Appendix 2.6A: Notes on Methods for Analyzing Below Detection (Censored) Data

Section 1

Introduction: CBP Data Censoring and Possible Solutions

Excerpt from original document:
Notes on Methods for Analyzing Below Detection (Censored) Data
Elgin Perry, 20Jan2016 originally: CB54_TestCensor_20Jan2016

In these notes I discuss some of the options available for analyzing data that were not fully reported by the laboratories that performed the chemical analysis. These are observations for which chemical analysis shows low concentrations and were recorded in the data base as below detection. In a statistical context, these data are called censored data.

The censoring mechanism that was employed for the early (1984-1998) part of the CBP data record is a commonplace practice of not reporting concentration results that are too small to be differentiated from zero with high probability. A threshold called the 'detection limit' was set and observations below this detection limit were not recorded in the data base. The detection limit was set by measuring the standard deviation of a number of laboratory splits of a sample with low concentration and setting the detection limit to a multiple of the standard deviation among these splits. For example, if analysis of splits resulted in a standard deviation of 0.01, then the detection limit might be set to 3 times this standard deviation which is 0.03. Three standard deviations is an upper bound such that if the true concentration were zero and the distribution of measurement error were normal, then the random error in the measurement process would produce an observation greater than this upper bound about 1 percent of the time. Therefore if the measurement process produces an observation greater than the detection limit, one can feel reasonably confident that the true concentration is greater than zero or the constituent being measured has been 'detected'. Observations below the detection limit lack this degree of certainty that the true concentration is greater than zero.

There are situations where it would be imprudent to report low measurements without some indication of the uncertainty that surrounds these data. It might cause undue concern if low values of cadmium were reported for a drinking water supply when due to the uncertainty of the measurement process these measurements cannot reliably be differentiated from zero. However, it should be emphasized that this type of censoring should only be implemented in reporting and not in recording data in the data base.

There are numerous reasons that the censoring of data in the data base is inappropriate. These include:

1. Detection limits apply to individual observations. Data interpretation is more often done

with means of observations.
2. Improving detection limits can result in spurious trends in censored data.
3. Analysis of censored data requires special statistical methods and consequently many powerful standard statistical tools are unavailable.
4. It is a common misconception that if data are recorded as below the detection limit, it means there is a high probability that the true concentration is below the detection limit.

In the late 1990's, the practice of censoring data in the data base was discontinued.

There are options for handling censored data which include:

1. In non-parametric methods based on ranks, set censored data to ties.
2. Set the censored data to a constant value at or below the detection limit.
3. Simulate random numbers from an appropriate distribution to replace censored data.
4. Use a maximum-likelihood method which has a composite likelihood for censored and un-censored data.

Historically analyses of trends for tidal data have been conducted using the non-parametric seasonal Kendall test for which the simple mechanism of setting censored data to ties was adequate. Even for this simple option, there are issues that must be considered. If the limit of detection is improving over time so that lower values are being recorded as observed data, this can result in a spurious decreasing trend in the data. To resolve this issue, data are re-censored to the worst case detection limit which results in some loss of information but does resolve the spurious trend issue. There is also the issue of how to propagate censoring into water quality concentrations that are calculated from several measurements. I do not believe this has ever been satisfactorily resolved.

If parametric analysis was conducted, censored data were typically set to ½ the detection limit. CBP studies (Alden et al. 2000) showed that this practice gave results that were similar to uncensored data results and maximum likelihood results for up to 20% censoring of data. These studies also found that least squares estimates with censored data set to ½ the detection limit was more robust than maximum likelihood if the data distribution (assumed for the statistical test) was mis-specified. Even though this simple data substitution coupled with least squares has some desirable properties, it is frequently criticized and the maximum likelihood method is favored.

As we move forward with the implementation of GAMs for trend analysis, we must develop some methodology for dealing with the legacy data that contain censored observations.

Section 2

Method: Expectation Maximization

Expectation for Left Censored Data

Follows logic in Liu, Lu, Kolpin and Meeker, 1997.

