (0-14 days), intermediate (15-364 days), and chronic (365 or more days) exposure durations. Specific examples of how EPCs vary with exposure duration are provided later in this section.

Environmental sampling reports use a wide range of statistics when summarizing results. These may include arithmetic means, geometric means, medians, upper confidence limits, percentiles, maximum

[2] When no sampling data are available to determine EPCs for exposure units, health assessors should note these data gaps and consider whether health effects evaluations can be based on other information sources (e.g., modeling results, health studies) and whether to recommend sampling to fill gaps.

values, and others. This section identifies the statistics health assessors should use as EPCs in their health evaluations. Note that if the exposure unit does not adequately reflect activity patterns that are expected to occur, then the exposure unit should be redefined to match what is expected to occur and evaluated for the appropriate time intervals.

2.1 Acute Exposures

When evaluating acute exposure durations, health assessors should use either the 95UCL of the arithmetic mean or the maximum detected concentration from an exposure unit's valid environmental samples. As shown in Figure 1, the decision of which statistic to use should be based on careful consideration of applicable toxicity data (e.g., acute Minimal Risk Levels [MRLs]), temporal variations in environmental contamination and exposure patterns, and the number of samples. Health assessors must use their judgement when deciding which statistic best represents exposure conditions and should contact their ADS group for additional guidance if there is uncertainty in deciding the appropriate EPC. General guidelines follow for assisting health assessors with their determinations.

1. *Applicable Toxicity Data:* Health assessors should examine substance-specific toxicity information to determine the appropriate acute exposure EPC statistic. The critical consideration is the exposure duration for the human or animal study that forms the basis of ATSDR's acute MRLs. Health assessors can find the MRL exposure duration by reviewing Appendix A of the corresponding chemical's Toxicological Profile. The exposure duration also is listed in the Levels of Significant Exposure (LSE) tables in the main text. Once the exposure duration is located, health assessors should consider the following:

   Study Duration. When a chemical's acute MRL is based on a toxicity study that involved exposures lasting at least several days (four or more days), health assessors should generally use the 95UCL of the arithmetic mean as the EPC for acute exposures (though some exceptions are presented later in this section). The 95UCL is preferred because it is unlikely that individuals will only and repeatedly be exposed to an exposure unit's maximum concentration for a week or two. A more realistic scenario is that individuals will be exposed to a range of contaminant levels throughout the exposure unit, for which the 95UCL is a more appropriate statistic. For example, the acute ingestion MRL for chlordane is based on effects observed in animals who were exposed to chlordane for seven days. Given this basis for the MRL, the 95UCL of the arithmetic mean should be used when evaluating acute ingestion exposures to chlordane if other criteria are met (e.g., sufficient sample size and others described below).

   However, for those chemicals where the acute MRL is based on shorter exposures (one to three days of exposure), it is reasonable to assume that someone might be exposed to the maximum concentration for this duration. In this situation, the maximum concentration should often be used as the EPC to evaluate acute exposures.

   The main consideration when evaluating study duration is whether the acute EPC for a given contaminant matches the window of exposure used in the applicable toxicity study. The

criteria described above should be followed as general guidelines and health assessors should use their best professional judgement when evaluating the underlying toxicity data.

Inhalation Studies Not Adjusted to a Daily Exposure. Some chemicals exist where the inhalation MRL is based on an exposure concentration that is not adjusted to a daily concentration. Most of these chemicals cause irritant or point of contact effects. For this reason, the study concentration is more important than the product of the concentration and the duration of exposure in determining harmful effects. Examples of chemicals for which acute MRLs are based on these unadjusted exposure concentrations include sulfur dioxide, chloroform, formaldehyde, ammonia, and hydrogen sulfide. For those chemicals where the exposure duration was not adjusted to a daily exposure, it is important that an EPC match the window of exposure from the study used as the basis for the MRL.

For example, you would not take a series of 10-minute sulfur dioxide concentrations and convert those concentrations to a 24-hour concentration. That approach could average away important peaks that might cause harmful effects. Similarly, if you only have 24-hour sulfur dioxide data and you use that as your EPC, the critical window of exposure used to derive the MRL (10-minute duration of exercising asthmatics) will not match your EPC and possibly result in an incorrect health call.

Developmental Effects. In addition, depending on the exposure window of concern, health assessors may want to use the maximum value as their EPC for acute exposures to chemicals where developmental effects might result from very brief *in utero* exposures, particularly if pregnant women could be exposed for several days to the maximum concentration.

2. *Temporal Variations in Environmental Contamination:* When an exposure unit has multiple samples collected from the same location periodically over a long-time frame, then use of the maximum concentration as the EPC is generally preferred. For example, if an indoor air sample is collected at a residence once a season over a 3-year time frame to evaluate vapor intrusion issues, the central tendency estimate of these data (whether calculated as the 95UCL or some other statistic) represents an EPC estimate over durations longer than the acute exposure window (<14 days). In these cases, acute exposures are best evaluated with the maximum concentration, because it is reasonable to assume that a resident could have been exposed to the maximum concentration for a duration of up to 2 weeks.

3. *Temporal Variations in Human Behavior:* Health assessors should also consider specific human activity patterns for the scenarios they are evaluating, as these can affect the decision of which statistic to use for acute EPC evaluations. One example is consumption of food, for which residents may be exposed over the entire acute exposure duration (<14 days) to contamination in a single animal collected for human consumption—an issue that is particularly common for large game (e.g., deer). In this instance, the maximum concentration is the more appropriate EPC for acute exposures. Another example is the soil pica scenario, where maximum concentrations should be used because a child with pica

behavior could ingest soil from the most highly contaminated part of a yard. See the yellow box below detailing how to estimate exposures for the soil-pica scenario.

Additional Information: EPCs for the Soil-Pica Scenario

Under this scenario, health assessors consider ingestion of unusually high amounts of soil, whether due to intentionally ingesting soil or unintentional exposure due to excessive mouthing behavior. Health assessors should use the maximum concentration as the EPC. If the maximum concentration is a concern, health assessors should enter site-specific parameters in ATSDR's Public Health Assessment Site Tool (PHAST) to determine whether a single pica event is a health concern. Additional details are provided in the soil/sediment EDG.

4. *Influence of Atypical Conditions:* If the maximum concentration is selected as the EPC, health assessors should ensure that the maximum concentration in a sampling report represents environmental conditions for the exposure unit. This is a concern when uncharacteristic events or sampling accuracy may influence maximum concentrations. Health assessors should confirm that the maximum concentration is not influenced by an exceptional event (e.g., a nearby wildfire burning on the day an ambient air sample was collected), an analytical artifact[3] (e.g., a false positive detection for methylene chloride or other common laboratory chemicals), or even a reporting error (e.g., data presented with incorrect units)—all of which have been observed in previous ATSDR public health assessments. If a maximum concentration at a site represents an unusual condition that is not likely to be representative of a reasonable maximum exposure, choose the highest concentration that is representative of a reasonable maximum exposure as the EPC.

5. *Statistical Considerations When Using Maximum Concentrations:* In some cases, especially studies involving relatively few samples, the maximum concentration among environmental data is a poor indicator of the upper end of the distribution of actual contamination levels. For instance, if only three surface soil samples are available for an exposure unit the size of a football field, it is highly unlikely that the maximum concentration from the three samples falls above the 90th percentile of the distribution of actual contamination levels. In cases like this, use of the maximum concentration as the EPC for acute exposures could understate actual exposures by a considerable margin. When exposure units have large enough data sets such that health assessors are highly confident that maximum concentrations fall within the higher end of the distribution of actual contamination levels, health assessors may not need to recommend additional sampling for higher confidence in their findings. On the other hand, when smaller sample sets prevent health assessors from having the same level of confidence, they should still use the maximum concentration as the EPC when evaluating acute exposures (where warranted), but they should acknowledge the data limitations, describe the uncertainty in their findings, and should consider recommending additional sampling.

[3] Data quality narratives in sampling reports typically identify analytical artifacts, whether due to blank contamination or other factors, but this is not always the case—particularly for older studies.

6. *Statistical Considerations When Using 95UCLs:* The 95UCL may not adequately characterize contamination levels throughout an exposure unit when the statistic is based on only a few observations. As described in Section 2.2 under the second bulleted item, 95UCLs should not be used as EPCs for exposure units with fewer than 8 samples. In such cases, health assessors should use the maximum detected concentration as the EPC, acknowledge the data limitations, describe the uncertainty in their findings, and should consider recommending additional sampling. Refer to Section 2.2 for further details and considerations when working with a small dataset.

7. *Magnitude of Acute EPC with Respect to Health Guidelines.* Another important consideration in the acute evaluation is the margin separating the acute EPC and the corresponding health guideline. If a health assessor is not confident that the maximum concentration represents the higher end of the contamination distribution in the exposure unit (e.g., when only one or two samples are available) and if the calculated exposure dose from a maximum concentration is only marginally lower than health guidance values (e.g., acute MRLs), then a strong case can be made for recommending additional sampling—especially if community members have reported health effects consistent with those documented in the literature for acute exposures to the chemical being evaluated. Conversely, when calculated exposure doses from a maximum concentration are considerably lower than health guidance values, then additional sampling may not be necessary, even if the original data set is limited. However, as with all limited datasets, health assessors should be careful to make sure the maximum value may reasonably represent the true maximum value to evaluate acute exposures. Health assessors should consult with their ADS group before deciding if additional sampling should be recommended to increase confidence in EPCs used for acute exposure evaluations.

Overall, the previous discussion outlines many different factors that health assessors must consider when determining if the maximum concentration or the 95UCL is the most appropriate EPC statistic for acute exposure evaluations. One of the most critical factors is to make sure that the study duration is considered when deciding whether to use the maximum concentration or 95UCL as the EPC. Health assessors should carefully weigh these factors before deciding which statistic is most appropriate for their site-specific exposure scenario and environmental data set. Health assessors should defend their choice of the 95UCL or maximum concentration from an exposure unit's valid environmental samples as the EPC for acute exposure durations. ADS groups should be consulted if health assessors have questions regarding their selection.

2.2 Intermediate or Chronic Exposures

ATSDR's EPC guidance is based on the philosophy of erring on the side of caution and recommends approaches that minimize the possibility of health assessors underestimating actual exposures. In general, as Figure 1 shows, the *major factor* to consider when determining the most appropriate EPC for intermediate and chronic exposure durations is the number of valid environmental samples from the exposure unit (see Section 3.1 for additional factors to consider). Further details on the two scenarios follow:

Key Point: EPCs for intermediate or chronic exposures

Use the following statistics as EPCs based on the number of valid samples:

- *95UCL concentrations* for exposure units with 8 or more samples (see Section 3.1 for additional factors to consider when calculating a 95UCL for exposure units with 8 or more samples)
- *Maximum concentrations* for exposure units with <8 samples
- *Arithmetic mean* and *geometric mean* concentrations should not be used as EPCs for intermediate or chronic exposures

However, other factors may need to be considered when determining whether to use the 95UCL or maximum and health assessors should therefore use their best professional judgement when determining EPCs for intermediate/chronic durations. For example, health assessors may want to err on the side of caution and use a maximum value if sampling data are not available to match the exposure duration used to derive the health guideline. For example, for chemicals like trichloroethylene, where the exposure window of concern for developmental effects may be as low as three weeks, health assessors should consider using the maximum value if enough data are not available to determine an EPC that matches the exposure window of concern. As with the acute scenario above, health assessors should carefully weigh these factors before deciding which statistic is most appropriate for their site-specific intermediate/chronic exposure scenario and environmental data set. Health assessors should provide scientific justification for their choice of the 95UCL or maximum concentration from an exposure unit's valid environmental samples as the EPC for intermediate/chronic exposure durations. ADS groups should be consulted if health assessors have questions regarding their selection.

- *Exposure units with 8 or more samples.* Health assessors should use the 95UCL as the EPC, because this value offers a health-protective estimate of the actual mean concentration of contamination levels in the exposure unit (see "Why the 95UCL?" text box below). Use of the 95UCL greatly reduces the likelihood of health assessors failing to identify a public health hazard from environmental data when a hazard actually exists. Section 3.0 presents step-by-step procedures for calculating 95UCLs for many different site-specific scenarios and exceptions to the general rule of calculating a 95UCL for exposure units with 8 or more samples.

- *Exposure units with <8 samples.* When too few samples are available to adequately characterize contamination levels throughout an exposure unit, health assessors cannot quantify exposures with confidence. Based on a statistical review of typical environmental data sets, ATSDR has determined that 95UCLs should not be used as EPCs for exposure units with fewer than 8 environmental samples. In these cases, health assessors should use their professional judgement to determine whether the maximum detected concentration is appropriate to use as the EPC, acknowledge all related limitations of using this approach, and consider recommending additional sampling. This guidance cannot describe the full range of site-specific scenarios that health assessors may encounter, but the following examples illustrate key issues that health assessors should consider when determining whether it is appropriate to use the maximum detected concentration to evaluate intermediate or chronic exposures.

  One important consideration is the representativeness of the samples to the exposure unit. Health assessors should only use the maximum as the EPC when they are highly confident that the maximum concentration falls within the higher end of the distribution of actual

contamination levels in the exposure unit and therefore adequately captures "worst-case" exposures. Health assessors should carefully review their data when making this determination and consult with their ADS group if there is any uncertainty. As an example, consider two exposure units where seven samples of metals contamination in vegetables are available: (1) a 20-square foot residential gardening plot and (2) a large community garden that is the size of a typical baseball field. For the small gardening plot, the health assessor should have much greater confidence that use of the maximum concentration is a reasonable basis for exposure assessment—and possibly without the need for additional sampling data. For the community garden, on the other hand, the health assessor would not have the same level of confidence that the maximum falls within the higher end of the distribution of actual contamination levels. In this scenario, the health assessor would need to carefully review the sampling data and site history to ensure that the maximum concentration reflects "worst-case" exposures across the entire community garden, include strong caveats about the small sample size in the limitations section of their health assessment, and possibly recommend additional sampling to ensure that exposures have not been underestimated.

The health assessor should also consider the margin between the EPC (i.e., the maximum concentration, in this case) and the corresponding health guideline value (e.g., MRLs). In the case of the gardening plot from the previous example, if the maximum concentration is more than an order of magnitude below the intermediate or chronic MRL, then the health assessor may reach health conclusions and decide that no further sampling is warranted. However, if the maximum concentration is only marginally below the chronic MRL, a stronger case can be made for recommending additional sampling—especially if community members report health effects consistent with those known to occur for the chemical of interest.

In general, health assessors should use the maximum concentration as the EPC when evaluating intermediate or chronic exposures for exposure units with fewer than 8 samples; and they should acknowledge limitations of this approach and consider recommending additional sampling. As the previous paragraphs suggest though, some professional judgement is required to confirm that the maximum concentration from the data set does not underestimate "worst-case" exposures in the exposure unit. Site-specific conditions ultimately determine the appropriateness of using the maximum, the significance of related limitations, and the importance of additional sampling. Health assessors are encouraged to consult with their ADS group when deciding how best to present results for exposure units with limited number of samples. In some cases, ADS groups may recommend use of advanced statistical methods (e.g., Bayesian approaches) to make inferences for exposure units with limited sample size or additional sampling.

Additional Information: Why the 95UCL for intermediate or chronic exposures?

The average contamination level throughout an exposure unit is the best metric for characterizing exposures that occur over longer durations (>14 days). However, the actual or true average concentration throughout an exposure unit is rarely known because discrete sampling programs only provide an estimate of the actual average concentration. For example, the arithmetic mean of seven fish tissue samples collected from a large lake only approximates the true average contamination level across all of the lake's fish. Average concentrations from discrete sampling programs may understate or overstate actual average contamination levels—sometimes by considerable margins. The inherent concern with discrete sampling programs is uncertainty associated with characterizing the actual average concentrations in exposure units.

Statistical confidence limits are the classical tool for addressing uncertainties when estimating the average of a distribution from a discrete number of samples. For a collection of samples, the 95UCL is defined as a value that equals or exceeds the actual average of a distribution 95 percent of the time, and the 95UCL has a very low likelihood (5 percent probability) of understating the actual average of a distribution. In the context of environmental data sets and the public health assessment process, using the 95UCL as the EPC is a health-protective estimate of the actual average concentration within an exposure unit, with little chance of understating exposures. ATSDR therefore recommends using the 95UCL as the EPC for intermediate and chronic exposure durations to ensure that important public health decisions are based on reasonable and health-protective assumptions. Health assessors should not confuse 95th percentiles of sampling data (i.e., point values for which there is a 5 percent chance of seeing a larger value) with 95UCLs. These are two different statistics; 95th percentiles refer to single values while 95UCLs refer to mean values.

Applied to environmental sampling, the 95UCL calculated from environmental sampling data provides an upper-bound estimate of an exposure unit's actual or true average contamination levels. However, the ability of various statistical methods to achieve this depends on unique characteristics of the data (e.g., the number of samples, data distribution, variability of the data, percentage of non-detect values), some of which may not always be accounted for. The magnitude of the 95UCL is affected most by the number of environmental samples and the variability in measured concentrations:

- *Number of samples.* For most environmental sampling applications, the 95UCL will decrease and approach the true or actual average contamination level as the number of samples increase. More specifically, the 95UCL will approach the true average at a rate proportional to the reciprocal of the square root of the sample size. This is to be expected because an increasing number of samples in an exposure unit will lead to greater confidence that the sampling results characterize the distribution of contamination levels. The figure on the next page provides a quantitative example to illustrate this concept. The arithmetic mean and 95th percentile of sampling data should remain about the same with increasing sample size.

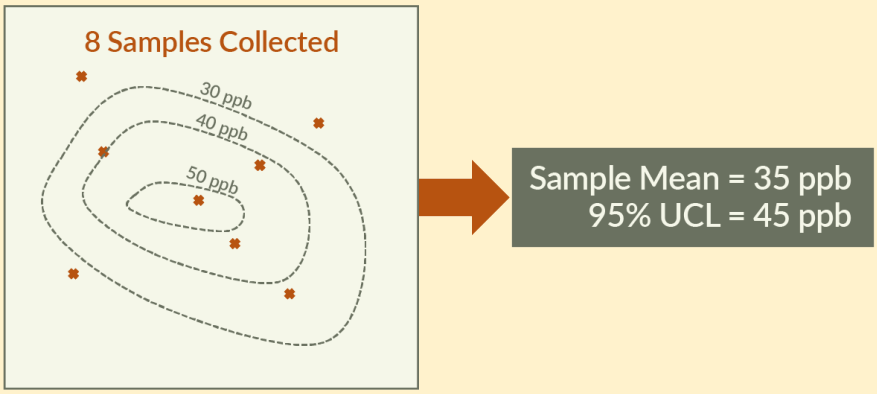
- *Variability.* 95UCLs also depend on data variability. For an exposure unit with highly uniform contamination levels, only a small number of samples may be needed to characterize exposures with confidence. Conversely, an exposure unit with extremely variable contamination levels will require many more samples to have the same level of confidence in characterizing exposures. All other factors considered equal, 95UCLs will increase with variability in contamination levels.

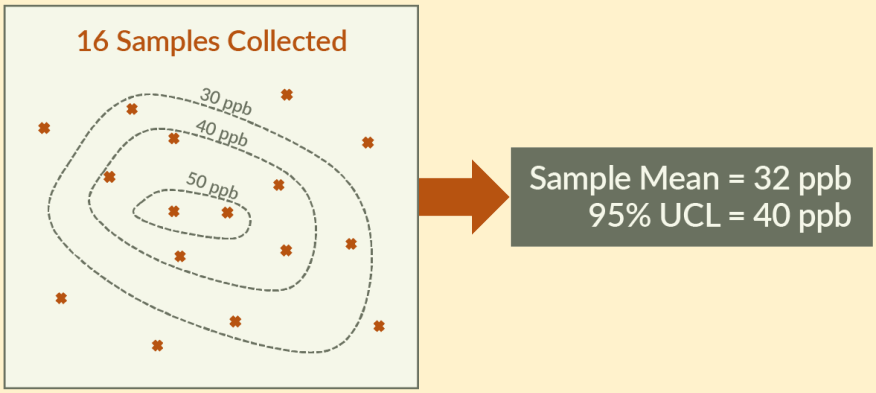
This text box provides a very general overview of 95UCLs, how they apply to environmental sampling data, and the factors that affect their magnitude; and Section 3.0 presents step-by-step guidance on how health assessors should calculate 95UCLs using different software programs and programming languages. Health assessors interested in learning more about the underlying statistical concepts and how they apply to the EPC calculation for the health assessment process should consult with their ADS group for further information.

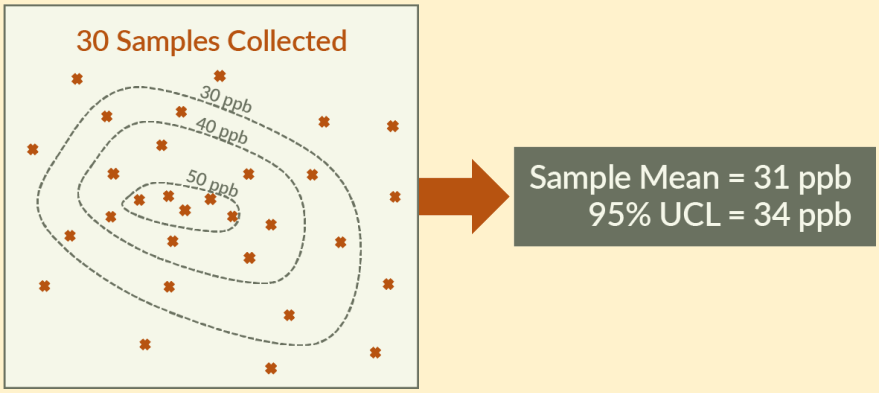
Additional Information: Why the 95UCL for intermediate or chronic exposures? (Continued)

The following illustrations show how 95UCLs can vary with number of samples in environmental applications. Each scenario shown below is the same square-shaped exposure unit, and the contours denote actual surface soil contamination levels. The illustrations show three hypothetical sampling strategies applied to the same exposure unit—the top illustration has 8 samples, the middle illustration has 16 samples, and the bottom illustration has 30 samples. Assume in this case that the exposure unit's actual mean contamination level is 30 ppb.

For the first scenario (8 samples), the data provide least confidence in characterizing the actual mean concentration (30 ppb), and this is reflected by the higher 95UCL. As the number of samples increases in the second and third scenarios, the 95UCL decreases. By further increasing sample size, the 95UCL would decrease further and eventually converge upon the value of the exposure unit's actual mean concentration. In all cases below, one can see that the 95UCL is a health-protective selection for the EPC and that increasing sample size results in 95UCL estimates that are closer to the true mean.







3.0 GUIDANCE RECOMMENDATION 2: CALCULATING THE 95UCL

This section shows health assessors how to calculate 95UCLs for health assessment purposes. As Section 2.2 explains, health assessors should use 95UCLs as EPCs when evaluating intermediate or chronic exposures for exposure units having at least 8 discrete environmental samples. There are also some scenarios where the 95UCL should be used as the EPC when evaluating acute exposures (Section 2.1). Although numerous academic texts describe methods for calculating 95UCLs, health assessors should not use such general references because they typically do not address nuances specific to environmental data (e.g., how to handle non-detect observations when calculating 95UCLs).

For a given data set, 95UCLs can be calculated in many ways. However, the preferred approach for calculating 95UCLs for health assessment purposes depends on many factors, including the number of environmental samples in the exposure unit, the distribution of data, whether the data includes non-detect observations, and the software application or programming language being used for the calculations. This section presents step-by-step guidance for calculating 95UCLs and describes ATSDR's preferred approaches based on the aforementioned factors. Although multiple computational approaches are presented, some common themes apply to every calculation.

Key Point: General strategy for calculating 95UCLs

To calculate 95UCLs for an exposure unit, health assessors should:

- Determine what software application or programming language to use for the calculations (see Section 3.2 for options)
- Use Figure 2 to determine what flow chart (Figure 3-6) to use for the calculations
- Follow the selected flow chart's instructions to calculate the EPC and refer to different parts of this section for more detailed guidance where directed and to the appendixes for sample calculations

The remainder of this section presents guidance on various aspects of the 95UCL calculations to explain concepts presented in the flow charts (Figure 1 through Figure 6). Some parts of this section apply to every 95UCL calculation, such as general considerations (Section 3.1), background on software applications and programming languages (Section 3.2), initial data processing steps (Section 3.3), quality control checks (Section 3.7), and advanced topics (Section 3.8). Health assessors should read these general sections before applying this guidance.

The remaining parts of this section apply to specific software applications and programming languages: Sections 3.4, 3.5, and 3.6 pertain to ProUCL software, the R statistical programming language, and all other software applications, respectively. Health assessors need only read through or apply the section that pertains to the application or language that they will use.

3.1 General Approaches

For measured concentrations in environmental samples, the 95UCL is an upper-bound estimate of the actual average concentration within an exposure unit. The 95UCL should always be higher than the arithmetic mean of an exposure unit's environmental sampling data; and, as noted previously, increasing numbers of discrete environmental samples in the exposure unit typically result in 95UCL values decreasing and approaching the actual average value. The following "key point" identifies the

preferred approach for calculating 95UCLs as a function of the number of environmental samples in an exposure unit.

