APPENDIX E:
MODIFIED DELTA-LOGNORMAL DISTRIBUTION

APPENDIX E:
MODIFIED DELTA-LOGNORMAL DISTRIBUTION

This appendix describes the modified delta-lognormal distribution and the estimation of the episode long-term averages and variability factors used to calculate the limitations and standards.[1] This appendix provides the statistical methodology that was used to obtain the results presented in Chapter 8.

E.1 BASIC OVERVIEW OF THE MODIFIED DELTA-LOGNORMAL DISTRIBUTION

EPA selected the modified delta-lognormal distribution to model pollutant effluent concentrations from the aquatic animals industry in developing the long-term averages and variability factors. A typical effluent data set from a sampling episode or self-monitoring episode (see Chapter 8 for a discussion of the data associated with these episodes) consists of a mixture of measured (detected) and non-detected values. The modified delta-lognormal distribution is appropriate for such data sets because it models the data as a mixture of measurements that follow a lognormal distribution and non-detect measurements that occur with a certain probability. The model also allows for the possibility that non-detect measurements occur at multiple sample-specific detection limits. Because the data appeared to fit the modified delta-lognormal model reasonably well, EPA has determined that this model is appropriate for these data.

The modified delta-lognormal distribution is a modification of the ‘delta distribution' originally developed by Aitchison and Brown.[2] While this distribution was originally developed to model economic data, other researchers have shown the application to environmental data.[3] The resulting mixed distributional model, which combines a continuous density portion with a discrete-valued spike at zero, is also known as the delta-lognormal distribution. The delta in the name refers to the proportion of the overall distribution contained in the discrete distributional spike at zero; that is, the proportion of zero amounts. The remaining non-zero, non-censored (NC) amounts are grouped together and fit to a lognormal distribution.

EPA modified this delta-lognormal distribution to incorporate multiple detection limits. In the modification of the delta portion, the single spike located at zero is replaced by a discrete distribution made up of multiple spikes. Each spike in this modification is associated with a distinct sample-specific detection limit associated with non-detected (ND) measurements in the database.[4] A lognormal density is used to represent the set of

[1] In the remainder of this appendix, references to ‘limitations' includes ‘standards.'

[2] Aitchison, J. and Brown, J.A.C. (1963) The Lognormal Distribution. Cambridge University Press, pages 87-99.

[3] Owen, W.J. and T.A. DeRouen. 1980. “Estimation of the Mean for Lognormal Data Containing Zeroes and Left-Censored Values, with Applications to the Measurement of Worker Exposure to Air Contaminants.” *Biometrics*, 36:707-719.

[4] Previously, EPA had modified the delta-lognormal model to account for non-detected measurements by placing the distributional “spike” at a single positive value, usually equal to the nominal method detection limit, rather than at zero. For further details, see Kahn and Rubin, 1989. This adaptation was used in

measured values. This modification of the delta-lognormal distribution is illustrated in Figure 1.

The following two subsections describe the delta and lognormal portions of the modified delta-lognormal distribution in further detail.

