Paper SDA-09

A Macro for Calculating Summary Statistics on Left Censored Environmental Data using the Kaplan-Meier Method

Dennis Beal, Ph.D.
Science Applications International Corporation, Oak Ridge, Tennessee

ABSTRACT

Calculating summary statistics such as the mean, standard deviation, and an upper confidence limit on the mean is straightforward when the data values are known. However, environmental data often are reported from the analytical laboratory as left censored, meaning the actual concentration for a given contaminant was not detected above the method detection limit. Therefore, the true concentration is known only to be between 0 and the reporting limit. The nonparametric Kaplan-Meier product limit estimator has been widely used in survival analysis on right censored data, but only recently has this method been applied to left censored data. Kaplan-Meier can be used on censored data with multiple reporting limits with minimal assumptions. This paper presents a SAS® macro that calculates the mean, standard deviation, and standard error of the mean of a left censored environmental data set using the nonparametric Kaplan-Meier method. Kaplan-Meier has been shown to provide more robust estimates of the mean and standard deviation of left censored data than other methods such as simple substitution and maximum likelihood estimates. This paper is for intermediate SAS users of SAS/BASE.

Key words: Kaplan-Meier, nonparametric, left censoring, environmental data

INTRODUCTION

Statisticians, environmental scientists, and risk assessors collect environmental data for many purposes, including characterization, quantification of risks and hazards, waste disposal, and compliance to applicable environmental regulations. Environmental analytical data that are left censored at the reporting limit are called “non-detects” and often qualified with a laboratory or validation qualifier of “<” or “U”. Unlike detected data that are reported as measured concentrations and are uncensored, for chemicals the estimated concentration for non-detects is known only to be within the interval from 0 to the reporting limit provided by the laboratory. For example, the laboratory may report a detected (and hence uncensored) lead concentration of 12 mg/kg in soil from sample 1 and <5 (or 5 U) mg/kg from sample 2. The censored concentration X in sample 2 is known only to be in the interval $0 \le X < 5$ mg/kg. However, for radionuclides that are reported as background corrected, negative and zero concentrations can be reported with “U” qualifiers, so the lower bound of the censoring interval may extend below 0. This paper will focus on applying the nonparametric Kaplan-Meier method to chemicals.

The problem is how are summary or descriptive statistics calculated with data that are a mix of censored and uncensored data? Many publications have been written on the subject of environmental data analysis. Gilbert (1987) describes some simple substitution methods and maximum likelihood estimators such as Cohen to estimate a mean and standard deviation of censored environmental data. Helsel (1990) compares several methods for calculating summary statistics using censored data. Helsel and Cohn (1988) estimate summary statistics on water quality data. Helsel and Gilliom (1986) estimate distributional parameters for water quality data.

Helsel (2005) compares the usual methods for handling censored data such as simple substitution, maximum likelihood estimators (MLE), regression on order statistics (ROS), and nonparametric methods. For less than 50% non-detects, Helsel recommends the Kaplan-Meier method. For 50 to 80% non-detects, Helsel recommends the robust MLE or ROS for the number of samples $n \le 50$ and the MLE or $n > 50$. For more than 80% non-detects, Helsel recommends high sample percentiles such as the 90^{th} or 95^{th} percentiles from the highly censored data set. Other methods such as Cohen’s assume an underlying normal distribution and only a single reporting limit. Since the Kaplan-Meier is nonparametric, it is more robust with fewer assumptions than Cohen’s, simple substitution, or MLE when at least half the samples are detected. The SAS code presented in this paper uses the SAS System for personal computers version 9.2 running on a Windows XP Professional platform.

KAPLAN-MEIER METHOD

The Kaplan-Meier (KM) method was first introduced in the literature by Kaplan and Meier (1958) as a nonparametric product limit estimator based upon a statistical distribution function estimate that adjusts for right censoring. KM has historically been applied in survival analysis and medical applications. The U.S. Environmental Protection Agency

software ProUCL 4.0 (U.S. EPA 2007) uses the KM as one method for calculating upper confidence limits (UCLs) on the mean. The equations used by KM as described in the ProUCL Technical Guide are as follows.