Key Point: How number of samples determines preferred statistical approaches

The preferred approaches based on number of samples within an exposure unit are:

- Fewer than 8 samples—do not calculate the 95UCL (see Section 2.2)
- 8 to 19 samples*—determine whether the data best fit a normal distribution, a lognormal distribution, or a gamma distribution and calculate the 95UCL using the recommended parametric statistical approaches shown in Sections 3.4 and 3.5 for ProUCL and R, respectively.
- 20 or more samples*—calculate the 95UCL using non-parametric "bootstrapping" techniques shown in Sections 3.4 and 3.5 for ProUCL and R, respectively.

**For 95UCL calculations, the following three conditions must also be met:*

- *There must be at least four samples with detected values of the contaminant.*
- *No more than 80 percent of the samples can be non-detects for the contaminants (or, alternately, at least 20 percent of the samples must have detected values of the contaminant)—see section text below for more details.*
- *There must be at least three unique detected values.*

Another important consideration for UCL calculations is how to handle non-detect observations. Non-detects are valid measurements in which the concentration of the contaminant of interest is too low to be measured with confidence. Sampling reports typically present non-detects either as being less than a specified limit (e.g., "<50 mg/kg"), with that limit being either a method detection limit or a quantitation limit (see Appendix A for definitions). In these cases, health assessors can only conclude that the actual contamination level is somewhere between 0 and the specified limit, but the actual value is not known.

Many environmental and public health agencies have published guidance on how to handle non-detect observations for risk assessment purposes. Health assessors should apply the following rules when calculating 95UCLs for data sets containing non-detects:

- *Do not delete non-detect observations from data sets.* Although actual environmental concentrations are not known for non-detect observations, these samples are valid measurements and must be included in 95UCL calculations. Health assessors should not delete non-detect observations from their data sets, because doing so will generally remove the lowest contamination levels from the data set, thus introducing a positive bias to the calculated 95UCLs.

- *Do not consider non-detect observations with extremely high detection limits.* As the one exception to the previous rule, health assessors should delete from data sets any non-detects reported for relatively insensitive methods. For example, if the majority of garden vegetable samples from an exposure unit have detected metal concentrations between 1 and 10 µg/kg but two samples are reported as "<10,000 µg/kg," health assessors should exclude the latter samples from the EPC calculation because they offer no informational value. All non-detect results with detection limits above the highest detected concentration in an exposure unit should not be considered when calculating 95UCLs.

- *Do not replace non-detect observations with a single surrogate value.* In some evaluations of environmental sampling data, health assessors may notice that non-detect observations have been replaced with surrogate values (i.e., concentrations of zero, one-half the detection limit, or the detection limit). When calculating 95UCLs, health assessors should never do this. Among other problems, replacing non-detects with the same number multiple times will generally underestimate the variability (i.e., standard deviation) of the data, which then underestimates the 95UCL. The preferred statistical approaches presented later in this section (particularly in Sections 3.4.2 and 3.5.2) were developed specifically for computing 95UCLs for data sets including non-detect observations. These methods address non-detect values without the need for them to be substituted with a surrogate value.

- *Do not calculate 95UCLs for data sets containing less than four samples with detected values and more than 80 percent non-detects.* Uncertainty in the 95UCLs increases with increasing numbers of non-detect observations in an exposure unit's sampling data. Health assessors should not calculate 95UCLs when these data sets do not contain at least four detected values and have more than 80 percent non-detect observations. Otherwise stated, when at least 20 percent of the samples have detected values after the minimum four detected values rule-of-thumb is met, health assessors may use the maximum detected concentration as the EPC. If there is any uncertainty around this approach, health assessors should consult with the ADS group about preferred computational approaches and consider recommending additional sampling using more sensitive methods, if available.

- *Do not calculate 95UCLs for data sets containing less than three unique detected values*. For example, consider a dataset with five detections at a concentration of 1 ppb and four non-detects at concentrations of <0.5 ppb. The lack of variability in the detected concentrations will lead to computational issues in some of the statistical approaches proposed in this document. As a result, health assessors should assume that the EPC is equal to the unique detected value anytime the dataset includes only one unique detected value. At least three unique detected values are needed in order to calculate a 95UCL using the procedures outlined in this document. If there are fewer than three unique detected values, use the maximum detected concentration as the EPC.

In addition to the aforementioned rules for processing non-detects, health assessors should be aware that laboratories report "estimated values" for environmental sampling, typically when laboratory equipment quantifies a chemical concentration—but at levels below the detection limit. These estimated values are usually flagged with a "J-qualifier" in sampling reports (e.g., arsenic concentration = 0.4 J µg/L). Health assessors should treat J-qualified data as valid sampling results and use the reported numerical values in UCL calculations. However, health assessors should recognize that J-qualified results have greater uncertainty than measurements without data qualifiers, and 95UCLs will be increasingly uncertain for data sets with large numbers of J-qualified results. For further insights into qualifiers for a given site's environmental data, health assessors should consider contacting the laboratory that analyzed the samples.

In addition, health assessors should note that it is not appropriate to compare one 95UCL to another. If a calculated 95UCL is greater than another 95UCL it does not necessarily follow that the exposures were higher where the higher UCL was estimated.

3.2 Software Applications and Programming Languages

Health assessors can use many different software applications and programming languages to calculate 95UCLs. ATSDR recommends the use of its recently developed web-based EPC Tool for calculating EPCs. Health assessors who use the ATSDR EPC Tool should refer to the Users' Guide on the EPC Tool Resources page for additional information and examples. For questions or assistance accessing the ATSDR EPC Tool, please contact the ATSDR technical team at PHAST@cdc.gov and include the phrase "ATSDR EPC Tool" in the subject line.

The remainder of this guidance focuses largely on how 95UCLs may be calculated using the ProUCL software application and the R programming language, due to their current widespread use among health assessors. Nonetheless, the following list reviews general features of a broader range of software applications and programming languages that health assessors may use for 95UCL calculations:

- ProUCL is a software program developed by EPA with a primary function of calculating 95UCLs for environmental data sets. From EPA's ProUCL website (see: https://www.epa.gov/land-research/proucl-software), health assessors can download the software for free, access the User Guide and Technical Guide, and view training sessions. Section 3.4 describes how health assessors should use ProUCL to calculate 95UCLs for health assessment purposes.

- R is a free programming language that can be downloaded from https://www.r-project.org/. While R is a powerful tool for visualizing and characterizing data sets and has many applications for conducting statistical analyses on environmental data, it is a programming language (as opposed to a software program with a graphical user interface) with its own syntax that requires a greater level of effort to learn and master, when compared to ProUCL. References are available with technical guidance on using R for environmental applications (e.g., Helsel 2012, Millard 2013). Section 3.5 describes how health assessors should use R to calculate 95UCLs for health assessment purposes.

- Spreadsheet programs, like Microsoft Excel, are useful for calculating descriptive statistics and manipulating data. However, many statistical tests and methods for 95UCL calculations that are readily accessible to users of ProUCL and R are not readily accessible to spreadsheet users, except through coding of macros and use of add-in functionalities. ATSDR therefore recommends that health assessors not use spreadsheet programs when calculating 95UCLs for environmental data.

- Many other statistical software packages (e.g., Minitab, SAS, and SPSS) can be used to perform the statistical tests and supplemental computations necessary for calculating 95UCLs. Section 3.6 describes how health assessors should use other statistical software packages to calculate 95UCLs for health assessment purposes. However, this option is only recommended for health assessors who are highly proficient with the software packages and who have advanced backgrounds in statistics.

3.3 Initial Data Processing Steps

As Figure 3-6 show, health assessors should conduct the same initial data processing steps for every 95UCL calculation, regardless of the flow chart that is followed. Those steps include:

- *Performing a data quality review.* Before using environmental data for EPC calculations, health assessors should carefully review the exposure unit's environmental sampling data and confirm that those data are of a known and high quality and meet data quality objectives for the health evaluation in question. Rejected data should never be used in EPC calculations and health assessors should carefully review the data following the guidance on data validation and usability outlined in chapter 5 of the ATSDR Public Health Assessment Guidance Manual.

- *Identifying and processing non-detect observations.* Section 3.1 presents several guidelines health assessors should follow when calculating 95UCLs for data sets with non-detect observations. When preparing numerical input files for ProUCL, R, or other software packages, health assessors must remember to properly differentiate non-detects from other sampling results. This is typically achieved by using two columns of data to compile the exposure unit's sampling data, where one column has a data flag for detections and the second column has a quantitative value (e.g., the detection limit or reporting limit for a non-detect and the measured concentration for all other samples). Health assessors should refer to Chapter 2 of the ProUCL User Guide and selected publications (e.g., Helsel 2012, Millard 2013) for the R programming language for instructions on compiling data input files, including specifications for identifying non-detects. Additionally, the computational examples in the appendixes to this guidance show how health assessors should prepare numerical input files for ProUCL and R.

  While processing non-detect observations, health assessors should confirm that the dataset contains at least four samples with detected values and that no more than 80 percent of the sample results are non-detect (refer to Section 3.1 for further details).

- *Identifying and processing lower-bound concentration data points.* Just as environmental contamination levels can fall below the measurement sensitivity of a sampling and analytical method (i.e., a non-detect), environmental contamination levels can also exceed a method's measurement range. For instance, an air monitoring device might only be capable of measuring a chemical between concentrations of 1 and 100 parts per billion (ppb); and when concentrations exceed this level, the instrument outputs a reading of ">100 ppb." When processing this type of data for UCL calculations, health assessors should replace lower-bound concentration data points with the upper-end of the sampling range (i.e., a reading of ">100 ppb" should be replaced with 100 ppb). Health assessors should acknowledge the potential bias introduced by this approach: 95UCLs might understate actual average concentrations, especially when environmental data include many lower-bound concentration results.

- *Identifying and processing duplicate samples and replicate analyses.* Two commonly used approaches for characterizing measurement precision are through analyzing duplicate samples (i.e., two environmental samples collected from the exact same place and time) or conducting replicate analyses of a single environmental sample. Health assessors should not use both measurements from duplicate samples or replicate analyses in their EPC calculations, because

doing so artificially assigns greater weight to these samples or analyses. Before starting EPC calculations, health assessors should identify all duplicate samples and replicate analyses among the exposure unit's environmental data and reduce paired results into single values. Health assessors may either choose the higher of the two paired values or use the average of the two values when processing the duplicate samples and replicate analyses. In cases where paired values include a detected concentration and a non-detect result, health assessors should use the detected value for the EPC calculations. *Plotting data.* Health assessors should plot an exposure unit's environmental data to visualize the distribution of measurements. Health assessors may use graphic features in ProUCL and R to visualize data, or they may choose to use spreadsheet programs for this purpose. Plotting histograms or box-and-whisker plots can indicate whether data are clustered within certain concentration ranges, whether data are skewed to lower or higher concentrations, and whether the data set includes outliers. These insights do not affect the process for calculating 95UCLs presented in Figure 3-6, but they are important because health assessors need to be highly familiar with an exposure unit's environmental sampling data.

3.4 Using ProUCL to Calculate 95UCLs

ATSDR's recommended approaches for using ProUCL to calculate 95UCLs differ for environmental data sets without non-detect observations (see Section 3.4.1) and with non-detect observations (see Section 3.4.2). For both scenarios, Figure 3 and Figure 4 outline the general computational approaches and Appendixes B and C present sample calculations using actual environmental data sets.

EPA's technical documentation for ProUCL includes useful instructions and tips for calculating 95UCLs. However, because EPA originally developed ProUCL for very specific applications (i.e., estimating background concentrations), not all information in the ProUCL technical documentation applies to calculating 95UCLs for public health assessment purposes, and important examples of how EPA and ATSDR approaches differ are noted below.

Key Point: ATSDR's preferred approaches sometimes differ from ProUCL default approaches
While health assessors are encouraged to read ProUCL technical documentation to learn how to use the software program and understand its features, they should always follow this ATSDR guidance for using ProUCL to calculate 95UCLs.

3.4.1 Data with No Non-detects

Figure 3 shows how health assessors should use ProUCL to calculate 95UCLs for data sets without non-detect observations. The figure's first and third steps involve pre- and post-processing of data, and health assessors conduct this work outside of ProUCL. The second step outlines specifically how health assessors should use ProUCL to calculate the 95UCLs, with procedures depending on the number of samples within the exposure unit:

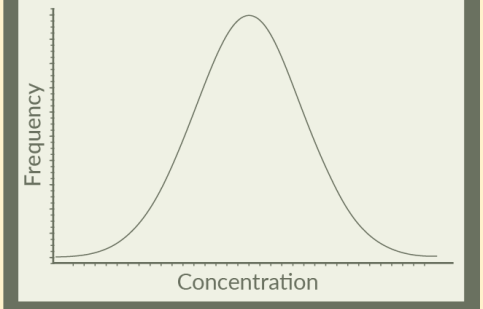
- *Exposure units with 8 to 19 samples.* The general approach for smaller data sets is to first determine whether data are best represented by normal, lognormal, or gamma distributions (see text box on next page for more information on these distributions). With two exceptions, health assessors should initially select the data distribution—normal, lognormal, or gamma—found to have a correlation coefficient (or R value) closest to 1. This is done with the "Goodness-

of-Fit Statistics" option under ProUCL's "Statistical Tests" menu. The output from this function provides correlation coefficients for the three distributions. Correlation coefficients calculated for this purpose measure agreement between observed data (i.e., sampling results) and a fitted distribution (i.e., normal, lognormal, or gamma distributions), with higher values indicating better fit to the specified distribution.

As the first exception, if the highest correlation coefficient is less than 0.80, health assessors should not calculate 95UCLs and should contact their ADS group for further direction. In these cases, the data do not adequately fit to a normal, lognormal, or gamma distribution and the statistical methods prescribed in this guidance are not appropriate for calculating 95UCLs. As the second exception, if data are initially found to best fit a normal distribution, health assessors must confirm that the normal distribution is physically realistic for environmental data applications. In ProUCL, this is done by generating a "Q-Q plot" from the "Graphs" menu. Upon making these menu selections, the software will generate and display a "Q-Q plot" that includes a line representing the normal distribution (see "Q-Q plot" text box for an example). The key issue to evaluate is the value of the "theoretical quantile" (i.e., the x-variable) at the point where the concentration value (i.e., the y-variable) equals zero. If that value is less than -3, health assessors should use the normal distribution for the 95UCL calculations. If that value is greater than -3, then the normal distribution fit to the data contains too many negative values to be realistic for environmental applications; and when this occurs, health assessors should instead base their 95UCL calculations on the lognormal distribution or the gamma distribution, whichever was found to have the higher correlation coefficient with ProUCL's "Goodness-of-Fit Statistics," as described above. Refer to the "Q-Q plot" text box for an example of using ProUCL's graphing functions to confirm whether normal distributions should be used in 95UCL calculations.

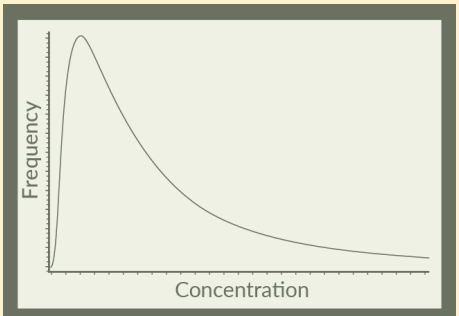
Additional Information: What are normal, lognormal, and gamma distributions?

When calculating 95UCLs for exposure units that have between 8 and 19 samples, health assessors must first determine whether the data best fit a normal distribution, a lognormal distribution, or a gamma distribution. This first step is performed regardless of the software program or programming language used in the analysis. The diagrams in this text box display key features of these three distributions. Once the best-fitting distribution is identified, health assessors should follow the preferred parametric computational approaches for the distribution of interest. For exposure units that have 20 or more samples, health assessors do not need to identify the best-fitting distribution. That is because 95UCLs can be calculated using non-parametric methods, which do not require the underlying data set to be represented by the shape of a particular distribution.



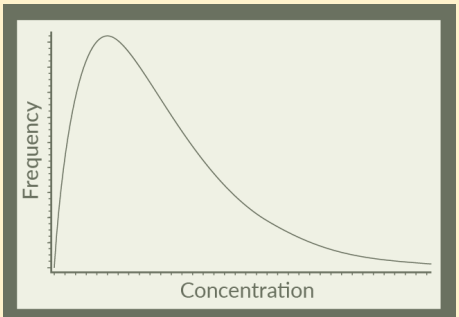
Normal

Data are symmetric. A majority of observations are clustered around the central value with equal porportions above and below that value.



Lognormal

Data are right skewed. A greater proportion of observations fall above the central value or in the right tail of the distribution.

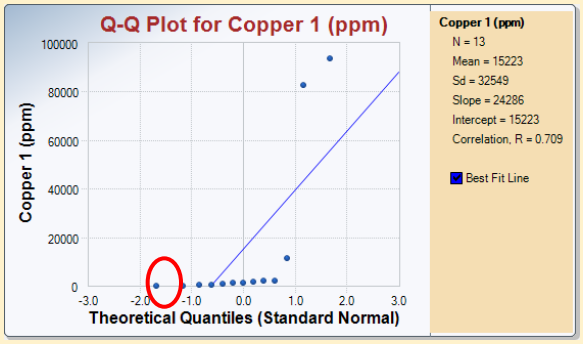


Gamma

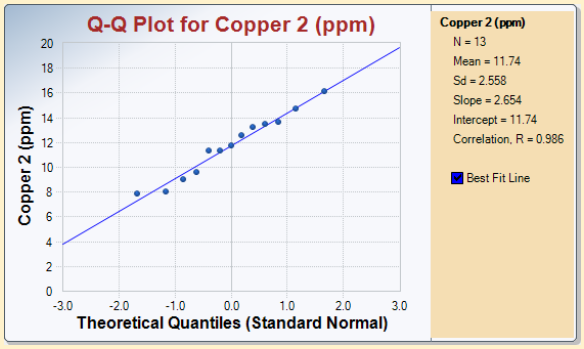
Data are right skewed. Similar to the lognormal distribution but with less of a tail on the right.

Additional Information: Using Q-Q Plots in ProUCL to Evaluate Normal Distributions

Health assessors should generate Q-Q plots in ProUCL under the following circumstance: when evaluating exposure units with between 8 and 19 samples and after initially identifying the normal distribution as the best-fitting distribution.



The first graph shown here is the Q-Q plot generated by ProUCL for a sample copper dataset (Copper 1). The red circle shown in the figure is added to illustrate the "theoretical quantile" value when the concentration equals zero. In this case, that value is approximately -0.6. Because this value is greater than -3.0 (i.e., the cut-off presented earlier in this section), the appropriate interpretation is that the normal distribution fit to the data contains an unrealistic number of negative values and should therefore not be used for the EPC calculation. In this case, health assessors would refer to their original "Goodness-of-Fit Statistics" and select a lognormal distribution or gamma distribution for the EPC calculation—using the distribution that has the higher correlation coefficient. The second graph shows another sample copper dataset (Copper 2). In this case, the "theoretical quantile" value when the concentration equals zero is far less than -3.0 (beyond the bounds of the axis). For these data, it would be appropriate to base 95UCL statistics on the normal distribution.



Note: The axes of ProUCL's Q-Q plots can be manually altered by right clicking on the x- or y-axis and selecting properties, thus enabling health assessors to create a plot from which they can quickly identify the value of interest (i.e., the theoretical quantile when the concentration is equal to zero). The R code provided here does not always create a horizontal grid line where the concentration equals zero, and health assessors should be mindful of this when using R to generate Q-Q plots.

At this point, health assessors should have their distribution selection, with which they will use ProUCL to determine the 95UCL for their data set. This function is performed by selecting the "UCLs/EPCs" menu option for the selected distribution. Step 2B (Option A) in Figure 3 lists the specific ProUCL output values that should be used for 95UCLs for the three different types of distributions.

By applying this approach, ATSDR expects that nearly all environmental sampling data sets with between 8 and 19 samples will be adequately fit by one of the three candidate distributions. However, it is possible that health assessors will find that sampling data do not adequately fit the normal, lognormal, or gamma distributions. In this unlikely case, health assessors should consult with their ADS group for guidance on computing 95UCLs.

Note: When using ProUCL to calculate 95UCLs for lognormally distributed data, health assessors should use Chebyshev's method (see Figure 3).[4] However, for lognormal data, Cox's method tends to produce better estimates of the 95UCL than Chebyshev's method, but Cox's method currently is not available in the ProUCL software package. In this case, health assessors who use ProUCL should either (a) proceed with the Chebyshev's method estimates (e.g., 95% Chebyshev UCL) but recognize that this method sometimes calculates relatively high 95UCLs or (2) use the R programming language or other software programs such that Cox's method can be applied to data best fit by lognormal distributions. This limitation of ProUCL is only of temporary concern as the agency will use the best available statistics when these methods are transferred to ATSDR's Public Health Assessment Site Tool (PHAST).

- *Exposure units with 20 or more samples.* With larger data sets, health assessors need not identify best-fitting distributions, because calculated 95UCLs are relatively insensitive to the different distribution shapes and types. In these cases, health assessors should select "Non-parametric" from ProUCL's "UCLs/EPCs" menu and use the "95% percentile bootstrap UCL" as the 95UCL.

3.4.2 Data with Non-detects

Figure 4 shows how health assessors should use ProUCL to calculate 95UCLs for data sets with non-detect observations. The figure's first and third steps involve pre- and post-processing of data, and health assessors conduct those steps outside of ProUCL. The second step outlines specifically how health assessors should use ProUCL to calculate the 95UCLs for data sets with non-detects. The procedures are conceptually similar to those outlined in Section 3.4.1 and again vary by the number of samples in the exposure unit:

- *Exposure units with 8-19 samples.* The general approach for smaller data sets is to first determine whether data are best represented by normal, lognormal, or gamma distributions. This determination should be made using the "Goodness of Fit Statistics" for data sets "With NDs" from the "Statistical Tests" menu function in ProUCL. With two exceptions, 95UCLs should be calculated for the type of data distribution—normal, lognormal, or gamma—found to have a correlation coefficient (or R value) closest to 1. ProUCL provides several correlation coefficients for each distribution, with each estimated while processing non-detects in a different way (e.g., excluding non-detects, setting non-detects equal to the detection limit, setting non-detects equal to ½ the detection limit, and regression on order statistics [ROS] imputation). In this case, health assessors should base their selection solely on correlation coefficients estimated with ROS imputation and ignore all other values.

  As the first exception, if the highest correlation coefficient is less than 0.80, health assessors should not calculate 95UCLs and should contact their ADS group for further direction. In these cases, the data do not adequately fit to a normal, lognormal, or gamma distribution and the

[4] Chebyshev 95UCLs are based on Chebyshev's inequality theorem, which assumes that 75 percent of values lie within two standard deviations of the mean and 89 percent of the values lie within three standards. This is different than the normal distribution, which assumes 95 and 99.7 percent of results lie within two and three standard deviations of the mean, respectively.

statistical methods prescribed in this guidance are not appropriate for calculating 95UCLs. As the second exception, if data are initially found to best fit a normal distribution, health assessors must confirm that the normal distribution is physically realistic for environmental data applications. This confirmation is done with the following sequence of menu selections in ProUCL: "Stats/Samples Sizes" then "Imputed NDs using ROS methods" then "Normal ROS" option (see Appendix C for an example). This selection will generate a new column of numbers in the ProUCL input data file under the header "Normal ROS." If any negative numbers appear in this new column of data, then the normal distribution is not a realistic fit, and health assessors should base their 95UCL calculations on the lognormal distribution or the gamma distribution, whichever was found to have the higher correlation coefficient (estimated with ROS imputation), as described above. If only positive numbers appear in this new column of data, then the normal distribution should be considered the best fitting distribution.

Once health assessors have determined that an exposure unit's sampling data are best fit by a normal, lognormal, or gamma distribution, they should determine 95UCLs using the preferred computational approaches shown in Step 2B (under Option A) of Figure 4. Briefly, Kaplan-Meier approaches should be used for data best fit by normal distributions and gamma distributions, and a lognormal ROS bootstrapping method should be used for data best fit by lognormal distributions (see Figure 4 for details).

- *Exposure units with 20 or more samples.* With larger data sets, health assessors need not identify best-fitting distributions, because calculated 95UCLs are relatively insensitive to the different distribution shapes and types. The preferred approach does however depend on whether the non-detect observations all have the same detection (or reporting) limit. For data sets with different values for the detection limits, the following sequence of ProUCL options should be selected: "UCLs/EPCs," "With NDs," and "Non-parametric," and Kaplan-Meier estimates should then be used to determine the 95UCL. For data sets with the same values for the detection limits, the following sequence of ProUCL options should be selected: "UCLs/EPCs," "With NDs," and "Lognormal," and a bootstrapping approach with lognormal ROS should be used to determine the UCL. Refer to Step Two (under Option B) in Figure 4 for the specific names of the two preferred approaches.

3.5 Using R to Calculate 95UCLs

ATSDR's recommended approaches for using R to calculate 95UCLs differ for environmental data sets without non-detect observations (see Section 3.5.1) and with non-detect observations (see Section 3.5.2). For both scenarios, Figure 5 and Figure 6 outline the general computational approaches and Appendix D and Appendix E present sample calculations using actual environmental data sets.