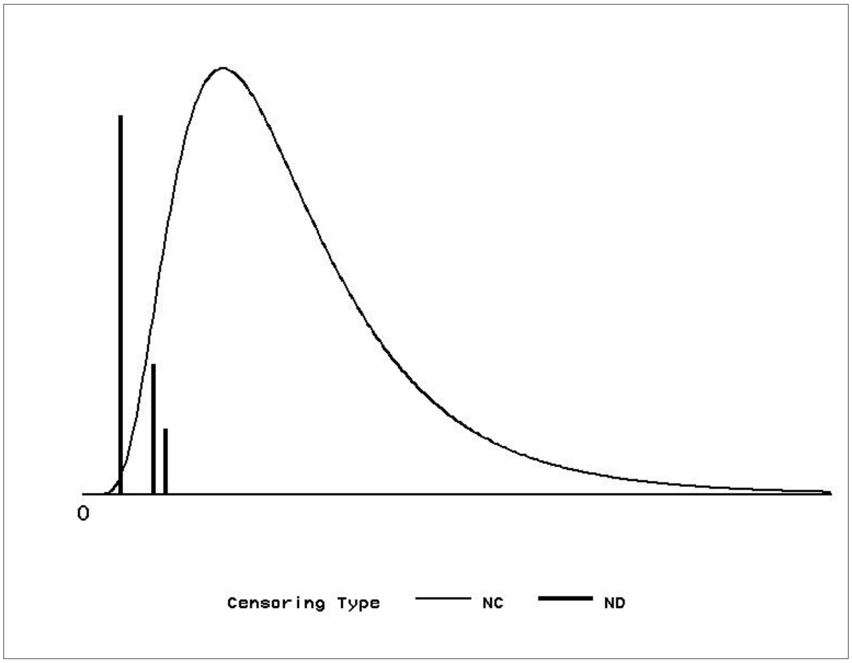

Figure E-1. Modified Delta-Lognormal Distribution

E.2 Continuous and Discrete Portions of the Modified Delta-Lognormal Distribution

The discrete portion of the modified delta-lognormal distribution models the non-detected values corresponding to the k reported sample-specific detection limits. In the model, δ represents the proportion of non-detected values in the dataset and is the sum of smaller fractions, δ_i, each representing the proportion of non-detected values associated with each distinct detection limit value. By letting D_i equal the value of the i^{th} smallest distinct detection limit in the data set and the random variable X_D represents a randomly chosen non-detected measurement, the cumulative distribution function of the discrete portion of the modified delta-lognormal model can be mathematically expressed as:

$$\Pr\left(X_D \le c\right) = \frac{1}{\delta} \sum_{i:D_i \le c} \delta_i \qquad 0 < c \qquad \text{(E-1)}$$

developing limitations and standards for the organic chemicals, plastics, and synthetic fibers (OCPSF) and pesticides manufacturing rulemakings. EPA has used the current modification in several, more recent, rulemakings.

The mean and variance of this discrete distribution can be calculated using the following formulas:

$$E(X_D) = \frac{1}{\delta}\sum_{i=1}^{k} \delta_i \, D_i \tag{E-2}$$

$$Var(X_D) = \frac{1}{\delta}\sum_{i=1}^{k} \delta_i \left(D_i - E(X_D)\right)^2 \tag{E-3}$$

The continuous, lognormal portion of the modified delta-lognormal distribution was used to model the detected measurements from the aquatic animals industry database. The cumulative probability distribution of the continuous portion of the modified delta-lognormal distribution can be mathematically expressed as:

$$\Pr\left[X_C \le c\right] = \Phi\left[\frac{\ln(c) - \mu}{\sigma}\right] \tag{E-4}$$

where the random variable X_C represents a randomly chosen detected measurement, Φ is the standard normal distribution, and μ and σ are parameters of the distribution.

The expected value, $E(X_C)$, and the variance, $Var(X_C)$, of the lognormal distribution can be calculated as:

$$E\left(X_C\right) = \exp\left(\mu + \frac{\sigma^2}{2}\right) \tag{E-5}$$

$$Var\left(X_C\right) = \left[E\left(X_C\right)\right]^2 \left(\exp\left(\sigma^2\right) - 1\right) \tag{E-6}$$

E.3 Combining the Continuous and Discrete Portions

The continuous portion of the modified delta-lognormal distribution is combined with the discrete portion to model data sets that contain a mixture of non-detected and detected measurements. It is possible to fit a wide variety of observed effluent data sets to the modified delta-lognormal distribution. Multiple detection limits for non-detect measurements are incorporated, as are measured ("detected") values. The same basic framework can be used even if there are no non-detected values in the data set (in this case, it is the same as the lognormal distribution). Thus, the modified delta-lognormal distribution offers a large degree of flexibility in modeling effluent data.

