USING THE REVERSE KAPLAN-MEIER TO ESTIMATE POPULATION DISTRIBUTIONS WITH DATA BELOW A LIMIT OF DETECTION

Gillespie B W[1,2], Reichert H[2], Chen Q[1], Franzblau A[3], Lepkowski J[4], Adriaens P[5], Demond A[5], Luksemburg W[6] and Garabrant D[3]

[1]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109-2029; [2]Center for Statistical Consultation and Research, University of Michigan, Ann Arbor, MI 48109-1070; [3]Department of Environmental Health Sciences, University of Michigan, Ann Arbor, MI 48109-2029; [4]Institute for Social Research, University of Michigan, Ann Arbor, Michigan 48109-1248 USA; [5]Department of Civil and Environmental Engineering, University of Michigan College of Engineering, Ann Arbor, Michigan 48109-2135 USA; [6]Vista Analytical Laboratory, El Dorado Hills, California 95762 USA

Abstract: *Introduction*: Data below a limit of detection (LOD) can be analyzed using methods of survival analysis for left-censored data. The reverse Kaplan-Meier (KM) estimator provides an effective method of estimating population percentiles for such data. *Materials and methods*: The reverse KM estimator is illustrated using serum dioxin data from both the University of Michigan Dioxin Exposure Study (UMDES) and the National Health and Nutrition Examination Survey (NHANES). The reverse KM can be calculated using commonly available software for the KM estimator. SAS, JMP, R and Minitab will calculate the reverse KM directly using their Turnbull estimator routines. (The Turnbull estimator, developed for the more general interval censoring case, is equivalent to the reverse KM for left-censored data.) *Results and Discussion*: In this setting, the reverse KM estimator is the recommended method in the statistical literature. For estimation of the distribution function and population percentiles, the reverse KM estimator is recommended in preference to commonly used methods such as substituting LOD/2 or LOD/√2 for values below the LOD, assuming a known parametric distribution, or using single or multiple imputation to replace the left-censored values.

Introduction

In toxicology and environmental science, estimates of population percentiles for environmental contaminants are often needed where the data have some values below an analytical limit of detection (LOD). The reverse Kaplan-Meier (KM) estimator,[1,2,3] also known as the (more general) Turnbull estimator, was first introduced in the 1970s, and provides an effective method of estimating population percentiles for such left-censored data. The reverse KM estimator is recommended in preference to commonly used methods such as substituting LOD/2 or LOD/√2 for values below the LOD, assuming a known parametric distribution, or using single or multiple imputation to replace the left-censored values.

Our goal is to facilitate broader use of the reverse KM estimator by describing its desirable properties, illustrating its use with real data, and showing how it can be calculated using standard software. Examples are based on serum dioxin concentrations from two sources: (1) the University of Michigan Dioxin Exposure Study (UMDES),[4] using subjects from the control region of Jackson and Calhoun counties in Michigan (i.e., subjects assumed to have background levels of exposure, measured in 2004-2005), and (2) the National Health and Nutrition Examination Survey (NHANES), a population-based sample of the non-institutionalized U.S. population, measured in 2003-2004.[5]

Materials and Methods

The Limit of Detection (LOD): Values below the LOD, sometimes called “non-detects,” are common in environmental measurements.[6,7,8,9] The LOD can depend on the precision of the assay and the volume of the sample. The LOD may be a batch-specific, or may vary widely from sample to sample.

Human serum concentrations of two furan congeners with substantial proportions below the LOD were chosen to illustrate the reverse KM method: 2,3,7,8 TCDF and 2,3,4,6,7,8 HxCDF, measured in UMDES (n=251) and NHANES participants (n=1792 for TCDF, n=1789 for HxCDF). In the UMDES, 167 (67%) of TCDF and 187 (75%) of HxCDF samples were below LOD. In NHANES, 1739 (97%) of TCDF and 1701 (95%) of HxCDF samples were below LOD. Estimation was only possible in these cases of >90% below LOD because of the large NHANES sample size, with at least 50 observed values for each congener.