3.5.1 Data with No Non-detects

Figure 5 shows how health assessors should use R to calculate 95UCLs for data sets without non-detect observations. The figure's first and third steps involve pre- and post-processing of data, and health assessors conduct these steps outside of R. The second step outlines specifically how health assessors should use R to calculate 95UCLs, with procedures depending on the exposure unit's number of samples. Health assessors will need to install and load the EnvStats and Bootstrap packages to R in order

to estimate 95UCLs with the methods described below. Appendix D presents an example 95UCL calculation in R for a small dataset with no non-detect observations.

- *Exposure units with 8-19 samples.* The general approach for smaller data sets is to first determine whether data are best represented by normal, lognormal, or gamma distributions. Health assessors should initially select the data distribution—normal, lognormal, or gamma—found to have the highest W test statistic from the Shapiro-Wilk GOF hypothesis test. The following commands[5] in R provide the W test statistic for the three distributions of interest (and see Appendix D for a specific example):

```
> plot(gofTest(Data, dist="norm"))
> plot(gofTest(Data, dist="lnorm"))
> plot(gofTest(Data, dist="gamma"))
```

  If the normal distribution has the highest W test statistic, health assessors must then use Q-Q probability plots to ensure that the distribution does not include an unrealistic amount of negative values. Health assessors can generate the Q-Q plot for normally distributed data using the following command:

```
> qqPlot(Data,dist="norm",add.line=T)
```

  Health assessors should inspect the graph that R displays to determine if the normal distribution is appropriate for 95UCL calculations. This is done by determining the value of the "quantile of normal" (i.e., the x-variable) at the point where the concentration quantile (i.e., the y-variable) equals zero. If this point is less than -3, then health assessors should use the normal distribution for the 95UCL calculations. On the other hand, if this value is greater than -3, the normal distribution includes an unacceptable amount of negative values and should not be used for calculating 95UCLs. When this occurs, health assessors should instead base their 95UCL calculations on the lognormal or gamma distribution for the 95UCL calculations, whichever has the next highest W test statistic from the Shapiro-Wilk test.

  For all three distributions, health assessors should also check to ensure that the model-estimated mean is less than the maximum detected value. If the model-estimated mean is greater than the maximum detected value in the dataset, reject it and try the distribution with the next highest W test statistic instead. If the next highest distribution has a model-estimated mean beneath the maximum detected value, use it as the distribution to calculate the 95UCL. If not, continue this process until all three distributions have been tried. If all three distributions have model-estimated means greater than the maximum detected value, report the maximum detected value as the EPC.

  Once health assessors have determined that an exposure unit's sampling data are best fit by a normal, lognormal, or gamma distribution, they then calculate 95UCLs for the selected

[5] For the R commands presented throughout this section, "Data" represents the input column of chemical concentrations, and health assessors should replace "Data" with whatever column heading they use for site-specific data sets. For example, if the data column is titled "Copper," health assessors should replace "Data" with "Copper" in each R command.

distribution. Refer to step 2B (under Option A) in Figure 5 for the applicable R commands that should be used to calculate 95UCLs for the three different types of distributions. For all three distributions, the 95UCL and the model-estimated mean can be obtained using the following R commands. In these commands, the variable "output" represents the output of the R commands shown in Figure 5.

```
> output[["interval"]][["limits"]][["UCL"]]
> output[["parameters"]][["mean"]]
```

- *Exposure units with 20 or more samples.* With larger data sets, health assessors need not identify best-fitting distributions, because calculated 95UCLs are relatively insensitive to the different distribution shapes and types. In these cases, health assessors use non-parametric percentile bootstrap estimates of the 95UCL with 5,000 bootstrap repetitions and a 90% confidence interval). Refer to Step 2 (Option B) in Figure 5 for the R command that should be used to estimate 95UCLs for data sets with more than 20 samples and no non-detects.

3.5.2 Data with Non-detects

Figure 6 shows how health assessors should use R to calculate 95UCLs for data sets with non-detect observations. The figure's first and third steps involve pre- and post-processing of data, and health assessors conduct these steps outside of R. The second step outlines specifically how health assessors should use R to calculate 95UCLs, with procedures depending on the number of samples within the exposure unit. Note that health assessors need to install and load the EnvStats and NADA (non-detects and data analysis for environmental data) packages to R to estimate 95UCLs with the methods described below. Appendix E shows an example 95UCL calculation in R for a small dataset with non-detects.

- *Exposure units with 8-19 samples.* The general approach for smaller data sets is to first determine whether data are best represented by normal, lognormal, or gamma distributions. Health assessors can determine which of these three distributions best fits the exposure unit's sampling data by using R's boxcoxCensored command.[6] This command computes Probability Plot Correlation Coefficient (PPCC) statistics for lambda values between 0 and 1 at increments of 0.1. A lambda of 1 represents a normal distribution, a lambda of 0 represents a lognormal distribution, and a lambda of approximately 0.3 represents a gamma distribution. With one exception, health assessors should review the PPCC statistics at these three values of lambda and select the distribution with the highest PPCC statistic.

  The command below will estimate the PPCC statistic at different values of lambda and is demonstrated in Appendix E. Note that "Data" represents the input column of concentrations and "Cen" is the censoring indicator column; both may need to be updated to reflect the column headings of site-specific data sets.

```
> boxcoxCensored(Data,Cen,lambda=seq(0,1,0.1))
```

[6] Health assessors who are interested in additional details on the boxcoxCensored command are referred to background information published on the EnvStats package for R (Millard, 2018).

As the exception, if data are initially found to best fit a normal distribution, health assessors must confirm that the normal distribution is physically realistic for environmental data applications and that the lower end of the distribution does not include an unreasonable amount of negative values. This is done by creating a normal Q-Q plot with the qqplot Censored command (see below) and determining the "quantile of normal" value (i.e., the x-variable) at the point where the concentration value (i.e., the y-variable) equals zero. If this value is higher than -3, an unreasonable amount of negative numbers is being estimated by the normal distribution, and health assessors should select the lognormal or gamma distribution, whichever has the next highest PPCC statistic. The following code is used to create a normal Q-Q plot for data with non-detect observations and is demonstrated in Appendix E.

```
> qqPlotCensored(Data,Cen,dist="norm",add.line=TRUE)
```

For all three distributions, health assessors should check to ensure that the model-estimated mean is less than the maximum detected value. If the model-estimated mean is greater than the maximum detected value in the dataset, reject it and try the distribution with the next highest PPCC statistic instead. If the next highest distribution has a model-estimated mean beneath the maximum detected value, use it as the distribution to calculate the 95UCL. If not, continue this process until all three distributions have been tried. If all three distributions have model-estimated means greater than the maximum detected value, report the maximum detected value as the EPC.

Once health assessors have determined the best-fitting distribution (i.e., normal, lognormal, or gamma), they then calculate 95UCLs using the commands shown in Step 2B (under Option A) in Figure 6. For all three distributions, the 95UCL and the model-estimated mean can be obtained using the following R commands. In these commands, the variable "output" represents the output of the R commands shown in Figure 6.

```
> output[["interval"]][["limits"]][["UCL"]]
> output[["parameters"]][["mean"]]
```

- *Exposure units with 20 or more samples.* With larger data sets, health assessors need not identify best-fitting distributions, because calculated 95UCLs are relatively insensitive to the different distribution shapes and types. In most cases, health assessors should use non-parametric percentile bootstrap estimates of the 95UCL with 5,000 bootstrap repetitions. The R command discussed in Option B under Step Two of Figure 6 implements this approach using the Kaplan-Meier method to impute the censored data. However, in some cases the Kaplan-Meier method is not appropriate; in these instances health assessors should use bootstrap sampling of a lognormal distribution to determine the 95UCL. Health assessors should use bootstrap sampling of a lognormal distribution when their data are singly censored (have only one censoring limit), or when their data are mixed-censored but the percent of the data exceeding the various censoring limits is the same. The following function from the NADA R package can be used to identify the censoring properties of a dataset.

```
> dataCensoringSummary <- NADA::censummary(Data, Cen)
```

If either of the following two statements are true, health assessors should use bootstrap sampling of the lognormal distribution to estimate the 95UCL. Otherwise, they should use non-parametric percentile bootstrap estimates. Refer to Step Two (under Option B) in Figure 6 for the R commands to estimate 95UCLs in both scenarios.

```
> length(dataCensoringSummary$limits$pexceed) == 1
> var(dataCensoringSummary$limits$pexceed) == 0
```

3.6 Using Other Software Programs to Calculate 95UCLs

As noted in Section 1.0, ATSDR has developed a web-based EPC Tool that automates the calculations described in this document. For questions or assistance accessing and using the ATSDR EPC Tool, please contact the ATSDR technical team at PHAST@cdc.gov and include the phrase "ATSDR EPC Tool" in the subject line.

As Section 3.2 acknowledges, Minitab, SAS, SPSS, and additional software programs other than ProUCL and R can also be used to calculate 95UCLs for environmental data sets. This document does not include step-by-step guidance for these software programs, because ATSDR anticipates that most health assessors would prefer to use either the ProUCL software program or the R programming language for 95UCL calculations. When other software programs are used, health assessors should follow the same conceptual approaches outlined in Section 3.4. For example, for exposure units with between 8 and 19 samples, health assessors should use a parametric approach by first identifying the best-fitting distribution (normal, lognormal, or gamma) and then applying the proper computational technique; and for 20 or more samples, non-parametric approaches should be used. Health assessors who plan to use any of these software programs should first consult with their ADS group about methods.

3.7 Quality Control Checks

After calculating 95UCLs for exposure units, health assessors should perform the following four quality control checks to ensure that the calculated values are reasonable:

- *Check #1: Compare the 95UCL to the maximum concentration.* For very small and highly variable data sets, it is not uncommon for 95UCLs to exceed the maximum data points. However, health assessors should only calculate 95UCLs for exposure units with at least 8 valid environmental samples. For this sample size, a 95UCL greater than the maximum concentration will sometimes occur (e.g., for highly variable data), but also could be a sign of computational error. If health assessors determine that a 95UCL is calculated correctly and indeed exceeds the maximum concentration from the exposure unit's environmental samples, they should select the appropriate EPC statistic (i.e., the 95UCL or the maximum detected concentration) based on the number of samples included in the data set and their own professional judgement.

    - *Exposure units with 8 to 19 samples.* When the 95UCL exceeds the maximum concentration for data sets of this size, ATSDR recommends using the 95UCL as the EPC. In this case and where limited data are available, the maximum concentration may underestimate the true mean of contaminant concentrations in the exposure unit.

- *Exposure units with 20 or more samples.* For larger data sets, health assessors should not expect a 95UCL to exceed the maximum detected concentration. When this occurs, there is likely a high degree of variability in the sample data and not enough data to support a reasonable 95UCL calculation. ATSDR therefore recommends using the maximum as the EPC so long as that concentration is assumed to fall within the higher end of the distribution of actual contamination levels

- *Check #2: Compare the 95UCL to the average concentration.* Health assessors should make two comparisons between a 95UCL and the average concentration of the exposure unit's environmental samples. First, health assessors should confirm that each calculated 95UCL is higher than the average concentration calculated for the same set of samples; this should always be the case, and health assessors need to double-check calculations if 95UCLs are less than the corresponding average concentrations. Second, health assessors should evaluate whether 95UCLs are unexpectedly higher than the corresponding average concentrations. When a 95UCL is more than three times higher than the average concentration calculated from the same data set, the calculated value requires further scrutiny. When this occurs, health assessors should first double-check the accuracy of their calculations. If the calculations are correct, the result may be influenced by a potential outlier or there may not be enough samples to support a precise estimate of exposure. Health assessors should therefore evaluate whether the 95UCL is influenced by data outliers following the directions in the next bulleted item. In some cases, it may also be useful to conduct a sensitivity analysis to see if the average concentration and 95 percent lower confidence limit around the average support any health conclusions.

- *Check #3: Consider the influence of data outliers.* If health assessors suspect that data outliers are influencing the magnitude of the 95UCL, they should consider removing the potential outliers from the data set, recalculating the 95UCL, and comparing the original and recalculated values. This decision should be science-based and not rely solely on the measured concentrations. In most cases, particularly for larger data sets, 95UCLs should be relatively insensitive to removal of one or two data points, even those on the higher end of the distribution. However, some 95UCLs may be found to be highly sensitive to the presence of data outliers. This becomes particularly important if an exposure dose calculated from a 95UCL exceeds a health guideline value (e.g., a Minimal Risk Level or Reference Dose) with all measurements considered but falls below a health guideline value when apparent outliers are removed. Such cases warrant closer evaluation and additional sampling may be necessary to have greater confidence that the exposure unit's contamination has been adequately characterized. This step should only be completed as a way to evaluate the influence of a potential outlier on calculated 95UCLs, and not as justification to remove an outlier from the dataset entirely. Health assessors should only ever remove outliers from a dataset if the results are judged to be invalid or not representative of site conditions.

Additional Information: How to handle outliers?

Environmental contamination levels within an exposure unit can be highly variable. In some cases, a sampling result from an exposure unit may be dramatically higher than all other measurements, and such a result is typically referred to as an outlier. Statistical tests have been developed to objectively identify outliers among data, but health assessors should not apply those tests for 95UCL determinations. Health assessors should base the decision of whether to remove an outlier only on science-based criteria, not the concentration value itself. With one exception, health assessors should use all data—including potential outliers—in their 95UCL calculations. Section 3.8 presents additional considerations for interpreting 95UCLs when an exposure unit's environmental data include suspected outliers.

As the one exception, health assessors should remove any outliers that are not valid results or are not representative of the exposure scenario being evaluated. They should always check potential outliers against the original sampling report to ensure that no transcription errors were made. Additionally, they should consider whether a potential outlier is influenced by atypical conditions (e.g., a surface water sample collected during a rare algal bloom that is not site-related); in such cases, health assessors may be justified in removing sampling results prior to the 95UCL calculation, but they should also document this data processing step and the rationale for excluding any data when determining EPCs.

Health assessors should only remove outliers from a dataset if the results are judged to be invalid or not representative of site conditions.

- *Check #4: Consider having a colleague double-check calculations.* It is always good practice to have colleagues review calculations that support public health assessment conclusions, especially 95UCL calculations used in health effects evaluations. To facilitate this review, health assessors should briefly document their calculations (e.g., identify the software package or programming language used, provide the sampling data from the exposure unit, and specify the statistical method used to calculate the 95UCL) and ask an experienced colleague to replicate the results—an ADS group may help with this review as well.

3.8 Special Considerations and Advanced Topics

While ATSDR developed this guidance to apply to a broad range of site-specific scenarios, some sites and environmental data sets will present unique challenges for calculating 95UCLs. This section identifies a few special considerations and advanced topics for 95UCL calculations. In general, health assessors should consult with their ADS group should they encounter any of the following site-specific scenarios or other circumstances not sufficiently covered by the general guidance presented earlier in this section.

- *Approaches for highly skewed or highly variable data.* In some cases, environmental data sets will have sampling results that are highly skewed to large concentration values. For purposes of this guidance, a data set with geometric standard deviation greater than 4.5 is considered "highly skewed." In these cases, health assessors should still calculate 95UCLs using the approaches described earlier in this section. However, before using the calculated 95UCLs in the health effects evaluation, health assessors should examine data for extreme outliers, determine whether the data set includes stratified results that could argue for splitting the exposure unit into multiple sub-units (see below), and consider the need for additional sampling to better characterize the distribution of contamination levels.

- *Approaches for evaluating data that best fit other types of distributions.* This section presents 95UCL computational approaches for environmental sampling data found to best fit normal, lognormal, and gamma distributions. In some cases, health assessors might encounter site reports suggesting that environmental data are best fit by other types of distributions (e.g., Weibull). In these cases, health assessors should confirm that the alternate distribution types are indeed better fitting and, if so, consult with their ADS group about preferred 95UCL computational approaches. In some cases, the alternate distribution form might provide further insights into contamination levels. For instance, should an exposure unit have environmental sampling data that appear to fit a bimodal distribution, the health assessor should investigate whether the exposure unit was properly defined and consider whether the data support a decision to split the exposure unit into sub-units.

- *Approaches for evaluating weighted data.* As described in Section 1.1, the computational approaches outlined throughout this guidance assume that an exposure unit's environmental sampling data are random and independent, which therefore supports the "unweighted" statistical methods for calculating 95UCLs. These methods assume random or equal-area sampling within an exposure unit. However, at some contaminated sites, sampling is focused on areas of suspected high contamination. In such cases where sampling is not random, it is important to avoid bias in determining the EPC, and each data point needs to be weighted by the area it represents. When many measurements are taken within a small area, weights for those measurements will be small. When few measurements are taken over large areas, their weights will be large.

  Under a judgmental or biased sampling plan that deliberately collects more samples in highly contaminated areas, an unweighted UCL95 should not be used as the EPC. However, 95UCLs can be calculated using weighted areas that take these uneven sampling designs into account if in the professional judgement of the health assessor, the samples can be assumed to be representative of the areas or times they were collected in. The general idea is that exposure units are split into smaller sub-units, and an area proportional method is then used to weigh each sub-unit's sampling data. For example, if an exposure unit has three sub-units that cover 15 percent, 20 percent, and 65 percent of the total area, the health assessor would weight data (or their mean) in the first sub-unit by 0.15, in the second sub-unit by 0.20, and in the third by 0.65, to compute the overall mean or EPC for the entire EU. The weighted 95UCL is a bit more complicated to calculate but follows the same general principles. Appendix F outlines the process for calculating weighted 95UCLs in R for datasets without and with non-detect results. Note that the current version of ProUCL cannot be used to calculate weighted 95UCLs. In these cases, the health assessor's judgment becomes essential, and health assessors should consult with their ADS group before applying 95UCL weighted computational approaches to sub-units.

4.0 REFERENCES

[ATSDR] Agency for Toxic Substances and Disease Registry. 2022a. Public health assessment guidance manual. Atlanta: US Department of Health and Human Services. Available from: https://www.atsdr.cdc.gov/pha-guidance/index.html .

[ATSDR] Agency for Toxic Substances and Disease Registry. 2022b. Toxic Equivalents Guidance for Dioxin and Dioxin-like Compounds. Atlanta, GA: U.S. Department of Health and Human Services, Public Health Service.

[ATSDR] Agency for Toxic Substances and Disease Registry. 2022c. Guidance for Calculating Benzo(a)pyrene Equivalents for Cancer Evaluations of Polycyclic Aromatic Hydrocarbons. Atlanta, GA: U.S. Department of Health and Human Services, Public Health Service.

[ATSDR] Agency for Toxic Substances and Disease Registry. 2022d. Exposure Point Concentration Guidance for Non-Discrete Sampling. Atlanta, GA: U.S. Department of Health and Human Services, Public Health Service.

[ATSDR] Agency for Toxic Substances and Disease Registry. 2020. Identifying Exposure Units for the Public Health Assessment Process. Atlanta, GA: U.S. Department of Health and Human Services, Public Health Service.

Helsel, D. 2012. Statistics for Censored Environmental Data Using Minitab and R, 2nd Edition. John Wiley and Sons, Inc.: Hoboken, New Jersey. 291 p.

Millard, S. 2013. EnvStats: An R Package for Environmental Statistics. Springer Science and Business Media: New York, New York. 960 p.

Millard, S. 2018. Package 'EnvStats', Package for Environmental Statistics, Including US EPA Guidance. Available at: https://cran.r-project.org/web/packages/EnvStats/EnvStats.pdf

[US EPA] US Environmental Protection Agency. 2015. ProUCL Version 5.1. Technical Guide. Statistical Software for Environmental Applications for Data Sets with and without Nondetect Observations. EPA/600/R-07/041.

[US EPA] US Environmental Protection Agency. ProUCL Software. Available for download at: https://www.epa.gov/land-research/proucl-software.

[US EPA] US Environmental Protection Agency US EPA, 2016. Lead at Superfund Sites: Guidance. Available at: https://www.epa.gov/superfund/lead-superfund-sites-guidance

Figure 1. Selection of EPC Statistics for Different Exposure Durations

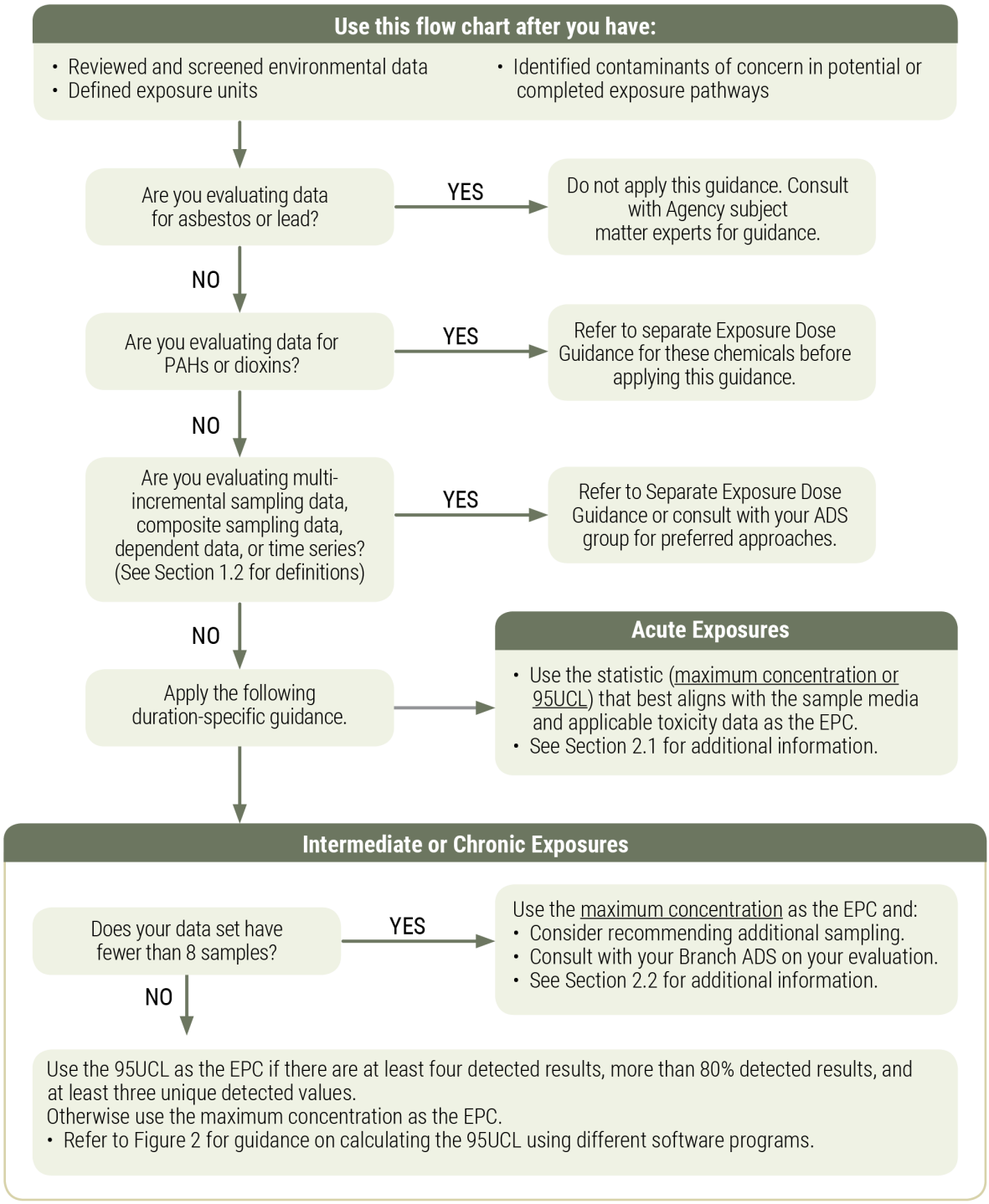


Figure 2. Where to Find Guidance on Calculating 95UCLs with Different Software Programs