The modified delta-lognormal random variable U can be expressed as a combination of three other independent variables, that is,

$$U = I_u X_D + \left(1 - I_u\right) X_C \tag{E-7}$$

where X_D represents a random non-detect from the discrete portion of the distribution, X_C represents a random detected measurement from the continuous lognormal portion, and I_u is an indicator variable signaling whether any particular random measurement, u, is non-detected or non-censored (that is, I_u=1 if u is non-detected; I_u=0 if u is non-censored). Using a weighted sum, the cumulative distribution function from the discrete portion of the distribution (equation 1) can be combined with the function from the continuous portion (equation 4) to obtain the overall cumulative probability distribution of the modified delta-lognormal distribution as follows,

$$\Pr\left(U \le c\right) = \sum_{i:D_i \le c} \delta_i + \left(1 - \delta\right) \Phi\left[\frac{\ln(c) - \mu}{\sigma}\right] \tag{E-8}$$

where D_i is the value of the ith sample-specific detection limit.The expected value of the random variable *U* can be derived as a weighted sum of the expected values of the discrete and continuous portions of the distribution (equations 2 and 5, respectively) as follows

$$E\left(U\right) = \delta E(X_D) + \left(1 - \delta\right) E\left(X_C\right) \tag{E-9}$$

In a similar manner, the expected value of the random variable squared can be written as a weighted sum of the expected values of the squares of the discrete and continuous portions of the distribution as follows

$$E\left(U^2\right) = \delta E\left(X_D^2\right) + \left(1 - \delta\right) E\left(X_C^2\right) \tag{E-10}$$

Although written in terms of U, the following relationship holds for all random variables, U, X_D, and X_C.

$$E\left(U^2\right) = Var\left(U\right) + \left[E\left(U\right)\right]^2 \tag{E-11}$$

So using equation 11 to solve for Var(U), and applying the relationships in equations 9 and 10, the variance of *U* can be obtained as

$$\mathrm{Var}\left(U\right) = \delta\left(\mathrm{Var}\left(X_D\right) + \left[\mathrm{E}\left(X_D\right)\right]^2\right) + \left(1 - \delta\right)\left(\mathrm{Var}\left(X_C\right) + \left[\mathrm{E}\left(X_C\right)\right]^2\right) - \left[\mathrm{E}\left(U\right)\right]^2 \tag{E-12}$$

E.4 Episode Estimates Under the Modified Delta-Lognormal Distribution

In order to use the modified delta-lognormal model to calculate the limitations, the parameters of the distribution are estimated from the data. These estimates are then used to calculate the limitations.The parameters $\hat{\delta}_\iota$ and $\hat{\delta}$ are estimated from the data using the following formulas:

$$\hat{\delta}_i = \frac{1}{n}\sum_{j=1}^{n_d} I\left(d_j = D_i\right)$$
$$\hat{\delta} = \frac{n_d}{n} \tag{E-13}$$

where n_d is the number of non-detected measurements, d_j, j = 1 to n_d, are the detection limits for the non-detected measurements, n is the number of measurements (both detected and non-detected) and I(…) is an indicator function equal to one if the phrase within the parentheses is true and zero otherwise. The "hat" over the parameters indicates that they are estimated from the data.

The expected value and the variance of the delta portion of the modified delta-lognormal distribution can be calculated from the data as:

$$\hat{E}\left(X_D\right) = \frac{1}{\hat{\delta}}\sum_{i=1}^{k} \hat{\delta}_i D_i \tag{E-14}$$

$$\hat{V}ar\left(X_D\right) = \frac{1}{\hat{\delta}}\sum_{i=1}^{k} \hat{\delta}_i \left(D_i - \hat{E}\left(X_D\right)\right)^2 \tag{E-15}$$

The parameters of the continuous portion of the modified delta-lognormal distribution, $\hat{\mu}$ and $\hat{\sigma}^2$, are estimated by

$$\hat{\mu} = \sum_{i=1}^{n_c} \frac{\ln\left(x_i\right)}{n_c}$$
$$\hat{\sigma}^2 = \sum_{i=1}^{n_c} \frac{\left(\ln\left(x_i\right) - \hat{\mu}\right)^2}{n_c - 1} \tag{E-16}$$

where x_i is the i$^{\text{th}}$ detected measurement value and n_c is the number of detected measurements. Note that $n = n_d + n_c$.

