

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

Dennis J. Beal, Leidos, Oak Ridge, Tennessee

ABSTRACT

Calculating percentiles such as the median and quartiles 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 recently this method has also 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 percentiles such as the median 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, standard deviation, and percentiles 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, percentiles

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) arsenic concentration of 10 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 $0 \leq X < 5$ 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 for calculating percentiles on censored distributions. This method is not appropriate for radionuclides since the actual reported concentration (whether positive, negative, zero, non-detected, or detected) can be used as reported from the laboratory.

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. Beal (2010) wrote a SAS macro that calculates Kaplan-Meier estimates of the mean, standard deviation and standard error of the mean.

Helsel (2005) compares the 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 $n \leq 50$ and the MLE or $n > 50$. For more than 80% non-detects, Helsel recommends high sample percentiles such as the 90th or 95th 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 9.4 running on a Windows 7 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 5.0.00 (U.S. EPA 2013) 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 $x_1, x_2, \dots, x_n$ 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 $y_1, y_2, \dots, y_k$ denote the k distinct values at which detects are observed so that $2 \leq k \leq n$. For $j = 1, 2, \dots, k$, let m_j denote the number of detects at y_j and let n_j denote the cumulative number of $x_i \leq y_j$. Define $F(x)$ in Eqn. 1.

$$\begin{aligned}
 F(x) &= 1 & x \geq y_k \\
 F(x) &= \prod_{j: y_j > x}^k \frac{n_j - m_j}{n_j} & y_1 \leq x \leq y_{k-1} \\
 F(x) &= F(y_1) & x_1 \leq x \leq y_1 \\
 F(x) &= 0 & 0 \leq x \leq x_1
 \end{aligned} \tag{1}$$

An estimate of the sample percentile p ($0 < p < 1$) using the KM method is the concentration y_j such that $F(y_j) \geq p$ and $F(y_{j-1}) < p$ for $1 \leq j \leq k$. For example, the sample median is the concentration y_j such that $F(y_j) \geq 0.5$ and $F(y_{j-1}) < 0.5$. The first quartile is the concentration y_j such that $F(y_j) \geq 0.25$ and $F(y_{j-1}) < 0.25$.

KAPLAN-MEIER CALCULATION EXAMPLE

Suppose the project has collected $n = 29$ discrete environmental soil samples from a site suspected to be contaminated with arsenic. These 29 samples were collected from a statistical sampling design and are representative of the site for characterization. The 29 arsenic 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 concentration. 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 5.0 confirm this. Note that 10 of the 29 samples (34%) are non-detects.

Table 1 shows the values of the variables used in the calculation of the estimated percentiles.

Table 1. Calculations used for Kaplan-Meier example

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

Table 2 shows the calculated first decile (D_1) or 10th percentile, first quartile (Q_1) or 25th percentile, median (Q_2) or 50th percentile, third quartile (Q_3) or 75th percentile and ninth decile (D_9) or 90th percentile using the KM method. Table 2 also compares the KM results with three simple substitution methods:

- Substitute 0 for each non-detect (ND),
- Substitute $\frac{1}{2}$ the reporting limit (RL) for each ND,
- Substitute the RL for each ND.

Table 2 shows the KM percentiles D_1 , Q_1 and Q_2 lie between the percentiles calculated for methods that substitute 0 and the reporting limit for each ND, as expected. These percentiles also lie between the methods of substituting 0 and half the RL. The KM Q_3 agrees with two of the three methods, while D_9 agrees with all other substitution methods since the left censoring occurs below the 90th percentile. Note that Q_1 , Q_2 , Q_3 and D_9 are detected concentrations, while D_1 is a non-detect. 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 percentile estimates compared to substitution methods

Method	D_1	Q_1	Q_2	Q_3	D_9
Kaplan-Meier	1	2.5	4.5	6.9	49
Substitute 0 for NDs	0	0	3.9	6.9	49
Substitute RL/2 for NDs	1.5	2.5	4.9	6.9	49
Substitute RL for NDs	2	3.9	5.5	9.5	49

KM estimates of Q_1 and Q_3 can then be used for producing box and whisker plots to visually show the distribution of the data.

SAS CODE TO CALCULATE KAPLAN-MEIER

The following SAS code will calculate the KM percentiles for any number of chemicals.

SAS INITIAL MACROS

Before running the SAS code, some initial macro variables 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 following macro variable definitions determine the number of chemicals that will be run and a `SORTBY` macro variable where additional variables (such as units) 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 Eqn. 1. 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;

  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 coll-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;
    %do i = 1 %to &N_DISTINCT_LEVELS_ALL.; ** calculate KM median;
      %let im1 = %eval(&i. - 1);
      if cdf&i. >= 0.5 and cdf&im1. < 0.5 then KM_MEDIAN = col&i.;
    %end;
  %end;

```

```

    %end;
    %do i = 1 %to &N_DISTINCT_LEVELS_ALL.; ** calculate KM 10th percentile;
        %let im1 = %eval(&i. - 1);
        if cdf&i. >= 0.1 and cdf&im1. < 0.1 then KM_P10 = col&i.;
    %end;
    %do i = 1 %to &N_DISTINCT_LEVELS_ALL.; ** calculate KM 25th percentile;
        %let im1 = %eval(&i. - 1);
        if cdf&i. >= 0.25 and cdf&im1. < 0.25 then KM_P25 = col&i.;
    %end;
    %do i = 1 %to &N_DISTINCT_LEVELS_ALL.; ** calculate KM 75th percentile;
        %let im1 = %eval(&i. - 1);
        if cdf&i. >= 0.75 and cdf&im1. < 0.75 then KM_P75 = col&i.;
    %end;
    %do i = 1 %to &N_DISTINCT_LEVELS_ALL.; ** calculate KM 90th percentile;
        %let im1 = %eval(&i. - 1);
        if cdf&i. >= 0.9 and cdf&im1. < 0.9 then KM_P90 = col&i.;
    %end;
    run;
%mend calc_ni;

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

data allkm;
    set calc_cdf1-calc_cdf&N_POPNS;
    keep &SORTBY n_distinct_detect_levels km_median km_p10 km_p25 km_p75 km_p90;

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 summary statistics such as the mean, standard deviation, standard error of the mean, and percentiles of left censored environmental data sets. Kaplan-Meier makes minimal 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 for percentiles that can process hundreds of chemicals quickly.

REFERENCES

- Beal, Dennis J. 2010. "A Macro for Calculating Summary Statistics on Left Censored Environmental Data using the Kaplan-Meier Method," *Proceedings of the 18th Annual Conference of the SouthEast SAS Users Group*.
- 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. 2013. *ProUCL Version 5.0.00 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 J. Beal, Ph.D.
Senior Statistician/Risk Scientist
Leidos
301 Laboratory Road
Oak Ridge, Tennessee 37831

phone: 865-481-8736
e-mail: beald@leidos.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are registered trademarks or trademarks of their respective companies.