Extended ABSTRACT

SYMPOSIUM OF MATHEMATICS

Fundamental Science Congress 2010

Scientific Innovations Transgress Time

(FSC2010)

"Scientific Innovations Transgress Time"

FUNDAMENTAL SCIENCE CONGRESS 2010

Editors:
ADEM KILICMAN
MALIK ABU HASSAN
LEONG WAH JUNE
SITI HASANA SAFAR

Organized by:
FACULTY OF SCIENCE
Universiti Putra Malaysia

Fundamental Science Congress 2010: Mathematics Symposium

"Scientific Innovations Transgress Time"

Editors:

Adem Kilicman
Malik Abu Hassan
Leong Wah June

ISBN 978-983-2519-07-2

1. Science Congress
2. Mathematics Symposium

Printed by:
Syarikat Perniagaan Weng Sing
Lot 452-A Jalan 18/2A,
Taman Seri Serdang, 43400
Serdang, Selangor D.E.
MALAYSIA

CONTENTS

Piecewise Exponential Estimator of Exceedance Probability of Environmental Data subject to Limits of Detection

[1]Abdul Kudus, [2]Noor Akma Ibrahim
[1] *Institute for Mathematical Research, UPM, Serdang 43400 and Department of Statistics, UNISBA, Bandung 40116 (email:akudus69@yahoo.com)*
[2] *Institute for Mathematical Research and Department of Mathematics, UPM, Serdang 43400 (email:nakma@upm.edu.my)*

Introduction

Survival analysis methods were designed initially for handling survival time data which subject to censoring. It's common that, in the end of the study, not all of patients have experience the event, such as death or relapse. Hence survival times of those patients are (right) censored. Environmental data are also often censored, with a number of nondetect values included in the data set. They are often multiply-censored, as detection limit thresholds change over time or with varying sample characteristics or among different laboratories (Helsel, 2005).

Piecewise exponential estimator (PEXE) was proposed by Kim and Proschan (1991) for the estimation of survivor function. Survivor function gives information on probability of patient who will survive beyond a specified time. Two classes of estimator are nonparametric and parametric. Kaplan-Meier (KM) estimator was the most popular of nonparametric estimator. Whereas, parametric estimator was developed based on some common distribution, such as exponential and Weibull.

PEXE is an alternative to KM estimator. Not like KM, which is of type step function with discontinuities at the observed failure time, PEXE has continuous function. Even though, PEXE was not parametric estimator.

Adopting PEXE for handling environmental data is not straight forward. There must be some modifications on handling nondetect (left-censored) data. The objective of this paper is twofold. First, we modify PEXE for handling nondetect data. Second, by applying the proposed method to environmental pollutant data, we explore the possibility of using PEXE along with comparison with KM and parametric estimator.

PEXE

The primary difference between environmental and medical data is that the environmental data are dominantly left-censored. Uncertainty occurs at the lower-end, for values typically plot on the left side of a graph. Left-censored data must be transformed into right-censored data before the available procedure can be used.

Here is the precise and general description of the PEXE of the survival function based on right-censored data. For each subject, we assume the time that observation ends is the minimum of 2 statistically independent random variables X and Y; X is governed by a continuous life distribution function $F(t)$ and Y by a continuous life distribution function $G(t)$. The random variable X represents the time to death (or failure) and the random variable Y represents the time to withdrawal. Let $f(t)$ and $g(t)$ be the corresponding probability density functions. Then the each subject in the study is observed to

- death at time t with likelihood $f(t)[1 - G(t)]$,
- be withdrawn at time t with likelihood $g(t)[1 - F(t)]$, or
- experience neither death nor withdrawal during $[0, t]$ with likelihood $[1 - F(t)][1 - G(t)]$.

