Ann. Occup. Hyg., Vol. 54, No. 3, pp. 257–262, 2010

doi:10.1093/annhyg/mep092

Commentary

Much Ado About Next to Nothing: Incorporating Nondetects in Science

DENNIS HELSEL*

Practical Stats, 9278 Lark Sparrow Drive, Highlands Ranch, CO 80126, USA

Received 29 October 2009; in final form 19 November 2009; published online 23 December 2009

A great many papers and one textbook have been published on the topic of how to incorporate 'nondetects', low-level values reported only as below a detection limit, into statistical analyses. This is of interest not only in occupational hygiene but also in environmental sciences and astronomy, among other fields. Here, the literature is reviewed from the earliest known publication on the topic >40 years ago and recommendations contrasted. I have tried to pull some unifying conclusions out of the mix, ending with four suggestions I believe all can agree on. See if you agree with me.

Keywords: censored data; detection limit; nondetects; risk assessment; statistics

Much Ado About Next to Nothing: Incorporating Nondetects in Science

This title is not original. I first saw 'Much Ado About Next to Nothing' as the title of a conference presentation given by a US Environmental Protection Agency (USEPA) scientist around 1980 on handling data below detection limits. I cannot find an official reference to the talk or be more specific. The remoteness of the reference should remind us of how long the discussion has continued on which methods can incorporate low-level left-censored data into scientific studies. A more easily referenced document is the US Geological Survey report by Al Miesch (Miesch, 1967). He stated that substituting a constant for values (now called 'nondetects') below the detection limit created unnecessary errors, instead recommending Cohen's Maximum Likelihood procedure. Cohen's procedure was published in the statistical literature in the late 1950s and early 1960s (Cohen, 1957, 1961), so its movement into an applied field by 1967 is a credit indeed to Miesch. Miesch read the literature of other disciplines. His recommendation has consistently been ignored and substitution of one-half (or one over the square root of two) times the detection limit remains the most common method to date in the environmental sciences for performing all manner of statistical procedures on low-level data. That is both unfortunate and potentially dangerous.

In addition to the environmental sciences where I work, the issue of correctly handling nondetect data has been of great interest in astronomy (Feigelson and Nelson, 1985) and in occupational health (Succop *et al.*, 2004; Hewett and Ganser, 2007). This journal has published several articles dealing with it (Hewett and Ganser, 2007; Finkelstein, 2008; Krishnamoorthy *et al.* 2009; Flynn, 2010, among others). We all deal with information overload, barely having time to read the relevant literature of our own discipline. It is next to impossible to keep up with work in other disciplines, even when they encounter the same issues as we do. Handling nondetect data is one example. So let me summarize several decades of work in environmental studies and then relate it to a few recent papers in your discipline.

In the 1980s, I published three papers along with co-workers on how to treat nondetects for the field of environmental water chemistry (Gilliom and Helsel, 1986; Helsel and Gilliom, 1986; Helsel and Cohn, 1988). These were simulation studies,

*Author to whom correspondence should be addressed.
Tel: +1-303-8704921; e-mail: dhelsel@practicalstats.com

generating data from multiple distributions because water quality data rarely follow any distribution very closely. We then censored the data to varying degrees. We estimated the mean and other descriptive statistics using several estimation methods, including substituting zero, one-half the detection limit, or the limit itself, as well as maximum likelihood estimation (MLE) and a method employing regression on a probability plot. The point was to see how well these methods could reproduce the correct value for the statistics, particularly when an underlying distribution is unknown. We found that the probability plot method performed best if only one method was to be applied to all statistics estimated, distributions, and censoring levels. If we separated the estimation of percentiles from moment statistics (mean, standard deviation), maximum likelihood performed best for the estimation of percentiles, as long as the data distribution was not too far away from that assumed by the MLE. However, if the distribution was badly misspecified, MLE could produce an estimate that was very far off the mark. The probability plot method still performed best for estimating moment statistics, as the mean and variance are quite sensitive to errors at the upper end of the distribution. The probability plot method uses the recorded data at the upper end rather than a distributional model. Substitution methods performed poorly across the board, except for the instance of estimating a mean with one detection limit, where substituting one-half the detection limit was not too bad. Sanford *et al.* (1993) later determined that substituting one over the square root of two was better than using one-half the detection limit to estimate the mean of lognormal data with one detection limit.

