Analysis of data with below detection limit values

**General Books / Book Chapters**

  Focuses on analysis of groundwater well data (repeated samples from fixed points) and regulatory statistics. Chapter 5 is a very nice summary of different ways to compute the d.l. Chapter 10 describes analysis of below dl values, emphasizing P[est. mean > action level].

  My favorite introductory book, but now a bit out of date. Clearly written. Focuses on useful techniques and avoids being an encyclopedia. Chapter 14 describes < dl values.

  Applied, emphasizing how to carry out various methods, and comparison between methods. More details on computations than Helsel and Hirsch. I disagree with a few issues, e.g., se’s of means.


  Diverse methods in environmental statistics, same technical level as Helsel. Chapter 10 deals with < dl values.

**Definition of detection limits**

  Source of the 3σ and 10σ limits.


  Alt. method of computing detection limit, allowing error variance to vary with the mean concentration.

  Radionuclide analyses are based on counts of events and Poisson distributions. Here is an entry to the literature on detection limits for Poisson events.
Overview papers

Emphasizes statistical theory underlying various estimation and test methods.

An attempt to convince environmental scientists to use methods other than substitution.

Argues for Kaplan-Meier based approaches. I don’t agree with all Dennis’s conclusions.

Helsel, D.R. 2006. Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere* 65:2434-2439.
A recent attempt to convince environmental scientists to use methods other than substitution.


Maximum likelihood estimation

Cohen, A.C. 1959. Simplified estimators for the normal distribution when samples are singly censored or truncated. *Technometrics* 1:217-237
Original derivation of the m.l.e. Includes a “simple” adjustment method for hand calculation using tables (Cohen 1961).


Book length treatment, emphasizing normal populations

mle of correlation coefficient when both X and Y include <dl values.

Multiply censored obs often arise from a lab measurement with a single dl divided by sample weight that varies between observations. Here’s an ingeneous way to incorporate that info. I haven’t read the paper in detail, but I suspect this would be a very promising MS or perhaps PhD topic.


Describes tobit regression, i.e. mle

**Robust order statistics estimator**


Original paper describing the ROS (robust order statistics) method


Simulation evaluation of ROS and various other estimators


Method for determining plotting positions (quantiles) when multiple detection limits. Motivating example is flood height, but method applies to detection limits.

**Log normal distributions**


Derived the UMVUE of the mean of a lognormal distribution.


Derivation of the distribution of sum of normal + k*Chi-square random variables. Useful for confidence intervals for the log normal mean.


Tables of coefficients used to calculate quantiles and confidence intervals.


Uses data sets from large regional surveys to argue that environmental chemical data are neither normally or lognormally distributed. “Distributions are usually skewed, have outliers, and originate from more than one process”.

Note: I suspect the last phrase above is the key: how much of their conclusion arises because a regional survey includes many different environments, each with its own distribution of the chemical.


Both Shumway and Stoline argue that log() is too restrictive and Box-Cox family should be used instead.

**Tests**


Compares empirical type I error rate and power for various score tests, especially when censoring patterns are not the same in the two groups.


Original paper applying score tests to two sample environmental problems. Actual method proposed by M and D uses normal scores, a slightly different score function than the one I discussed in class.


A biomedical application where distribution is assumed a mixture of a log normal and extra zeros.

**Confidence intervals**


Develops the mle for data with multiple censoring limits. Develops confidence interval estimators that account for the correlation between mean and variance.


Develops an asymptotic interval and bootstrap intervals. Uses the bootstrap sampling distribution to decide whether to trust the asymptotic interval.
Evaluation of methods

There are many, many of these. All use numerical simulation to evaluate performance (bias, variance, mse, type I error rate, power, or perhaps something else) for a collection of estimators and/or tests. Helsel (2012), p. 87-92, summarizes many of these.

Spatial analysis of below-detection limit data


Fridley, B.L. and Dixon, P. 2007. Data augmentation for a Bayesian spatial model involving censored observations *Environmetrics* 18:107-123.