Summing Nondetects

Estimating Total Mass of Low-Level Contaminants

Dennis Helsel

Practical Stats
The Problem

- Components (congeners) are to be summed to obtain a total mass
- or Regular (monthly) values are used to estimate a (yearly) total
- Some individual values are below the reporting limit of the laboratory (nondetect)
- How can values like these be summed?

12 <10 15 22 <5 18 9 ?
Methods for computing mass with nondetects

• Estimate the mean for data with nondetects
  – Substitution (0, 1/2 DL, DL)
  – Maximum Likelihood
  – Kaplan-Meier
  – Regression on Order Statistics

• Total = mean * n

• Correlation with other variables could provide better methods. Assuming none known here.
Methods 1-3. Substitution produces Invasive Data

- Substitution of one-half or \( \frac{1}{\sqrt{2}} \) times the RL for nondetects is common
- Produces an invasive pattern alien to the concentrations actually in samples
- Results in generally poor estimates and incorrect statistical tests
Why invasive data are produced

• The reporting limit can be a function of the concentration chosen for method calibration, of dilution or other lab prep, or of interferences from other analytes
• When these change, using a fraction of the RL adds a pattern that was not in the sample itself
• The pattern reflects lab conditions, protocols, and arbitrary choices. It obscures the original data.
Method 4. Maximum likelihood (MLE)

• You choose a specific distribution (lognormal most common for environmental data)
• MLE determines mean, std dev most likely to produce the observed data
• Information used:
  – detected values
  – proportion of values below each DL
MLE fits distribution to data

Values to left of blue line are nondetects
Methods 5. Kaplan-Meier
(nonparametric method)

• K-M estimates the survival function S, the probability of <= each detected value
• S estimates the empirical CDF (percentile function) of the original data
• K-M percentiles (y-axis) change at detected obs.
• No model, so area below the lowest DL not estimated. For 1 DL, K-M = DL sub.
Kaplan-Meier survival curve

S, the survival curve, is the set of percentiles (empirical cdf).

Here drawn right to left (backwards)

For the 75th percentile of S,
Prob (T > 2950) = 0.75

Prob (conc < 50) = 0.75
(the 75th percentile of Concentration)
K-M mean is area under survival curve
Method 6. Regression on Order Statistics (ROS)

Regression on a probability plot
How (robust) ROS works

1. Pick values off the regression line to estimate nondetect data. If in log units, retransform individual estimates by exponentiation.

2. Combine with detected observations to compute descriptive statistics.
Example – Robust ROS

(Invisible imputed estimates, never seen by user)
Used collectively. Should not be assigned to individual observations

Estimated Summary Stats:
mean 163.2  median 90.5
sd 393.1  IQR 69.6
The data

Lognormal distribution

Mean = 1.8
Median = 1
Bias in estimating the mean (mass or sum)

- 1000 reps 12 values 3 DLs at 0.5, 1 and 3
- Approx. 40% of 12 data values censored

<table>
<thead>
<tr>
<th>Method</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub zero</td>
<td>-0.246628</td>
</tr>
<tr>
<td>Sub half DL</td>
<td>0.0617677</td>
</tr>
<tr>
<td>Sub Dl</td>
<td>0.370164</td>
</tr>
<tr>
<td>ROS</td>
<td>-0.00294584</td>
</tr>
<tr>
<td>MLE</td>
<td>0.150299</td>
</tr>
<tr>
<td>K-M</td>
<td>0.05244434</td>
</tr>
</tbody>
</table>
Results

• 0 = no bias
• ROS, K-M have least bias
• Zero (-), DL(+) bias
• Large + errors for MLE not shown
• Sub 1/2DL about 20x bias as ROS
Estimating the mean (or sum) of 5 data values

ROS doesn’t work unless there are at least 3 or 4 detected values to perform regression

With small data sets, ROS (the best method with larger data sets) is not an option
Estimating the mean (or sum) of 5 data values

- 1000 reps  5 values  3 DLs at 0.5, 1 and 5
- Approx. 40% of 5 data values censored

<table>
<thead>
<tr>
<th>Method</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub zero</td>
<td>-0.138016</td>
</tr>
<tr>
<td>Sub half DL</td>
<td>0.108034</td>
</tr>
<tr>
<td>Sub Dl</td>
<td>0.354084</td>
</tr>
<tr>
<td>MLE</td>
<td>0.185547</td>
</tr>
<tr>
<td>K-M</td>
<td>0.057624</td>
</tr>
</tbody>
</table>
Results

Small data sets

- K-M has least bias
- Zero (-), DL(+) bias
- Large + errors for MLE not shown
- Sub 1/2DL about twice the bias as K-M
Estimating the mean (or sum) of 5 data values - one DL

- 1000 reps 5 values 1 DL = 1
- Approx. 50% of 5 data values censored

- Sub zero: -0.0847019
- Sub half DL: -0.0031019
- Sub DL: 0.0784981
- MLE: 0.102749
- K-M: 0.0784981

(For 1 DL, Kaplan-Meier sets all NDs = DL)
Results

- For small data sets and 1 DL, sub 1/2DL is last method standing
- With 1 DL, KM = DL sub
- Large + errors for MLE not shown
Summary: Which method best computes a sum?

<table>
<thead>
<tr>
<th></th>
<th>2 or more DLs</th>
<th>1 DL</th>
</tr>
</thead>
<tbody>
<tr>
<td>More than 3 detects</td>
<td>ROS, KM</td>
<td>ROS</td>
</tr>
<tr>
<td>3 or fewer detects</td>
<td>KM</td>
<td>Substitute half</td>
</tr>
</tbody>
</table>
For more info:

Nondetects and Data Analysis

Statistics for Censored Environmental Data

Helsel (2005)

www.PracticalStats.com / nada

And pick up the CD at our booth in the exhibit hall