Finding the expectation for applying the EM algorithm

Let

Y follow the lognormal distribution

Then Y* = log(Y) follows normal distribution mean = μ and variance = σ^2, N(μ,σ^2)

Z = (log(Y) - μ) / σ follows standard normal distribution, N(0,1)

By algebra Y = exp(Z σ + μ)

Let c_y = the detection limit on the lognormal scale of Y and let c_z be the detection limit rescaled to the Standard Normal scale of Z such that $c_z = (\log(c_y) - \mu)/ \sigma$.

Now consider the expectation of Y given that Y is less than c_y.

$E[Y \mid Y < c_y] = E[\exp(Z\sigma + \mu) \mid \exp(Z\sigma + \mu) < c_y]$ by substitution of $\exp(Z\sigma + \mu)$ for Y

$= E[\exp(Z\sigma + \mu) \mid Z < (\log(c_y) - \mu) / \sigma]$ by applying algebra to the condition

$= E[\exp(Z\sigma + \mu) \mid Z < c_z]$ by substituting $c_z = (\log(c_y) - \mu)/ \sigma$

Now let $\Phi(x)$ be the Standard Normal cumulative distribution function such that $\Phi(c_z)$ is the probability that Z is less that c_z which is equal to the probability that Y is less than c_y. The function $\Phi(x)$ is written

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x)^2}{2}} dx$$

Now the conditional density for Z given that $Z < c_z$ is given by

$$f(z;\mu,\sigma \mid z < c_z) = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{(z)^2}{2}}}{\Phi(c_z)}$$

Appendix 2.6A

$$E[Y \mid Y < c_y] = \int_{-\infty}^{c_z} f(z;\mu,\sigma \mid z < c_z)dz = \int_{-\infty}^{c_z} e^{(\sigma z+\mu)} \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{(z)^2}{2}}}{\Phi(c_z)} \quad dz$$

Because $\Phi(c_z)$ and e^{μ} are constants, they can be brought out of the integral, and the terms left in the integral can be rearranged.

$$E[Y \mid Y < c_y] = \frac{e^{\mu}}{\Phi(c_z)} \int_{-\infty}^{c_z} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z^2-2\sigma z)}{2}} \quad dz$$

At this point we note that the integral is starting to look like $\Phi(\ )$ except that we need to add a factor of $e^{-\sigma^2}$. Because this is a constant, we can simply add this term inside the integral as long as we add a term outside that cancels it.

$$E[Y \mid Y < c_y] = \frac{e^{\mu} e^{\sigma^2/2}}{\Phi(c_z)} \int_{-\infty}^{c_z} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z^2-2\sigma z+\sigma^2)}{2}} \quad dz$$

$$E[Y \mid Y < c_y] = \frac{e^{\mu} e^{\sigma^2/2}}{\Phi(c_z)} \int_{-\infty}^{c_z} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\sigma)^2}{2}} \quad dz$$

Now let $z^* = z-\sigma$, then the limit of integration c_z is changed to $(c_z - \sigma)$ and $dz = dz^*$. By change of variable we have

$$E[Y \mid Y < c_y] = \frac{e^{\mu} e^{\sigma^2/2}}{\Phi(c_z)} \int_{-\infty}^{c_z-\sigma} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z^*)^2}{2}} \quad dz^* = \frac{e^{\mu+\sigma^2/2}\Phi(c_z-\sigma)}{\Phi(c_z)}$$

Using $c_z = (\log(c_y) - \mu)/\ \sigma$.

$$E[Y \mid Y < c_y] = \frac{e^{\mu+\sigma^2/2}\Phi(\frac{\log(c_y)-\mu}{\sigma}-\sigma)}{\Phi(\frac{\log(c_y)-\mu}{\sigma})}$$

