Proceedings of the Korean Nuclear Society Spring Meeting

Cheju, Korea, May 2001

The Development of A New Algorithm to Calculate A Survival Function in Non-parametric Ways

Kwang-Won Ahn, Yoonik Kim and Chang-Hyun Chung

Seoul National University

San 56-1 Shilim-dong Kwanak-gu

Seoul, 151-742, Korea

Kil Yoo Kim

Korea Atomic Energy Research Institute

Taejon, Korea, 306-600 Integrated Safety Assessment Team

Abstract

In this study, a generalized formula of the Kaplan-Meier method is developed. The idea of this algorithm is that the result of the Kaplan-Meier estimator is the same as that of the redistribute-to-the-right algorithm. Hence, the result of the Kaplan-Meier estimator is used when we redistribute to the right. This can be explained as the following steps, at first, the same mass is distributed to all the points. At second, when you reach the censored points, you must redistribute the mass of that point to the right according to the following rule; to normalize the masses, which are located to the right of the censored point, and redistribute the mass of the censored point to the right according to the ratio of the normalized mass. Until now, we illustrate the main idea of this algorithm. The meaning of that idea is more efficient than PL-estimator in the sense that it decreases the mass of after that area. Just like a redistribute to the right algorithm, this method is enough for the probability theory.

I. Introduction

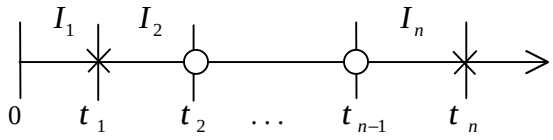
There are generally two kinds of methods to estimate reliability that focuses on system or component failure in engineering, or survival function that focuses on patient death in medicine. They are parametric and non-parametric methods. As we know, there are few data related to severe accidents in nuclear power plants. Therefore, the parametric method has been used. But if we have lots of data concerned with system failure or patient death, the non-parametric method could be used to estimate reliability or survival function. In this paper we focus on the latter in which case some problems can arise when we treat the

data. That is to say, when we treat the data, not only uncensored data but also the censored data included in them. Therefore, it is important to estimate correct reliability or survival function that reflects the information of the censored data, when we use the non-parametric method.

There are several methods e. g, to estimate survival function in non-parametric ways, reduced sample method, actuarial method and Kaplan-Meier (Product-Limit) method. In this paper, the comparisons are conducted between the PL-estimator and the new one. Because the reduced sample method and actuarial method do not reflect censoring information correctly. Therefore, PL-estimator is used to compare the result of the new algorithm.

II. Product-limit estimator

Let t_i, the right end point of I_i, be the i-th ordered censored or uncensored observation.



○ =censored and ╳=uncensored

Let $\Re(t)$ denote the risk set at time t, which is the set of subjects still alive at time t-, and let

n_i =# in $\Re(Y_{(i)})$ =# alive at time $Y_{(i)}-$,

d_i =# died at time $Y_{(i)}$,

$p_i = P$ {surviving through I_i | alive at beginning of I_i },

$= P\{T > t_i \mid T > t_{i-1}\}$,.

From the estimates

$$\hat{q}_i = \frac{d_i}{n_i},$$

$$\hat{p}_i = 1 - \hat{q}_i = \begin{cases} 1 - \frac{1}{n_i} & d_{(i)} = 1, (uncensored) \\ 1 & d_{(i)} = 0, (censored) \end{cases}$$

the PL estimate when no ties are present is

$$\hat{S}(t) = \prod_{y_{(i)} \le t} \hat{p}_i = \prod_{y_{(i)} \le t} \left(\frac{n-i}{n-i+1}\right)^{d_{(i)}}. \qquad (1)$$

(i) For tied uncensored observations, suppose just before time t, there are m individuals alive, and at time t, d death occur. Split the time of the d deaths infinitesimally so that the factor for the d deaths in the product-limit estimator is

$$(1-\frac{1}{m})(1-\frac{1}{m-1})\cdots(1-\frac{1}{m-d+1})=\frac{m-d}{m}=1-\frac{d}{m}.$$

(ii) If censored and uncensored observations are tied, consider the uncensored observations to occur just before the censored observations.

