APPENDIX C

STATISTICAL METHODOLOGY

The primary objective of the statistical analysis of the TNSSS data was to generate national estimates of the mean, standard deviation, and selected percentiles of analyte concentrations (i.e., 50th, 90th, 95th, 98th, 99th percentiles) in biosolids forthe survey's target population. By accounting for survey weights (Section 2.5), the statistical analysis yields estimates that are represented of EPA's target population (Section 2.2).

Because both detected and non-detected outcomes can occur among the collected samples for a given analyte, the statistical approach must account for non-detects. In addition, if a data review concluded that the data distribution is well approximated by a lognormal distribution (i.e., a distribution with known statistical properties), then this distribution can be used as the basis for estimating the statistics of interest.

EPA's primary statistical approach assumed an underlying lognormal distribution for the pollutant concentrations across POTWs. However, to handle situations among the different analytes in which the collected data do not support this assumption, EPA also used an alternative method that does not require this assumption. This alternative statistical approach also allows us to evaluate the sensitivity of the estimates to any particular approach. This helps evaluate the robustness of the underlying assumptions relative to the outcome of the procedures. The two statistical approaches are:

- Lognormal approach: This approach uses maximum likelihood estimation (MLE) techniques, where all data are assumed to originate from a common lognormal distribution, and non-detected data are treated as "left-censored" at the sample-specific detection limit. That is, for non-detects, the actual concentration cannot be reported but is known to be something less than the sample-specific detection limit.
- Nonparametric (distribution-free) approach: This approach yields simple stratified estimates, where not detects are represented by their sample-specific detection limits.

The distributional checks and data reviews in Section 4.2 and Appendix B help determine which technique is more appropriate for a given analyte. Each method, which is able to handle multiple detection limits, is discussed in detail below.

C.1 Lognormal Approach

The lognormal approach, which assumes an underlying lognormal distribution in the (aggregated) analyte concentrations, takes into account the survey weights assigned to each sample POTW and the survey's stratified sample design. Details of the lognormal approach differ between when all data for a given analyte are detected versus when non-detects are present.

C.1.1 Method for Handling 100% Detected Data. When all data for a given analyte were categorized as detected, the lognormal approach took the following form:

- First, the procedure calculated an estimate for the overall mean of the distribution of log-transformed data. Call this estimate $\hat{\gamma}$. The formula for calculating $\hat{\gamma}$ is as follows (Lohr, 1999):

$$\hat{\gamma} = \frac{\sum_{i=1}^{3} \sum_{j=1}^{n_i} w_{ij} y_{ij}}{\sum_{i=1}^{3} \sum_{j=1}^{n_i} w_{ij}}$$

where n_i denotes the number of POTWs sampled in the i^{th} stratum, and y_{ij} and w_{ij} denote the log-transformed data value and the survey weight, respectively, for the j^{th} sampled POTW within the i^{th} stratum. This formula is well-recognized as the formula for a weighted mean, which either the MEANS or SURVEYMEANS procedure in SAS® 9.1.3 is able to calculate. The SURVEYMEANS procedure is also capable of calculating the standard error associated with this estimated mean $\hat{\gamma}$, using a Taylor series approach to perform the calculation.

- Second, the procedure estimated the variance of the log-transformed data. This variance represented deviation in these data values from their estimated mean $\hat{\gamma}$. Call this estimate $\hat{\tau}^2$. Using the same notation as above, the formula for calculating $\hat{\tau}^2$ is as follows:

$$\hat{\tau}^2 = \frac{\sum_{i=1}^{3}\sum_{j=1}^{n_i} w_{ij}(y_{ij}-\hat{\gamma})^2}{\sum_{i=1}^{3}\sum_{j=1}^{n_i} w_{ij}} = \frac{\sum_{i=1}^{3}\sum_{j=1}^{n_i} w_{ij} y_{ij}^2}{\sum_{i=1}^{3}\sum_{j=1}^{n_i} w_{ij}} - \hat{\gamma}^2$$

  This estimate was calculated using the MEANS procedure in the SAS® system.

- Finally, the two estimates $\hat{\gamma}$ and $\hat{\tau}^2$ were substituted into the following equations. These equations serve as estimates of the mean, variance, and percentiles of the lognormal distribution.

$$\text{Mean} = \hat{\mu} = \exp(\hat{\gamma} + 0.5\hat{\tau}^2)$$

$$\text{Variance} = \hat{\sigma}^2 = \exp(2\hat{\gamma} + \hat{\tau}^2)\cdot(\exp(\hat{\tau}^2) - 1)$$

$$\text{p}^{\text{th}}\ \text{percentile} = \exp(\hat{\gamma} + z_p\hat{\tau})$$

In the equation for the percentile, the value z_p represents the p^{th} percentile of the standard normal distribution:

- $z_p = 0$ for the 50th percentile;
- $z_p = 1.282$ for the 90th percentile;
- $z_p = 1.645$ for the 95th percentile;
- $z_p = 2.054$ for the 98th percentile;
- $z_p = 2.328$ for the 99th percentile.

