Avoiding Costs of Mis-Analyzed Environmental Data(Newsletter)*

May 29, 2012

Every business requiring a discharge permit under the Clean Air or Clean Water Acts must monitor discharges and report analytical results to regulators. Excluding, or improperly including, concentrations below analytical detection limits can be expensive and impose unnecessary constraints on operations. For example, mercury in air, water, and fish is a nationwide concern with concentrations often below reporting levels.

To comply with environmental regulations (such as TMDLs), an operator must show that the daily mean mercury concentration in its wastewater or air discharges do not exceed the legal standard. Yet many of the analyzed samples have concentrations below the laboratory's reporting limit. These "less-thans" (also known as "censored" data) make it impossible to compute a simple mean concentration. When the operator substitutes a zero for each less-than, the standard is not exceeded. When the regulatory agency substitutes a value equal to the reporting limit, the standard is exceeded. Which is correct? Has the law been violated? Controls to meet the regulator's analytical approach can cost millions of dollars to develop, install, and maintain. Similar situations exist with arsenic, nitrates, magnesium, sulfates and many other chemicals, particularly in water.

Often, ground water chemistry is measured both upgradient and downgradient of a waste disposal site. Comparison of the two groups of data is performed to determine whether ground water is contaminated. Usually t-tests are employed for this purpose, yet the t-test requires estimates of means and standard deviations which are impossible to obtain unless numerical values are fabricated to replace any "less-thans" ("censored") data. By substituting one number, the two groups appear the same; substituting a second number causes the null hypothesis to be rejected, and the two groups to be declared different. Which is correct?

As part of environmental assessments or permit compliance monitoring biotic data (macroinvertebrates and fish) are collected. Organisms are not always present, observed, or captured during these surveys and the data contains many instances of no animals to record. Statisticians call an excessive num-

ber of zeros "zero-inflated data." The word "inflation" emphasize the probability of the number of instances with zero organisms in the population exceeding that allowed under a standard parametric family of distributions. Zero-inflated data are abundant in environmental data. If not properly modeled, the presence of excess zeros can invalidate the distributional assumptions of the data analyses and their interpretation. High variability in biotic data as well as taxonomic identifications cause other errors when not correctly accommodated. Unnecessary restrictions on operations or permit issuance might result with loss to operators, regulators, and the public.

Proper analyses of water, air, and soil chemical concentrations includes the censored data. Both parametric and non-parametric methods may be used. The best approaches use either maximum likelihood estimation (MLE) to find an empirical distribution best fitting the uncensored ("detected") values and used to describe the unseen censored values or survival analyses (developed in medical and industrial time-to-failure studies). Imputation, or inserting specific numbers for censored values, may be done with the appropriate approach but is less robust than are MLE or survival analyses.

Zero-inflated biological data can be validly analyzed using Bayesian methods (including Markov Chain Monte Carlo simulations), MLE using Poisson or Lognormal distributions, and mixed additive general regression models. The most technically sound and legally defensible method for biotic data depends on the specific data set; different approaches must be applied and their robustness and ability to explain observed variability quantified. There are well-established methods to make these assessments.

After properly describing and comparing data operators and regulators should determine why the specific results were obtained. Among the potential explanatory factors are location, chemical and biological interactions, time, and methods of collection and analysis. Only when observed data are robustly explained and solidly understood can informed decisions be made by operators and regulators.