(iii) If the last(ordered) observation $y_{(n)}$ is censored, then for $\hat{S}(t)$ as defined above

$$\lim_{t\to\infty}\hat{S}(t)>0$$

Sometimes it is preferable to redefine $\hat{S}(t)=0$ for $t\geq y_{(n)}$ or to think of it as being undefined for $t\geq y_{(n)}$ if $d_{(n)}=0$.

From notes (i) and (ii), by letting

$$y'_{(1)}<y'_{(2)}<\cdots<y'_{(r)}$$

denote the distinct survival times and

$$d'_{(j)}=\begin{cases}1 & \textit{if the observations at time } y'_{(j)} \textit{ are uncensored},\\ 0 & \textit{if censored},\end{cases}$$

n'_i =# *in* $\Re(y'_{(j)})$,

d'_i =# *died at time* $y'_{(j)}$,

the PL estimate allowing for ties is

$$\hat{S}(t)=\prod_{u:y'_{(j)}\leq t}(1-\frac{d_j}{n_j})=\prod_{y'_{(j)}\leq t}(1-\frac{d_j}{n_j})^{d'_{(j)}}. \qquad (2)$$

III. Redistribute-to-the-Right Algorithm

Efron introduced another method of computing the PL estimator. Generally speaking, the redistribute-to-the-right algorithm gives the Kaplan-Meier product-limit estimator. Assuming no ties, there are two principal ways of proving this result.

(1) With the redistribute-to-the-right algorithm, all points $y_{(i)}$, censored or uncensored, initially have equal mass 1/n. The algorithm moves from left to right through the order statistics. When it reaches $y_{(i)}-$, all the remaining points $y_{(i)}$, $y_{(i+1)}$, $\cdots$, $y_{(n)}$ have equal mass on them due to the way the algorithm operates. Suppose the total remaining mass is $\tilde{S}(y_{(i)}-)$. By the equality of the mass $y_{(i)}$ has $\frac{\tilde{S}(y_{(i)}-)}{n-i+1}$ assigned to it, which it will keep if it is uncensored. If it is censored, this mass is distributed to the right.

Since the PL estimator $\hat{S}$ starts at 1 as dose $\tilde{S}$ and jumps of sizes $\frac{\tilde{S}(y_{(i)}-)}{n-i+1}$ at the uncensored observations and zero at the censored observations, the two estimators are identical.

(2) For the Kaplan-Meier estimator

$$\begin{aligned}
\tilde{\Delta}_{(i)} &= \tilde{S}(y_{(i)}-) - \tilde{S}(y_{(i)}), \\
&= \prod_{j=1}^{i-1}\left(\frac{n-j}{n-j+1}\right)^{d_{(j)}} - \prod_{j=1}^{i}\left(\frac{n-j}{n-j+1}\right)^{d_{(j)}}, \\
&= \prod_{j=1}^{i-1}\left(\frac{n-j}{n-j+1}\right)^{d_{(j)}} \frac{d_{(i)}}{n-j+1}, \\
&= \frac{d_{(i)}}{n}\prod_{j=1}^{i-1}\left(\frac{n-j+1}{n-j}\right)^{1-d_{(j)}}.
\end{aligned} \tag{3}$$

Let $j_1 < \cdots < j_i$ be the indices of the censored observations which precede $y_{(i)}$. For the redistribute-to-the-right algorithm the mass assigned to $y_{(i)}$ if $d_{(i)} = 1$ is

$$\begin{aligned}
\tilde{\Delta}_{(i)} &= \frac{1}{n}\left(1+\frac{1}{n-j_1}\right)\left(1+\frac{1}{n-j_2}\right)\cdots\left(1+\frac{1}{n-j_i}\right), \\
&= \frac{1}{n}\prod_{j=1}^{i-1}\left(\frac{n-j+1}{n-j}\right)^{1-d_{(j)}}.
\end{aligned} \tag{4}$$

and if $d_{(i)} = 0$, $\tilde{\Delta}_{(i)} = 0$. This is identical to $\hat{\Delta}_{(i)}$, so the redistribute-to-the-right algorithm gives the PL estimator.

IV. The advanced algorithm of computing the survival function

We introduced another method of computing the survival function. At first, we should formulize the PL estimator as the basic mass of our estimator. Assume no ties.