The estimated standard deviation for the lognormal distribution ($\hat{\sigma}$) equals the square root of the estimated variance (i.e., $\sqrt{\hat{\sigma}^2}$).

C.1.2 Method for Handling a Mixture of Detected and Non-Detected Data

For selected analytes, the laboratory method was not sufficiently sensitive to permit a measurement from being quantified within the method's detectable range for some samples. This led to some samples being classified as non-detected. The lognormal approach assumed that concentrations within these samples originate from the same underlying lognormal distribution as the concentrations associated with samples

classified as containing detected levels of the analyte. However, because the method cannot quantify measurements for samples that it classified as non-detected, the approach treats these outcomes as "left censored" at the sample-specific detection limit. This means that while the method cannot quantify the measurement, it is known to be some number below the sample-specific detection limit. The censored lognormal distribution approach estimates the mean, variance, and percentiles of a lognormal distribution, for those analytes having some non-detects.

The censored lognormal distribution model assumes that all data, classified as either detected or non-detected (i.e., left-censored), originate from a common lognormal distribution within each flow group stratum. For each analyte, estimates of the mean and variance under the censored lognormal model were calculated within each stratum using the CENMLE module, written in the R programming language and obtained from within the "NADA" software package (Helsel, 2005). NADA, which stands for "Nondetects and Data Analysis," is available from the Comprehensive R Archive Network (CRAN) (Helsel and Lee, 2006).

Unlike the approach outlined in Section C.1.1, the CENMLE module was unable to handle the weighting of data for individual POTWs by their survey weights. Therefore, the module was used to calculate the mean and variance of the lognormal distribution within each flow group stratum, then a weighted average of these statistics across the three strata was calculated. The procedure consisted of the following steps.

- First, assume that within the i^{th} flow group stratum, the log-transformed concentrations originated from a normal distribution with mean γ_i and variance τ_i^2. We obtain estimates of γ_i and variance τ_i^2 by maximizing the following log-likelihood function:

$$\log(L_i) = \prod_{j=1}^{n_i} [\delta_{ij} \cdot \log(f(y_{ij})) + (1-\delta_{ij}) \cdot \log(F(y_{ij}))]$$

  where $f(y_{ij})$ denotes the probability density function of a lognormal distribution with parameters γ_i and τ_i^2, $F(y_{ij})$ denotes the cumulative distribution function of the normal distribution with mean γ_i and variance τ_i^2, and δ_{ij} is an indicator function equal to one if a detected measurement is reported for the j^{th} sampled POTW within the i^{th} stratum, and zero otherwise. The variable y_{ij} corresponds to the log-transformed concentration if δ_{ij} =1, and to the sample-specific detection limit if δ_{ij}=0. The CENMLE module uses an iterative computational search technique to find estimates of γ_i and τ_i^2 that maximize the log-likelihood function L_i. The resulting estimates, $\hat{\gamma}_i$ and $\hat{\tau}_i^2$, are called "maximum likelihood estimates" (hence, reference to "MLE" in the module's name). This step is repeated for each flow group stratum.

- Second, for the i^{th} stratum, estimates of μ_i and σ_i^2 are obtained, which represent the mean and variance, respectively, of the (untransformed) concentrations. These estimates are calculated as follows:

$$\hat{\mu}_i = \exp(\hat{\gamma}_i + 0.5\hat{\tau}_i^2)$$

$$\sigma_i^2 = \exp(2\hat{\gamma}_i + \tau_i^2) \cdot (\exp(\hat{\tau}_i^2) - 1)$$

- An estimate is calculated for the overall mean (μ) of the lognormal distribution. The estimation approach involves calculating a weighted average of the stratum-specific

$\hat{\mu}$, with each estimate weighted by the survey weight associated with its respective flow group stratum:

$$\hat{\mu} = \frac{\sum_{i=1}^{3} N_i \hat{\mu}_i}{\sum_{i=1}^{3} N_i}$$

where N_i denotes the size of the i^{th} stratum (i.e., the sum of the survey weights across all surveyed POTWs in the i^{th} stratum).

- Now the overall variance (σ^2) of the lognormal distribution is estimated. The overall variance equals of the sum of "within-stratum" variance and "between-stratum" variance. "Within-stratum" variance represents variability among data within a stratum. "Between-stratum" variance represents variability among the different strata, or equivalently, variability among the stratum-specific means. An estimate of within-stratum variance is obtained by calculating a weighted average of the stratum-specific variance estimates under the censored lognormal model. Between-stratum variance is obtained by calculating the weighted variance of the stratum-specific mean estimates under the censored lognormal model, relative to their deviation from the overall mean of the lognormal distribution. Thus, the variance estimate is calculated as follows:

$$\hat{\sigma}^2 = (\textit{within-stratum variation}) + (\textit{between-stratum variation})$$

$$= \frac{\sum_{i=1}^{3} N_i \hat{\sigma}_i^2}{\sum_{i=1}^{3} N_i} + \frac{\sum_{i=1}^{3} N_i (\hat{\mu}_i - \hat{\mu})^2}{\sum_{i=1}^{3} N_i}$$

The estimated standard deviation for the lognormal distribution ($\hat{\sigma}$) equals the square root of the estimated variance (i.e., $\sqrt{\hat{\sigma}^2}$).