Let x_1, x_2, …, x_n represent the n concentrations (either detected concentrations or non-detects) obtained from environmental samples. The n concentrations are assumed to be statistically independent and representative samples from the environmental population being measured. Let y_1, y_2, …, y_p denote the p *distinct* values at which detects are observed so that $p \le n$. For $j = 1, 2, \ldots, p$, let m_j denote the number of detects at y_j and let n_j denote the cumulative number of $x_i \le y_j$. Define F(x) in Eqn. 1.

$$\begin{aligned} F(x) &= 1 && x \ge y_p \\ F(x) &= \prod_{j \ni y_j > x}^{p} \frac{n_j - m_j}{m_j} && y_1 \le x \le y_{p-1} \\ F(x) &= F(y_1) && x_1 \le x \le y_1 \\ F(x) &= 0 && 0 \le x \le x_1 \end{aligned} \tag{1}$$

An estimate of the population mean μ using the KM method is shown in Eqn. 2,

$$\hat{\mu} = \sum_{j=1}^{p} y_j \left[F(y_j) - F(y_{j-1}) \right] \tag{2}$$

where $F(y_0) = 0$.

An estimate of the standard error (SE) of the mean is shown in Eqn. 3,

$$\hat{\sigma}_{SE}^2 = \frac{n-k}{n-k-1} \sum_{j=1}^{p-1} a_j^2 \frac{m_{j+1}}{n_{j+1}(n_{j+1} - m_{i+1})} \tag{3}$$

where k = number of non-detects. The a_j are shown in Eqn. 4,

$$a_j = \sum_{i=1}^{j} (y_{i+1} - y_i) F(y_i) \tag{4}$$

for $j = 1, 2, \ldots, p - 1$.

An estimate of the variance σ^2 of the censored data set is shown in Eqn. 5.

$$\hat{\sigma}^2 = \sum_{j=1}^{p} (y_j - \hat{\mu})^2 \left[F(y_j) - F(y_{j-1}) \right] \tag{5}$$

The standard deviation of the censored data set is shown in Eqn. 6.

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} \tag{6}$$

KAPLAN-MEIER CALCULATION EXAMPLE

Suppose the project has collected n = 29 discrete environmental soil samples from a site suspected to be contaminated with lead. These 29 samples were collected from a statistical sampling design and are representative of the site for characterization. The 29 lead concentrations (mg/kg) sorted from smallest to largest are:

< 1, <1, 2, 2.5, 2.8, < 3, 3.4, 3.9, < 4, < 4, < 4, 4.5, 4.9, 5.5, 5.5, 5.5, < 6, 6.7, 6.9, 7.4, < 9, 9.5, < 10, < 10, 10, 15, 49, 200, 9060

where "<" denotes a non-detect or left censored result. This data set is typical for environmental data as it has several distinct censoring levels or reporting limits, detected concentrations tied with non-detects, is heavily skewed to the right, and is not normally, lognormally, or gamma distributed. A censored probability plot of the data, the Shapiro-Wilk test, and ProUCL 4.0 confirm this. Note that k = 10 of the 29 samples (34%) are non-detects.

Table 1 shows the values of the variables used in the calculation of the estimated mean, standard deviation, and standard error of the mean.

Table 1. Calculations used for Kaplan-Meier example

j	y_j	$y_{j-1} < x_i \leq y_j$	m_j	n_j	F(y_j)	a_j
1	2	<1, <1, 2	1	3	0.2077	0.1039
2	2.5	2.5	1	4	0.2770	0.1870
3	2.8	2.8	1	5	0.3462	0.3947
4	3.4	<3, 3.4	1	7	0.4039	0.5967
5	3.9	3.9	1	8	0.4616	0.8736
6	4.5	<4, <4, <4, 4.5	1	12	0.5036	1.0751
7	4.9	4.9	1	13	0.5456	1.4024
8	5.5	5.5, 5.5, 5.5	3	16	0.6715	2.2082
9	6.7	<6, 6.7	1	18	0.7110	2.3504
10	6.9	6.9	1	19	0.7505	2.7256
11	7.4	7.4	1	20	0.7900	4.3846
12	9.5	<9, 9.5	1	22	0.8276	4.7984
13	10	<10, <10, 10	1	25	0.8621	9.1087
14	15	15	1	26	0.8966	39.59
15	49	49	1	27	0.9310	180.2
16	200	200	1	28	0.9655	8735
17	9060	9060	1	29	1	--

Table 2 shows the calculated mean, standard deviation, and standard error of the mean using the KM method. Table 2 also compares the KM results with three simple substitution methods:

- Substitute 0 for each non-detect (ND),
- Substitute ½ the reporting limit (RL) for each ND,
- Substitute the reporting limit (RL) for each ND.

Table 2 shows the KM mean of 325.34 mg/kg lies between the means calculated for methods that substitute 0 and the reporting limit for each ND, as expected. However, the KM standard deviation of 1651.09 mg/kg is smaller than any of the other three standard deviations from simple substitution methods. This is because the KM method maximizes the information available from the censored data instead of assigning fixed proxy values from substitution methods. The KM standard error of 315 mg/kg is slightly higher than the other standard errors from the substitution methods because the KM method recognizes the larger uncertainty of a mean that is calculated using censored data. Helsel (1990) and other publications referenced in this paper have consistently shown that using simple substitution methods introduces unnecessary biases in the summary statistics and are not recommended.