In 1990, I stated that techniques of survival analysis, statistical methods for handling right-censored 'greater-thans' in medical and industrial applications, could be turned around and applied to censoring on the low end (Helsel, 1990). The Kaplan–Meier (KM) method, standard in medical sciences since the late 1950s, joined the stable of possible methods for dealing with nondetects. However, there is an incredibly strong pull for doing something that is simple and cheap, not to mention familiar. My 1990 survey clearly states that substitution is generally a bad idea. The article has since been referenced a myriad of times to justify using substitution! The fact that I mention it there seemed to give others license to support the inferior practice. As I said, there is an incredibly strong pull for doing something simple and cheap.

The problem with substitution is what I have come to call 'invasive data'. Substituted values possess a pattern that is alien to the pattern of the original data. The effect of the artificial substituted pattern often dominates that of the original values. Consider the data of Fig. 1, a straight-line relationship between two variables, Concentration (y) versus Distance (x). The slope of the relationship is significant, with a strong correlation between the variables. What happens when the data are reported using two detection limits of 1 and 3, and one-half the limit is substituted for the nondetects? The result (Fig. 2) includes horizontal lines of substituted values, changing the slope, and dramatically decreasing the correlation coefficient between the variables.

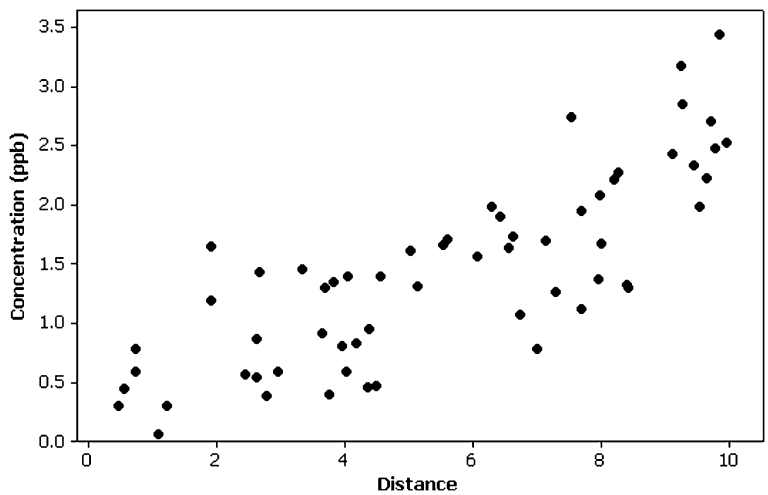

Fig. 1. Original data prior to censoring. True correlation equals 0.81.

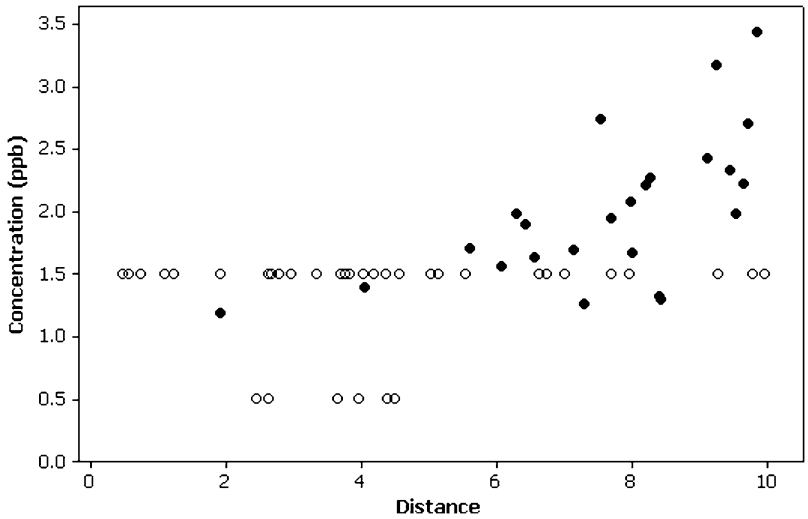

Fig. 2. Data from Fig. 1 after censoring at detection limits of 1 and 3 p.p.b. and substituting. Substituted values of half dl are shown as open circles. These invasive data form flat lines at one-half the detection limits, lowering the correlation to 0.55.

