Comparison of the Cumulative-Hazard and Kaplan-Meier Estimators of the Survivor Function

G. A. Bohoris
University of Birmingham, Birmingham

Key Words — **Survivor function, Kaplan-Meier, product limit, cumulative hazard, non-parametric reliability estimate, incomplete data, censored data.**

Reader Aids —
General purpose: Compare two non-parametric estimators
Special math needed for explanations: Probability & statistics
Special math needed to use results: Same
Results useful to: Reliability analysts and theoreticians

Abstract — **The survival-probability estimates calculated from the *cumulative hazard* approach are proved to be larger than the ones calculated from the *Kaplan-Meier* (product limit) approach.**

INTRODUCTION

When both the product-limit (Kaplan-Meier) and the cumulative-hazard estimates of the Sf are calculated for the same data set, the survival probabilities obtained using the cumulative-hazard approach are generally slightly larger than the ones obtained using the product-limit (Kaplan-Meier) approach. A typical example is the data set in table 1.

This heuristic inter-relation can not be specific to particular data sets; thus this paper shows that it holds consistently. More specifically, this paper proves this empirical relation of a regular deviation between the two main non-parametric reliability estimators.

Assumptions

1. The available multiply-censored data set is a random sample of the population; the sample consists of several i.i.d. lifetimes.

2. There are no failures at the initial time (t_0).

Notation

n	number of i.i.d. lifetimes
k	number of distinct i.i.d. lifetimes
t_0	0 (initial time)
t_j	ordered, distinct lifetimes j, $j=1,\ldots,k$
t_{k+1}	∞
d_j	number of failures at t_j; $d_0 \equiv 0$, $d_j \geq 1$, for $j \geq 1$
e_j	number of right-censored observations in $[t_j, t_{j+1})$
n_j	number of items *at risk of failure* at t_j, $j=0,\ldots,k$ $n_j = \sum_{l=j}^{k} (d_l+e_l)$, $n_0 = n$
x_j	d_j/n_j
$R(t)$	survivor (reliability) function
$h(t)$, $H(t)$	hazard rate function, cumulative hazard function
PL	product limit (Kaplan-Meier)
CH	cumulative hazard
$\hat{R}_{XX}(t_j)$	'XX' estimate of the reliability function, XX = CH or PL
PLE, CHE	[PL, CH] estimate.

Other, standard notation is given in "Information for Readers & Authors" at the rear of each issue.

2. REVIEW OF RELIABILITY ESTIMATORS

In reliability practice, very often the description and study of lifetime data are facilitated by estimating specific Cdf's. Some very efficient modelling techniques exist for this purpose; in particular, several non-parametric inference procedures have been proposed. A limited & sparse selection of these is found in a small number of reliability related papers and textbooks. A good summary of some important non-parametric methods is in [1].

This paper discusses only the product-limit (Kaplan-Meier) and cumulative-hazard estimates because they apply to all types of censored data and they provide the most accurate estimates. Therefore, and quite understandably, they dominate the non-parametric estimation in survival analysis.

2.1 Product-Limit Estimator

A very popular estimator of the Sf, particularly in medical statistics, is the PLE due to Kaplan-Meier [2]. PLE has played a central role in the analysis of clinical survival data and has been the standard procedure for estimating the Sf in most biomedical and in some of the statistical (*eg*, SAS) computer programs. The PLE is a key quantity in several more complicated survival analysis models like proportional-hazards and some of the 2-sample tests for censored data [3].

$$\hat{R}_{PL}(t_j) \equiv \prod_{l=1}^{j} (n_l-d_l)/n_l = \prod_{l=1}^{j} (1-x_l). \qquad (1)$$

The properties of the PLE have been studied by many people. The PLE is a step function, often referred to as the Kaplan-Meier estimate, and is the non-parametric maximum likelihood estimator of the Sf [1]. The PLE is essentially the same as the life-table estimate [4 - 6]. The only differences between the two are:

0018-9529/94/$4.00 ©1994 IEEE

TABLE 1
Estimates of the Reliability Function

i	t_i	j	t_j	n_j	d_j	e_j	$\hat{h}(t_j)$	$\hat{H}(t_j)$	$\hat{R}_{CH}(t_j)$	$\hat{R}_{PL}(t_j)$
1	69	1	69	21	1	0	0.048	0.048	0.953	0.952
2	176	2	176	20	1	1	0.050	0.098	0.907	0.905
3	196c	–	–	–	–	–	–	–	–	–
4	208	3	208	18	1	0	0.056	0.153	0.858	0.854
5	215	4	215	17	1	0	0.059	0.212	0.809	0.804
6	233	5	233	16	1	0	0.063	0.274	0.760	0.754
7	289	6	289	15	1	0	0.067	0.341	0.711	0.704
8	300	7	300	14	1	0	0.071	0.413	0.662	0.653
9	384	8	384	13	1	0	0.077	0.490	0.613	0.603
10	390	9	390	12	1	1	0.083	0.573	0.564	0.553
11	393c	–	–	–	–	–	–	–	–	–
12	401	10	401	10	1	0	0.100	0.673	0.510	0.498
13	452	11	452	9	1	0	0.111	0.784	0.457	0.442
14	567	12	567	8	1	2	0.125	0.909	0.403	0.387
15	617c	–	–	–	–	–	–	–	–	–
16	718c	–	–	–	–	–	–	–	–	–
17	782	13	782	5	1	0	0.200	1.109	0.330	0.310
18	783	14	783	4	1	0	0.250	1.359	0.257	0.232
19	806	15	806	3	1	2	0.333	1.692	0.184	0.155
20	1000c	–	–	–	–	–	–	–	–	–
21	1022c	–	–	–	–	–	–	–	–	–

c denotes: censored observation

- PLE is always based on individual survival times, while in the Life-Table method, survival data are often grouped.
- For the life-table estimate, the denominator of (1) would be $(n_j - \frac{1}{2}e_j)$ instead of n_j.[1]