Table 2. KM parameter estimates compared to substitution methods

Method	Mean	Std. Dev.	SE of Mean
Kaplan-Meier	325.34	1651.09	315.00
Substitute 0 for NDs	324.31	1680.53	312.07
Substitute RL/2 for NDs	325.21	1680.35	312.03
Substitute RL for NDs	326.10	1680.17	312.00

KM estimates of the mean and standard error of the mean can then be used in equations for calculating upper confidence limits (UCLs) on the mean.

SAS CODE TO CALCULATE KAPLAN-MEIER

The following SAS code will calculate the KM statistics for any number of chemicals. The SAS program has a huge advantage over the ProUCL software since the ProUCL software generally handles only one chemical at a time. The

SAS code uses the same equations as ProUCL, and therefore, produces the same results, but in a more efficient manner.

SAS INITIAL MACROS

Before running the SAS code, some initial macros must be executed. The following macro OBSNVARS simply stores the number of records and variables in a data set into SAS macro variables.

```
%macro obsnvars(ds);
   %*  this macro returns the number of variables and observations from a data set;
 %global dset nvars nobs;
 %let dset=&ds;
 %let dsid = %sysfunc(open(&dset));
 %if &dsid %then %do;
   %let nobs =%sysfunc(attrn(&dsid,NOBS));
   %let nvars=%sysfunc(attrn(&dsid,NVARS));
   %let rc = %sysfunc(close(&dsid));
   %end;
 %else
    %put Open for data set &dset failed - %sysfunc(sysmsg());
%mend obsnvars;
```

The DS_NAMES2 macro sets a number of SAS data sets together.

```
%macro ds_names2(name, num);
  %do zz = 1 %to &num;
    &name&zz.
    %end;
%mend ds_names2;
```

The following macro definitions are determining the number of chemicals that will be run and a SORTBY macro variable where additional variables can be added for sorting the data.

```
%let N_POPNS = 1;
%let SORTBY = chemical;
```

SAS CODE FOR KM

The macro calc_km calculates the KM method using Eqns. 1 – 6. The data set b4 contains the data, where the variable HIT = 0 for censored results (non-detects), and HIT = 1 for uncensored detected results. The variable RESULTS is the reported measurement from the laboratory for detects or the reporting limit for non-detects.

```
%macro calc_km;
  %do POPNUM = 1 %to &N_POPNS;
    data one&POPNUM.;
      set b4;
        where popn_num=&POPNUM.; run;

    %obsnvars(one&POPNUM.); %let TOTN = &nobs;

    proc sort data=one&POPNUM.; by &SORTBY results;

    proc summary data=one&POPNUM.;
      var results;
      where hit=0;  ** nondetects only;
      output out=nondets(drop=_type_ _freq_) n=nds;  run;

    %obsnvars(nondets); %let N_NONDETECTS = &nobs;

    proc summary data=one&POPNUM.;
      var results;
      where hit=1;  ** detects only;
      by &SORTBY results;
      output out=distinct&POPNUM.(drop=_type_ _freq_) n=mi;
    proc print data=distinct&POPNUM.; run;
```

```
  %obsnvars(distinct&POPNUM.); %let N_DISTINCT_LEVELS = &nobs;

  data hitsonly;
    set one&POPNUM.;
      where hit=1; run;

  %obsnvars(hitsonly); %let N_DETECTS = &nobs;

%macro calc_ni;
   proc transpose data=distinct&POPNUM. out=outt&POPNUM.; var results; run;
      ** detected distinct results only;

   data a&POPNUM.;
     set one&POPNUM.;
     if _N_=1 then set outt&POPNUM.;
     %do i = 1 %to &N_DISTINCT_LEVELS;
          if results <= col&i. then N&i.=1; else N&i.=0;
          if results = col&i. and hit=1 then M&i.=1; else M&i.=0;
          %end; run;

     proc summary data=a&POPNUM.;
       var N1-N&N_DISTINCT_LEVELS. M1-M&N_DISTINCT_LEVELS.;
          by &SORTBY col1-col&N_DISTINCT_LEVELS;
          output out=calc_ni_out&POPNUM.(drop=_type_ _freq_)
       sum=N1-N&N_DISTINCT_LEVELS. M1-M&N_DISTINCT_LEVELS.;
          proc print data=calc_ni_out&POPNUM.; run;

     proc summary data=a&POPNUM.;
       var hit;
          output out=hitsum&POPNUM.(drop=_type_ _freq_) sum=N_DETECTS;

     data calc_cdf&POPNUM.;
       set calc_ni_out&POPNUM.;
          if _N_=1 then set hitsum&POPNUM.;
          N_DISTINCT_DETECT_LEVELS = &N_DISTINCT_LEVELS;
          CDF&N_DISTINCT_LEVELS = 1;
          CDF0 = 0;
          %do i = &N_DISTINCT_LEVELS %to 2 %by -1;
            %let N_DISTINCT_LEVELS_M1 = %eval(&i. - 1);
             CDF&N_DISTINCT_LEVELS_M1 = cdf&i. * (n&i. - m&i.) / n&i.;
          %end;
          KM_MEAN=0;
          %do i = 1 %to &N_DISTINCT_LEVELS;  ** calculate KM mean ;
          %let im1 = %eval(&i. - 1);
            KM_MEAN = KM_MEAN + col&i. * (cdf&i. - cdf&im1.);
            %end;
          KM_VAR=0;
          %do i = 1 %to &N_DISTINCT_LEVELS;  ** calculate KM standard deviation;
          %let im1 = %eval(&i. - 1);
            KM_VAR = KM_VAR + (col&i. - km_mean)**2 * (cdf&i. - cdf&im1.);
            %end;
          KM_STD = sqrt(km_var);
          %do j = 1 %to &N_DISTINCT_LEVELS - 1;  ** calculate KM SE of the mean;
            a&j. = 0;
            %do i = 1 %to &j;
              %let ip1 = %eval(&i. + 1);
               a&j. = a&j. + (col&ip1. - col&i.) * cdf&i.;
               %end;
            %end;
          sum = 0;
          %do i = 1 %to &N_DISTINCT_LEVELS - 1;
             %let ip1 = %eval(&i. + 1);
             sum = sum + a&i.**2 * m&ip1. / (n&ip1. * (n&ip1. - m&ip1.));
            %end;
          if n_detects > 1 then KM_STDERR2 = n_detects * sum / (n_detects - 1);
            KM_STDERR = sqrt(km_stderr2);
            drop sum; run;
```

```
%mend calc_ni;

 %if &N_DISTINCT_LEVELS > 1 and &N_NONDETECTS > 0 %then %do;
   %calc_ni
   %end;
 %end;

 data allkm;
   set %ds_names2(calc_cdf, &N_POPNS);
   keep &SORTBY n_distinct_detect_levels km_mean km_var km_std km_stderr;  run;

 proc print data=allkm; run;

 %mend calc_km;

 %calc_km
```