As an exploratory step, it is useful to show the distributions of censored and non-censored values separately. Figure 1 shows the boxplot distributions of both the LODs (for left-censored values) and the uncensored values. In both the UMDES and NHANES data, the distributions show wide variability in LOD values, and the distributions of the LOD values substantially overlap the distributions of the respective observed values.

The Reverse Kaplan-Meier Estimator: The reverse KM estimator for left-censored data is described below, where F(x) = P(X≤x) is the right-continuous cumulative distribution function. The estimate of F(x) is a step function with a "jump" at each uncensored value, and constant between uncensored values. With no censoring, the "jump" has size 1/n at each point. If left-censored values are intermixed with observed values, the "jump" size increases from right to left after each censored value. If the smallest value is observed (not censored), then the estimate of F(x) is zero to the left of that point, and F(x) is fully defined. However, if the smallest value is censored, then the estimate of F(x) is undefined to the left of that value. Standard errors of reverse KM estimates are readily available using the Greenwood formula[2,10] that was developed for the KM estimator.[11]

The reverse KM estimator can be intuitively explained as a redistribution-to-the-left algorithm. For each left-censored observation, the associated probability (1/n, where n is the total number of all observations, censored or not) is spread equally to all observations to its left. The idea is that the distribution of the left-censored value is best estimated by the distribution of data values less than that value. In contrast, the conventions of replacing left-censored values with LOD/2 or LOD/√2 deposit all probability from a left-censored observation at a single point.

The Turnbull is the nonparametric maximum likelihood estimator of the population distribution function. It is consistent, asymptotically normally distributed, asymptotically efficient,[13] and satisfies the self-consistency principle.[12] In the statistical literature, the Turnbull estimator is commonly accepted as the best nonparametric estimator of F(x).[13,14]

The reverse KM estimates for the four example congeners are presented in Figure 3. For all four, the min(LOD) is smaller than x_1, so the reverse KM estimate is "hanging" (undefined) at the left end of the curve. However, the curves must connect to the origin, so one could imagine an interpolating line between the origin and the first estimated point of each curve. These examples were chosen to illustrate the surprising amount of information available in spite of some extremely high percents below LOD (e.g., 97% for TCDF - NHANES). The UMDES concentrations tend to be higher than the NHANES concentrations, as shown by the UMDES curves (darker lines) rising more slowly than the NHANES curves. In both UMDES and NHANES, the concentrations for HxCDF tend to be higher than those for TCDF.

Estimation of population percentiles: In Figure 2, although the median cannot be precisely estimated for any of the congeners, the estimates are bounded above by min(LOD), which is 0.1 ppt for TCDF (UMDES), 0.4 ppt for TCDF (NHANES), 0.2 ppt for HxCDF (UMDES), and 0.6 ppt for HxCDF (NHANES). Thus, we can report, for example, that the median serum concentration of TCDF in UMDES was <0.1 ppt. The median and 75th percentile estimates for both congeners for UMDES and NHANES are given Table 1. The respective medians estimated using LOD/2 and LOD/√2 were uniformly higher than those estimated using the reverse KM (Table 1). These results highlight the possibility of substantial differences between the estimates from the different methods.

Estimation of population arithmetic and geometric means: The arithmetic mean is estimated as the sum of each x value multiplied by its associated probability, or equivalently, as the area between F(x) and 1.0, for x>0. If the estimate of F(x) is "hanging" at the left end, then this area is bounded by the lower and upper limits for "completing" F(x) in the undefined region. The upper bound for the mean assumes F(x)=0 for x<min(LOD) and the lower bound assumes F(x)=F(min(LOD)) for x<min(LOD). The uncertainty due to the undefined region is the difference between the lower and upper bounds for the mean, i.e., min(LOD) * F(min(LOD)). This value

will be small if min(LOD) and/or F(min(LOD)) is small. One could add half the area between the lower and upper bound of F(x) to obtain a single estimate for the mean, i.e., ½ min(LOD) * F(min(LOD).