The expected value and the variance of the lognormal portion of the modified delta-lognormal distribution can be calculated from the data as:

$$\hat{E}\left(X_C\right) = \exp\left(\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right) \tag{E-17}$$

$$\hat{V}ar\left(X_C\right) = \left[\hat{E}\left(X_C\right)\right]^2 \left(\exp\left(\hat{\sigma}^2\right) - 1\right) \tag{E-18}$$

Finally, the expected value and variance of the modified delta-lognormal distribution can be estimated using the following formulas:

$$\hat{E}\left(U\right) = \hat{\delta}\,\hat{E}\left(X_D\right) + \left(1 - \hat{\delta}\right)\hat{E}\left(X_C\right) \tag{E-19}$$

$$\hat{V}ar\left(U\right) = \hat{\delta}\left(\hat{V}ar\left(X_D\right) + \left[\hat{E}\left(X_D\right)\right]^2\right) + \left(1 - \hat{\delta}\right)\left(\hat{V}ar\left(X_C\right) + \left[\hat{E}\left(X_C\right)\right]^2\right) - \left[\hat{E}\left(U\right)\right]^2 \tag{E-20}$$

Equations 17 through 20 are particularly important in the estimation of episode long-term averages and variability factors as described in the following sections. These sections are preceded by a section that identifies the episode data set requirements.

Example:

Consider a facility that has 10 samples with the following concentrations:

Sample number	Measurement Type	Concentration (mg/L)
1	ND	10
2	ND	15
3	ND	15
4	ND	20
5	NC	25
6	NC	25
7	NC	30
8	NC	35
9	NC	35
10	NC	40

The ND components of the variance equation are:

D_1 = 10, $\hat{\delta}_1$ = 1/10D_2 = 15, $\hat{\delta}_2$ = 1/5D_3 = 20, $\hat{\delta}_3$ = 1/10.

Since $\hat{\delta}$ = 2/5, the expected value and the variance of the discrete portion of the modified delta-lognormal distribution are

$$\hat{E}\left(X_D\right) = \frac{1}{2/5}\left(\frac{1}{10} \times 10 + \frac{1}{5} \times 15 + \frac{1}{10} \times 20\right) = 15,$$

$$\hat{V}ar\left(X_D\right) = \frac{1}{2/5}\left(\frac{1}{10} \times (10-15)^2 + \frac{1}{5} \times (15-15)^2 + \frac{1}{10} \times (20-15)^2\right) = 12.5.$$

The mean and variance of the log NC values are calculated as follows: $\hat{\mu} = \frac{\sum_{i=1}^{n_c} \ln\left(x_i\right)}{n_c} = \frac{\left(2 \times \ln(25) + \ln(30) + 2 \times \ln(35) + \ln(40)\right)}{6} = 3.44$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n_c}\left(\ln\left(x_i\right) - \hat{\mu}\right)^2}{n_c - 1} = \frac{\left(2 \times \left(\ln(25) - 3.44\right)^2\right) + \left(\ln(30) - 3.44\right)^2 + \left(2 \times \left(\ln(35) - 3.44\right)^2\right) + \left(\ln(40) - 3.44\right)^2}{5} = 0.0376$$

Then, the expected value and the variance of the lognormal portion of the modified delta-lognormal distribution are

$$\hat{E}\left(X_C\right) = \exp\left(3.44 + \frac{0.0376}{2}\right) = 31.779$$

$$\hat{V}ar\left(X_C\right) = [31.779]^2\left(\exp(0.0376) - 1\right) = 38.695.$$

The expected value and variance of the modified delta-lognormal distribution are

$$\hat{E}(U) = \frac{2}{5} \times 15 + \left(1 - \frac{2}{5}\right) \times 31.779 = 25.063$$

$$\hat{V}ar(U) = \frac{2}{5} \times \left(12.5 + 15^2\right) + \left(1 - \frac{2}{5}\right) \times \left(38.695 + 31.779^2\right) - 25.067^2 = 95.781.$$

E.4.1 Episode Data Set Requirements

Estimates of the necessary parameters for the lognormal portion of the distribution can be calculated with as few as two distinct detected values in a data set. (In order to calculate the variance of the modified delta-lognormal distribution, two distinct detected values are the minimum number that can be used and still obtain an estimate of the variance for the distribution.)