There are many published articles where substitution has been employed for an environmental contaminant and the correlation computed. A low correlation coefficient is cited as evidence that a proposed causative agent (plotted on the horizontal axis) is not the likely cause of contamination. All such conclusions should be considered suspect, due to the use of substitution. There are better ways.

A number of simulation studies over the years have evaluated methods for estimating descriptive statistics of censored data. Their findings do not always seem to agree. Shumway *et al.* (2002) used MLE and probability plot methods on data from lognormal and gamma distributions; both are skewed distributions commonly used to model water quality data. Their acronym for the probability plot methods, Regression on Order Statistics (ROS), has stuck in the environmental sciences. The same method has been called LPR—log-probit regression (Hewett and Ganser, 2007), when a lognormal distribution is assumed. Shumway *et al.* (2002) found that for estimating the mean, both performed similarly, with the tradeoff that bias is a problem for MLE while confidence intervals around the mean are wider for ROS with highly-skewed data. They did not consider the KM estimator and used the best fit of three candidate distributions (normal, square root, or lognormal) in ROS estimation. Singh *et al.* (2006) conducted perhaps the most comprehensive evaluation to date of methods to compute the 95% upper confidence limit on the mean (UCL95) of data with multiple detection limits. They found that the KM estimate of mean and standard deviation, followed by either a Chebyshev, *t*-interval, or bootstrap estimate of the UCL95 provided better coverage than MLE or ROS methods, and far better than substitution methods. Quality of the UCL bound determined by the two nonparametric (Chebyshev and bootstrap) and one parametric method (*t*-interval, requiring that the Central Limit Theorem be invoked) depended on sample size and other considerations. They state that 'contrary to the general rule of thumb, it should be noted that the DL/2 does not perform well even for low censoring levels ... such as 10%, 20%, and 30%.' Antweiler and Taylor (2008) evaluated KM, ROS, MLE, and substitution methods for estimating the mean and other statistics. Instead of generating 'true' concentrations from one or more statistical distributions, they used a precise research-grade laboratory technique resulting in no nondetects to determine the true value and used a less-precise typically used technique on the same samples to provide the censored data. After applying KM, MLE, and the other methods to the censored data to estimate descriptive statistics, the results were compared to the statistic of the true technique. They found that KM was the overall best method and MLE to be 'far inferior to all other treatments except substituting zero or the detection limit value'. They also found that using the machine readings from the less-precise technique did not work well, arguing against the common user request to 'just give me the numbers'

rather than censoring low-level values. Laboratory data are censored when machine readings fall into the range where any (small) signal is obscured by the noise. Some readings may be negative. Chemists consider these individual numbers to have low reliability, and hence the censoring. MLE and KM methods represent nondetects by the proportion of values falling below each detection limit, without attributing any individual value to them. Attributing noisy individual values to nondetects apparently results in less accuracy than using the proportional information.