- Finally, estimates of percentiles are obtained by first calculating the overall estimates of the mean (γ) and variance (τ^2) of the log-transformed data as follows:

$$\hat{\gamma} = \log(\hat{\mu}) - 0.5 \cdot \log(1 + \hat{\sigma}^2 / \hat{\mu}^2)$$

$$\hat{\tau}^2 = \log(1 + \hat{\sigma}^2 / \hat{\mu}^2)$$

Then, the p^{th} percentile was calculated as follows:

$$\text{p}^{\text{th}} \text{ percentile} = \exp(\hat{\gamma} + z_p \hat{\tau})$$

where z_p represents the p^{th} percentile of the standard normal distribution:

- $z_p = 0$ for the 50th percentile;
- $z_p = 1.282$ for the 90th percentile;
- $z_p = 1.645$ for the 95th percentile;
- $z_p = 2.054$ for the 98th percentile;
- $z_p = 2.328$ for the 99th percentile.

C.2 Nonparametric (Distribution Free) Approach

The nonparametric, or distribution-free, approach does not place an underlying assumption on the distributional form of the data, such as lognormality. Thus, this approach calculates estimates of means, standard deviations, and percentiles solely through information available from the collected data. Under this approach, each sample collected in this survey, including non-detects, needed to be represented by some data value. When a sample was classified as detected, its measured value as contained within the survey data set was used in the analysis. For non-detects, two different representations (or "substitution approaches") were considered, each resulting in a distinct set of parameter estimates for each analyte. One representation was the sample-specific detection limit, which is the largest concentration value that one would assign to the sample. (Any larger value would suggest that the analyte was detected within the sample.) The second representation was one-half of this detection limit. EPA used the sample-specific detection limit when reporting final estimates under this approach.

Under the nonparametric approach, the UNIVARIATE procedure in SAS® 9.1.3 was used to calculate estimates of the mean, standard deviation, and percentiles, accounting for the assigned survey weights. the overall mean (μ) was calculated as the weighted mean of the untransformed data (Lohr, 1999):

$$\hat{\mu} = \frac{\sum_{i=1}^{3} \sum_{j=1}^{n_i} w_{ij} x_{ij}}{\sum_{i=1}^{3} \sum_{j=1}^{n_i} w_{ij}}$$

where n_i denotes the number of POTWs sampled in the i^{th} stratum, and x_{ij} and w_{ij} denote the reported concentration value and the survey weight, respectively, for the j^{th} sampled POTW within the i^{th} stratum. The data value x_{ij} represents either the reported measured value (for detected outcomes) or the substituted value (for not detected outcomes) and is aggregated across multiple samples if EPA collected multiple samples at the given POTW. The data value is not transformed in any way (e.g., no log-transformation is made, as was done in the lognormal approach). Similarly, we calculated the overall variance (σ^2) as the weighted variance of the untransformed data:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{3} \sum_{j=1}^{n_i} w_{ij} (x_{ij} - \hat{\mu})^2}{\sum_{i=1}^{3} \sum_{j=1}^{n_i} w_{ij}}$$

The estimated standard deviation ($\hat{\sigma}$) equals the square root of the estimated variance (i.e., $\sqrt{\hat{\sigma}^2}$).

The p^{th} percentile was estimated by first placing the n data values $\{x_{ij}\}$ in increasing order and denoting this ordered list as $\{z_1, z_2, \ldots, z_n\}$. Let w_j denote the survey weight assigned to the POTW associated with

data value z_j (j=1, …, n). Then, using this notation, the formula for calculating the p^{th} percentile was as follows:

$$p^{th}\ percentile = \quad 0.5(z_i + z_{i+1}) \quad if \quad p = \frac{\sum_{j=1}^{i} w_j}{\sum_{j=1}^{n} w_j}$$

$$p^{th}\ percentile = \quad z_{i+1} \quad if \quad \frac{\sum_{j=1}^{i} w_j}{\sum_{j=1}^{n} w_j} < p < \frac{\sum_{j=1}^{i+1} w_j}{\sum_{j=1}^{n} w_j}$$

References

Helsel, D.R., and L. Lee. 2006. Analysis of Environmental Data with Nondetects: Statistical Methods for Censored Environmental Data. Continuing Education Workshop at the 2006 Joint Statistical Meetings, American Statistical Association, Seattle, WA.

Helsel, D.R. 2005. *Nondetects and Data Analysis: Statistics for Censored Environmental Data.* Hoboken, NJ: John Wiley & Sons, Inc.

Lohr, S.L. 1999. *Sampling: Design and Analysis.* Pacific Grove, CA: Duxbury Press.