If an episode data set for a pollutant contained three or more observations with two or more distinct detected concentration values, then EPA used the modified delta-lognormal distribution to calculate long-term averages and variability factors. If the episode data set for a pollutant did not meet these requirements, EPA used an arithmetic average to calculate the episode long-term average and excluded the dataset from the variability factor calculations (because the variability could not be calculated).

In statistical terms, each measurement was assumed to be independently and identically distributed from the other measurements of that pollutant in the episode data set.

The next two sections apply the modified delta-lognormal distribution to the data for estimating episode long-term averages and variability factors for the aquatic animals industry.

E.4.2 Estimation of Episode Long-Term Averages

If an episode dataset for a pollutant mets the requirements described in the last section, then EPA calculated the long-term average using equation 19. Otherwise, EPA calculated the long-term average as the arithmetic average of the daily values where the sample-specific detection limit was used for each non-detected measurement.

E.4.3 Estimation of Episode Variability Factors

For each episode, EPA estimated the daily variability factors by fitting a modified delta-lognormal distribution to the daily measurements for each pollutant. In contrast, EPA estimated monthly variability factors by fitting a modified delta-lognormal distribution to the monthly averages for the pollutant at the episode. EPA developed these averages using the same number of measurements as the assumed monitoring frequency for the pollutant. EPA is assuming that all pollutants will be monitored weekly (approximately four times a month).[5]

[5] Compliance with the monthly average limitations will be required in the final rulemaking regardless of the number of samples analyzed and averaged.

E.4.3.1 Estimation of Episode Daily Variability Factors

The episode daily variability factor is a function of the expected value, and the 99th percentile of the modified delta-lognormal distribution fit to the daily concentration values of the pollutant in the wastewater from the episode. The expected value, was estimated using equation 19 (the expected value is the same as the episode long-term average).

The 99th percentile of the modified delta-lognormal distribution fit to each data set was estimated by using an iterative approach. First, the pollutant-specific detection limits were ordered from smallest to largest. Next, the cumulative distribution function, p, for each detection limit was computed. The general form, for a given value c, was:

$$p = \sum_{i:D_i \le c} \hat{\delta}_i + \left(1 - \hat{\delta}\right)\Phi\left[\frac{\ln(c) - \hat{\mu}}{\hat{\sigma}}\right] \qquad \text{(E-21)}$$

where Φ is the standard normal cumulative distribution function. Next, the interval containing the 99th percentile was identified. Finally, the 99th percentile of the modified delta-lognormal distribution was calculated. The following steps were completed to compute the estimated 99th percentile of each data subset:

Step 1 Using equation 21, k values of p at $c=D_m$, m=1,...,k were computed and labeled p_m.

Step 2 The smallest value of m (m=1,...,k), such that $p_m \geq 0.99$, was determined and labeled as p_j. If no such m existed, steps 3 and 4 were skipped and step 5 was computed instead.

Step 3 Computed $p^* = p_j - \hat{\delta}_j$.

Step 4 If $p^* < 0.99$, then $\hat{P}99 = D_j$ else if $p^* \geq 0.99$, then

$$\hat{P}99 = \exp\left(\hat{\mu} + \hat{\sigma}\Phi^{-1}\left[\frac{0.99 - \sum_{i=1}^{j-1} \hat{\delta}_i}{1 - \hat{\delta}}\right]\right) \qquad \text{(E-22)}$$

where Φ^{-1} is the inverse normal distribution function.