Here in the Annals, Hewett and Ganser (2007) evaluated methods for estimating the mean and the 95th percentile of left-censored data, both with one and multiple detection limits. After evaluating MLE, ROS, KM, and substitution methods, they found that MLE consistently outperformed the others when root mean-square error was the index of performance. Similar to Shumway *et al.* (2002), they found that a robust method performed better when bias was used as the index of performance. They also found that KM did not work well for data with one detection limit (dl). This is a known characteristic of KM—as a distribution-free procedure, it will not estimate down below the lowest dl. It only sees what the data tells it—free of models for extrapolation beyond the data range. All values recorded as below the dl will be assigned either the dl itself (the Efron bias correction) or the lowest detected value (the standard practice). A positive bias results. Hewett and Ganser found that KM performed much better for data with multiple dls. As scientists become more familiar with these methods, we can choose their uses appropriately. Their remaining findings were fairly similar to those of Gilliom and Helsel (1986) and Helsel and Cohn (1988). MLE was the best procedure when data were close to the assumed distribution, in this case the lognormal, and lognormal distributions with some mild (at least for environmental studies) contamination. MLE methods should work best in these situations. ROS methods would be expected to be the second best in situations where data follow a known distribution. In this issue of the Annals, Flynn (2010) solves for descriptive statistics by maximizing the Shapiro–Wilk statistic. This approach is something like the Shumway *et al.* (2002) procedure in that the Shapiro–Wilk statistic is essentially the *r*-squared of data plotted on a probability plot. Determination of the best Box–Cox transformation follows from the best fit of the statistic. Following selection of the appropriate distribution (Flynn considered only normal and lognormal, but other transformed-normal distributions should be possible), statistics are computed by varying estimated values for nondetects, constrained to be between zero and their detection limit, until the maximum Shapiro–Wilk statistic is obtained. He notes that this is much like the 'robust' estimation process of ROS, avoiding transformation bias. Enabling this within Excel encourages use of the method by people who would otherwise substitute numbers, those who avoid using a commercial statistics package in favor of the more familiar Excel spreadsheet.

I have always been a lumper rather than a splitter, attempting to find broad statements that generally hold across many studies. Evaluations of the quality of methods for estimating descriptive statistics appear to differ based on at least three characteristics of the simulation studies: sample size, the number of detection limits, and the magnitude of departure from the assumed distribution. Larger departures from the assumed distribution will favor nonparametric methods such as KM and robust methods like the robust ROS. In environmental studies, this is key, as field data rarely follow any known distribution. In more controlled studies, data more likely follow a known distribution where MLE methods and the distributional ROS work very well. In simulations where data are generated using a known distribution or with small departures, as with Hewett and Ganser (2007), MLE methods win out. In other studies using more diverse data, either using multiple distributions (Singh *et al.*, 2006) or censoring observed field data (Antweiler and Taylor, 2008; Helsel and Gilliom, 1986), KM or ROS perform better. In environmental studies, we rarely encounter data today with only one dl. KM often works well with multiple dls. It is biased when there is only one dl, and it is best not to use it in that situation. Finally, all methods have lower errors with more data, obviously, but the amount of data interacts with the other two factors. If data follow a known distribution, MLE may work well for small data sets because it is using correct distributional information that KM and robust ROS do not. If data depart from the assumed distribution, however, the penalty for using a misspecified MLE can be large when there is little data to go on. In short, trying to reconcile studies that state 'the standard MLE method consistently outperformed the [other methods]' (Hewett and Ganser, 2007) with those finding 'the best technique overall for determination of summary statistics was the nonparametric Kaplan–Meier technique . . . . Maximum likelihood techniques were found to be far inferior to all other treatments except substituting zero or the detection limit value' (Antweiler and Taylor, 2008) is only

possible when considering the differing data types and conditions under which the methods were tested.

Finally, estimation of descriptive statistics is just one of the tasks where nondetects must be incorporated. There is also a need for methods to incorporate these data into hypothesis tests, correlation and regression, and multivariate procedures. Finkelstein (2008) presents a strong case for performing hypothesis tests with methods drawn from survival analysis to compare control versus test groups, rather than only the heuristic comparison of group medians for censored data. He used MLE procedures—nonparametric methods for censored data are also available. The message of his paper is entirely consistent with 'Nondetects And Data Analysis' (Helsel, 2005), ignoring methods that incorporate-censored data lead to wrong decisions both economically and for human or ecosystem health. In my 2005 book, I used the flawed decision to launch the Challenger shuttle as the example. Finkelstein's example of missing the effects of asbestos in the lungs of brake mechanics is equally compelling.

Software is often a hurdle to use these techniques. MLE procedures in commercial statistics software are sometimes coded with the ability to use left-censored values. Nonparametric methods are not. Implementations within simpler software are becoming available—the KM method has been embedded into an Excel worksheet, available at http://www.practicalstats.com/nada. Excel worksheets to compute ROS methods are on the Internet. Flynn (2010) provides an optimization approach for computing descriptive stats using Excel's Solver routine. Better methods than substitution will hopefully be used more frequently as the software to perform them becomes more easily available.

