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 ≤ 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 Gilling (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 ≤ 50 and the MLE or ROS when 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...
Let \( x_1, x_2, \ldots, x_n \) 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 \( y_1, y_2, \ldots, y_k \) denote the \( k \) distinct values at which detects are observed so that \( 2 \leq k \leq n \). For \( j = 1, 2, \ldots, k \), let \( m_j \) denote the number of detects at \( y_j \) and let \( n_i \) denote the cumulative number of \( x_i \leq y_j \). Define \( F(x) \) in Eqn. 1.

\[
\begin{align*}
F(x) &= 1 & x \geq y_k \\
F(x) &= \prod_{j=1}^{k} \frac{n_j - m_j}{n_j} & y_j \leq x \leq y_{k-1} \\
F(x) &= F(y_j) & x_j \leq x \leq y_j \\
F(x) &= 0 & 0 \leq x \leq x_j
\end{align*}
\]

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:

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

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>
<thead>
<tr>
<th></th>
<th>( y_j )</th>
<th>( y_{j-1} &lt; x_j \leq y_j )</th>
<th>( m_j )</th>
<th>( n_i )</th>
<th>( F(y_j) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>&lt;1, &lt;1, 2</td>
<td>1</td>
<td>3</td>
<td>0.2077</td>
</tr>
<tr>
<td>2</td>
<td>2.5</td>
<td>2.5</td>
<td>1</td>
<td>4</td>
<td>0.2770</td>
</tr>
<tr>
<td>3</td>
<td>2.8</td>
<td>2.8</td>
<td>1</td>
<td>5</td>
<td>0.3462</td>
</tr>
<tr>
<td>4</td>
<td>3.4</td>
<td>&lt;3, 3.4</td>
<td>1</td>
<td>7</td>
<td>0.4039</td>
</tr>
<tr>
<td>5</td>
<td>3.9</td>
<td>3.9</td>
<td>1</td>
<td>8</td>
<td>0.4616</td>
</tr>
<tr>
<td>6</td>
<td>4.5</td>
<td>&lt;4, &lt;4, &lt;4</td>
<td>1</td>
<td>12</td>
<td>0.5036</td>
</tr>
<tr>
<td>7</td>
<td>4.9</td>
<td>4.9</td>
<td>1</td>
<td>13</td>
<td>0.5456</td>
</tr>
<tr>
<td>8</td>
<td>5.5</td>
<td>5.5, 5.5</td>
<td>3</td>
<td>16</td>
<td>0.6715</td>
</tr>
<tr>
<td>9</td>
<td>6.7</td>
<td>&lt;6, 6.7</td>
<td>1</td>
<td>18</td>
<td>0.7110</td>
</tr>
<tr>
<td>10</td>
<td>6.9</td>
<td>6.9</td>
<td>1</td>
<td>19</td>
<td>0.7505</td>
</tr>
<tr>
<td>11</td>
<td>7.4</td>
<td>7.4</td>
<td>1</td>
<td>20</td>
<td>0.7900</td>
</tr>
<tr>
<td>12</td>
<td>9.5</td>
<td>&lt;9, 9.5</td>
<td>1</td>
<td>22</td>
<td>0.8276</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
<td>&lt;10, &lt;10</td>
<td>1</td>
<td>25</td>
<td>0.8621</td>
</tr>
<tr>
<td>14</td>
<td>15</td>
<td>15</td>
<td>1</td>
<td>26</td>
<td>0.8966</td>
</tr>
<tr>
<td>15</td>
<td>49</td>
<td>49</td>
<td>1</td>
<td>27</td>
<td>0.9310</td>
</tr>
<tr>
<td>16</td>
<td>200</td>
<td>200</td>
<td>1</td>
<td>28</td>
<td>0.9655</td>
</tr>
<tr>
<td>17</td>
<td>9060</td>
<td>9060</td>
<td>1</td>
<td>29</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 2 shows the calculated first decile (D₁) or 10th percentile, first quartile (Q₁) or 25th percentile, median (Q₂) or 50th percentile, third quartile (Q₃) or 75th percentile and ninth decile (D₉) 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 ½ the reporting limit (RL) for each ND,
- Substitute the RL for each ND.

Table 2 shows the KM percentiles D₁, Q₁ and Q₂ lie between the percentiles calculated for methods that substitute 0 and half the RL. The KM Q₃ agrees with two of the three methods, while D₉ agrees with all other substitution methods since the left censoring occurs below the 90th percentile. Note that Q₁, Q₂, Q₃ and D₉ are detected concentrations, while D₁ 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

<table>
<thead>
<tr>
<th>Method</th>
<th>D₁</th>
<th>Q₁</th>
<th>Q₂</th>
<th>Q₃</th>
<th>D₉</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kaplan-Meier</td>
<td>1</td>
<td>2.5</td>
<td>4.5</td>
<td>6.9</td>
<td>49</td>
</tr>
<tr>
<td>Substitute 0 for NDs</td>
<td>0</td>
<td>0</td>
<td>3.9</td>
<td>6.9</td>
<td>49</td>
</tr>
<tr>
<td>Substitute RL/2 for NDs</td>
<td>1.5</td>
<td>2.5</td>
<td>4.9</td>
<td>6.9</td>
<td>49</td>
</tr>
<tr>
<td>Substitute RL for NDs</td>
<td>2</td>
<td>3.9</td>
<td>5.5</td>
<td>9.5</td>
<td>49</td>
</tr>
</tbody>
</table>

KM estimates of Q₁ and Q₃ 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.

```sas
%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.

```sas
%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.
A Macro for Calculating Percentiles on Left Censored Environmental Data using the Kaplan-Meier Method, continued

%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 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.;
  %let N_DISTINCT_DETECT_LEVELS = &N_DISTINCT_LEVELS;
  %let CDF&N_DISTINCT_LEVELS = 1;
  %let CDF0 = 0;
  %do i = &N_DISTINCT_LEVELS %to 2 %by -1;
    %let N_DISTINCT_LEVELS_M1 = %eval(&i. - 1);
    %let 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;
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


**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.