Results and Discussion

Figure 3 illustrates the situation of small min(LOD) but large F(min(LOD)) for TCDF and HxCDF. For TCDF (UMDES), min(LOD)=0.10, F(min(LOD))=0.57, and their product equals 0.057, which is the difference between upper and lower bounds. The lower bound is 0.41, and the upper bound is then 0.41+0.06 = 0.47. In this case, the impact of the left-censoring on estimation of the mean will be small, even though F(min(LOD)) is large. The estimate of the mean using LOD/2 is 0.54, and using LOD/√2 it is 0.61, both of which are larger than the upper bound of the estimate based on the reverse KM estimator. Table 1 shows these values for both congeners for UMDES and NHANES. For both NHANES congeners, where the percent below LOD is ≥95%, the range between the lower and upper bound is substantial (e.g., 0.11 to 0.50 for TCDF-NHANES). However, in all cases, the estimates of the mean using either LOD/2 or LOD/√2 are larger than upper bound using the reverse KM.

The geometric mean (GM) is sometimes used to provide an estimate of central tendency that is not as strongly affected by outliers as the arithmetic mean. For a lognormal distribution, the GM is exp(μ), where μ is the mean of the corresponding normal distribution. For other distributions, the interpretation of the GM is less intuitive. Compared to the GM, the median has a simple interpretation for all distributions, is also stable in the presence of outliers, and is easily estimated using the reverse KM method. Consequently, it may be a better choice than the GM in the situation of skewed distributions.

Software Issues: Most statistical software packages (SPSS, SAS, JMP, Stata, Minitab, Splus/R) include procedures to calculate the KM estimator. Alternatively, a few statistical packages (JMP, Minitab, SAS, and Splus/R,[15] but not Stata or SPSS) have procedures to calculate the Turnbull estimator directly. A new feature that allows the R software to be incorporated and used inside of Excel will make the Turnbull estimator more accessible for non-statisticians, particularly as R is available as a free download.[16] Q-Q plots for checking the fit of a parametric distribution using the reverse KM estimator are available in both SAS and Minitab.

Discussion: This paper illustrates the reverse KM method for estimating the population distribution function in the presence of left-censored (below LOD) data. In the statistical literature, the reverse KM estimator is considered to be the best nonparametric estimator of the distribution function in the setting of left-censored data,[13,14] and at least one author in the environmental literature has recommended this method.[17] However, the reverse KM remains in limited use for below-LOD data in practice.

Given its strong statistical qualities and the availability of software, we recommend that the reverse KM estimator be used routinely in place of other methods for estimating population distributions, means and percentiles.

Financial disclosure: Financial support came from The Dow Chemical Company through an unrestricted grant to the University of Michigan.

Acknowledgements:
The authors acknowledge Dr. Donald Patterson Jr, Dr. John Gillespie, and Ms. Sharyn Vantine for their assistance, and Drs. Linda Birnbaum, Ronald Hites, Paolo Boffetta and Marie Haring Sweeney for their guidance as members of our Scientific Advisory Board.

References:

1. Peto, R. *Applied Statistics*, 1973; 22; 86-91.
2. Turnbull BW. *J Amer Statist Assoc* 1974; 69: 169-173.
3. Turnbull B. *J R Stat Soc (B)* 1976;38:290–5.
4. Hedgeman E, Chen Q, Hong B, Chang C-W, Olson K, LaDronka K, Ward B, Adriaens P, Demond A, Gillespie BW, Lepkowski J, Franzblau A, Garabrant DH. *Environ Health Persp*, (in press).
5. Patterson DG, Turner WE, Caudill SP, Needham LL. *Chemosphere* 2008; 73(1,Supp 1):S261-S277.
6. Hornung RW and Reed LD. Appl Occup Environ Hyg 1990; 5:46-51.
7. Pavuk M, Patterson DG, Jr, Turner WE, Needham LL, Ketchum NS. *Chemosphere* 2007; 68(1):62-68.
8. Niskar AS, Needham LL, Rubin C, Turner WE, Martin CA, Patterson DG, Jr, Hasty L, Wong L-Y, Marcus M. *Chemosphere* 2009; 74(7):944-949.
9. Caudill SP, Wong L-Y, Turner WE, Lee R, Henderson A, and Patterson DG, Jr. *Chemosphere* 2007; 68; 169-180.
10. Greenwood M. *Reports on Public health and Medical Subjects*, London: Her Majesty's Stationery Office 1926; 33:1-26.
11. Kaplan EL and Meier P. *J Amer Stat Assoc* 1958; 53: 457-481.
12. Efron B. Proc. *5th Berkeley Symp. IV.* 1966; 831-853.
13. Gu MG and Zhang C-H. *Annals of Statistics* 1993; 21(2);611-624.
14. Lindsey JC and Ryan LM. *Statistics in Medicine* 1998; 17: 219-238.
15. Giolo SR. Technical Report, August 2004, Department of Statistics, Federal University of Parana, Brazil. [www.est.ufpr.br/rt/suely04a.pdf]
16. Heiberger, RM and Neuwirth E. *R through Excel*. Springer-Verlag, New York (in press for 2009).
17. Antweiler RC and Taylor HE. *Environ Sci Technol* 2008; 42(10):3732-8.

Table 1. Comparison of median, 75th percentile, and mean estimates of serum concentrations in parts per trillion (ppt) of TCDF and HxCDF using the reverse Kaplan-Meier (KM) estimator compared with replacing values below the limit of detection (LOD) with either LOD/2 or LOD/√2. Data are from the University of Michigan Dioxin Exposure Study (UMDES) and the National Health and Nutrition Examination Survey (NHANES). Survey weights were not used.

Estimate	**Method**	**TCDF**		**HxCDF**	
		UMDES	**NHANES**	**UMDES**	**NHANES**
Median	Reverse-KM	<0.1	<0.4	<0.2	<0.6
	LOD/2	0.3	0.8	0.6	1.0
	LOD/√2	0.4	1.1	0.8	1.3
75th Percentile	Reverse-KM	0.6	<0.4	1.0	<0.6
	LOD/2	0.7	1.0	1.1	1.3
	LOD/√2	0.7	1.4	1.3	1.8
Mean					
Min(LOD)*		0.10	0.40	0.21	0.60
F(min(LOD))*		0.57	0.96	0.70	0.93
Product*		0.06	0.39	0.14	0.56
Lower bound	Reverse-KM	0.41	0.11	0.60	0.19
Upper bound	Reverse-KM	0.47	0.50	0.74	0.75
	LOD/2	0.54	0.91	0.93	1.12
	LOD/√2	0.61	1.24	1.09	1.52

*Min(LOD)=the smallest LOD, which in these examples is smaller than all uncensored values. F(min(LOD)) is the reverse-KM estimate of F(x) at min(LOD). The product, min(LOD) * F(min(LOD)), is the area of the rectangle enclosing the "undefined region" of the estimate of F(x), and is the difference between the lower and upper bounds of the mean.