Appendix 2.6A

Interval censored data

$$E[Y \mid c_{1y} < Y < c_{2y}] = \int_{c_{1y}}^{c_{2y}} f(y;\mu,\sigma \mid c_{1y} < c_{2y})dy = \int_{c_{1z}}^{c_{z2}} e^{(\sigma z+\mu)} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z)^2}{2}} \Big/ \left[\Phi(c_{z2}) - \Phi(c_{z1})\right] \quad dz$$

$$E[Y \mid c_{1y} < Y < c_{2y}] = \frac{e^{u} e^{\sigma^2/2}}{\left[\Phi(c_{z2}) - \Phi(c_{z1})\right]} \int_{c_{1z}}^{c_{z2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z^2 - 2\sigma z + \sigma^2)}{2}} \quad dz$$

$$E[Y \mid c_{1y} < Y < c_{2y}] = \frac{e^{u} e^{\sigma^2/2}}{\left[\Phi(c_{z2}) - \Phi(c_{z1})\right]} \int_{c_{1z}}^{c_{z2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\sigma)^2}{2}} \quad dz$$

$$E[Y \mid c_{1y} < Y < c_{2y}] = \frac{e^{u+\sigma^2/2}}{\left[\Phi(c_{z2}) - \Phi(c_{z1})\right]} \int_{c_{1z}-\sigma}^{c_{z2}-\sigma} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z^*)^2}{2}} \quad dz^*$$

$$E[Y \mid c_{1y} < Y < c_{2y}] = \frac{e^{u+\sigma^2/2}\left[\Phi(c_{z2} - \sigma) - \Phi(c_{z1} - \sigma)\right]}{\left[\Phi(c_{z2}) - \Phi(c_{z1})\right]}$$

Right Censored data

For right censored data the conditional density for Z given that Z $> c_z$ is given by

$$f(z;\mu,\sigma \mid z > c_z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(z)^2}{2}} \Big/ \left(1 - \Phi(c_z)\right)$$

$$E[Y \mid Y > c_y] = \int_{c_z}^{\infty} f(z;\mu,\sigma \mid z > c_z)dz = \int_{c_z}^{\infty} e^{(\sigma z+\mu)} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z)^2}{2}} \Big/ \left(1 - \Phi(c_z)\right) \quad dz$$

Because $\Phi(c_z)$ and e^{μ} are a constants, they can be brought out of the integral, and the terms left in the integral can be rearranged.

$$E[Y \mid Y > c_y] = \frac{e^{\mu}}{\left(1 - \Phi(c_z)\right)} \int_{c_z}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z^2 - 2\sigma z)}{2}} \quad dz$$

At this point we note that the integral is starting to look like $\Phi(\ )$ except that we need to add a factor of $e^{-\sigma^2}$. Because this is a constant, we can simply add this term inside the integral as long as we add a term outside that cancels it.

$$E[Y \mid Y > c_y] = \frac{e^{\mu} e^{\sigma^2/2}}{(1-\Phi(c_z))} \int_{c_z}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z^2-2\sigma z+\sigma^2)}{2}} \quad dz$$

$$E[Y \mid Y > c_y] = \frac{e^{\mu} e^{\sigma^2/2}}{(1-\Phi(c_z))} \int_{c_z}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\sigma)^2}{2}} \quad dz$$

Now let $z^* = z-\sigma$, then the limit of integration c_z is changed to $(c_z - \sigma)$ and $dz = dz^*$. By change of variable we have

$$E[Y \mid Y > c_y] = \frac{e^{\mu} e^{\sigma^2/2}}{(1-\Phi(c_z))} \int_{c_z-\sigma}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z^*)^2}{2}} \quad dz^* = \frac{e^{\mu+\sigma^2/2}(1-\Phi(c_z-\sigma))}{(1-\Phi(c_z))}$$

Using $c_z = (\log(c_y) - \mu)/\ \sigma$.

$$E[Y \mid Y > c_y] = \frac{e^{\mu+\sigma^2/2}\left(1-\Phi(\frac{\log(c_y)-\mu}{\sigma}-\sigma)\right)}{\left(1-\Phi(\frac{\log(c_y)-\mu}{\sigma})\right)}$$

References

Alden, R.W. III, E.S. Perry and M.F. Lane. 2000. A Comparison of Analytical Techniques for Determining Trends in Chesapeake Bay Water Quality Monitoring Program Data. AMRL Technical Report #3114. Applied Marine Research Laboratory, Norfolk, VA.

Liu, S; Lu, J; Kolpin, D; and Meeker, W. (1997) "Analysis of Environmental Data with Censored Observations" USGS Staff -- Published Research. Paper 71. http://digitalcommons.unl.edu/usgsstaffpub/71