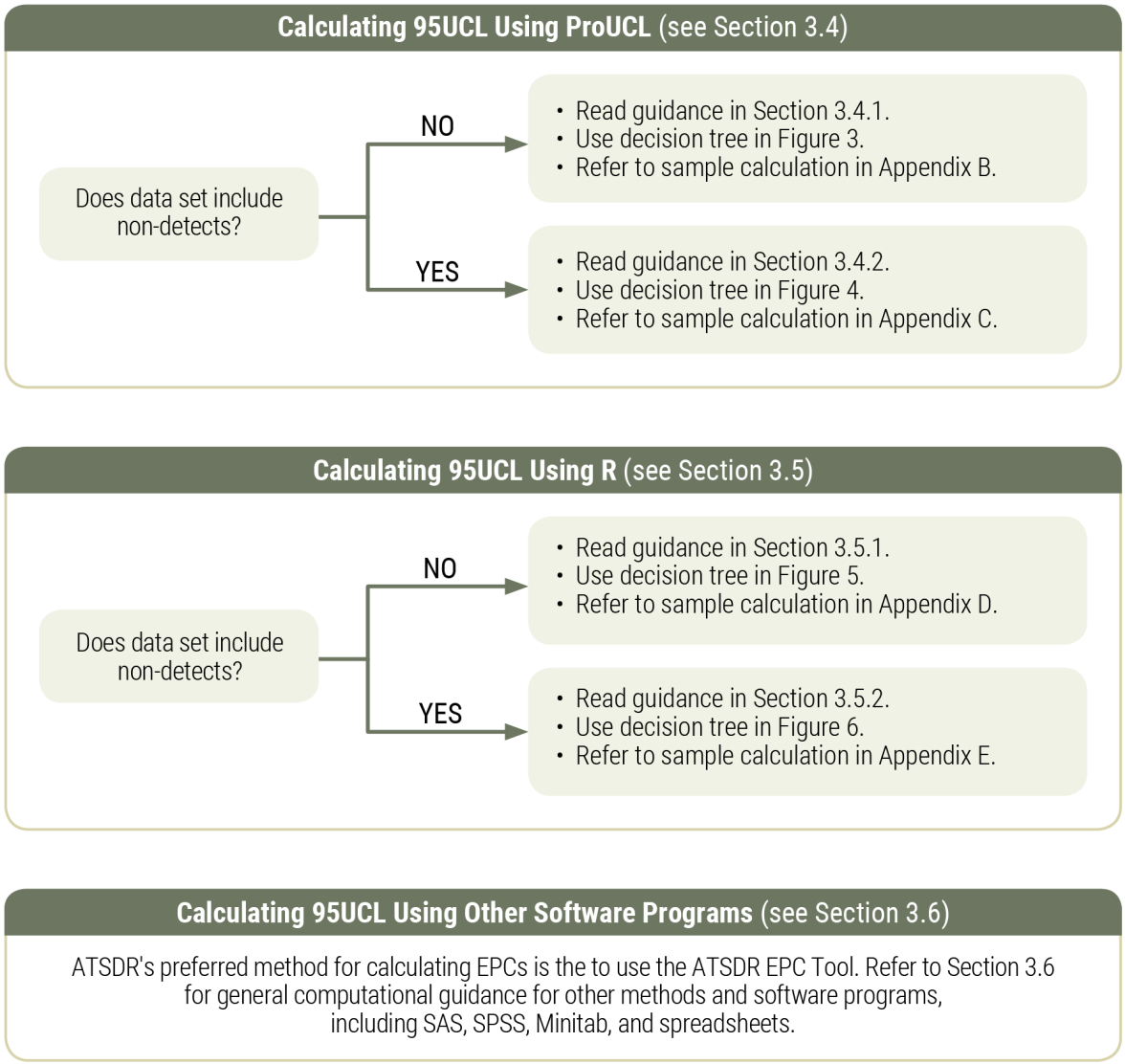
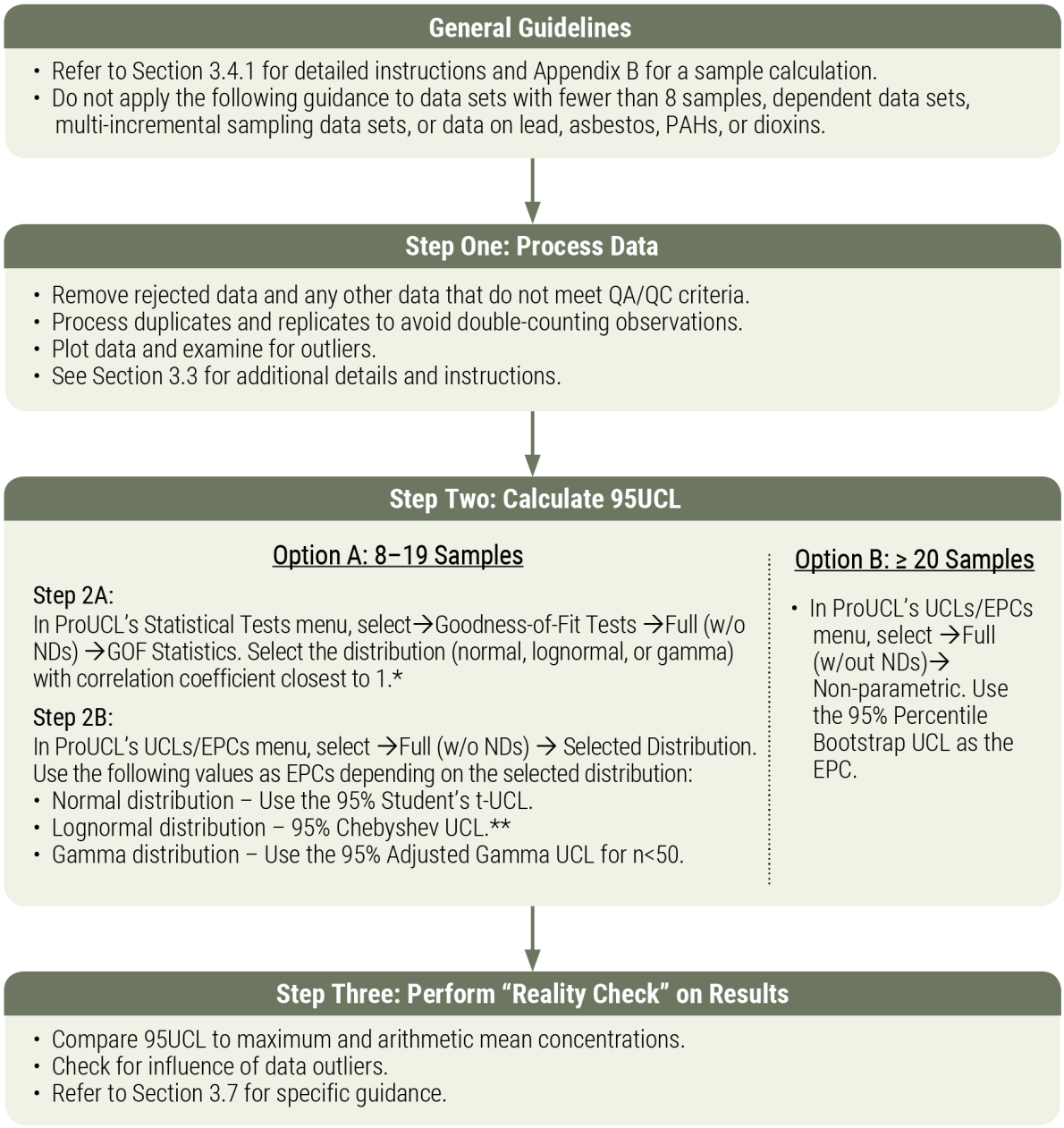


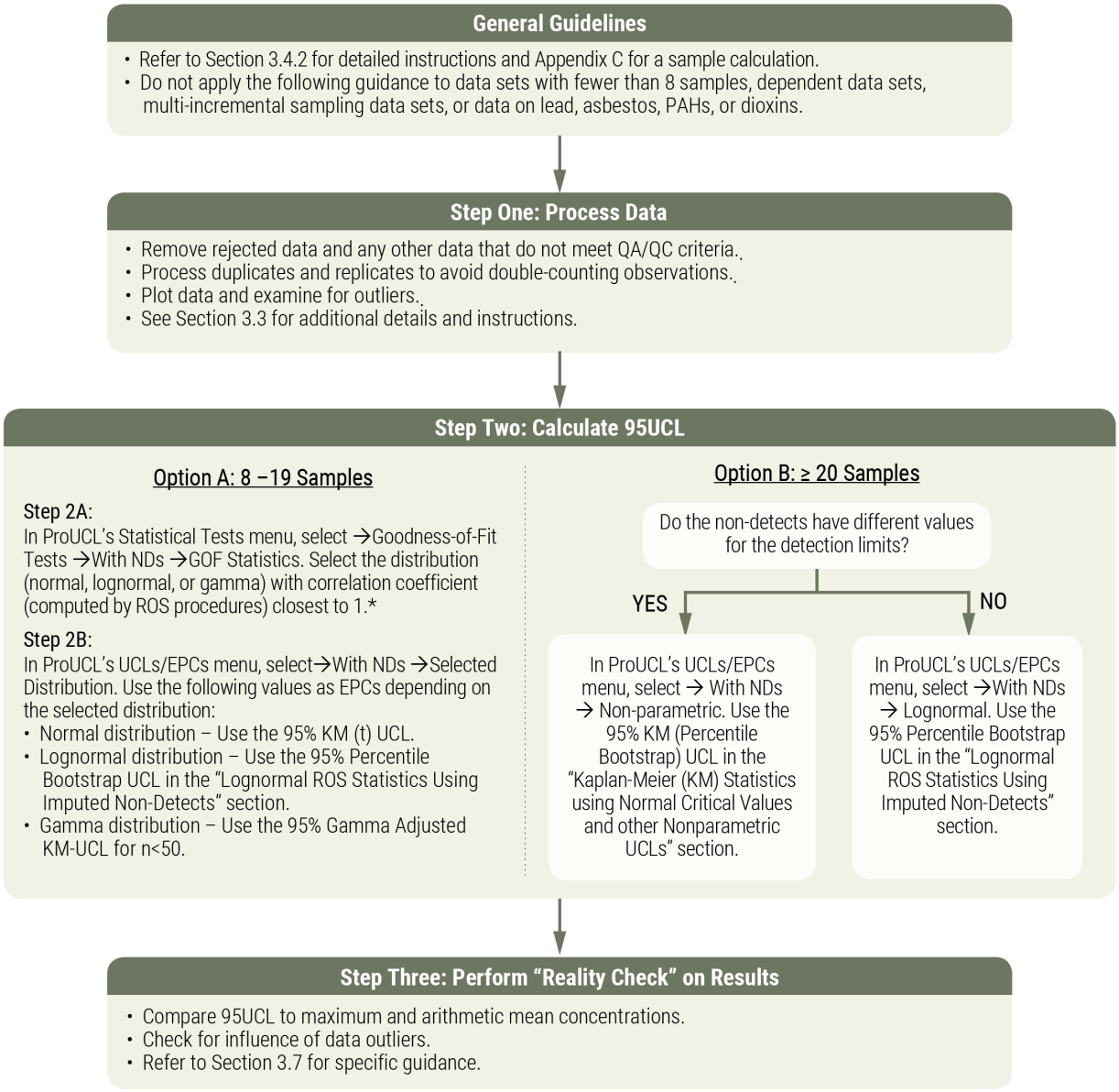
Figure 3. Using ProUCL to Calculate 95UCLs for Data Sets with at Least 8 Samples and No Non-Detects



* *If the highest correlation coefficient is less than 0.80, do not calculate a 95UCL and contact your ADS group for further direction. If the normal distribution is initially selected at this stage, first check to ensure that the left-end of the distribution does not include an unreasonable amount of negative values. If it does, then do not use the normal distribution and instead use whichever other distribution (lognormal or gamma) has a correlation coefficient closest to 1. See Section 3.4.1 for specific details.*

** *While the preferred method for lognormal distributions, the Chebyshev method sometimes gives very high estimates of the 95UCL. Read guidance in the corresponding text box in Section 3.4.1 before applying this method.*

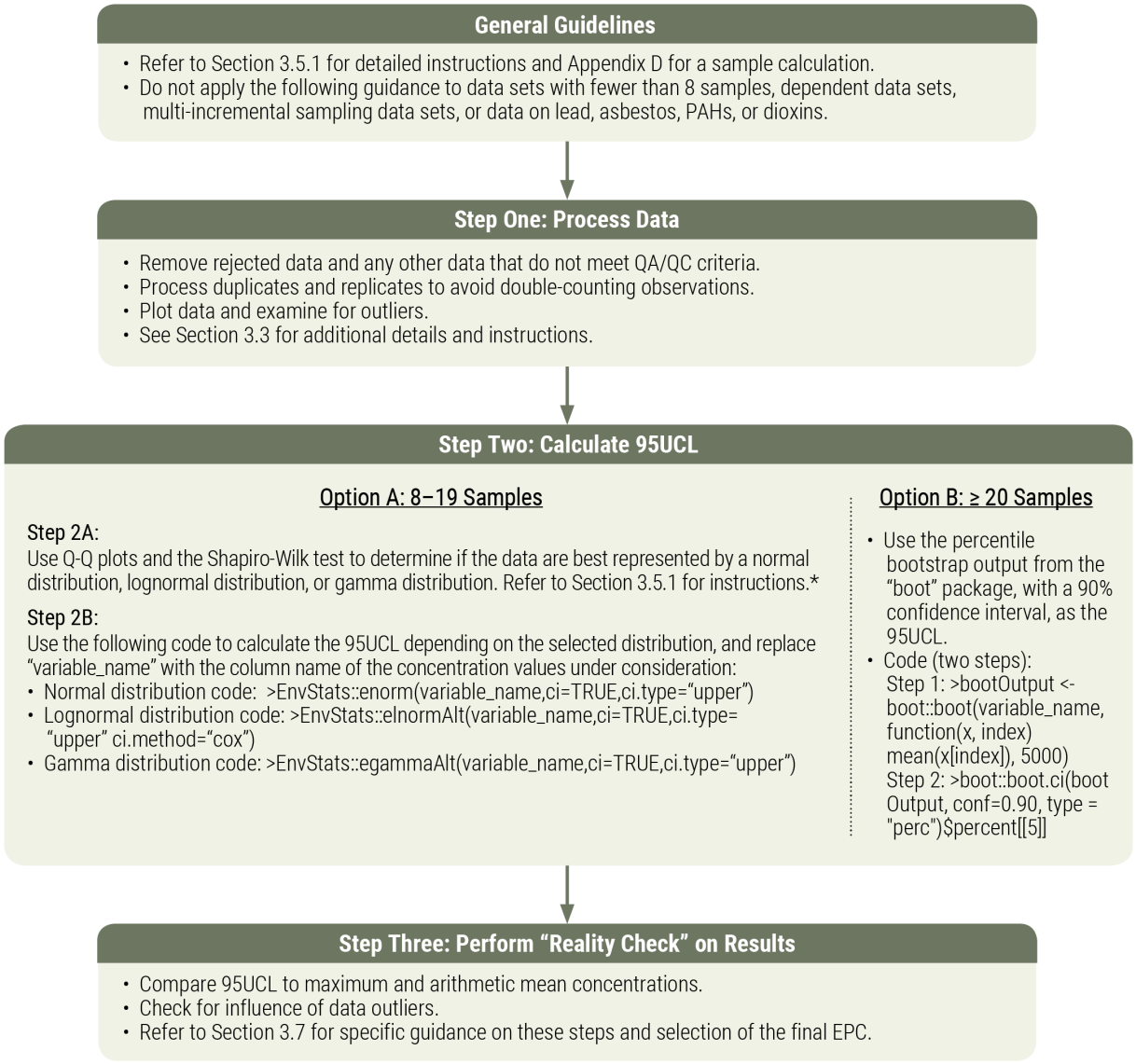
Figure 4. Using ProUCL to Calculate 95UCLs for Data Sets with at Least 8 Samples and Non-Detects



* *If the highest correlation coefficient is less than 0.80, do not calculate a 95UCL and contact your Branch Associate Director for Science for further direction. If the normal distribution is initially selected at this stage, first check to ensure that the left-end of the distribution does not include an unreasonable amount of negative values. If it does, then do not use the normal distribution and instead use whichever other distribution (lognormal or gamma) has a correlation coefficient closest to 1. See Section 3.4.2 for specific details.*

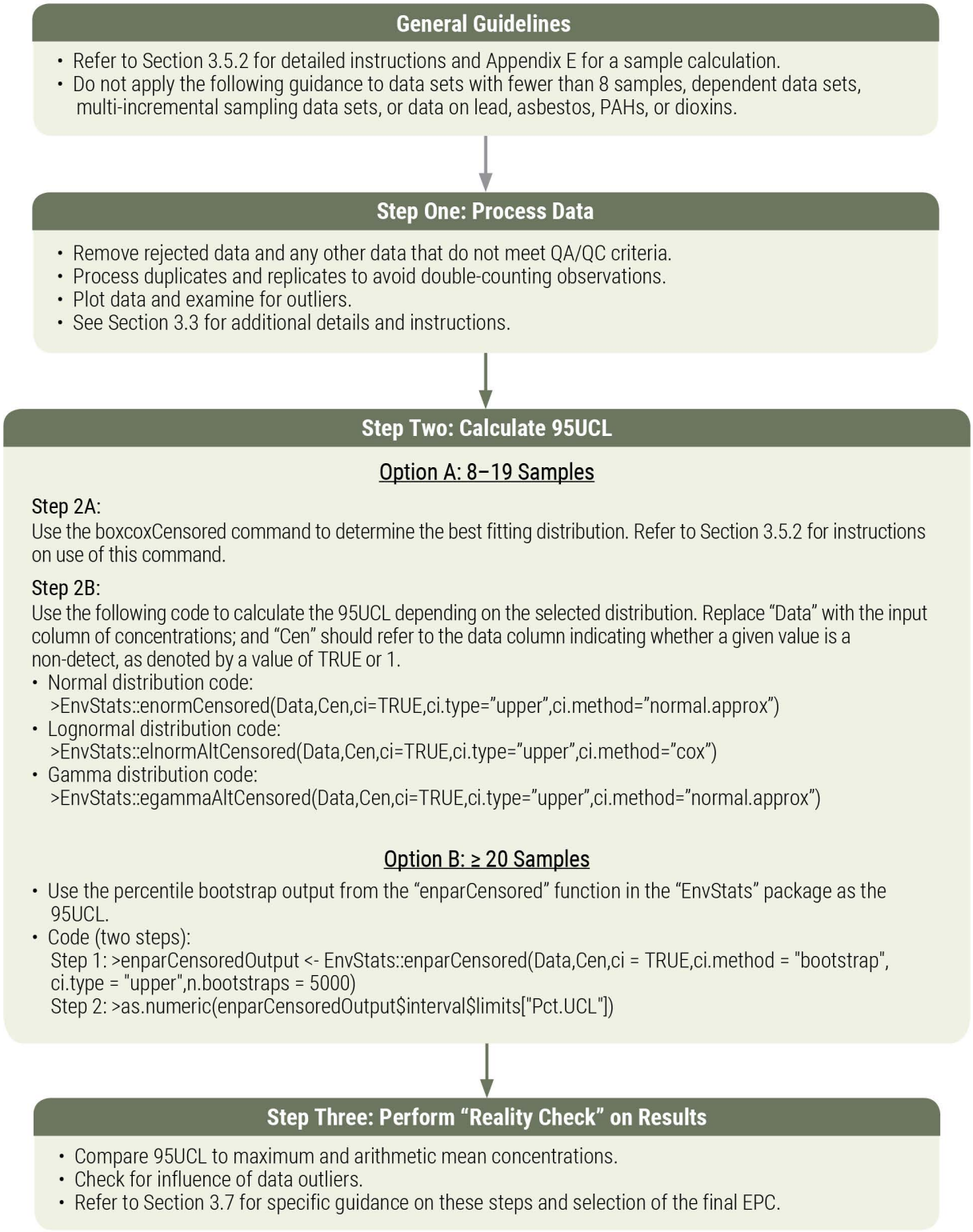
As a reminder, EPCs should only be calculated if there are at least four detected results, if 80% or less of the records are non-detects, and if there are at least three unique detected values. When any of these criteria are not met, use the maximum detected values as the EPC instead. ATSDR'S EPC Tool automates the second and third steps shown in this figure.

Figure 5. Using R to Calculate 95UCLs for Data Sets with at Least 8 Samples and No Non-Detects



** If the normal distribution is initially selected at this stage, first check to ensure that the left-end of the distribution does not include an unreasonable amount of negative values. If it does, then do not use the normal distribution and instead use whichever other distribution (lognormal or gamma) has a correlation coefficient closest to 1. See Section 3.5.1 for specific details.*

Figure 6. Using R to Calculate 95UCLs for Data Sets with at Least 8 Samples and Non-Detects



As a reminder, EPCs should only be calculated if there are at least four detected results, if 80% or less of the records are non-detects, and if there are at least three unique detected values. When any of these criteria are not met, use the maximum detected values as the EPC instead. ATSDR'S EPC Tool automates the second and third steps shown in this figure.

Appendix A. Glossary

95 Percent Upper Confidence Limit of the Arithmetic Mean (95UCL): The 95UCL is a calculated value that equals or exceeds an exposure unit's actual arithmetic mean of site concentrations 95 percent of the time. For a given number of discrete environmental samples in an exposure unit, the calculated arithmetic mean may be lower or higher than the actual arithmetic mean. However, it is highly unlikely (i.e., no more than 5 percent probability) that the 95UCL will be lower than the exposure unit's actual arithmetic mean. As the number of environmental samples in an exposure unit increases, the difference between the 95UCL and the sample arithmetic mean decreases. The 95UCL should not be confused with the 95^{th} percentile.

Arithmetic Mean: For environmental sampling, the arithmetic mean is the average of a set of sampling results. It is calculated by adding the measured concentrations from individual samples together and dividing the sum by the number of samples.

Bootstrapping: A statistical technique used to estimate statistics (e.g., 95UCLs) for data distributions. The technique relies on repeated random sampling with replacement from a collection of data points.

Censored Data: Censored data have partially unknown values. The censored value can be known to be less than a value (e.g. non-detect), greater than some value (e.g. above an instrument's response range), or between two values.

An example of a lower censoring boundary is the recording of pollutants in water. The researcher's instruments may not be able to reliably detect the level of a pollutant if it falls below a certain threshold (e.g., .005 parts per million). In this case, any pollutant level below .005 ppm is reported as "<.005 ppm." The <0.005 is a censored value because we are certain that the value is less than 0.005 ppm, but we are uncertain what the actual concentration is precisely.

Detection Limit: For environmental sampling, detection limits (often referred to as method detection limits) are thresholds below which measured concentrations are not significantly different from a blank signal, at a specified level of probability. Measurements above detection limits are evidence of a nonzero signal at a given probability, confirming that the analyte of interest is present in the sample.

Environmental Sample: A collected quantity of air, water, soil, food, or other media in which contamination levels are measured, whether directly in the field or at a laboratory.

Exposure Unit: Also called "area of exposure" or "exposure area," an exposure unit is a location and time where people may come into contact with contaminants in soil, water, air, or food. It is the result of exposure pathway evaluation and is a person's point of contact with contamination in a specific environmental medium. The general assumption is that all points and times within an exposure unit contribute equally to a person's or a group's exposure.

Exposure Pathway: An exposure pathway is the link between environmental releases and populations that might become exposed to environmental contamination. Refer to Chapter 6 in PHAGM for the five elements of an exposure pathway.

Exposure Point Concentration (EPC): The representative contaminant concentration within an exposure unit or area in an exposure pathway to which receptors are exposed for acute, intermediate, or chronic durations during the past, present, or future.

Gamma distribution: A probability distribution often used to characterize distributions of environmental contamination levels, in part because the distribution does not allow for negative concentration values. When fit to environmental data, gamma distributions typically have less of a "tail" at higher concentration values when compared to lognormal distributions.

Geometric Mean: The n^{th} root of the product of n observations. For lognormal distributions, the geometric mean estimates the population median and is less than the arithmetic mean for right-skewed distributions (Perkins 1997). Geometric means—or the 95 percent upper confidence limit of the geometric mean—are generally less protective than 95UCLs of arithmetic means and therefore should not be used for EPCs.

Goodness-of-fit: A term typically applied to statistical models to characterize how well the models match observed values. Common goodness-of-fit applications include statistical hypothesis tests and tests to assess whether data fit a given type of distribution.

Kaplan-Meier (KM): Kaplan-Meier is a standard method for calculating statistics (e.g., 95UCLs) for data sets including censored data. In the context of this guidance, the Kaplan-Meier method should primarily be applied when calculating 95UCLs for environmental sampling data sets that have non-detect observations with different values of the detection limit.

Lognormal distribution: A probability distribution in which the logarithms of the observed values are normally distributed. These distributions are often fit to environmental data sets, because the distribution does not allow for negative concentration values.

Non-parametric methods: Non-parametric methods are statistical methods that do not assume data have a theoretical shape or distribution. These methods use the shape of observed data to represent the underlying distribution, rather than a theoretical distribution such as the normal distribution. These methods are also called "distribution-free methods." In this guidance, health assessors are to use non-parametric methods to calculate 95UCLs for exposure units with 20 or more samples.

Normal distribution: A symmetrical probability distribution centered on the mean value, with the standard deviation indicating the spread of data. Normal distributions have finite probabilities for all numerical values, including negative values. Therefore, these are typically applied to environmental sampling data when the bulk of the distribution includes positive concentrations.

Outlier: An outlier is a single observation that appears markedly different (e.g., higher or lower) than a majority of the other observations in a given dataset and may be due to experimental error or extreme variability in measurements.

Parametric methods: Parametric methods are statistical methods based on an assumed shape or distribution of the underlying data. One must select a distribution that best fit available data before applying these methods. In this guidance, health assessors are to use parametric methods to calculate 95UCLs for exposure units having between 8 and 19 samples.

Regression on Order Statistics (ROS): ROS is a standard method for calculating summary statistics using least-squared regression of a probability plot. In the context of this guidance, the ROS method should primarily be applied when calculating 95UCLs for environmental sampling data sets that have non-detect observations with a single value for the detection limit.

