Development Document for the Proposed Effluent Limitations Guidelines and Standards for the Meat and Poultry Products Industry Point Source Category (40 CFR 432)

EPA-821-B-01-007

January 2002

U.S. Environmental Protection Agency
Office of Water (4303T)
Washington, DC 20460

Complete proposed document available at:
http://www.epa.gov/ost/guide/mpp/
The Final Development Document is available as well.

APPENDIX G

MODIFIED DELTA-LOGNORMAL DISTRIBUTION

This appendix describes the modified delta-lognormal distribution and the estimation of the episode-specific long-term averages and variability factors used to calculate the proposed limitations and standards.[1] This appendix provides the statistical methodology that was used to obtain the results presented in Section 13.

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

EPA selected the modified delta-lognormal distribution to model pollutant effluent concentrations from the meat products industry in developing the long-term averages and variability factors. A typical effluent data set from a sampling episode or self-monitoring episode (see Section 13 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.

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.

---

[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.

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

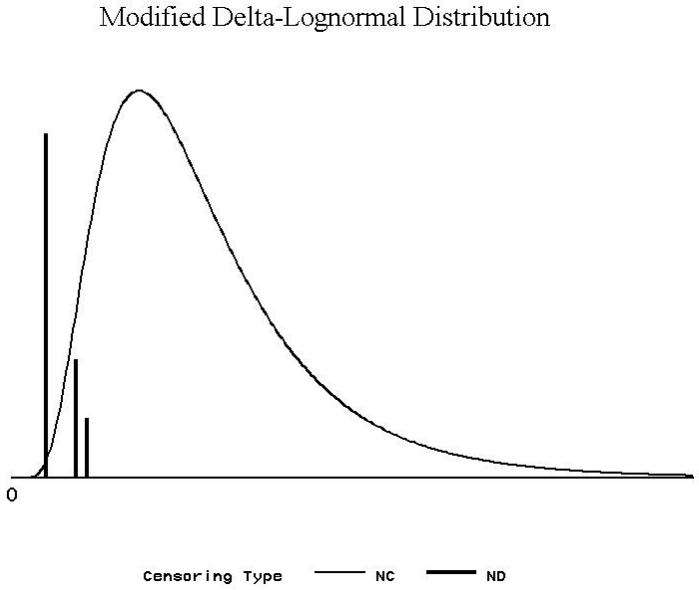

Figure G-1.

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

G.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

[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 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.

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 represent 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:

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

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

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

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

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

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

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:

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

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

G.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,

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

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,

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

where D_i is the value of the i^{th} 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

$$E(U) = \delta E(X_D) + (1-\delta)E(X_C) \qquad \text{(G-9)}$$

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

$$E(U^2) = \delta E(X_D^2) + (1-\delta)E(X_C^2) \qquad \text{(G-10)}$$

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

$$E(U^2) = Var(U) + [E(U)]^2 \qquad \text{(G-11)}$$

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

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

G.4 Episode-specific Estimates Under the Modified Delta-Lognormal Distribution

In order to use the modified delta-lognormal model to calculate the proposed limitations, the parameters of the distribution are estimated from the data. These estimates are then used to calculate the proposed limitations.

The parameters $\hat{\delta}_i$ and $\hat{\delta}$ are estimated from the data using the following formulas:

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

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 lognormal portion of the modified delta-lognormal distribution can be calculated from the data as:

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

$$\hat{Var}\left(X_D\right) = \frac{1}{\hat{\delta}}\sum_{i=1}^{k} \hat{\delta}_i \left(D_i - E\left(X_D\right)\right)^2 \qquad \text{(G-15)}$$

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

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

where x_i is the ith 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:

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

$$\hat{Var}(X_C) = \left[\hat{E}(X_C)\right]^2 \left(\exp\left(\hat{\sigma}^2\right) - 1\right) \qquad \text{(G-18)}$$

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

$$\hat{E}(U) = \hat{\delta}\hat{E}(X_D) + \left(1 - \hat{\delta}\right)\hat{E}(X_C) \qquad \text{(G-19)}$$

$$\hat{Var}(U) = \hat{\delta}\left(\hat{Var}(X_D) + \left[\hat{E}(X_D)\right]^2\right) + \left(1 - \hat{\delta}\right)\left(\hat{Var}(X_C) + \left[\hat{E}(X_C)\right]^2\right) - \left[\hat{E}(U)\right]^2 \qquad \text{(G-20)}$$

Equations 17 through 20 are particularly important in the estimation of episode-specific 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.