While differences in objectives and data characteristics might lead to using different methods, there are at least four things that I think we should be able to agree on:

1. In general, do not use substitution. Journals should consider it a flawed method compared to the others that are available and reject papers that use it. The lone exception might be when estimating the mean for data with one censoring threshold, but not for any other situations or procedures. Substitution is NOT imputation, which implies using a model such as the relationship with a correlated variable to impute (estimate) values. Substitution is fabrication.
2. Method evaluations for estimating a mean do not necessarily carry over to the more difficult issues of how to compute interval estimates, upper percentiles, a correlation coefficient, a regression slope and intercept, or a multidimensional surface when left censoring is present. There are many interesting issues still to be evaluated.
3. We should all become more familiar with the literature on censored data from the survival/reliability analysis discipline. There is no need to reinvent the wheel for tires on the left side if the wheel already exists on the right. There should be more widespread training in survival/reliability methods within university programs in both our disciplines.
4. Commercial software should more easily incorporate left-censored data into its survival/reliability routines. For example, plots and hypothesis tests of whether censored data fit a normal and other distributions, as requested by Hewett and Ganser (2007), exist in these packages. They are usually coded to handle only right-censored data. Users in both environmental sciences and occupational hygiene should loudly request that this be changed.

REFERENCES

Antweiler RC, Taylor HC. (2008) Evaluation of statistical treatments of left-censored environmental data using coincident uncensored data sets: I. Summary statistics. Env Sci Technol; 42: 3732–8.

Cohen AC. (1957) On the solution of estimating equations for truncated and censored samples from normal populations. Biometrika; 44: 225–36.

Cohen AC. (1961) Tables for maximum likelihood estimates: singly truncated and singly censored samples. Technometrics; 3: 535–41.

Feigelson E, Nelson P. (1985) Statistical methods for astronomical data with upper limits. I—Univariate distributions. Astrophy J; 293: 192–206.

Finkelstein MM. (2008) Asbestos fibre concentrations in the lungs of brake workers: another look. Ann Occup Hyg; 52: 455–61.

Flynn M. (2010) Analysis of censored exposure data by constrained maximization of the Shapiro-Wilk W statistic. Ann Occup Hyg; 54: 263–27.

Gilliom RJ, Helsel DR. (1986) Estimation of distributional parameters for censored trace level water quality data 1. Estimation techniques. Water Resour Res; 22: 135–46.

Helsel DR. (1990) Less than obvious: statistical treatment of data below the detection limit. Env Sci Technol; 24: 1766–74.

Helsel DR. (2005) Nondetects and data analysis. New York, NY: John Wiley.

Helsel DR, Cohn TA. (1988) Estimation of descriptive statistics for multiply censored water quality data. Water Resour Res; 24: 1997–2004.

Helsel DR, Gilliom RJ. (1986) Estimation of distributional parameters for censored trace level water quality data 2. Verification and applications. Water Resour Res; 22: 147–55.

Hewett P, Ganser GH. (2007) A comparison of several methods for analyzing censored data. Ann Occup Hyg; 51: 611–32.
Krishnamoorthy K, Mallick A, Matthew T. (2009) Model-based imputation approach for data analysis in the presence of non-detects. Ann Occup Hyg; 53: 249–63.
Miesch A. (1967) Methods of computation for estimating geochemical abundance. Washington, DC: US Geological Survey Professional Paper 574-B.
Sanford RT, Pierson CT, Crovelli RA. (1993) An objective replacement method for censored geochemical data. Math Geol; 25: 59–80.
Shumway RH, Azari RS, Kayhanian M. (2002) Statistical approaches to estimating mean water quality concentrations with detection limits. Env Sci Technol; 36: 3345–53.
Singh A, Maichle R, Lee SE. (2006) On the computation of a 95% upper confidence limit of the unknown population mean based upon data sets with below detection limit observations. Washington, DC: U.S. Environmental Protection Agency EPA/600/R-06/022.
Succop PA, Clark S, Chen M. (2004) Imputation of data values that are less than a detection limit. J Occup Environ Health; 1: 436–41.