Appendix B. Sample 95UCL Calculation: Using ProUCL for Data with No Non-detects

This example demonstrates the preferred approach for calculating 95UCLs using ProUCL for a data set with no non-detects. The data set considered in this example includes 13 surface soil samples collected from a single exposure unit that were analyzed for concentrations of copper. These data are presented in the table below. Health assessors who are using ProUCL for the first time are encouraged to replicate this example before using the software to calculate 95UCLs for site-specific scenarios. The remainder of this example follows the process outlined in Figure 3. For simplicity, only the steps completed in ProUCL are demonstrated here. All figures shown in this example are screen shots of actual ProUCL outputs, and the calculations were performed using ProUCL version 5.1.002 (5.1) (downloaded on August 1, 2016).

Example Data Set: Copper concentrations (ppm) in surface soil

Cu
93,500
1,050
1,210
82,700
118
1,320
11,300
194
2,110
246
1,800
2,130
217

Step One: Process Data

1. Remove rejected data and any other data that do not meet QA/QC criteria. Health assessors should do this prior to loading data to ProUCL.

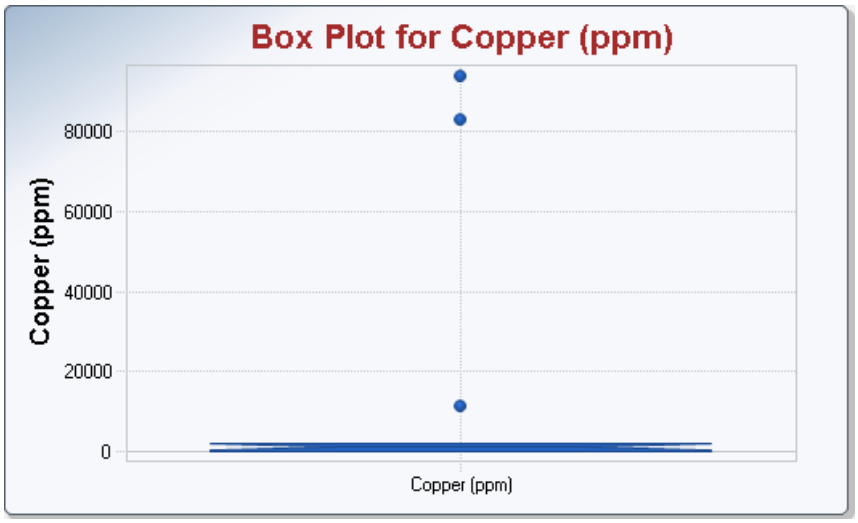
2. Process duplicates and replicates to avoid double-counting observations. Health assessors should do this prior to loading data to ProUCL.

3. Plot data and examine for outliers in ProUCL. Health assessors should do this after loading data to ProUCL.

   Health assessors can manually enter Site-specific data into ProUCL or import data from a Microsoft Excel file (.xls or .xlsx). To manually enter data, click on ProUCL's File menu, select → New, and click OK. A blank worksheet will appear in a new window where data can be entered. To import data from Excel, click on ProUCL's File menu, select → Open Single File Sheet, and

click OK. A worksheet with the data from Excel will appear in a new window. Users may also import Excel files with multiple worksheets through this same process.

In ProUCL's Graphs menu, select → Boxplot. Select the variable Copper from the left side, using the >> to move it to the right side. Click OK. The resulting boxplot is placed into a new window (as shown below).



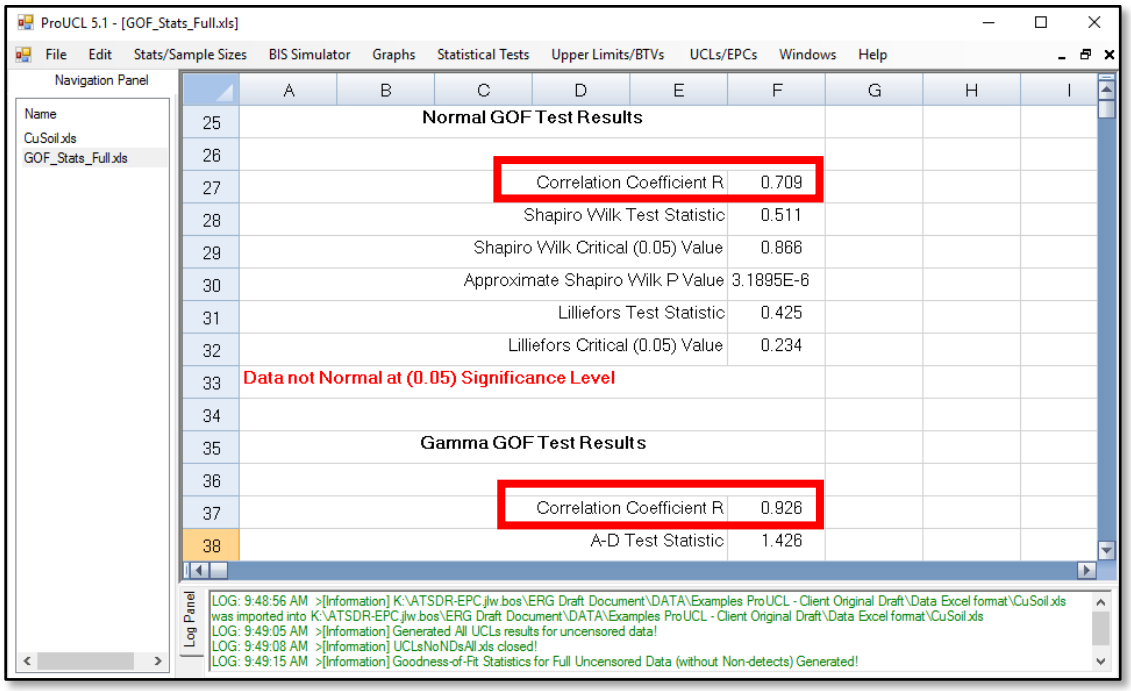
Plotting the data should help health assessors gain some insights into the range and variability among the environmental sampling data.[7] In this case, 10 of the 13 samples are clustered around concentrations of 1,000 ppm and three sampling results have considerably higher concentrations (93,500 ppm, 82,700 ppm, and 11,300 ppm). These concentrations are notably higher than the other 10 measurements, but they are still valid results. Without a science-based reason to exclude them, these higher values should be kept in the dataset for 95UCL calculations.

Step Two: Calculate 95UCL (Note: This example includes 13 samples and therefore follows the directions for Option A on Figure 3.)

Step 2A. Determine which distribution best fits the data

In ProUCL's Statistical Tests menu, select → Goodness-Of-Fit Tests → G.O.F. Statistics. Select the variable "Copper" from the left side, using the >> to move it to the right. Click OK. The resulting GOF output appears in a new window, and a partial copy of the GOF output appears in the figure below. Red boxes have been added to this figure to highlight key outputs.

[7] Health assessors may consider plotting logarithms of the original data if the data are found to be highly skewed and the health assessors would like to better understand the distribution of results. However, they are still encouraged to plot sampling results as an initial step for visualizing the full range of measured concentration.



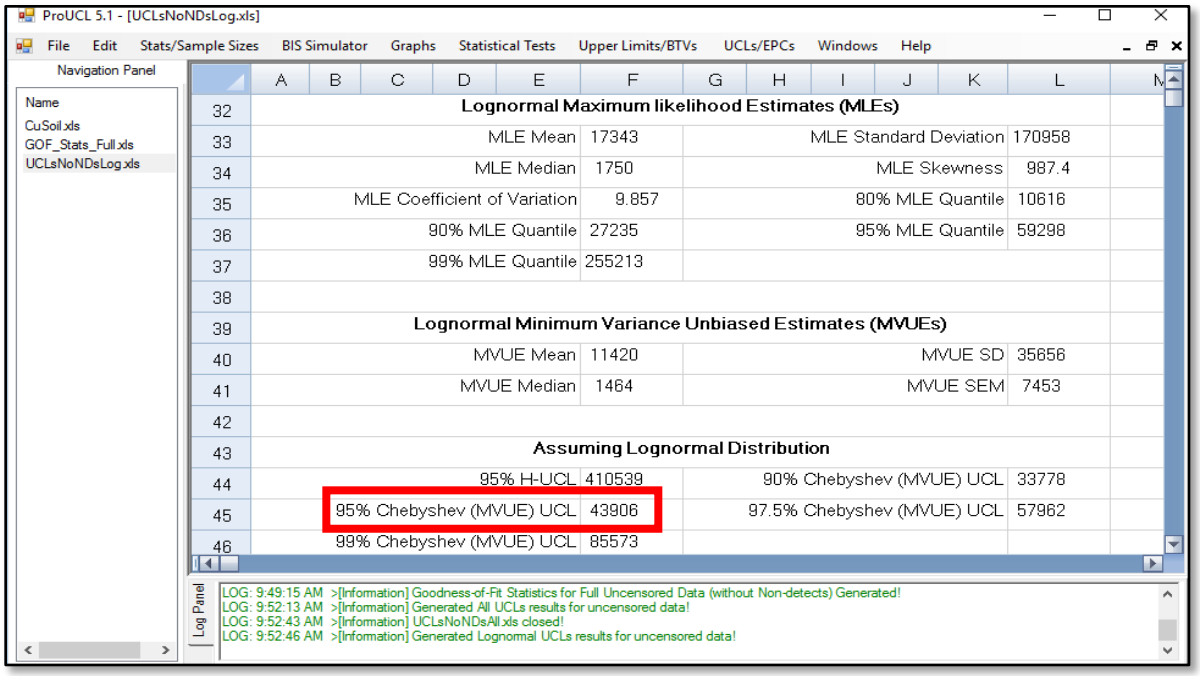
For this example, ProUCL reports the following correlation coefficients for the three different distributions of interest:

Correlation coefficient for the normal distribution = 0.709
Correlation coefficient for the gamma distribution = 0.926
Correlation coefficient for the lognormal distribution = 0.950

In this case, health assessors should select the lognormal distribution for the 95UCL calculation, because that distribution has the highest correlation coefficient. In addition, since this correlation coefficient is greater than 0.80, the data sufficiently fit to the selected distribution.

Step 2B. Select the 95UCL for the appropriate distribution

In ProUCL's UCLs/EPCs menu, select → selected distribution (in this example, lognormal). Select the variable Copper from the left side, using the >> to move it to the right. Click OK. The resulting UCLs/EPCs output is placed into a new window, and a partial copy of the GOF output appears in the figure below. Select the EPC based on the previously identified best-fitting distribution.



For exposure units having between 8 and 19 samples with no non-detects that are best fit by lognormal distributions, ATSDR's guidance (see Figure 3) is for health assessors using ProUCL to select the 95% Chebyshev UCL. In this example, the selected EPC is equal to 43,906 ppm.

Step Three: Perform "Reality Check" on Results

1. Compare the estimated 95UCL to maximum and arithmetic mean concentrations to ensure the calculated value is reasonable. To do this, health assessors should confirm that (1) the 95UCL is below the maximum concentration and (2) the 95UCL is greater than the mean, but not more than three times the mean.

   In this example, the estimated 95UCL (43,906 ppm) is below the maximum concentration of 93,500 ppm. The 95UCL (43,906 ppm) is greater than the arithmetic mean of 15,223 ppm, as it should be, and less than three times so. These comparisons confirm that the selected 95UCL is reasonable for use.

2. Check for influence of data outliers. Because the 95UCL is less than three times the mean, the health assessor need not investigate influence of outliers.

Appendix C. Sample 95UCL Calculation: Using ProUCL for Data with Non-detects

This example demonstrates the preferred approach for calculating 95UCLs using ProUCL for a dataset with non-detects. The data set considered in this example is 19 groundwater samples collected from a single exposure unit that were analyzed for concentrations of arsenic. The data set includes 11 detections and 8 non-detects. These data are presented in the table below. Health assessors who are using ProUCL for the first time are encouraged to replicate this example before using the software to calculate 95UCLs for site-specific scenarios. The remainder of this example follows the process outlined in Figure 4. For simplicity, only the steps completed in ProUCL are demonstrated here. All figures shown in this example are screen shots of actual ProUCL outputs, and the calculations were performed using ProUCL version 5.1.002 (downloaded on August 1, 2016).

Example Data Set: Arsenic concentrations (µg/L) in groundwater

Arsenic	d_Arsenic
4.0	0
4.20	1
0.55	1
4.0	1
3.50	1
0.83	1
4.0	0
1.56	1
4.0	0
4.0	0
4.0	0
0.5	0
2.00	1
3.56	1
4.0	0
4.0	0
4.85	1
5.10	1
1.25	1

Note that when working with data including non-detects in ProUCL, health assessors must code each result as a detect or non-detect. Chemical concentrations should be listed in a column with the chemical name as a header. A second column is added to identify detected concentrations and non-detects. The required approach is to enter “d_[chemical name]” as the header—in this case, the header would be “d_Arsenic.” This column then has entries of 0 for non-detects and 1 for detected results. This coding is shown in the table above.

Step One: Process Data

1. Remove rejected data and any other data that do not meet QC/QC criteria. Health assessors should do this prior to loading data to ProUCL.

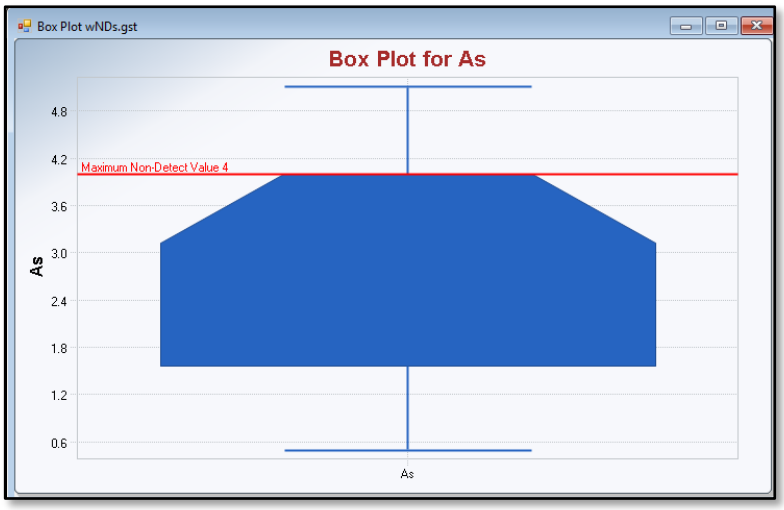
2. Process duplicates and replicates to avoid double-counting observations. Health assessors should do this prior to loading data to ProUCL.

3. Confirm that there are at least four detected values and that no more than 80 percent of the sample results are non-detects.

4. Plot data and examine for outliers. Health assessors should do this after loading data to ProUCL.

   Health assessors can manually enter site data into ProUCL or import data from a Microsoft Excel file (.xls or .xlsx). To manually enter data, click on ProUCL's File menu, select → New, and click OK. A blank worksheet will appear in a new window where data can be entered. To import data from Excel, click on ProUCL's File menu, select → Open Single File Sheet, and click OK. A worksheet with the data from Excel will appear in a new window. Users may also import Excel files with multiple worksheets through this same process.

   Once the data are entered, click ProUCL's Graphs menu, select → Boxplot → With NDs. Select the variable "Arsenic" from the left side, using the >> to move it to the right. Click OK. The resulting boxplot is placed into a new window (as shown below).

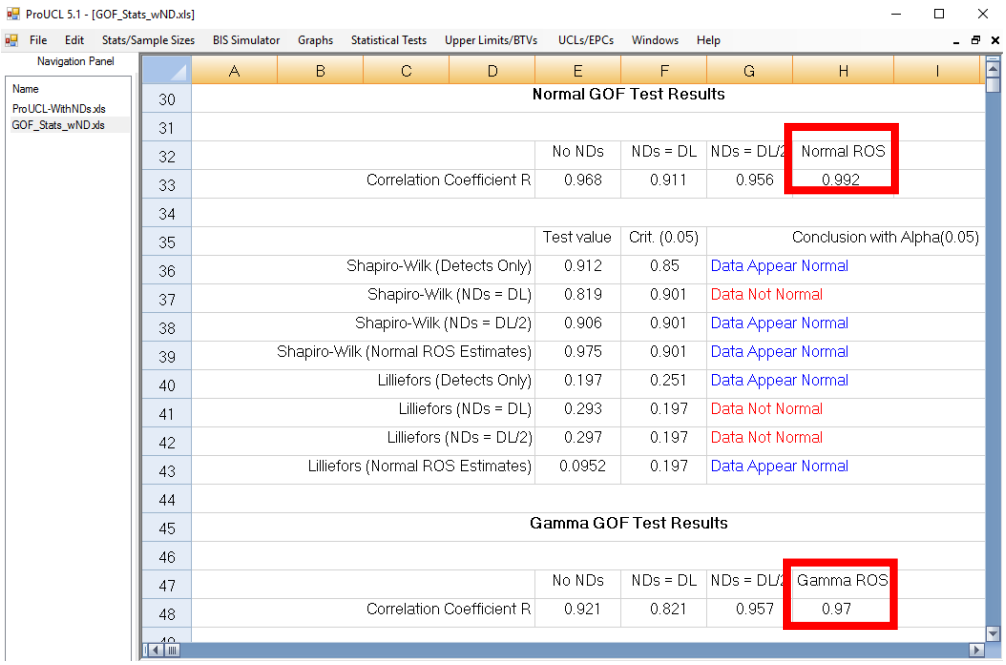


   Plotting the data should help health assessors gain some insights into the range and variability among the environmental sampling data. In this example, the highest detection limit is drawn as the horizontal line at 4 µg/L. There are no outliers above the box whisker and the data are not highly skewed.

Step Two: Calculate 95UCL (Note: This example includes 19 samples and follows the directions for Option A).

Step 2A. Determine which distribution best fits the data

In ProUCL's Statistical Tests menu, select → Goodness-Of-Fit Tests → With NDs → G.O.F. Statistics. Select the variable "Arsenic" from the left side, using the >> to move it to the right. Click OK. The resulting GOF output is placed into a new window and a partial copy of the GOF output appears in the figure below. Red boxes have been added to this figure to highlight key outputs. Select the distribution from this output with the highest correlation coefficient based on ROS imputation.



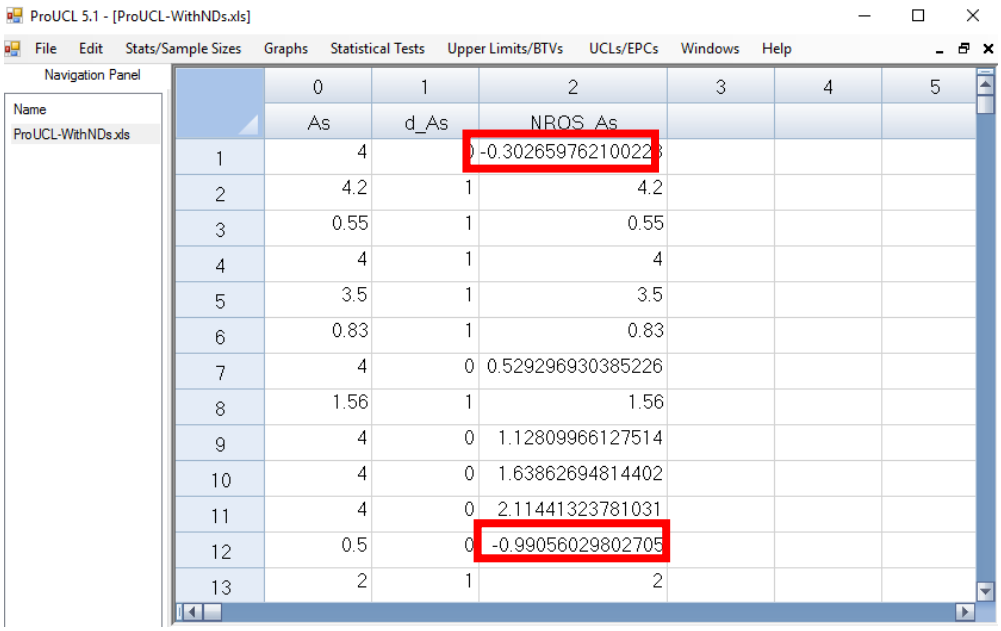
	A	B	C	D	E	F	G	H	I
30	Normal GOF Test Results								
31									
32					No NDs	NDs = DL	NDs = DL/2	Normal ROS	
33				Correlation Coefficient R	0.968	0.911	0.956	0.992	
34									
35					Test value	Crit. (0.05)	Conclusion with Alpha(0.05)		
36				Shapiro-Wilk (Detects Only)	0.912	0.85	Data Appear Normal		
37				Shapiro-Wilk (NDs = DL)	0.819	0.901	Data Not Normal		
38				Shapiro-Wilk (NDs = DL/2)	0.906	0.901	Data Appear Normal		
39				Shapiro-Wilk (Normal ROS Estimates)	0.975	0.901	Data Appear Normal		
40				Lilliefors (Detects Only)	0.197	0.251	Data Appear Normal		
41				Lilliefors (NDs = DL)	0.293	0.197	Data Not Normal		
42				Lilliefors (NDs = DL/2)	0.297	0.197	Data Not Normal		
43				Lilliefors (Normal ROS Estimates)	0.0952	0.197	Data Appear Normal		
44									
45	Gamma GOF Test Results								
46									
47					No NDs	NDs = DL	NDs = DL/2	Gamma ROS	
48				Correlation Coefficient R	0.921	0.821	0.957	0.97	

For this example, ProUCL reports the following correlation coefficients for the three different distributions of interest:

Correlation coefficient for the normal distribution with ROS imputation = 0.992
Correlation coefficient for the gamma distribution with ROS imputation = 0.970
Correlation coefficient for the lognormal distribution with ROS imputation = 0.983

In this case, health assessors should initially select the normal distribution for the 95UCL calculation, because the normal distribution has the highest correlation coefficient (with ROS imputation for non-detects) for the arsenic sampling data. In addition, since this correlation coefficient is greater than 0.80, the data sufficiently fit to the selected distribution.

Since the normal distribution was initially selected, health assessors must then verify that the left-end of the distribution does not include an unreasonable amount of negative values. In ProUCL's Stats/Sample Sizes menu select → Imputed NDs with using ROS Methods → Normal ROS. Select the variable "Arsenic" from the left side, using the >> to move it to the right. Click OK. The resulting Normal ROS estimates are written to a new column in the existing worksheet, along with the original observations. A partial view of these estimates is shown in the figure below, and red boxes have been added to highlight key points.

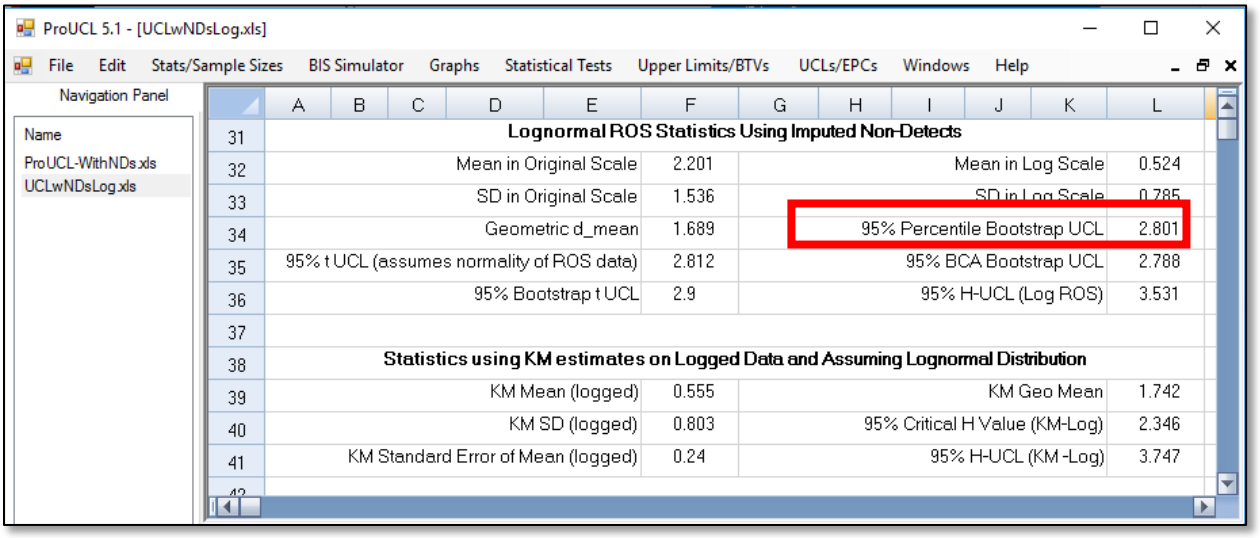


	0	1	2	3	4	5
	As	d_As	NROS_As			
1	4	0	-0.30265976210022			
2	4.2	1	4.2			
3	0.55	1	0.55			
4	4	1	4			
5	3.5	1	3.5			
6	0.83	1	0.83			
7	4	0	0.529296930385226			
8	1.56	1	1.56			
9	4	0	1.12809966127514			
10	4	0	1.63862694814402			
11	4	0	2.11441323781031			
12	0.5	0	-0.99056029802705			
13	2	1	2			

In this example, two of the estimated values are negative. Based on the guidance presented earlier in this document, the normal distribution should not be used for 95UCL calculations when any of the estimated values are negative, and the distribution with the next highest correlation coefficient should be selected. For this example, the lognormal distribution has the next highest correlation coefficient and is therefore used to determine the 95UCL.