$n = total\ \#\ of\ observations.$

$y_{(1)} < y_{(2)} \cdots < y_{(n)} : ordered\ statistics\ of\ observations$

$y_1 < y_2, \cdots, < y_j \ (i = 1,2,\cdots, j) : observations\ which\ censoring\ censoring\ occur$

$mass\ of\ y_{(k)} = w_{(k)}$

$$w_{(k)} = \frac{1}{n}\ (mass\ at\ start)$$

$$w_{(k)} = \frac{1}{n}\ (k < i_1)$$

$$w_{(k)} = 0 \ \ (k = i_1)$$

After first redistribution,

$$w_{(k)} = \frac{1}{n} + \frac{1}{n-i_1} \cdot \frac{1}{n} \quad (k > i_1)$$

$$w_{(k)} = 0 \quad (k = i_2) .$$

After second redistribution,

$$w_{(k)} = \frac{1}{n} + \frac{1}{n-i_1} \cdot (\frac{1}{n}) + \frac{1}{n-i_2} \{\frac{1}{n} + \frac{1}{n-i_1} \cdot (\frac{1}{n})\} \quad (k > i_2)$$

. . . .

After k-th redistribution, the mass of $y_{(k)} (= w_{(k)})$ is defined as following equations.

$$w_{(k)} = w_{(k-1)} + \frac{1}{n-i_h} \cdot w_{(k-1)} \quad (k > i_h) \tag{5}$$

Therefore, the mass after last redistribution is defined as following equations.

$$w_{(k)} = \frac{1}{n} \ (k < i_1) \tag{6}$$

$$w_{(k)} = 0 \quad (k = i) \tag{7}$$

$$w_{(k)} = w_{(k-1)} + \frac{1}{n-i_h} w_{(k-1)} \quad (i_h < k < i_{h+1}) \tag{8}$$

$$w_{(k)} = w_{(k-1)} + \frac{1}{n-i_j} w_{(k-1)} \ (k > i_j) \tag{9}$$

If we consider the ties, the results are the same as following equations.

$$w_{(k)} = \frac{d_1}{n} \ (k < i_1) \tag{10}$$

$$w_{(k)} = 0 \quad (k = i) \tag{11}$$

$$w_{(k)} = w_{(k-1)} + \frac{d_{i_k}}{n-i_h} w_{(k-1)} \quad (i_h < k < i_{h+1}) \tag{12}$$

$$w_{(k)} = w_{(k-1)} + \frac{d_{i_j}}{n-i_j} w_{(k-1)} \ (k > i_j) \tag{13}$$

$d_i = the \ \# \ of \ \ i-th \ observations$

Until now, we derive a mass function, which will be used in the following theorem. We use this mass function when we redistribute to the right of censoring point. This theorem will support the PL-estimator. Because Kaplan-Meier (PL) estimator is rational if the survival tendency is fixed. But in the real world, the events do not occur with same tendency. In this theorem, we distribute a same mass at first. This assumption will be sufficient in the statistical point of view.

F(last observation)=1

S(last observation)=0

But it is rational considering the weighting of mass, which is located in the right of censoring point when we redistribute the mass of censoring point to the right. From now, let' s derive a formula of estimating a survival function, reflecting a tendency of data. The purpose of this paper is to estimate a correct survival function. Therefore we develop a model that is more realistic.

$w_{(i,k)} = mass\ of\ \ y_{(k)}\ \ according\ \ to\ the\ i-th\ redistribu\,tion$

$d_k = \#\ of\ \ ties\ at\ \ y_{(k)}$

$$w_{(i_0,k)} = \frac{d_k}{n} \tag{14}$$

$$\begin{aligned} w_{(i_j,k)} &= w_{(i_{j-1},k)} && (k < i_j) \\ &= 0 && (k = i_j) \\ &= w_{(i_{j-1},k)} + w_{i_j} \cdot \frac{w_{(k)}}{\sum_{k>i_j}^{n} w_{(k)}} && (k > i_j) \end{aligned} \tag{15}$$