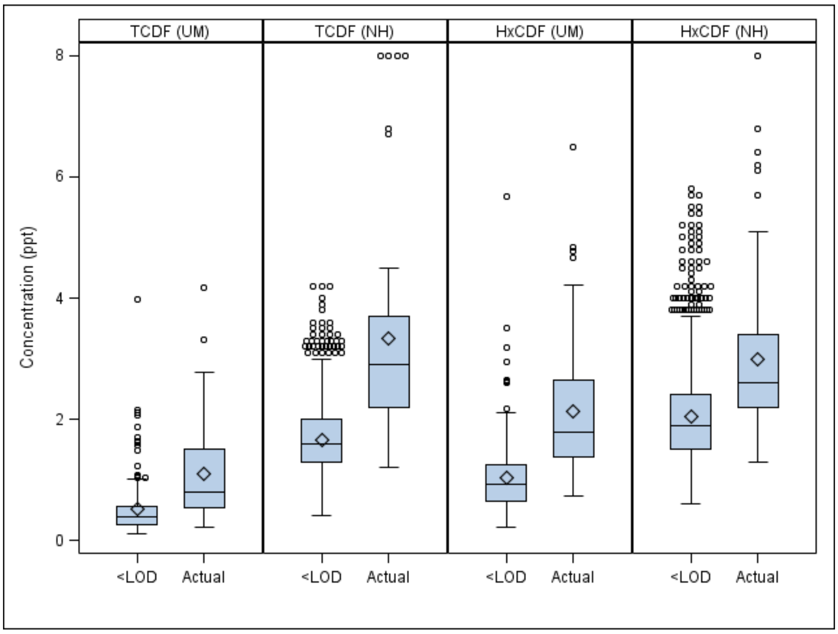

Figure 1. Boxplots showing the distributions of the LOD (for values <LOD) and the actual value (for values above the LOD) for serum concentrations of two congeners (2,3,7,8 TCDF and 2,3,4,6,7,8 HxCDF) from both UMDES and NHANES. Boxes extend from the 25th to 75th percentiles; whiskers extend to the largest value within 1.5*inter-quartile range of each end of the box. The line across each box represents the median, and the diamond represents the mean.

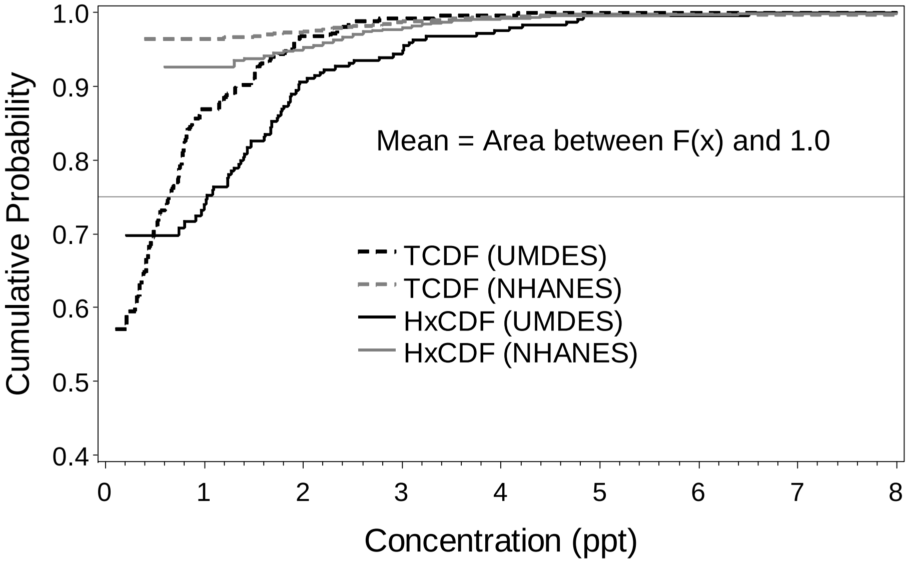

Figure 2. Reverse KM estimates of the distribution functions for serum concentrations of two furan congeners: 2,3,7,8 TCDF and 2,3,4,6,7,8 HxCDF in UMDES and NHANES. The 75th percentiles are shown by the x-values where the reference line at F(x)=0.75 intersects each curve (0.64 and 1.02 for TCDF and HxCDF, respectively, in UMDES). Units are parts per trillion (ppt).