Step 2B. Select the 95UCL for the appropriate distribution

In ProUCL's UCLs/EPCs menu, select → With NDs → selected distribution (in this example, lognormal). Select the variable "Arsenic" from the left side, using the >> to move it to the right. Click OK. The resulting UCLs/EPCs output is placed into a new window and a partial copy of this output is shown in the figure below. Health assessors should determine the EPC based on the previously selected distribution and the guidance shown in Figure 4.



For exposure with 8 to 19 samples and non-detects that are best fit by lognormal distributions, ATSDR's guidance is for health assessors to select the 95% Percentile Bootstrap UCL (see Figure 4). In this example, the selected EPC is 2.801 µg/L. Note that because this 95UCL is estimated with bootstrapping simulations, this value will vary slightly each time the procedure is run.

Step Three: Perform "Reality Check" on Results

1. Compare the 95UCL to the maximum and arithmetic mean concentrations. To do this, health assessors should confirm that (1) the 95UCL is below the maximum concentration and (2) the 95UCL is greater than the mean, but not more than three times the mean.

   In this example, the estimated 95UCL (2.801 µg/L) is below the maximum detected concentration of 5.1 µg/L. The 95UCL (2.801 µg/L) is greater than the estimated mean of 2.201 µg/L shown on the ProUCL output (see entry for "Mean in Original Scale") but not more than three times so. These comparisons confirm that the selected 95UCL is reasonable for use.

2. Check for influence of data outliers. Because the 95UCL is less than three times the mean, the health assessor need not investigate influence of outliers.

Appendix D. Sample 95UCL Calculation: Using R for Data with No Non-detects

This example demonstrates the preferred approach for calculating 95UCLs using R for a data set with no non-detects. The data set considered in this example includes 13 surface soil samples collected from a single exposure unit that were analyzed for concentrations of copper. These data are presented in the table below. Health assessors who are using R for the first time are encouraged to replicate this example before using R software to calculate 95UCLs for site-specific scenarios. The remainder of this example follows the process outlined in Figure 5. For simplicity, only the steps completed in R are demonstrated here. All figures shown in this example are screen shots of actual R outputs, and the calculations were performed using R for Windows Version 3.3.1 (downloaded on August 15, 2016). Health assessors must install and load the EnvStats and Bootstrap packages to R to estimate 95UCLs for data sets without non-detects.

Details on the EnvStats R package can be found here:
https://cran.r-project.org/web/packages/EnvStats/EnvStats.pdf

Details on the Bootstrap R package can be found here:
https://cran.r-project.org/web/packages/bootstrap/bootstrap.pdf

Example Data Set: Copper concentrations (ppm) in surface soil

Copper
93,500
1,050
1,210
82,700
118
1,320
11,300
194
2,110
246
1,800
2,130
217

Step One: Process Data

1. Remove rejected data and any other data that do not meet QA/QC criteria. Health assessors should do this prior to loading data to R.

2. Process duplicates and replicates to avoid double-counting observations. Health assessors should do this prior to loading data to R.

3. Plot data and examine for outliers in R. Health assessors should do this after loading data to R.

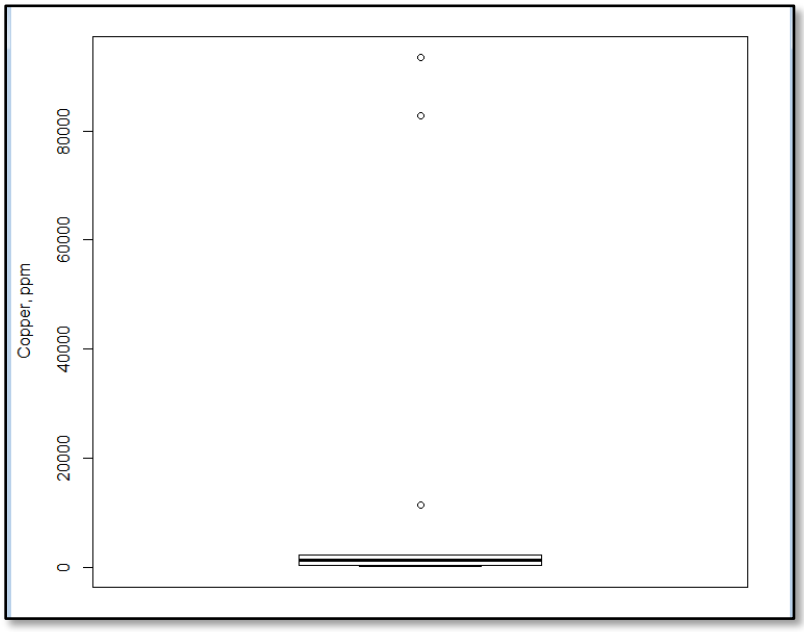
Health assessors can automatically load environmental sampling data in CSV or TXT file formats into R. In this example, a CSV file is used. Use the following code to read the CSV file into R (note that in this example, the CSV file is titled Cu.csv):

```
> Cu=read.csv(file=file.choose())
> head(Cu)
> attach(Cu)
```

Once the data are loaded into R, the following code will create a boxplot of the measured copper concentrations (note that in this example, the chemical heading in the CSV file is "Copper"):

```
> boxplot(Copper,ylab="Copper, ppm")
```

The resulting boxplot is opened in a new window and shown in the figure below.



Plotting the data should help health assessors gain some insights into the range and variability among the environmental sampling data. In this case, most of the samples are clustered around concentrations of 1,000 ppm and three sampling results had considerably higher concentrations (93,500 ppm, 82,700 ppm, and 11,300 ppm). These concentrations are notably higher than the other 10 measurements, but they are still valid results. Without a science-based reason to exclude them, these higher values should be kept in the dataset for 95UCL calculations.

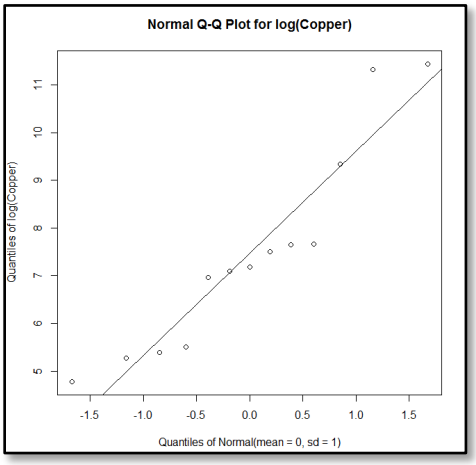
Step Two: Calculate 95UCL (Note: This example includes 13 samples and therefore follows the directions for Option A.)

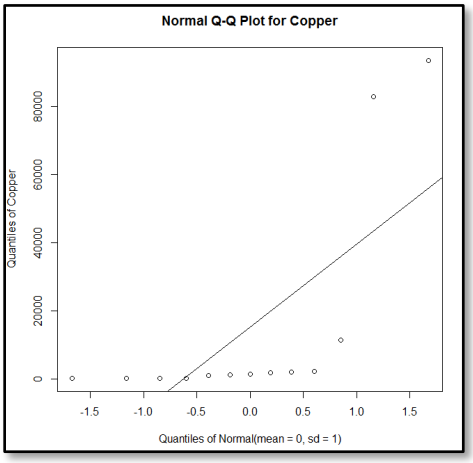
Step 2A. Use Q-Q plots and the Shapiro Wilk test to determine if the data are best represented by a normal distribution, lognormal distribution, or gamma distribution

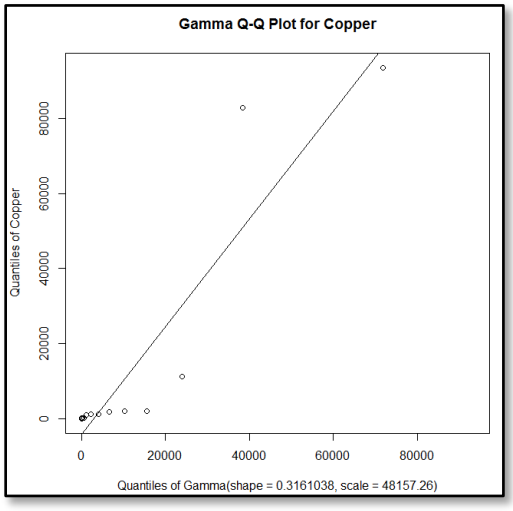
Q-Q Plots: Health assessors are encouraged to first visualize data by plotting Q-Q probability plots for the three distributions (normal, lognormal, and gamma distributions). This step is not mandatory for 95UCL calculation, but it is advised such that health assessors can see potentially important data trends. The following code will create Q-Q plots for each distribution (note that in this example, the chemical heading in the CSV file is "Copper"):

```
> qqPlot(Copper, dist="norm",add.line=T)
> qqPlot(log(Copper), dist="norm",add.line=T)
> qqPlot(Copper, dist="gamma", estimate.params=TRUE, add.line=T)
```

The resulting Q-Q plots will appear in new windows (see below). Based on visual inspection alone, the lognormal distribution appears to be the best fit because the data points are closest to the line in the Q-Q plot. However, health assessors should not select best-fitting distributions based on visual inspection, and should instead rely on statistical tests, as described below.





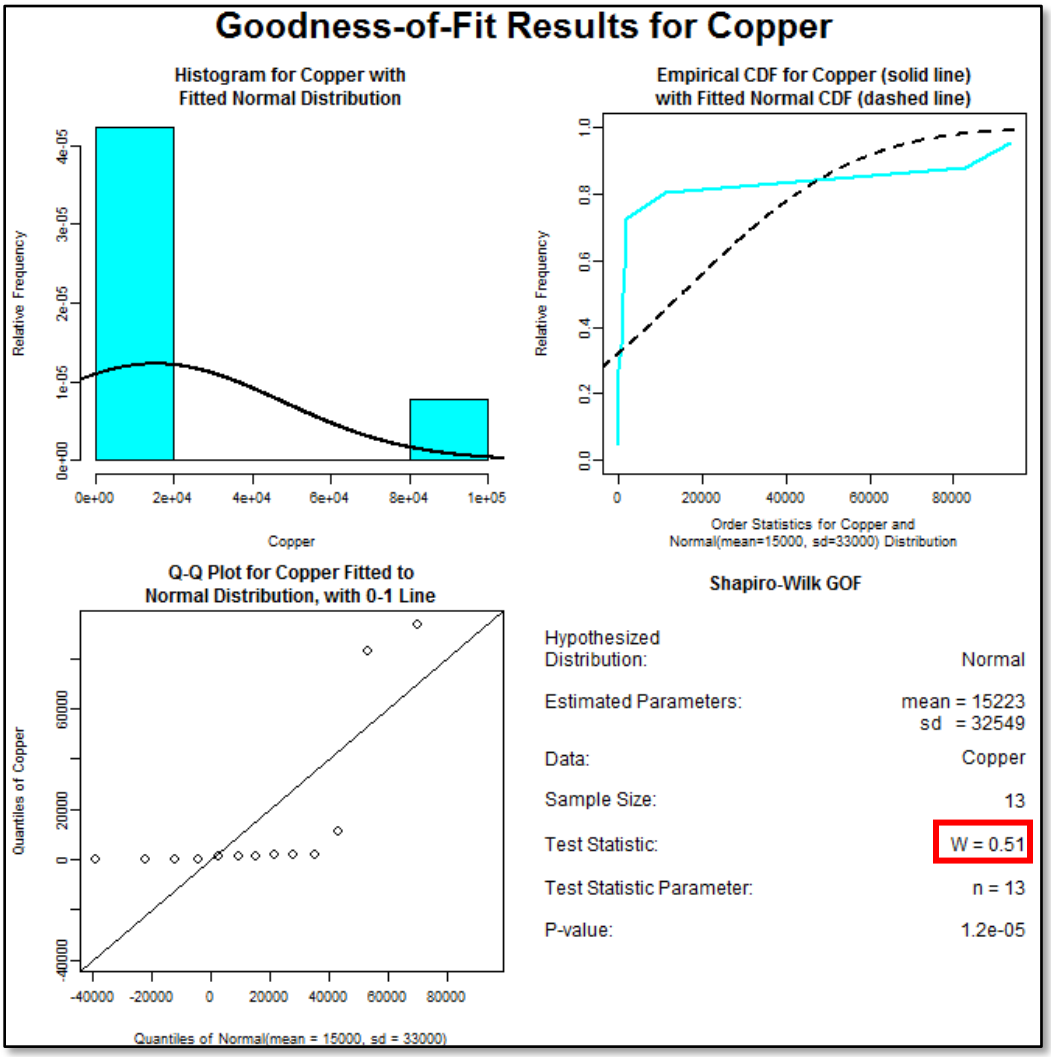


Shapiro-Wilk Test: Health assessors should select the best fitting distribution based on the W test statistic from the Shapiro-Wilk goodness-of-fit (GOF) test. Use the code below to run the Shapiro-Wilk hypothesis test for each distribution of interest:

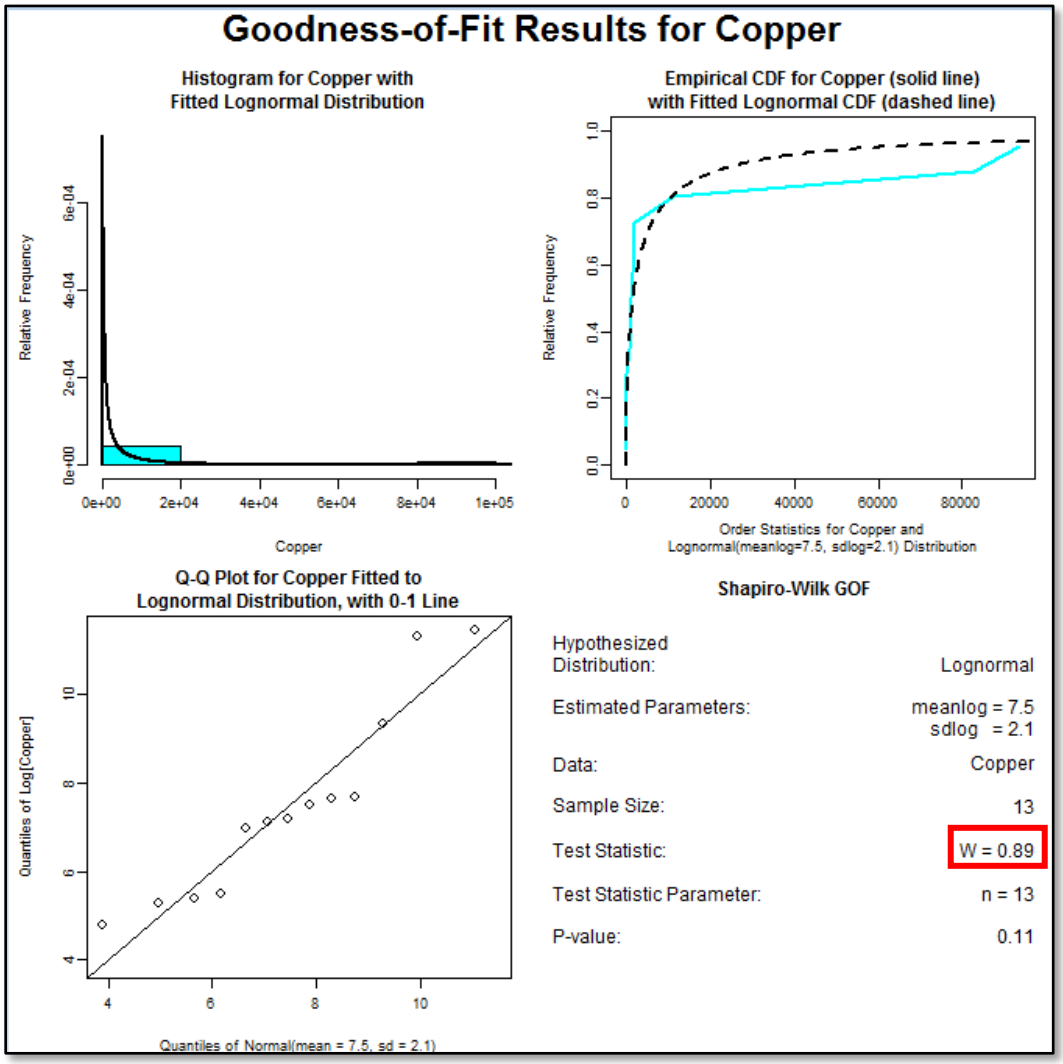
```
> plot(gofTest(Copper, dist="norm"))
> plot(gofTest(Copper, dist="lnorm"))
> plot(gofTest(Copper, dist="gamma"))
```

The results appear in separate windows for each distribution and are shown below. The W Test Statistic appears at the bottom right of each window. Red boxes have been added to these outputs to illustrate where health assessors can find the necessary information. Note that health assessors may omit the "plot" portion of the aforementioned commands if they are only interested in the text output of the Shapiro-Wilk hypothesis test.

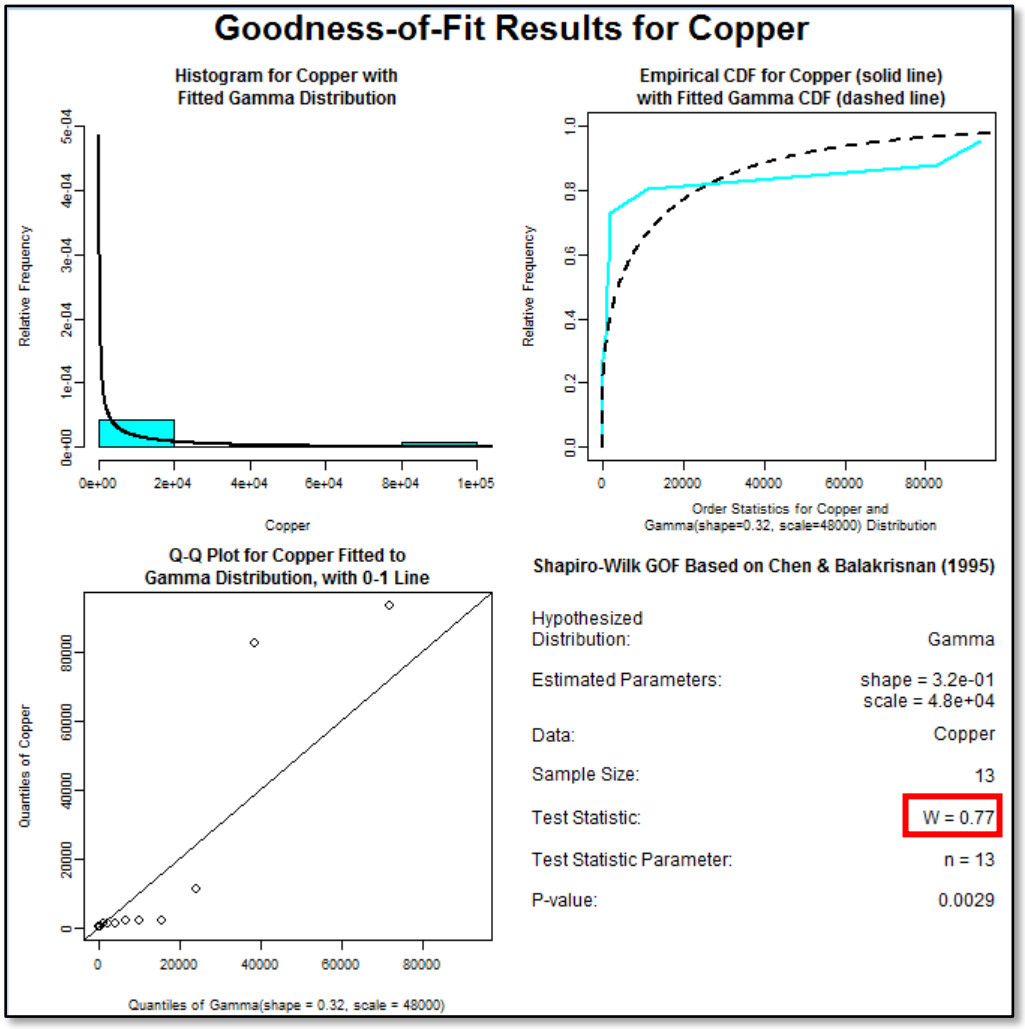
Output window for the normal distribution:



Output window for the lognormal distribution:



Output window for the gamma distribution:



For this example, R reports the following W test statistics from the Shapiro-Wilk GOF hypothesis test for the three distributions of interest:

W test statistic for the normal distribution = 0.51
W test statistic for the lognormal distribution = 0.89
W test statistic for the gamma distribution = 0.77

In this case, health assessors should select the lognormal distribution for the 95UCL calculation, because the lognormal distribution has the highest W test statistic for the copper sampling data.

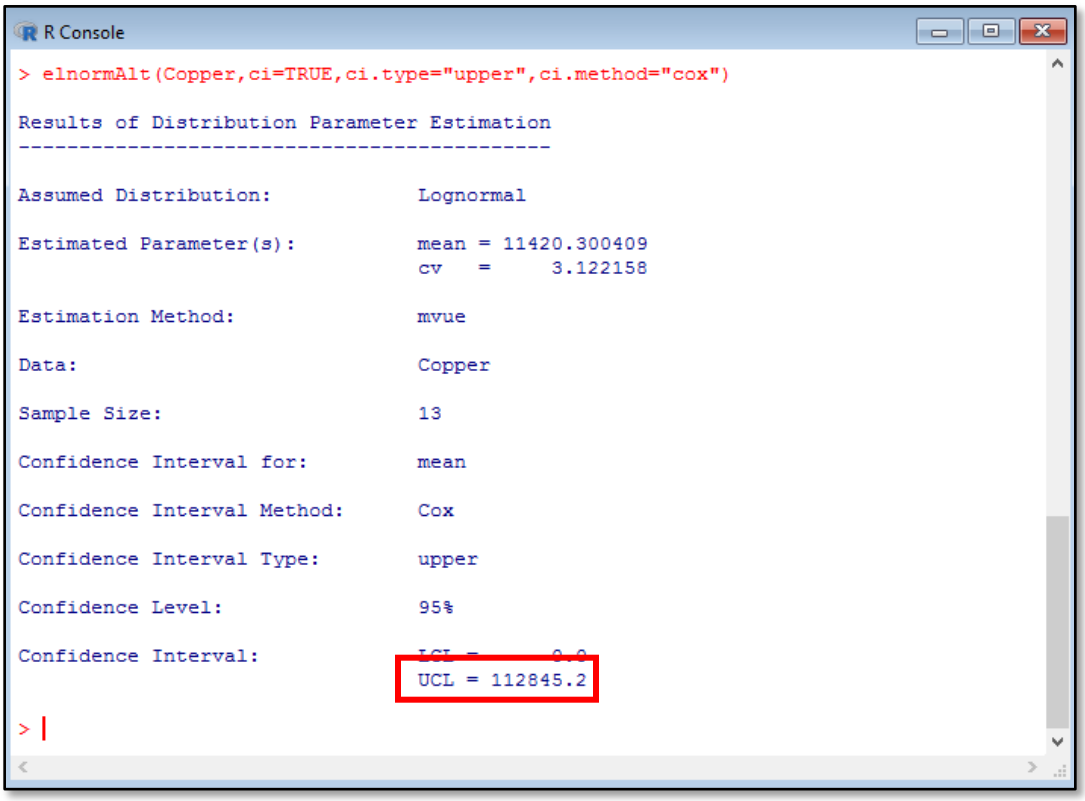
Step 2B. Calculate 95UCL depending on the selected distribution

For exposure units having between 8 and 19 samples with no non-detects that are best fit by lognormal distributions, ATSDR's guidance is for health assessors using R to calculate 95UCL

using Cox methods (see Figure 5). Use the following code to calculate the 95UCL for a lognormal distribution with "cox" methods:

```
> elnormAlt(Copper,ci=TRUE,ci.type="upper",ci.method="cox")
```

The results for this command appear in the same window as the written R code and are shown below.



The lognormal estimate of the 95UCL with Cox methods is 112,845 ppm, which is the value that should be used as the EPC.

Step Three: Perform "Reality Check" on Results

1. Compare the estimated 95UCL to maximum and arithmetic mean concentrations to ensure the calculated value is reasonable. To do this, health assessors should confirm that (1) the 95UCL is below the maximum concentration and (2) the 95UCL is greater than the mean, but not more than three times the mean.

   In this example, the estimated 95UCL (112,845 ppm) is above the maximum concentration of 93,500 ppm. As described in Section 3.7, when a data set includes 8 to 19 samples and the calculated 95UCL exceeds the maximum, health assessors should still use the 95UCL as the EPC.

   The 95UCL (112,845 ppm) is also greater than the arithmetic mean of 15,223 ppm, as it should be. However, the 95UCL is more than three times the estimated mean. These comparisons

suggest that the selected 95UCL may not be reasonable for use, because the calculated 95UCL is higher than every single measurement considered in this example. This outcome likely results from the relatively small sample size and high data variability.