Step 5 If no such m exists such that $p_m > 0.99$ (m=1,...,k), then

$$\hat{P}99 = \exp\left(\hat{\mu} + \hat{\sigma}\Phi^{-1}\left[\frac{0.99 - \hat{\delta}}{1 - \hat{\delta}}\right]\right) \tag{E-23}$$

The episode daily variability factor, VF1, was then calculated as:

$$VF1 = \frac{\hat{P}99}{\hat{E}(U)} \tag{E-24}$$

Example:

Since no such m exists such that $p_m > 0.99$ (m=1,...,k),

$$\hat{P}99 = \exp\left(3.44 + 0.194 \times \Phi^{-1}\left[\frac{0.99 - 0.4}{1 - 0.4}\right]\right) = 47.126.$$

The episode daily variability factor, VF1, was then calculated as:

$$VF1 = \frac{47.126}{25.067} = 1.880.$$

E.4.3.2 Estimation of Episode Monthly Variability Factors

EPA estimated the monthly variability factors by fitting a modified delta-lognormal distribution to the monthly averages. These equations use the same basic parameters, μ and σ, calculated for the daily variability factors. Episode monthly variability factors were based on 4-day monthly averages because the monitoring frequency was assumed to be weekly (approximately four times a month).

Before estimating the episode monthly variability factors, EPA considered whether autocorrelation was likely to be present in the effluent data. When data are said to be positively autocorrelated, it means that measurements taken at specific time intervals (such as 1 day or 1 week apart) are related. For example, positive autocorrelation would be present in the data if the final effluent concentration of TSS was relatively high one day and was likely to remain at similar high values the next and possibly succeeding days. Because EPA is assuming that the pollutants will be monitored weekly, EPA based the monthly variability factors on the distribution of the averages of four measurements. If concentrations measured on consecutive weeks were positively correlated, then the autocorrelation would have had an effect on the estimate of the variance of the monthly average and thus on the monthly variability factor. Adjustments for positive autocorrelation would increase the values of the variance and monthly variability factor. (The estimate of the long-term average and the daily variability factor are generally only slightly affected by autocorrelation.)

EPA has not incorporated an autocorrelation adjustment into its estimates of the monthly variability factors. In some industries, measurements in final effluent are likely to be similar from one day (or week) to the next because of the consistency from day-to-day in the production processes and in final effluent discharges due to the hydraulic retention time of wastewater in basins, holding tanks, and other components of wastewater treatment systems. To determine if autocorrelation exists in the data, a statistical evaluation is necessary and will be considered before the final rule. To estimate autocorrelation in the data, many measurements for each pollutant would be required with values for equally spaced intervals over an extended period of time. If such data are available for the final rule, EPA intends to perform a statistical evaluation of autocorrelation and if necessary, provide any adjustments to the limitations.

Thus, in calculating the monthly variability factors for the proposal, EPA assumed that consecutive daily measurements were not correlated. In order to calculate the 4-day variability factors (VF4), EPA further assumed that the approximating distribution of $\overline{U}_4$, the sample mean for a random sample of four independent concentrations, was derived from the modified delta-lognormal distribution.[6] To obtain the expected value of the 4-day averages, equation 19 is modified for the mean of the distribution of 4-day averages in equation 25:

$$\hat{E}\left(\overline{U}_4\right) = \hat{\delta}_4' \, \hat{E}\left(\overline{X}_4\right)_D + \left(1 - \hat{\delta}_4'\right)\hat{E}\left(\overline{X}_4\right)_C \tag{25}$$

where $\hat{\delta}_4'$ denotes the probability of detection of the 4-day average, $\left(\overline{X}_4\right)_D$ denotes the mean of the discrete portion of the distribution of the average of four independent concentrations, (i.e., when all observations are non-detected values), and $\left(\overline{X}_4\right)_C$ denotes the mean of the continuous lognormal portion (i.e., when any observations are detected).

First, it was assumed that the probability of detection (δ) on each of the four days was independent of the measurements on the other three days (as explained in Section E.4.1, daily measurements were also assumed to be independent) and therefore, $\delta'_4 = \delta^4$. Because the measurements are assumed to be independent, the following relationships hold:

[6] As described in Section 8.4, when non-detected measurements are aggregated with non-censored measurements, EPA determined that the result should be considered non-censored.