CONCLUSION

The nonparametric Kaplan-Meier product limit estimator has historically been used in survival analysis and medical studies for right censored data. However, Kaplan-Meier has been shown in recent literature to be the preferred method in many cases for estimating the mean, standard deviation, and standard error of the mean of left censored environmental data sets. Kaplan-Meier makes no underlying assumptions about the data, can be used with multiple reporting limits, and efficiently uses the censoring information from the data. SAS code was presented to calculate the Kaplan-Meier estimates that can process hundreds of chemicals much faster than the U.S. EPA ProUCL software and yet yield consistent results.

REFERENCES

Gilbert, Richard O. 1987. *Statistical Methods for Environmental Pollution Monitoring*. New York: Van Nostrand Reinhold.

Helsel, Dennis R. 1990. Less than obvious: “Statistical treatment of data below the detection limit.” *Environmental Science and Technology*. Vol. 24, 1766-1774.

Helsel, Dennis R. 2005. *Nondetects and Data Analysis: Statistics for Censored Environmental Data*. New Jersey: John Wiley & Sons, Inc.

Helsel, Dennis R. and T. A. Cohn. 1988. “Estimation of descriptive statistics for multiply censored water quality data.” *Water Resources Research*. Vol. 24, 1997-2004.

Helsel, Dennis R. and R. J. Gilliom. 1986. “Estimation of distributional parameters for censored trace level water quality data, verification and applications.” *Water Resources Research*. Vol. 22, 147-155.

Kaplan, E. L. and O. Meier. 1958. “Nonparametric estimation from incomplete observations.” *Journal of the American Statistical Association*. Vol. 53, 457-481.

U.S. Environmental Protection Agency. 2007. *ProUCL Version 4.0 Technical Guide*. EPA/600/R-07/041.

CONTACT INFORMATION

The author welcomes and encourages any questions, corrections, feedback, and remarks. Contact the author at:

Dennis Beal, Ph.D.
Senior Statistician/Risk Scientist
Science Applications International Corporation (SAIC)
P.O. Box 2501 — phone: 865-481-8736
151 Lafayette Drive — fax: 865-481-4757
Oak Ridge, Tennessee 37831 — e-mail: beald@saic.com