2. Check for influence of data outliers. Because the 95UCL is more than three times the average concentration, the health assessor should investigate how sensitive the 95UCL value is to apparent outliers. This is typically done by seeing how the 95UCL changes by removing unusually high measurements. To illustrate this point, the health assessor should consider recalculating the 95UCL by excluding the two highest copper sampling results (93,500 ppm and 82,700 ppm). This was done here for illustrative purposes only to show health assessors how highly variable data affect 95UCL calculations. The revised data set was still found to best fit a lognormal distribution (W test for normal distribution = 0.56, W test statistic for lognormal distribution = 0.92, and W test statistic for gamma distribution = 0.87). As expected, excluding these two data points resulted in a lower mean value of 1,972 ppm and a lower 95UCL from Cox methods of 6,096 ppm—more than a factor of 18 lower than the 95UCL calculated from the entire data set. In cases where the 95UCL is highly sensitive to a small number of measurements, health assessors should consider recommending additional sampling, especially if the original 95UCL was at an exposure level of potential health concern. Health assessors should explain in their report the thought process for the EPC calculation and any associated limitations (e.g., the fact that 95UCLs are highly sensitive to two measurements).

Appendix E. Sample 95UCL Calculation: Using R for Data with Non-detects

This example demonstrates the preferred approach for calculating 95UCLs using R for a data set with non-detects. The data set considered in this example is 19 groundwater samples collected from a single exposure unit that were analyzed for concentrations of arsenic. The data set includes 11 detections and 8 non-detects. These data are presented in the table below. Health assessors who are using R for the first time are encouraged to replicate this example before using R software to calculate 95UCLs for site-specific scenarios. The remainder of this example follows the process outlined in Figure 6. For simplicity, only the steps completed in R are demonstrated here. All figures shown in this example are screen shots of actual R outputs, and the calculations were performed using R for Windows Version 3.3.1 (downloaded on August 15, 2016). Health assessors must install and load the EnvStats and NADA packages to R to estimate 95UCLs for data sets with non-detects.

Details on the EnvStats R package can be found here:
https://cran.r-project.org/web/packages/EnvStats/EnvStats.pdf

Details on the NADA R package can be found here:
https://cran.r-project.org/web/packages/NADA/NADA.pdf

Example Data Set: Arsenic Concentrations (µg/L) in Groundwater (txt formatting)

As	AsCen	Ascen
4.00	1	TRUE
4.20	0	FALSE
0.55	0	FALSE
4.00	0	FALSE
3.50	0	FALSE
0.83	0	FALSE
4.00	1	TRUE
1.56	0	FALSE
4.00	1	TRUE
4.00	1	TRUE
4.00	1	TRUE
0.50	1	TRUE
2.00	0	FALSE
3.56	0	FALSE
4.00	1	TRUE
4.00	1	TRUE
4.85	0	FALSE
5.10	0	FALSE
1.25	0	FALSE

Data for R should be formatted as CSV or TXT files and include the information shown above. Chemical concentrations should be listed in a column with the chemical name as the header. A

second column must be added to identify censored data. This column should be titled with the chemical name and "Cen", and filled in with 1 for non-detects and 0 for detected results. A third column must be added to further identify censored data for different commands. This column should be titled with the chemical name and "cen," and filled in with TRUE for non-detects and FALSE for detected results.

Step One: Process Data

1. Remove rejected data and any other data that do not meet QA/QC criteria. Health assessors should do this prior to loading data to R.

2. Process duplicates and replicates to avoid double-counting observations. Health assessors should do this prior to loading data to R.

3. Confirm that there are at least four detected values and that no more than 80 percent of the sample results are non-detects.

4. Plot data and examine for outliers in R. Health assessors should do this after loading data to R.

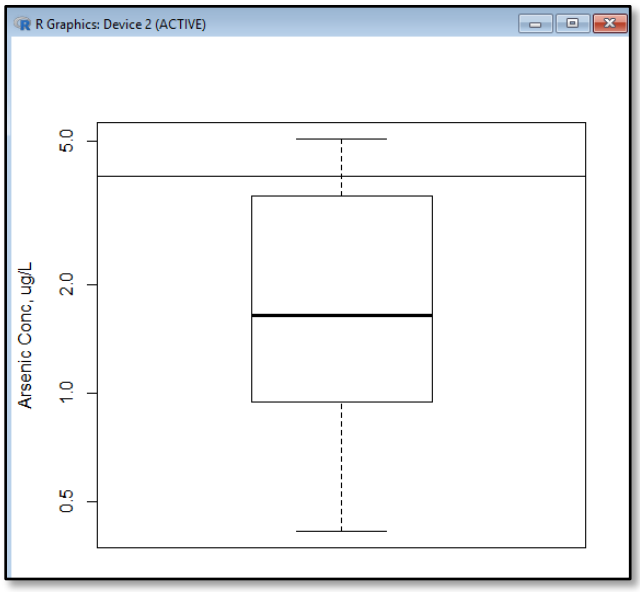
   Health assessors can load environmental sampling data into to R by importing CSV or TXT files. In this example, we demonstrate the use of TXT files, which can be loaded into R using the following code (note that in this example, the TXT file is titled "Example1.txt"):

```
> Example1=read.table(file=file.choose(),header=T)
> attach(Example1)
> head(Example1)
```

   Once the data are loaded into R, use the following command to create a boxplot of the measured arsenic concentrations (note that in this example, the chemical is titled "As" within the TXT file):

```
> cenboxplot(As,Ascen,ylab="Arsenic Conc, ug/L")
```

   The resulting boxplot is opened in a new window:



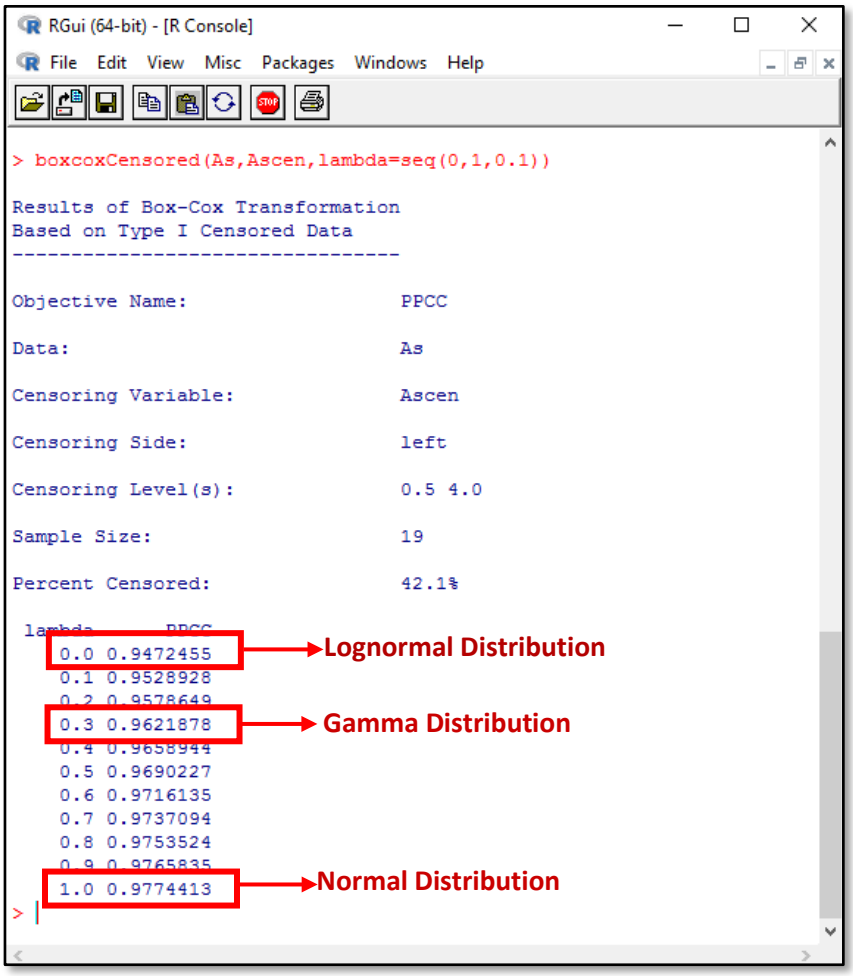
Plotting the data is encouraged so health assessors can visualize the range and variability among the sampling results. Note that the highest detection limit is drawn as the horizontal line at 4 µg/L. There are no outliers above the box-whisker and the data are therefore not highly skewed.

Step Two: Calculate 95UCL (Note: This example includes 19 samples and therefore follows the directions for Option A in Figure 6.)

Step 2A. Use the boxcoxCensored command to determine the best fitting distribution

Health assessors should use the boxcoxCensored command to determine whether the data best fit normal, lognormal, or gamma distributions. When using this command, a lambda of 1 represents a normal distribution, a lambda of 0 represents a lognormal distribution, and a lambda of approximately 0.3 represents a gamma distribution. Health assessors should select the distribution with the highest PPCC statistic. If the normal distribution is initially selected, health assessors must confirm that the normal distribution does not include an unreasonable number of negative values. Use the following command to obtain PPCC statistics for the three distributions of interest. The command instructs R to calculate PPCC statistics for lambda values between 0 and 1 at increments of 0.1.

```
> boxcoxCensored(As,Ascen,lambda=seq(0,1,0.1))
```



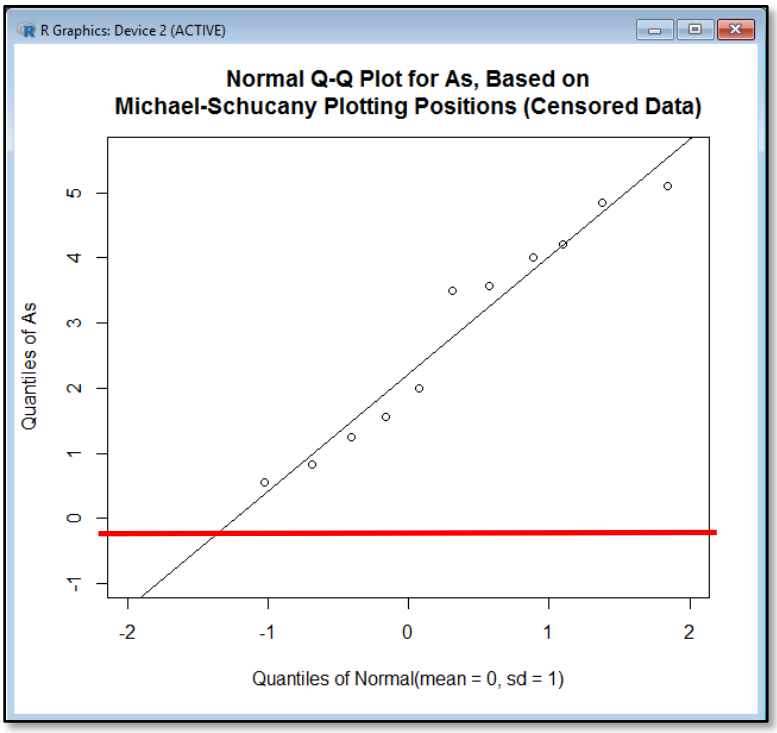
For this example, R reports the following PPCC statistic for the three distributions of interest:

PPCC statistic for the lognormal distribution (lambda at 0) = 0.95
PPCC statistic for the gamma distribution (lambda at 0.3) = 0.96
PPCC statistic for the normal distribution (lambda at 1) = 0.98.

In this case, health assessors should initially select the normal distribution for the 95UCL calculation, because the normal distribution has the highest PPCC statistic from the boxcoxCensored command.

Since the normal distribution was initially selected, health assessors must verify that the distribution does not include an unreasonable amount of negative values. As explained earlier in the guidance, this is done by creating a Q-Q plot and determining the "quantile of normal" value (i.e., the x-value) at the point where the "quantiles of As" value (i.e., the y-value) equals zero. If the value where this occurs is greater than -3, an unreasonable amount of negative numbers is being estimated by the normal distribution, and health assessors should choose the distribution with the next-highest PPCC statistic. Health assessors should use the following code to create a normal Q-Q plot, and the resulting plot is shown in the figure below for the arsenic data set.

```
> qqPlotCensored(As,Ascen,dist="norm",add.line=TRUE)
```



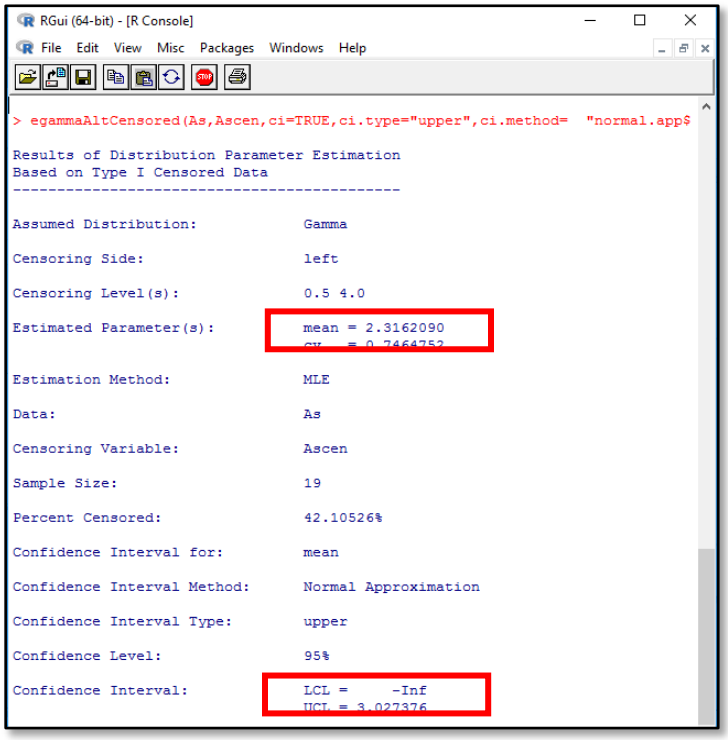
The previous figure is the graph that R generates for this scenario. While ProUCL's Q-Q plots typically have horizontal gridlines at the point where concentration values equal zero, this is not the case for R. The key point for evaluating the normal distribution is the value of "quantiles of normal" at the point where the "quantiles of As" equals zero. For illustrative purposes, a red line has been added to the previous figure to indicate where the "quantiles of As" point is zero. (Note: This red line does not appear in the figure generated by R.) The normal distribution line falls below an arsenic concentration of zero (i.e., y-value = 0) when the "quantiles of normal" is approximately -1.25. ATSDR guidance tells health assessors to not use the normal distribution when the intercept in the Q-Q plots is greater than -3, which is the case in this example. Therefore, the health assessor would not use the normal distribution to calculate the 95UCL and would instead select the distribution with the next highest PPCC statistic for the 95UCL calculation. In this example, the gamma distribution has the next highest PPCC statistic. Note that the ProUCL example shown in Appendix C for these same data identified the lognormal distribution as the best fit. This discrepancy is due to the different distributional tests available in the two software programs; R gives an exact test while ProUCL gives an approximate test. Most of the time the tests will select the same distribution, but they can sometimes (as with this example) provide a different ordering when two distributions are close in fit. In such cases the resulting 95UCL estimates are likely to be similar using "best" distribution. Health assessors should proceed with using whatever 95UCL values are output from following this guidance, whether generating numbers from ProUCL or R.

Step 2B. Calculate 95UCL depending on the selected distribution

For exposure units having between 8 and 19 samples with no non-detects that are best fit by gamma distributions, ATSDR's guidance is for health assessors using R to calculate 95UCLs with the following command:

```
> egammaAltCensored(As,Ascen,ci=TRUE,ci.type="upper",ci.method= "normal.approx")
```

The results for this command appear in the same window as the written R code and are shown below.



The gamma estimate of the 95UCL is 3.03 µg/L, which should be used as the EPC.

Step Three: Perform "Reality Check" on Results

1. Compare the estimated 95UCL to maximum and arithmetic mean concentrations to ensure the calculated value is reasonable. To do this, health assessors should confirm that (1) the 95UCL is below the maximum concentration and (2) the 95UCL is greater than the mean, but not more than three times the mean.

   In this example, the estimated 95UCL (3.03 µg/L) is below the maximum detected concentration of 5.1 µg/L. The 95UCL (3.03 µg/L) is greater than the estimated arithmetic mean of 2.32 µg/L, as it should be. These comparisons suggest that the selected 95UCL is reasonable for use.

2. Check for influence of data outliers. Because the 95UCL is less than three times the mean, the 95UCL is not likely to be influenced by unusually high concentrations and no further evaluation is necessary.

Appendix F. Sample Weighted 95UCL Calculations for Non-Random Samples in R

At some contaminated sites, sampling is focused on areas of suspected high contamination. To avoid bias in determining the EPC in such cases, weighted 95UCLs should be calculated. This appendix demonstrates approaches for calculating weighted 95UCLs using R for (1) data sets with no non-detects (Appendix F-1) and (2) data sets with non-detects (Appendix F-2). Weighted 95UCLs currently cannot be calculated with ProUCL.

Health assessors who are using R for the first time are encouraged to replicate these examples before using the R programming language to calculate weighted 95UCLs for site-specific scenarios. They also should consider consulting with their ADS group with any questions about evaluating weighted data sets. All figures shown in this example are screen shots of actual R outputs, and the calculations were performed using R for Windows Version 3.3.1 (downloaded on August 15, 2016).

Appendix F-1. Sample Weighted 95UCL Calculation: Using R for Data with No Non-detects

This example demonstrates the preferred approach for calculating weighted 95UCLs using R for a data set with no non-detects. The data set considered in this example includes 26 surface soil samples collected from three exposure units (i.e., subunit1, subunit2, and subunit3) that were analyzed for concentrations of cadmium. Twelve samples were collected in subunit1 (representing 40 percent of the EU), 10 samples were collected in subunit2 (representing 20 percent of the EU), and 4 samples were collected in subunit 3 (representing 40 percent of the EU). These data are presented in the table below. Health assessors must install and load the Bootstrap package to R to estimate weighted 95UCLs for data sets without non-detects.

Details on the Bootstrap R package can be found here:
https://cran.r-project.org/web/packages/bootstrap/bootstrap.pdf

Example Data Set: Cadmium concentrations (ppm) in surface soil

Cd	SubUnit	Area	Ninv	Weight
182	SU1	0.4	0.083	0.033
343	SU1	0.4	0.083	0.033
96	SU1	0.4	0.083	0.033
390	SU1	0.4	0.083	0.033
520	SU1	0.4	0.083	0.033
631	SU1	0.4	0.083	0.033
152	SU1	0.4	0.083	0.033
308	SU1	0.4	0.083	0.033
256	SU1	0.4	0.083	0.033
332	SU1	0.4	0.083	0.033
21.3	SU1	0.4	0.083	0.033
49.4	SU1	0.4	0.083	0.033
705	SU2	0.2	0.1	0.02

193	SU2	0.2	0.1	0.02
301	SU2	0.2	0.1	0.02
267	SU2	0.2	0.1	0.02
15.6	SU2	0.2	0.1	0.02
56.7	SU2	0.2	0.1	0.02
112	SU2	0.2	0.1	0.02
47.4	SU2	0.2	0.1	0.02
84.2	SU2	0.2	0.1	0.02
73.9	SU2	0.2	0.1	0.02
38.1	SU3	0.4	0.25	0.1
156	SU3	0.4	0.25	0.1
233	SU3	0.4	0.25	0.1
144	SU3	0.4	0.25	0.1

Data for R should be formatted as CSV or TXT files and include the information shown above. The five columns required to calculate weighted 95UCLs for datasets without non-detects are:

- Chemical concentrations should be listed in a column with the chemical name as the header.
- The subunit that each sample was collected from should be listed in column titled “SubUnit.”
- The fraction of the EU covered by that subunit should be listed in a column titled “Area.” In this example, subunit1 makes up 40 percent of the EU, and a value of 0.4 is therefore listed for all samples collected in subunit1.
- The inverse of the total number of samples collected in each subunit should be listed in a column titled “NInv.” In this example, 12 samples were collected in subunit1. The inverse of 12 (or the result of one divided by twelve) is equal to 0.083, and that value is listed for all samples collected in subunit1.
- The product of the values shown in the Area and Ninv columns, in a column titled “Weight.” In this example, the area for subunit1 is equal to 0.40 and the calculated Ninv is equal to 0.083. The product of these two values is 0.033, which is shown in the Weight column for all samples collected in subunit1.

Step One: Process Data and Load Data to R

1. Remove rejected data and any other data that do not meet QA/QC criteria. Health assessors should do this prior to loading data to R.
2. Process duplicates and replicates to avoid double-counting observations. Health assessors should do this prior to loading data to R.

3. Load data to R. Health assessors can load environmental sampling data in CSV or TXT file formats into R. In this example, a CSV file is used. Use the following code to read the CSV file into R:

```
> soilCd=read.csv(file=file.choose())

> attach(soilCd)
```

The 'head' function in R is a useful tool to preview the first few rows of the loaded data. Use the following code to preview your data in R:

```
> head(soilCd)
```

Step Two: Calculate Weighted Arithmetic Mean

The weighted mean for the entire EU is the sum of the area-weighted means of the individual three subunits. The area-weighted mean in this example is calculated as follows:

$$\text{EU mean} = (271.4 \times 0.4) + (185.6 \times 0.2) + (142.8 \times 0.4)$$
$$\text{EU mean} = 203 \text{ ppm}$$

The weighted mean can also be computed with the 'weighted.mean' function in R. In this example, the code shown below will return the same estimate of 203 ppm.

```
> weighted.mean(Cd,Weight)
```

Step Three: Calculate Weighted 95UCL

A percentile bootstrap estimate of the 95UCL for weighted samples is computed using the bcanon function, after first defining pbdat as a data frame consisting of just the two variables needed to compute the estimate and defining theta as the weighted mean function. The following code is used to estimate the weighted 95UCL for this example (e.g., a data set with no non-detects):

```
> CdDat=data.frame(Cd,Weight)
> theta=function(x,CdDat){weighted.mean(CdDat[x,1],CdDat[x,2]) }
> Cdboot=bootstrap(1:length(Cd),5000,theta,CdDat)
> uclboot=quantile(Cdboot$thetastar,probs=0.95)
> uclboot
```

The last line of the code above will return the percentile bootstrap 95UCL for the weighted data. In this example, the weighted 95UCL equals 262.1 ppm. As with other bootstrapped estimates, your estimate will differ slightly each time it is computed, and so will be slightly different than the one computed here.

Appendix F-2. Sample Weighted 95UCL Calculation: Using R for Data with No Non-detects

This example demonstrates the preferred approach for calculating weighted 95UCLs using R for a data set with non-detects. The data set considered in this example includes 38 surface soil samples collected from three exposure units (e.g. subunits) that were analyzed for concentrations of cadmium. Twelve samples were collected in subunit1, 120 samples were collected in subunit2, and 14 samples were collected in subunit3. Subunit1 covers 15 percent of the exposure unit, subunit 2 covers 65 percent of the exposure unit, and subunit3 covered 20 percent of the exposure unit. These data are presented in the table below.

Health assessors must install and load the EnvStats, NADA, and Bootstrap packages to R to estimate weighted 95UCLs for data sets with non-detects.

Details on the EnvStats R package can be found here:
https://cran.r-project.org/web/packages/EnvStats/EnvStats.pdf

Details on the NADA R package can be found here:
https://cran.r-project.org/web/packages/NADA/NADA.pdf

Details on the Bootstrap R package can be found here:
https://cran.r-project.org/web/packages/bootstrap/bootstrap.pdf

Area1Cd	Area1Cens	Area2Cd	Area2Cens	Area3Cd	Area3Cens
1.78	FALSE	1.55	FALSE	18.4	FALSE
2.82	FALSE	1.58	FALSE	102	FALSE
3.85	FALSE	1.74	FALSE	23.1	FALSE
1.51	FALSE	1.68	FALSE	57.69	FALSE
1.53	FALSE	1.37	FALSE	174	FALSE
2.18	FALSE	1.42	FALSE	40.9	FALSE
1.49	FALSE	1.51	FALSE	13	TRUE
2.12	FALSE	1.36	FALSE	10	TRUE
2.41	FALSE	1.37	FALSE	62.9	FALSE
2.86	FALSE	1.37	FALSE	15.3	FALSE
1.43	FALSE	1	TRUE	23.4	FALSE
3.98	FALSE	1	TRUE	41.2	FALSE
--	--	--	--	83.8	FALSE
--	--	--	--	12.7	FALSE

Data for R should be formatted as CSV or TXT files and include the information shown above. A column for chemical concentrations and a column to indicate whether those concentrations are censored (e.g., non-detect) should be created for each subunit. The chemical concentration columns should have the area/subunit number and the chemical name as the header. The censoring columns should be titled with the area/subunit number and "Cens," and filled in with "True" for non-detects and "False" for detected results.

Step One: Process Data and Load Data to R

1. Remove rejected data and any other data that do not meet QA/QC criteria. Health assessors should do this prior to loading data to R.

2. Process duplicates and replicates to avoid double-counting observations. Health assessors should do this prior to loading data to R.