Suppose n new patients joined a clinical trial at time 0. We observe each patient until death due to the illness that is under study or withdrawal from further observation, whichever occurs first. The observed outcomes are listed in order of occurrence as follows:

$$0 < w_{11} < \ldots < w_{1k1} < z_1 < w_{21} < \ldots < w_{2k2} < z_2 < \ldots < z_{i-1} < w_{m1} < \ldots < w_{mkm} < z_m < w_{m+1,1} < \ldots < w_{m+1,km+1}$$

We take $z_0 = 0$, $z_{m+1} = \infty$, and $k_i = 0$ when no withdrawals occur in (z_{i-1}, z_i). In the above expression, strict inequality occurs between successive observed values. This is a consequence of the assumption that the life distribution functions F and G are continuous. There are m observed deaths and $\sum_{i=1}^{m+1} k_i$ withdrawals. Because the total number of observations is n, we must have $m + \sum_{i=1}^{m+1} k_i = n$. Based on the ordered observations displayed above, we describe the PEXE for survival function $S(t)$. On the first failure interval $[0, Z_1]$, we estimate the failure rate of the

exponential survival function that best fits the data observed during [0, Z_1]; "best" in the sense of the MLE. We are operating as if the unknown survival function $S(t)$ is exponential on [0, Z_1]. It is well known that the MLE of the failure rate of an exponential distribution is of the following form:

$$\frac{\text{Observed number of deaths}}{\text{Observed total time on test}}$$

Applying this principle, we obtain the failure rate estimator r_1 on [0, Z_1],

$$\hat{r}_1 = \frac{1}{\sum_{i=1}^{k_1} w_{1i} + z_1 + (n - k_1 - 1)z_1}$$

The value 1 in the numerator is simply the number of deaths observed during [0, Z_1]. The denominator represents the total time on test observed during [0, Z_1] : One patient was observed until he withdrew from the study at time w_{11}. A second was observed until he withdrew at time w_{12}, etc., until the last patient withdrew during [0, Z_1], contributing w_{1k_1} years to total time on test observed during [0, Z_1]. In addition, the patient who died at time z_1 was observed for z_1 years. Finally, at time z_1, there were $n - k_1 - 1$ patients still remaining in the study that had been observed for z_1 years each. Thus, on the first failure interval, an estimator of survival function is:

$$S(t) = \exp(-\hat{r}_1 t), \text{ for } 0 \le t \le z_1$$

In the similar fashion, we estimate $\hat{r}_2$ the MLE of the failure rate of the exponential distribution that fits the data on (z_1, z_2]:

$$\hat{r}_2 = \frac{1}{\sum_{i=1}^{k_2} (w_{2i} - z_1) + (z_2 - z_1) + (n - k_1 - k_2 - 2)(z_2 - z_1)}$$

The exponential estimator of the survival function on the second failure interval (z_1, z_2] is

$$S(t) = \exp\{-\hat{r}_1 z_1 - \hat{r}_2 (t - z_2)\}, \text{ for } z_1 \le t \le z_2$$

At $t = z_1$, the two expressions $\hat{r}_1$ and $\hat{r}_2$ agree; that is, the exponential pieces on successive intervals between deaths are joined to form a continuous survival function curve. From the two survival functions, one can see that this is accomplished for the first two exponential pieces on [0, z_1] and (z_1, z_2] by inserting the term $-\hat{r}_1 z_1$ in the exponent of e in $\hat{r}_2$. We continue in this fashion obtaining in succession $\hat{r}_3, \ldots, \hat{r}_m$. We can now express the survival function estimator in general over the interval [0, z_m]:

$$\hat{S}(t) = \begin{cases} \exp(-\hat{r}_1 t) & \text{for } 0 \le t \le z_1 \\ \exp\{-[\hat{r}_1 z_1 + \ldots + \hat{r}_i (z_i - z_{i-1}) + \hat{r}_{i+1}(t - z_i)]\} & \text{for } z_i \le t \le z_{i+1}, i = 1,2,\ldots,m-1 \\ \text{No estimator} & t \ge z \end{cases}$$

Real Application

This example uses data of copper concentrations with 35 uncensored and 14 nondetect observations in the San Joaquin Valley, California, published by Millard and Deverel (1988).

