# APPENDIX F
# ALTERNATIVE STATISTICAL METHODS

# APPENDIX F: ALTERNATIVE STATISTICAL METHODS

This appendix describes statistical methods that EPA may consider for modeling the effluent data for developing the final limitations and standards for the concentrated aquatic animal production (CAAP) industry. A typical CAAP 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 concentrations and values reported as being less than some sample-specific detection limit (e.g., <10 mg/L) or "non-detected." In statistical terms, measured concentrations are "non-censored" and non-detected values are "left-censored." The distinction between non-censored and left-censored measurements is often important in modeling the data and each model described in this appendix has different underlying assumptions about the physical processes that generate non-censored and left-censored measurements. For example, the modified delta-lognormal distribution assumes that they are generated from different processes and models the non-detected values using a delta distribution, while the censored lognormal distribution assumes that all observations (non-censored and non-detected) are regarded as random measurements generated from a common underlying lognormal distribution. In the censored lognormal model, non-detect measurements are treated as left-censored observations in the lognormal distribution.

Section F.1 provides a brief summary of the modified delta-lognormal distribution that was used for the proposal and is described in Appendix E. The remaining sections discuss another modification of delta-lognormal distribution, the censored lognormal distribution, the probability regression method for the lognormal distribution, and nonparametric methods. Before the final rule, EPA will evaluate the appropriateness of these models for the CAAP industry effluent data. EPA also will evaluate whether the predicted values are consistent with the observed effluent values.

## F.1 MODIFIED DELTA-LOGNORMAL MODEL

For the proposed, EPA used the modified delta-lognormal distribution to model the effluent concentrations from the CAAP industry. As explained in Appendix E, this distribution models the data as a mixture of measurements that follow a lognormal distribution and non-detected measurements that occur with a certain probability (Aitchison and Brown (1963), Kahn and Rubin (1989), and U.S. EPA (1993)). By a modification to the delta portion of the distribution, this model also allows for the possibility that non-detected measurements can be observed at different sample-specific detection limits.

For some industries, different pollutant-generating mechanisms appear to act to produce non-censored and non-detected measurements at a facility. For example, non-detected measurements may indicate that the pollutant is not generated by a particular source or production practice, and non-censored values may be generated by different source, production, and/or wastewater treatment conditions. The modified delta-lognormal

distribution is appropriate for such data sets because each data type (i.e., non-censored measurements and non-detected measurements) is modeled separately with different distributional properties. For the final rule, EPA will evaluate whether this assumption is appropriate for CAAP data.

## F.2 ANOTHER MODIFICATION OF THE DELTA-LOGNORMAL MODEL

Another possible model for the CAAP effluent data is a further modification of the delta-lognormal distribution described in the previous section. This modification would incorporate left-censoring into the lognormal portion of the model while retaining the delta distribution for the non-detected measurements. This model would explicitly censor the lognormal distribution at some point, such as the minimum sample-specific detection limit observed in a data set. The lognormal distribution would be censored at this point because laboratory instruments would be incapable of measuring below that point and would be reported as non-detected values. Thus, non-censored values would be assumed to be observed only above this point. This modification is based upon an extension of the method developed by Moulton and Halsey (1995). EPA used a similar modification in developing the limitations for the pulp and paper industry (USEPA). Its implementation resulted in only minor differences from the values obtained from the model described in Section F.1.

## F.3 CENSORED LOGNORMAL DISTRIBUTION

In a censored lognormal model (see Cohen, 1959), all observations (non-censored and non-detected) are regarded as random measurements generated from a common underlying lognormal distribution. Estimates of the mean, variance, and upper percentiles, used as a basis of the limitations, can be computed from the estimated best-fitting lognormal distribution. These estimates are similar to those derived under the modified delta-lognormal model, except that in Cohen's procedure non-detected measurements are treated merely as one type of censored sample, namely left-censored. Thus, it is assumed that non-detects, if the true concentration or mass amounts were measurable, would follow the same lognormal pattern as the rest of the data set.

## F.4 PROBABILITY REGRESSION METHOD FOR THE LOGNORMAL DISTRIBUTION

The probability regression method assumes that the entire data set would follow a specific distributional model (e.g., the lognormal distribution) if concentrations of non-detected measurements could be observed. The basic idea behind the probability regression technique can be described by first considering the case with no censored measurements (for instance, a set of detected and precisely known observations). If it is assumed that the data were generated by an underlying lognormal distribution, then it would be expected that the logged values would plot on a probability plot in roughly a linear pattern when graphed against ordered quantiles from a standard normal distribution. In fact, it would be possible in this case to fit a linear regression to the points

on the probability plot and determine the slope and intercept of the regression equation. The slope and intercept of this regression equation allow the estimation of an "optimal" set of parameters for fitting a specific lognormal density to the observed data. When the censored data are non-detects exhibiting multiple detection limits, and the set of detection limits overlaps the set of detected values, the desired ordering of the data is more difficult to construct. However, Helsel and Cohn (1988) adapt the simpler probability regression method with a single detection limit to the more general case of multiple detection limits and overlapping of non-censored and non-detected measurements. This adaptation orders the data in terms of conditional probabilities. EPA will evaluate whether an ordering of the non-detected values is appropriate for the CAAP effluent data.

## F.5 NONPARAMETRIC METHODS

In contrast to the other statistical methods discussed in this appendix, nonparametric methods are not based on fitting a distribution to the data. The nonparametric estimate of the 99th percentile of an effluent concentration data set is the observed value that exceeds 99 percent of the data points. If a data set consists of fewer than 100 observations the best that can be done, using nonparametric methods, is to use the maximum value as an approximate nonparametric estimate of the 99th percentile, but this will underestimate the true value (in statistical expectation). Because most of the data sets analyzed in support of limitations development had fewer than 100 observations, it was prudent to adopt a parametric approach, such as the modified delta-lognormal distribution, to avoid underestimating the values used as a basis of the limitations. EPA will determine if these sample size constraints exist for the final rule.

## F.6 REFERENCES

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

Cohen, A.C., Jr. 1959. Simplified estimators for the normal distribution when samples are singly censored or truncated. *Technometrics*, vol. 1, pp. 217-237.

Helsel. D.R. and T.A. Cohn. 1988. Estimation of descriptive statistics for multiply censored water quality data. *Water Resources Research*, vol. 24, no. 12, pp. 1997-2004.

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, pp. 129-148.

Moulton, L.H. and N.A. Halsey. 1995. A mixture model with detection limits for regression analysis of antibody response to vaccine. *Biometrics*, vol. 51, pp. 1197-1205.

U.S. Environmental Protection Agency (USEPA). 1993. *Statistical Support Document for Proposed Effluent Limitations Guidelines and Standards for the Pulp, Paper, and Paperboard Point Source Category*. EPA 821-R-93-023. U.S. Environmental Protection Agency, Washington, DC.