V. Sample calculation

Table1. 21 patients receiving<5000 rad, Survival rate using Kaplan-Meier method

Patient, rank	Survival time, mo	Age at entry, yr	Gender	$\hat{S}_n(t)$
1	7	68	F	0.9524
2	9	69	F	0.9048
3	12	68	F	0.8096
4	12	71	F	0.7620
5	23	77	M	0.7144
6	24	70	F	0.5239
7	24	67	F	0.5239
8	24	68	M	0.5239
9	24	88	M	0.5239
10	29+	89	M	0.5239
11	34	28	M	0.4715
12	41	73	M	0.4191
13	54	60	F	0.3667
14	72+	60	F	0.3667
15	78	44	M	0.3056
16	80+	82	F	
17	83+	62	F	
18	92+	53	F	
19	139+	66	F	
20	139+	63	F	
21	139+	68	M	

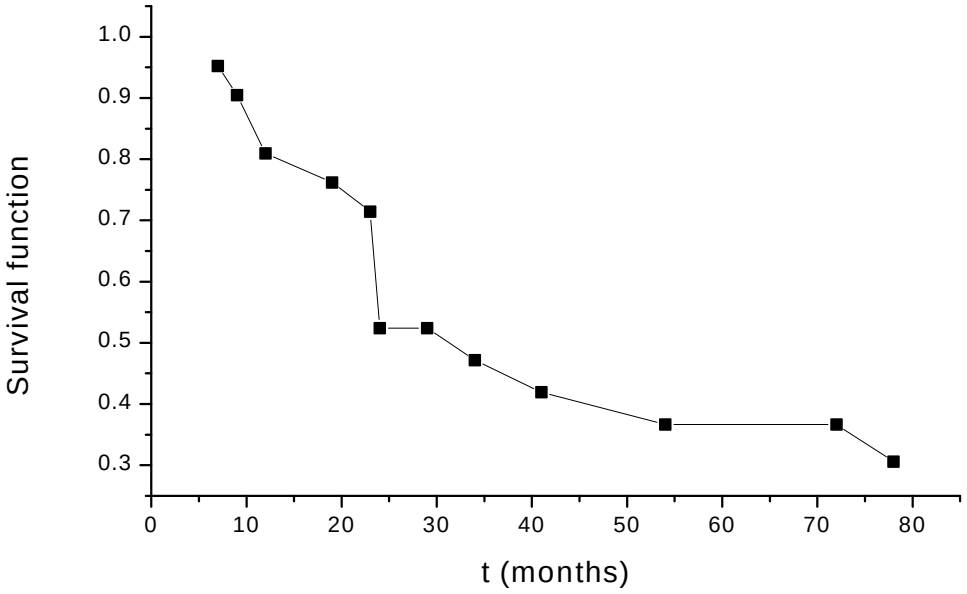


Fig1. Survival function using Kaplan-Meier method.

Table2. Estimating survival function using the advanced algorithm

Patient, rank	Survival time, mo	$w_{(i)}$	Mass at Start	Mass After First Redistribution	Mass After Second Redistribution	$\hat{S}(y_{(i)})$
1	7	0.047619	0.047619	0.047619	0.047619	0.952381
2	9	0.047619	0.047619	0.047619	0.047619	0.904762
3	12	0.047619	0.047619	0.047619	0.047619	0.809524
4	12	0.047619	0.047619	0.047619	0.047619	0.809524
5	23	0.047619	0.047619	0.047619	0.047619	0.761905
6	24	0.047619	0.047619	0.047619	0.047619	0.571429
7	24	0.047619	0.047619	0.047619	0.047619	0.571429
8	24	0.047619	0.047619	0.047619	0.047619	0.571429
9	24	0.047619	0.047619	0.047619	0.047619	0.571429
10	29+	0	0.047619	0	0	0.571429
11	34	0.051948	0.047619	0.051948	0.051948	0.519481
12	41	0.051948	0.047619	0.051948	0.051948	0.467533
13	54	0.051948	0.047619	0.051948	0.051948	0.415585
14	72+	0	0.047619	0.047619	0	0.415585
15	78	0.059369	0.047619	0.052566	0.059369	0.356215
16	80+	0.059369	0.047619	0.052566	0.059369	
17	83+	0.059369	0.047619	0.052566	0.059369	
18	92+	0.059369	0.047619	0.052566	0.059369	
19	139+	0.059369	0.047619	0.052566	0.059369	
20	139+	0.059369	0.047619	0.052566	0.059369	
21	139+	0.059369	0.047619	0.052566	0.059369	