Table 1: Groundwater concentrations of copper (cu) in micrograms per liter

No	x_i	*No*	x_i	*No*	x_i	*No*	x_i	*No*	x_i	*No*	x_i	*No*	x_i
1	2	8	4	15	2	22	<5	29	<15	36	8	43	3
2	2	9	<10	16	<10	23	17	30	<5	37	1	44	6
3	12	10	<1	17	3	24	23	31	4	38	15	45	3
4	2	11	1	18	<1	25	9	32	<5	39	3	46	4
5	1	12	<2	19	1	26	9	33	<5	40	3	47	5
6	<10	13	<2	20	1	27	3	34	<5	41	1	48	14
7	<10	14	1	21	3	28	3	35	4	42	6	49	4

Results and Discussions

For comparison purposes, we also provide KM and parametric log-normal estimator. Figure 1 showed these three probabilities of exceedance of groundwater concentration of copper. As concentration of copper increase, the probability of exceedance decrease up to about 20 micrograms per liter. The estimated median of copper concentration is in increasing order resulted by parametric, PEXE and KM method, respectively.

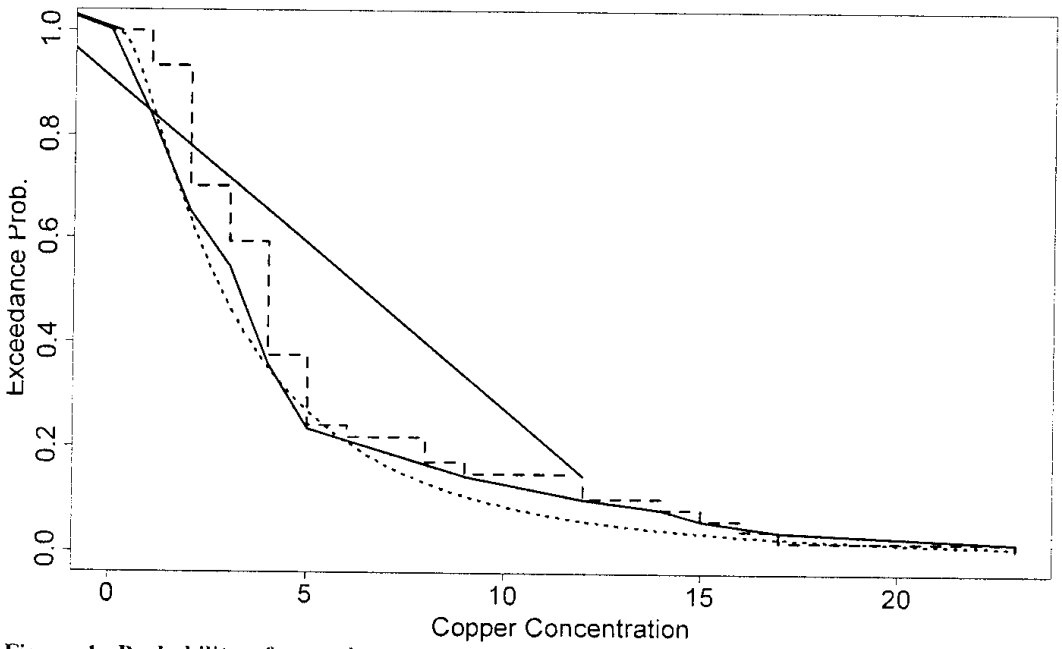

Figure 1: Probability of exceedance curve resulted from KM (dashed), PEXE (solid) and parametric log-normal (dotted) method

References

Helsel, D.R. (2005). *Nondetects and Data Analysis.* Wiley-Interscience, New jersey.

Kim, J.S. and Proschan, F. (1991). Piecewise Exponential Estimator of the Survivor Function. *IEEE Trans. On Reliab.* 40(2), 134 – 139.

Millard, S. P. and Deverel, S.J. (1988). Nonparametric statistical methods for comparing two sites based on data with multiple nondetect limits. *Water Resources Research* 24(12), 2087-2098.