G.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-specific 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-specific long-term averages and variability factors for the iron and steel industry.

G.4.2 Estimation of Episode-specific Long-Term Averages

If an episode dataset for a pollutant met 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.

G.4.3 Estimation of Episode-Specific 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 daily.[5]

G.4.3.1 Estimation of Episode-specific Daily Variability Factors

The episode-specific 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-specific 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:

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

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:

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

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

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

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

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

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

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

G.4.3.2 Estimation of Episode-Specific 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-specific monthly variability factors were based on 30-day monthly averages because the monitoring frequency was assumed to be daily (approximately thirty times a month). As explained in Section 13.6.2, EPA recognizes that small poultry facilities are unlikely to operate on weekends and is soliciting comment on whether their monthly limitations should be based upon 20 days. This section describes the calculations for monthly variability factors based upon 30-day averages. To calculate the monthly variability factors based upon 20 days, the same basic procedure is used except that 20-day averages are used instead of 30-day averages.

Before estimating the episode-specific 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 2 days apart) are related. For example, positive autocorrelation would be present in the data if the final effluent concentration of HEM 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 daily, EPA based the monthly variability factors on the distribution of the averages of 30 (or 20) measurements. If concentrations measured on consecutive days 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 many industries, measurements in final effluent are likely to be similar from one day 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 ponds, and other components of wastewater treatment systems. To determine if autocorrelation exists in the data, a statistical evaluation is necessary. However, the data used for the proposal were insufficient for the purpose of evaluating autocorrelation. To estimate autocorrelation in the data, many measurements for each pollutant would be required with values

for every single day 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.

In calculating the monthly variability factors, EPA assumed that consecutive daily measurements were not correlated, and therefore

$$\hat{E}\left(\overline{U}_{30}\right)=\hat{E}\left(U\right) \qquad and \qquad \hat{Var}\left(\overline{U}_{30}\right)=\frac{\hat{Var}\left(U\right)}{30} \tag{G-25}$$

where $\hat{E}\left(U\right)$ and $\hat{Var}\left(U\right)$ were calculated as shown in equations 19 and 20. Finally, because $\overline{U}_{30}$ is approximately normally distributed by the Central Limit Theorem, the estimate of the 95th percentile of a 30-day mean and the corresponding episode-specific 30-day variability factor (VF30) were approximated by

$$\hat{P}95_{30}=\hat{E}\left(\overline{U}_{30}\right)+\left[\Phi^{-1}\left(0.95\right)\right]\sqrt{\hat{Var}\left(\overline{U}_{30}\right)} \tag{G-26}$$

where $\Phi^{-1}(0.95)$ is the 95th percentile of the inverse normal distribution. By using the substitutions in equation 25, equation 26 simplified to

$$\hat{P}95_{30}=\hat{E}\left(U\right)+\left[\Phi^{-1}\left(0.95\right)\right]\sqrt{\tfrac{1}{30}\hat{Var}\left(U\right)} \tag{G-27}$$

Then

$$VF30=\frac{\hat{P}95}{\hat{E}\left(U\right)} \qquad because \qquad \hat{E}\left(\overline{U}_{30}\right)=\hat{E}\left(U\right) \tag{G-28}$$

G.4.3.3 Evaluation of Episode-Specific 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 G.4.1, EPA used the modified delta-lognormal distribution to develop episode-

specific 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-specific 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. All the episode-specific variability factors used for the proposed limitations and standards met these criteria.

G.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., New York.

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.

U.S. Environmental Protection Agency. 2000. *Development Document for Proposed Effluent Limitations Guidelines and Standards for the Iron and Steel Manufacturing Point Source Category*. EPA-821-B-99-011.