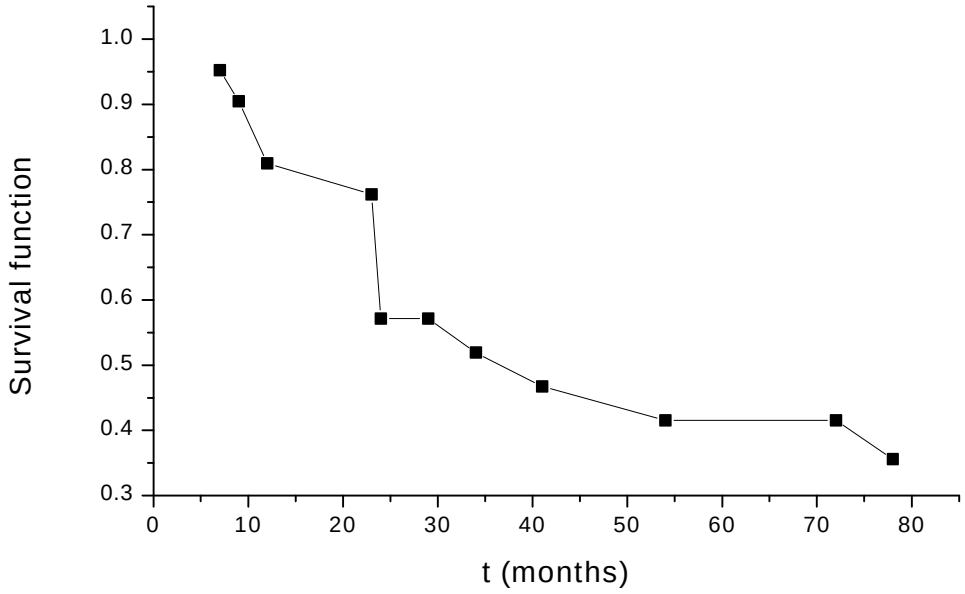


Fig2. Survival function using the advanced algorithm.

VI. Summary and conclusion

Until now, we have shown the results of a sample calculation. Table.1 is calculated using the Kaplan-Meier method. And Table.2 is calculated using the advanced algorithm. From the above result, it is possible to compare the results of both the Kaplan-Meier method and the advanced algorithm. Therefore, we know that the PL-estimator is available and agrees well with the results of the advanced algorithm. That is to say, the shape of the two graphs is similar. This means that the PL-estimator reflects the censoring information in it' s own ways. But if we think more strictly, it has overestimation near the last point. For example, the comparison of the two methods shows that PL-estimator weighs too much at t=78(months), so it' s survival function are less than the latter. However, the assumption of Kaplan-Meier method is good in the sense of the probability theory. In other words, if the tendency of the survival rate is same, it exactly corrects. But in the real phenomena, the pattern of the observations does not present the same tendency. Therefore, this model is applicable to many fields such as a semi-parametric model or a non-parametric model. Using this model, we will get more precise results. Of course, you may use another models as a mass function according to the purposes such as the reduced sample method or actuarial method. But adding to this algorithm, you will get a more exact result.

References

1. Chiang, *Stochastic Processes in Biostatistics* (1968), chapter 9.
2. Leiderman et al., Nature. (1973).
3. Berkson and Gage, *Proc. Staff Meet*. Mayo Clin. (1950).
4. Cutler and Ederer, J. chronic Dis. (1958).
5. Breslow and Crowley, Ann. Stat. (1974).
6. Thomas and Grunkemeier, JASA. (1975).
7. Kaplan and Meier, JASA. (1958).
8. Embury et al. , West. J. Med.(1977).
9. Peterson, JASA. (1977).
10. Efron, *Proc. Fifth Berkeley Symp. IV*. (1967).
11. Rupert G. Miller, Jr. Gail Gong. Alvaro Munoz. *Survival Analysis*. Stanford University. (1981).
12. David G, Kleinbaum. *Survival analysis: a self-learning text*. Springer-verlag New York, Inc. (1995).
13. Eugene K. Harris, Adelin Albert. *Survivorship analysis for clinical studies*. Marcel Dekker, Inc. (1991).