$$\begin{aligned} \hat{E}\left(\overline{U}_4\right) &= \hat{E}(U) \\ \hat{V}ar\left(\overline{U}_4\right) &= \frac{\hat{V}ar(U)}{4} \\ \hat{E}\left(\left(\overline{X}_4\right)_D\right) &= \hat{E}\left(X_D\right) \\ \hat{V}ar\left(\left(\overline{X}_4\right)_D\right) &= \frac{\hat{V}ar\left(X_D\right)}{4} \end{aligned} \tag{E-26}$$

Substituting into equation 26 and solving for the expected value of the continuous portion of the distribution gives:

$$\hat{E}\left(\overline{X}_4\right)_C = \frac{\hat{E}(U) - \hat{\delta}^4\, \hat{E}\left(X_D\right)}{1-\hat{\delta}^4} \tag{E-27}$$

Using the relationship in equation 20 for the averages of 4 daily measurements and substituting terms from equation 25 and solving for the variance of the continuous portion of $\overline{U}_4$ gives:

$$\hat{V}ar\left(\overline{X}_4\right)_C = \frac{\dfrac{\hat{V}ar(U)}{4} + \left[\hat{E}(U)\right]^2 - \hat{\delta}^4 \left(\dfrac{\hat{V}ar\left(X_D\right)}{4} + \left[\hat{E}\left(X_D\right)\right]^2\right)}{1-\hat{\delta}^4} - \left[\hat{E}\left(\overline{X}_4\right)_C\right]^2 \tag{E-28}$$

Using equations 17 and 18 and solving for the parameters of the lognormal distribution describing the distribution of $\left(\overline{X}_4\right)_C$ gives:

$$\hat{\sigma}_4^2 = \ln\left(\frac{\hat{V}ar\left(\overline{X}_4\right)_C}{\left(\hat{E}\left(\overline{X}_4\right)_C\right)^2} + 1\right)$$

and (E-29)

$$\hat{\mu}_4 = \ln\left(\hat{E}\left(\overline{X}_4\right)_C\right) - \frac{\hat{\sigma}_4^2}{2}$$

In finding the estimated 95th percentile of the average of four observations, four non-detects, not all at the same sample-specific detection limit, can generate an average that is not necessarily equal to D_1, D_2,..., or D_k. Consequently, more than k discrete points exist in the distribution of the 4-day averages. For example, the average of four non-detects at k=2 detection limits, are at the following discrete points with the associated probabilities:

i	D_i^*	δ_i^*
1	D_1	δ_1^4
2	$\left(3D_1 + D_2\right)/4$	$4\delta_1^3\delta_2$
3	$\left(2D_1 + 2D_2\right)/4$	$6\delta_1^2\delta_2^2$
4	$\left(D_1 + 3D_2\right)/4$	$4\delta_1\delta_2^3$
5	D_2	δ_2^4

When all four observations are non-detected values, and when k distinct non-detected values exist, the multinomial distribution can be used to determine associated probabilities. That is,

$$\Pr\left[\overline{U}_4 = \frac{\sum_{i=1}^{k} u_i D_i}{4}\right] = \frac{4!}{u_1!u_2!\ldots u_k!}\prod_{i=1}^{k}\delta_i^{u_i} \qquad \text{(E-30)}$$

where u_i is the number of non-detected measurements in the data set with the D_i detection limit. The maximum number of possible discrete points, k^*, for k=1,2,3,4, and 5 are as follows:

k	k*	1	1
2	5	3	15
4	35	5	70

To find the estimated 95th percentile of the distribution of the average of four observations, the same basic steps (described in Section E.4.3.1) as for the 99th percentile of the distribution of daily observations, were used with the following changes:

Step 1 Change P_{99} to P_{95}, and 0.99 to 0.95.

Step 2 Change D_m to D_m^*, the weighted averages of the sample-specific detection limits.

Step 3 Change δ_i to δ_i^*.

Step 4 Change k to k^*, the number of possible discrete points based on k detection limits.

Step 5 Change the estimates of δ, $\hat{\mu}$,and $\hat{\sigma}$ to estimates of δ^4, $\hat{\mu}_4$ and $\hat{\sigma}_4^2$ respectively.