3. Load data to R. Health assessors can automatically load environmental sampling data in CSV or TXT file formats into R. In this example, a CSV file is used Use the following code to read the CSV file into R:

```
> soilCd=read.csv(file=file.choose())

> attach(soilCd)
```

The 'head' function in R is a useful tool to previous the first few rows of the loaded data. Use the following code to preview your data in R:

```
> head(soilCd)
```

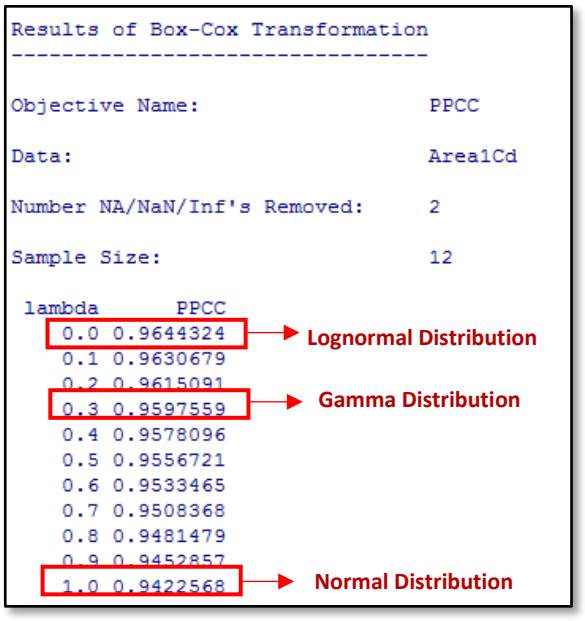
Step Two: Calculate Weighted Mean

Step 2A. Use the boxcoxCensored command to determine the best fitting distribution

Health assessors should use the boxcoxCensored (for datasets with non-detects) and boxcox (for datasets without non-detects) commands to determine whether data for each subunit best fit a normal, lognormal, or gamma distribution. Using either command, a lambda of 1 represents a normal distribution, a lambda of 0 represents a lognormal distribution, and a lambda of approximately 0.3 represents a gamma distribution. Health assessors should select the distribution with the highest PPCC statistic.

In this example, subunit1 does not include any nondetects and the following boxcox code was used:

```
> boxcox(Area1Cd,lambda=seq(0,1,0.1))
```



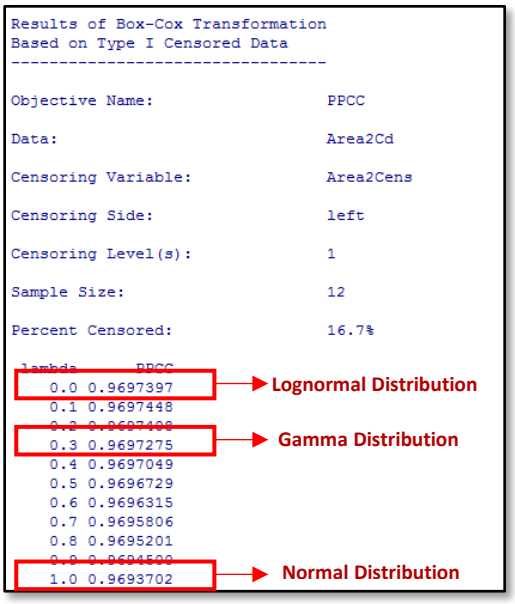
In this example, R reports the following PPCC statistic for the three distributions:

PPCC statistic for the lognormal distribution (lambda at 0) = 0.964
PPCC statistic for the gamma distribution (lambda at 0.3) = 0.960
PPCC statistic for the normal distribution (lambda at 1) = 0.942

In this case, health assessors should select the lognormal distribution for the 95UCL calculation, because this distribution has the highest PPCC statistic from the boxcox command.

Subunit2 does include nondetects and the following boxcoxCensored code was used:

> boxcoxCensored(Area2Cd,Area2Cens,lambda=seq(0,1,0.1))



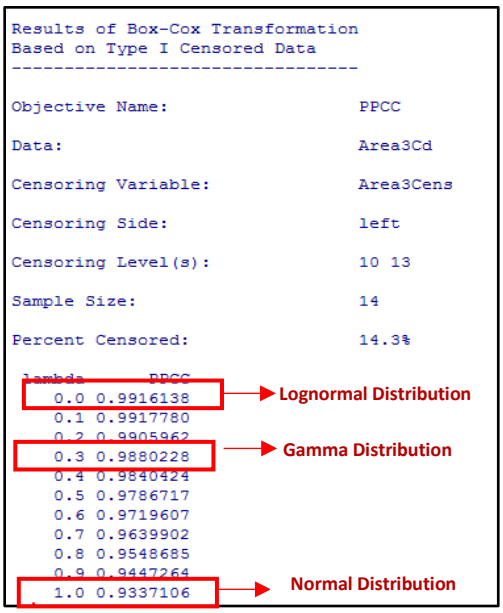
In this example, R reports the following PPCC statistic for the three distributions:

PPCC statistic for the lognormal distribution (lambda at 0) = 0.97
PPCC statistic for the gamma distribution (lambda at 0.3) = 0.97
PPCC statistic for the normal distribution (lambda at 1) = 0.97

In this case, health assessors should select the lognormal distribution for the 95UCL calculation, because this distribution has the highest PPCC statistic from the boxcoxCensored command.

Subunit3 does include nondetects and the following boxcoxCensored code was used:

```
> boxcoxCensored(Area3Cd,Area3Cens,lambda=seq(0,1,0.1))
```



In this example, R reports the following PPCC statistic for the three distributions:

PPCC statistic for the lognormal distribution (lambda at 0) = 0.99
PPCC statistic for the gamma distribution (lambda at 0.3) = 0.90
PPCC statistic for the normal distribution (lambda at 1) = 0.94

In this case, health assessors should select the lognormal distribution for the 95UCL calculation, because this distribution has the highest PPCC statistic from the boxcoxCensored command.

Step 2B. Calculated the Weighted Mean

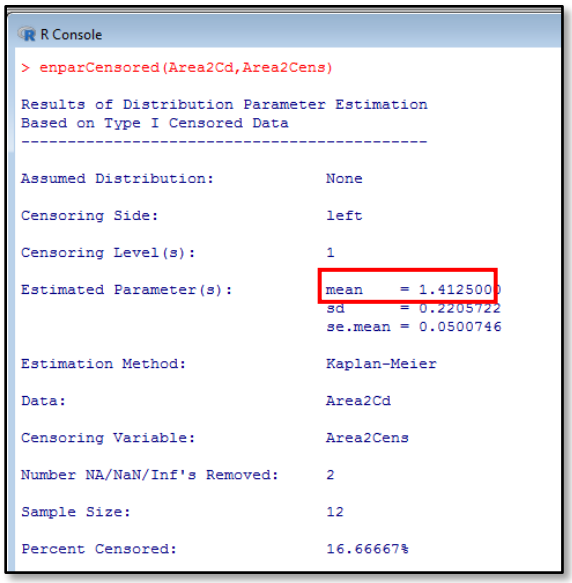
The weighted mean for the entire EU is calculated as the sum of each of the individual subunit means after multiplying by their relative weights. The individual means for the three subunits are calculated as follows:

Subunit1: This dataset does not include any censored values. The arithmetic mean is calculated at 2.33 ppm with the following code:

```
> mean(na.omit(Area1Cd))
```

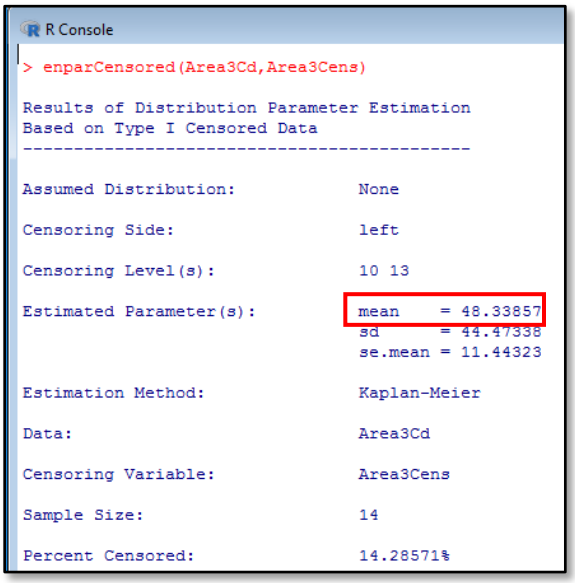
Subunit2: This dataset does include censored values. The KM mean is calculated at 1.41 ppm with the following code (screenshot shown below the code):

```
> enparCensored(Area2Cd,Area2Cens)
```



Subunit3: This dataset does include censored values. The KM mean is calculated at 48.3 ppm with the following code (screenshot shown below the code):

> enparCensored(Area3Cd,Area3Cens)



The weighted mean is calculated by summing the individual subunit means, weighted by the percentage of the total EU represented by each subunit (note that subunit1=15 percent, subunit2=65 percent, and subunit3=20 percent). In this example, the weighted mean is calculated as follows:

$$\textit{Weighted Mean} = (2.33 \times 0.15)+(1.41 \times 0.65)+(48.4 \times 0.20)$$

$$\textit{Weighted Mean} = 10.9 \textit{ ppm}$$

Step Three: Calculate Weighted 95UCL

The weighted 95UCL is calculated in three steps: (1) A lognormal distribution is assumed for all three subunits and 5,000 bootstrap estimates of the MLE mean are calculated for each area, (2) 5,000 weighted means are calculated from these estimates and the percent area of each subunit, and (3) the percentile bootstrap 95UCL of the 5,000-weighted means is estimated. Although subunit1 does not include any censored values and subunits2 and 3 do, the same code can be applied to all three subunits.

Step 1a: R code to calculate 5,000 bootstrap estimates of the MLE mean for subunit1:

```
> Cd1dat=data.frame(na.omit(Area1Cd),na.omit(Area1Cens))
> n1=length(Cd1dat[,1])
> theta1=function(x,Cd1dat){mean(cenmle(Cd1dat[,1],Cd1dat[,2]))[1]}
> results1=bootstrap(1:n1,5000,theta1,Cd1dat)
```

Step 1b: R code to calculate 5,000 bootstrap estimates of the MLE mean for subunit2:

```
> Cd2dat=data.frame(na.omit(Area2Cd),na.omit(Area2Cens))
> n2=length(Cd2dat[,1])
> theta2=function(x,Cd2dat){mean(cenmle(Cd2dat[,1],Cd2dat[,2]))[1]}
> results2=bootstrap(1:n2,5000,theta2,Cd2dat)
```

Step 1c: R code to calculate 5,000 bootstrap estimates of the MLE mean for subunit3:

```
> Cd3dat=data.frame(na.omit(Area3Cd),na.omit(Area3Cens))
> n3=length(Cd3dat[,1])
> theta3=function(x,Cd3dat){mean(cenmle(Cd3dat[,1],Cd3dat[,2]))[1]}
> results3=bootstrap(1:n3,5000,theta3,Cd3dat)
```

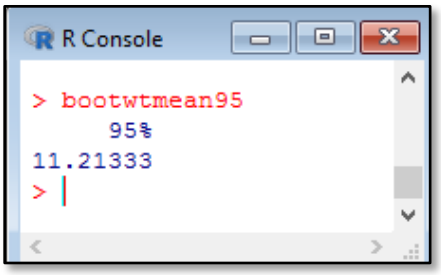
Step 2: R code to calculated weighted mean, using the correct weights for each subunit. Note that in this example subunit1=15 percent, subunit2=65 percent, and subunit3=20 percent:

```
>wtmean=results1$thetastar*0.15+results2$thetastar*0.65+results3$thetastar*0.20
```

Step 3: R code to calculate percentile bootstrap 95UCL of the 5,000-weighted means:

```
> bootwtmean95=quantile(wtmean,probs=.95,type=6)
> bootwtmean95
```

The final command from the code above will return the weighted 95UCL, as shown in the screenshot below.



In this example, the weighted 95UCL is calculated at a value of 11.21 ppm.

Appendix G: Alt Text Descriptions for Document Figures

Figure 1. Selection of EPC Statistics for Different Exposure Durations

The first box in this flow chart states to use this flow chart after you have: Reviewed and screened environmental data. Defined exposure units. Identified contaminants of concern in potential or completed exposure pathways.

An arrow points from this box to the next box which asks the question: Are you evaluating data for asbestos or lead? If the answer is yes, an arrow points to a new box which states: Do not apply this guidance. Consult with Agency subject matter experts for guidance.

If the answer is no, an arrow points to the next question: Are you evaluating data for PAHs or dioxins? If the answer is yes, an arrow points to a new box which states: Refer to separate Exposure Dose Guidance for these chemicals before applying this guidance. This is followed by a footnote which states: Guidance documents for these chemicals and topics are under development. Until they are issued, consult with your Branch ADS on preferred approaches.

If the answer is no, an arrow points to the next question: Are you evaluating multi-incremental sampling data, composite sampling data, dependent data, or times series? (See Section 1.2 for definitions). If the answer is yes, an arrow points to a new box which states: Consult with your Branch ADS for preferred approaches for these topics. This is followed by a footnote which states: Guidance documents for these chemicals and topics are under development. Until they are issued, consult with your Branch ADS on preferred approaches.

If the answer is no, an arrow points the next box which states: Apply the following duration specific guidance. This box has arrows pointing towards two new boxes. The first box is labeled Acute Exposures and states: Use the statistic (maximum concentration or 95UCL) that best aligns with the sample media and applicable toxicity data as the EPC. See Section 2.1 for additional information.

The second box is labeled Intermediate or Chronic Exposures and has a question box and two answers within it. The question box asks: Does your data have fewer than 8 samples? If the answer is yes, an arrow points to a new box which states: Use the maximum concentration as the EPC and: Consider recommending additional sampling. Consult with your Branch ADS on your evaluation. See Section 2.2 for additional information. If the answer is no, an arrow points to another new box which states: Use the 95UCL as the EPC and: Refer to Figure 2 for guidance on calculating the 95UCL using different software programs.

Figure 2. Where to Find Guidance on Calculating 95UCLs with Different Software Programs

The first major box in this flowchart is labeled Calculating 95UCL Using ProUCL (see Section 3.4). Within this major box, there is a question box which asks: Does data set include non-detects?

If the answer is no, an arrow points to a box which states: Read guidance in Section 3.4.1. Use decision tree in Figure 3. Refer to sample calculation in Appendix B.

If the answer is yes, an arrow points to a box which states: Read guidance in Section 3.4.2. Use decision tree in Figure 4. Refer to sample calculation in Appendix C.

The second major box in this flowchart is labeled Calculating 95 UCLs Using R (see Section 3.5). Within this major box, there is a question box which asks: Does data set include non-detects?

If the answer is no, an arrow points to a box which states: Read guidance in Section 3.5.1. Use decision tree in Figure 5. Refer to sample calculation in Appendix D.

If the answer is yes, an arrow points to a box which states: Read guidance in Section 3.5.2. Use decision tree in Figure 6. Refer to sample calculation in Appendix E.

The third major box in this flowchart is labeled: Calculating 95 UCL Using Other Software Programs (See Section 3.6). This box states: Refer to Section 3.6 for general computational guidance for other methods and software programs, including SAS, SPSS, Minitab, and spreadsheets.

Figure 3. Using ProUCL to Calculate 95UCLs for Data Sets with at Least 8 Samples and No Non-Detects

The first box in this flowchart is labeled General Guidelines and states: Refer to Section 3.4.1 for detailed instructions and Appendix B for a sample calculation. Do not apply the following guidance to data sets with fewer than 8 samples, dependent data sets, multi-incremental sampling data sets, or data on lead, asbestos, PAHs, or dioxins.

An arrow points from this box to the next box which is labeled Step One: Process Data and states: Remove rejected data and any other data that do not meet QA/QC criteria. Process duplicates and replicates to avoid double-counting observations. Plot data and examine for outliers. See Section 3.3 for additional details and instructions.

An arrow points from this box to the next box which is labeled Step Two: Calculate 95UCL and presents two options: Option A for 8-19 samples and Option B for greater than or equal to 20 samples. For Option A (8-19 samples) there are two steps stated: Step 2A: In ProUCL's Statistical Tests menu, select, Goodness-of-Fit Tests, Full (w/o NDs), GOF Statistics. Select the distribution (normal, lognormal, or gamma) with correlation coefficient closest to 1. This is followed by a footnote which states: If the highest correlation coefficient is less than 0.80, do not calculate a 95UCL and contact your Branch Associate Director for Science for further direction. If the normal distribution is initially selected at this stage, first check to ensure that the left-end of the distribution does not include an unreasonable amount of negative values. If it does, then do not use the normal distribution and instead use whichever other distribution (lognormal or gamma) has a correlation coefficient closest to 1. See Section 3.4.1 for specific details.

Step 2B: In ProUCL's UCLs/EPCs menu, select, Full (w/o NDs), Selected Distribution. Use the following values as EPCs depending on the selected distribution: Normal distribution – Use the 95% Student's t-UCL. Lognormal distribution – 95% Chebyshev UCL. Gamma distribution – Use the 95% Adjusted Gamma UCL for n<50. The instructions for the lognormal distribution are followed by a footnote which states: While the preferred method for lognormal distributions, the Chebyshev method sometimes gives very high estimates of the 95UCL. Read guidance in the corresponding text box in Section 3.4.1 before applying this method.

For Option B for greater than or equal to 20 samples, there is one step stated: In ProUCL's UCLs/EPCs menu, select, Full (w/out NDs), Non-parametric. Use the 95% Percentile Bootstrap UCL as the EPC.

An arrow points from this box to the next box which is labeled Step Three: Perform "Reality Check" on Results. This box states the following: Compare 95UCL to maximum and arithmetic mean concentrations. Check for influence of data outliers. Refer to Section 3.7 for specific guidance.

Figure 4. Using ProUCL to Calculate 95UCLs for Data Sets with at Least 8 Samples and Non-Detects

The first box in this flowchart is labeled General Guidelines and states: Refer to Section 3.4.2 for detailed instructions and Appendix C for a sample calculation. Do not apply the following guidance to data sets with fewer than 8 samples, dependent data sets, multi-incremental sampling data sets, or data on lead, asbestos, PAHs, or dioxins.

An arrow points from this box to the next box labeled Step One: Process Data and states: Remove rejected data and any other data that do not meet QA/QC criteria. Process duplicates and replicates to avoid double-counting observations. Plot data and examine for outliers. See Section 3.3 for additional details and instructions.

An arrow points from this box to the next box labeled Step Two: Calculate 95 UCL and presents two options: Option A for 8-19 samples and Option B for greater than or equal to 20 samples. For Option A (8-19 samples) there are two steps stated. Step 2A states: In ProUCL's Statistical Tests menu, select, Goodness-of-Fit Tests, With NDs, GOF Statistics. Select the distribution (normal, lognormal, or gamma) with correlation coefficient (computed by ROS procedures) closest to 1. This is followed by a footnote which states: If the highest correlation coefficient is less than 0.80, do not calculate a 95UCL and contact your Branch Associate Director for Science for further direction. If the normal distribution is initially selected at this stage, first check to ensure that the left-end of the distribution does not include an unreasonable amount of negative values. If it does, then do not use the normal distribution and instead use whichever other distribution (lognormal or gamma) has a correlation coefficient closest to 1. See Section 3.4.2 for specific details.

Step 2B states: In ProUCL's UCLs/EPCs menu, select, With NDs, Selected Distribution. Use the following values as EPCs depending on the selected distribution: Normal distribution – Use the 95% KM (t) UCL. Lognormal distribution – Use the 95% Percentile Bootstrap UCL in the "Lognormal ROS Statistics Using Imputed Non-Detects" section. Gamma distribution – Use the 95% Gamma Adjusted KM-UCL for n<50.

Option B for greater than or equal to 20 samples asks the following question: Do the non-detects have different values for detection limits? If the answer is yes, an arrow points to a box which states: In ProUCL's UCLs/EPCs menu, select, With NDs, Non-parametric. Use the 95% KM (Percentile Bootstrap) UCL in the "Kaplan-Meier (KM) Statistics using Normal Critical Values and other Nonparametric UCLs" section. If the answer is no, an arrow points to a box which states: In ProUCL's UCLs/EPCs menu, select, With NDs, Lognormal. Use the 95% Percentile Bootstrap UCL in the "Lognormal ROS Statistics Using Imputed Non-Detects" section.

An arrow points from this box to the next box which is labeled Step Three: Perform "Reality Check" on Results. This box states the following: Compare 95UCL to maximum and arithmetic mean concentrations. Check for influence of data outliers. Refer to Section 3.7 for specific guidance.

Figure 5. Using R to Calculate 95UCLs for Data Sets with at Least 8 Samples and No Non-Detects

The first box in this flowchart is labeled General Guidance and states: Refer to Section 3.5.1 for detailed instructions and Appendix D for a sample calculation. Do not apply the following guidance to data sets with fewer than 8 samples, dependent data sets, multi-incremental sampling data sets, or data on lead, asbestos, PAHs, or dioxins. An arrow points from this box to the next box which is labeled Step One: Process Data. This box states: Remove rejected data and any other data that do not meet QA/QC criteria. Process duplicates and replicates to avoid double-counting observations. Plot data and examine for outliers. See Section 3.3 for additional details and instructions.

An arrow points from this box to the next box labeled Step Two: Calculate 95UCL. Two options are presented in this box, Option A for 8-19 samples and Option B for greater than 20 samples. For Option A, there are two steps stated. The first step is Step 2A: Use Q-Q plots and the Shapiro-Wilk test to determine if the data are best represented by a normal distribution, lognormal distribution, or gamma distribution. Refer to Section 3.5.1 for instructions*. The second step is Step 2B: Use the following code to calculate the 95UCL depending on the selected distribution, and replace “variable_name” with the column name of the concentration values under consideration: Normal distribution code: >EnvStats::enorm(variable_name,ci=TRUE,ci.type=“upper”); Lognormal distribution code: >EnvStats::elnormAlt(variable_name,ci=TRUE,ci.type=“upper” ci.method=“cox”); Gamma distribution code: >EnvStats::egammaAlt(variable_name,ci=TRUE,ci.type=“upper”). For Option B (greater than 20 samples) there are two steps stated: 1. Use the percentile bootstrap output from the “boot” package, with a 90% confidence interval, as the 95UCL, and 2. Code (two steps):Step 1: >bootOutput <- boot::boot(variable_name, function(x, index) mean(x[index]), 5000). Step 2: >boot::boot.ci(bootOutput, conf=0.90, type = "perc")$percent[[5]].

An arrow points from this box to the next box labeled Step Three: Perform “Reality Check” on Results. This box states the following: Compare 95UCL to maximum and arithmetic mean concentrations. Check for influence of data outliers. Refer to Section 3.7 for specific guidance on these steps and selection of the final EPC.

The asterisk in box two is described below the flow chart: * If the normal distribution is initially selected at this stage, first check to ensure that the left-end of the distribution does not include an unreasonable amount of negative values. If it does, then do not use the normal distribution and instead use whichever other distribution (lognormal or gamma) has a correlation coefficient closest to 1. See Section 3.5.1 for specific details.

Figure 6. Using R to Calculate 95UCLs for Data Sets with at Least 8 Samples and Non-Detects

The first box in this flowchart is labeled General Guidelines and states: Refer to Section 3.5.2 for detailed instructions and Appendix E for a sample calculation. Do not apply the following guidance to data sets with fewer than 8 samples, dependent data sets, multi-incremental sampling data sets, or data on lead, asbestos, PAHs, or dioxins.

An arrow points from this box to the next box which is labeled Step One: Process Data and states: Remove rejected data and any other data that do not meet QA/QC criteria. Process duplicates and replicates to avoid double-counting observations. Plot data and examine for outliers. See Section 3.3 for additional details and instructions.

An arrow points from this box to the next box which is labeled Step Two: Calculate 95UCL and presents two options: Option A for 8-19 samples and Option B for Greater than 20 samples.

For Option A (8-19 samples) there are two steps stated: Step 2A: Use the boxcoxCensored command to determine the best fitting distribution. Refer to Section 3.5.2 for instructions on use of this command. Step 2B: Use the following code to calculate the 95UCL depending on the selected distribution. Replace "Data" with the input column of concentrations; and "Cen" should refer to the data column indicating whether a given value is a non-detect, as denoted by a value of TRUE or 1. Normal distribution code: >EnvStats::enormCensored(Data,Cen,ci=TRUE,ci.type="upper",ci.method="normal.approx"). Lognormal distribution code: >EnvStats::elnormAltCensored(Data,Cen,ci=TRUE,ci.type="upper",ci.method="cox"). Gamma distribution code: >EnvStats::egammaAltCensored(Data,Cen,ci=TRUE,ci.type="upper",ci.method="normal.approx").

Option B (greater than 20 samples) has a question asking if the data are singly censored or is the percent of data detected above all censoring limits the same?

If yes, complete the following steps: 1. Use the percentile bootstrap output from the "enparCensored" function in the "EnvStats" package as the 95UCL; and 2. Code (two steps): Step 1: >enparCensoredOutput <- EnvStats::enparCensored(Data,Cen,ci = TRUE,ci.method = "bootstrap", ci.type = "upper",n.bootstraps = 5000). Step 2: >as.numeric(enparCensoredOutput$interval$limits["Pct.UCL"]).

If no, complete the following steps: 1. Use the percentile bootstrap output from the "enparCensored" function in the "EnvStats" package as the 95UCL. 2. Code (two steps): Step 1: >enparCensoredOutput <- EnvStats::enparCensored(Data,Cen,ci = TRUE,ci.method = "bootstrap", ci.type = "upper", n.bootstraps = 5000). Step 2: >as.numeric(enparCensoredOutput$interval$limits["Pct.UCL"]).

An arrow points to the next box labeled Step Three: Perform "Reality Check" on Results. This box states the following: Compare 95UCL to maximum and arithmetic mean concentrations. Check for influence of data outliers. Refer to Section 3.7 for specific guidance on these steps and selection of the final EPC.