Then, using $\hat{E}\left(\overline{U}_4\right) = \hat{E}\left(U\right)$, the estimate of the episode 4-day variability factor, VF4, was calculated as:

$$VF4 = \frac{\hat{P}95}{\hat{E}\left(U\right)} \tag{E-31}$$

Example:

$$\hat{E}\left(\overline{U}_4\right) = 25.067$$

$$\hat{V}ar\left(\overline{U}_4\right) = \frac{95.781}{4} = 23.95$$

$$\hat{E}\left(\left(\overline{X}_4\right)_D\right) = 15$$

$$\hat{V}ar\left(\left(\overline{X}_4\right)_D\right) = \frac{12.5}{4} = 3.125$$

$$\hat{E}\left(\overline{X}_4\right)_C = \frac{25.067 - 0.4^4 \times 15}{1 - 0.4^4} = \frac{24.683}{0.974} = 25.331.$$

$$\hat{V}ar\left(\overline{X}_4\right)_C = \frac{23.95 + 25.067^2 - 0.4^4 \times \left(3.125 + 15^2\right)}{1 - 0.4^4} - 25.331^2 = 21.789$$

$$\hat{\sigma}_4^2 = \ln\left(\frac{21.789}{25.331^2} + 1\right) = 0.0334 \quad \hat{\delta}_4' = \hat{\delta}^4 = \left(\frac{2}{5}\right)^4 = 0.0256$$

$$\hat{\mu}_4 = \ln(25.331) - \frac{0.0334}{2} = 3.215.$$

$$\hat{P}95 = \exp\left(3.215 + 0.183 \times \Phi^{-1}\left[\frac{0.95 - 0.4^4}{1 - 0.4^4}\right]\right) = 33.683.$$

$$VF4 = \frac{33.683}{25.067} = 1.344.$$

E.4.3.3 Evaluation of Episode Variability Factors

Estimates of the necessary parameters for the lognormal portion of the distribution can be calculated with as few as two distinct measured values in a data set (in order to calculate the variance); however, these estimates can be unstable (as can estimates from larger data sets). As stated in Section E.4.1, EPA used the modified delta-lognormal distribution to develop episode variability factors for data sets that had a three or more observations with two or more distinct measured concentration values.

To identify situations producing unexpected results, EPA reviewed all of the variability factors and compared daily to monthly variability factors. EPA used several criteria to determine if the episode daily and monthly variability factors should be included in calculating the option variability factors. One criteria that EPA used was that the daily and monthly variability factors should be greater than 1.0. A variability factor less than 1.0 would result in a unexpected result where the estimated 99th percentile would be less than the long-term average. This would be an indication that the estimate of $\hat{\sigma}$ (the log standard deviation) was unstable. A second criteria was that the daily variability factor had to be greater than the monthly variability factor. A third criteria was that not all of the sample-specific detection limits could exceed the values of the non-censored values. All the episode variability factors used for the limitations and standards met these criteria.

E.5 References

Aitchison, J. and J.A.C. Brown. 1963. *The Lognormal Distribution.* Cambridge University Press, New York.

Barakat, R. 1976. Sums of Independent Lognormally Distributed Random Variables. *Journal of the Optical Society of America* 66: 211-216.

Cohen, A. Clifford. 1976. Progressively Censored Sampling in the Three Parameter Log-Normal Distribution. *Technometrics* 18:99-103.

Crow, E.L. and K. Shimizu. 1988. *Lognormal Distributions: Theory and Applications*. Marcel Dekker, Inc., NY.

Kahn, H.D., and M.B. Rubin. 1989. Use of Statistical Methods in Industrial Water Pollution Control Regulations in the United States. *Environmental Monitoring and Assessment,* vol. 12:129-148.

Owen, W.J. and T.A. DeRouen. 1980. Estimation of the Mean for Lognormal Data Containing Zeroes and Left-Censored Values, with Applications to the Measurement of Worker Exposure to Air Contaminants. *Biometrics* 36:707-719.

U.S. Environmental Protection Agency. 2000. *Development Document for Effluent Limitations Guidelines and Standards for the Centralized Waste Treatment Point Source Category*. Volume I, Volume II. EPA 440/1-87/009.