Although many people believe that the life-table approach is probably more accurate, this paper uses (1) to calculate the PLE.

2.2 Cumulative-Hazard Estimator

The CHE, proposed by Nelson [7, 8], estimates the Sf through the estimation of the hazard rate and cumulative hazard functions:

$$\hat{h}(t_j) = d_j/n_j = x_j, \tag{2a}$$

$$\hat{H}(t_j) = \sum_{l=1}^{j} \hat{h}(t_l) = \sum_{l=1}^{j} x_l. \tag{2b}$$

Calculation of the reliability is then a direct application of the relationship:

$$\hat{R}_{CH}(t_j) \equiv \exp\left(-\sum_{l=1}^{j} x_l\right)$$

$$= \prod_{l=1}^{j} \exp(-x_l), \text{ for all } j \in \{1, 2, \ldots, k\}. \tag{3}$$

[1]The reason for the reduced divisor is that, within the biomedical field, it is assumed that censored observations are uniformly distributed within each class interval; hence, censored items are considered to have been at risk on average for only half of the interval.

CHE is the method mainly used by the engineering world in analyzing multiply censored data. Its applications in the medical field are limited and only a few authors have considered this alternative approach [9] in studying biomedical data. The CHE is larger than the PLE.

3. EXAMPLE

Given

1. The multiply-censored data-set in table 1, column 2 [10] is used. Thus $n=21$.
2. Assumptions 1 - 2 are true. ◄

The events in column 2 are already ordered. Column 4 is calculated from column 2, for the $k=15$ distinct times to failure. The remaining part of table 1 presents the calculations involved in estimating the reliability probabilities at each of these distinct failure times.

4. DIFFERENCE IN THE RELIABILITIES

The objective of this paper is to show that the survival probabilities obtained using CHE are larger than those obtained using PLE [11]. A close observation of the last two columns of table 1, verifies (for the data-set) that:

$$\hat{R}_{CH}(t_j) > \hat{R}_{PL}(t_j) \Rightarrow \hat{R}_{CH}(t_j) - \hat{R}_{PL}(t_j) > 0$$

$$\text{for all } j \in \{1, 2, \ldots, k\}. \tag{4}$$

If (4) is to be true for any data set, then the following must be valid:

$$\hat{R}_{CH}(t_j) > \hat{R}_{PL}(t_j) \Rightarrow \prod_{l=1}^{j} \exp(-x_l) > \prod_{l=1}^{j} (1-x_l)$$

$$\text{for all } 1 \in \{1, 2, \ldots, k\}. \quad (5)$$

Since $e^{-x} > (1-x)$ for all $|x| > 0$, (4) is obviously valid for every single failure time in the combined data set, as long as all x_l in (5) are positive. Since, by definition (section 3.1):

$$\{d_j > 0\} \;\&\; \{n_j > 0\} \;\&\; \{n_j = \sum_{l=j}^{k} (d_l + e_l) \Rightarrow n_j \geq d_j\}$$

$$\Rightarrow 0 < x_j = d_j/n_j \leq 1,$$

the difference in (4) is always positive for every conforming data-set.

REFERENCES

[1] W.J. Padgett, D.T. McNicholls, "Nonparametric density estimation from censored data", *Communications in Statistics - Theory and Methods*, vol 13, num 13, 1984, pp 1581-1611.

[2] E.L. Kaplan, P. Meier, "Nonparametric estimation from incomplete observations", *J. American Statistical Assoc.*, vol 53, 1958, pp 457-481.

[3] G.A. Bohoris, D.M. Walley, "Comparative statistical techniques in maintenance management", *IMA J. Mathematics Applied in Business & Industry*, vol 3, 1992 Mar-Apr, pp 241-248.

[4] S.J. Cutler, F. Ederer, Bethesda, "Maximum utilisation of the Life Table method in analyzing survival", *J. Chronic Diseases*, vol 8, num 6, 1958, pp 699-712.

[5] J. Crowley, "Asymptotic normality of a new nonparametric statistic for use in organ transplant studies", *J. American Statistical Assoc.*, vol 69, num 348, 1974, pp 1006-1011.

[6] E.T. Lee, *Statistical Methods for Survival Data Analysis*, 1980, pp 27-52; Lifetime Learning Publications.

[7] W. Nelson, *How To Analyze Data with Simple Plots*, 1979; American Society for Quality Control.

[8] W. Nelson, *Applied Life Data Analysis*, 1982; John Wiley & Sons.

[9] B. Altshuler, "Theory for the measurement of competing risks in animal experiments", *Mathematical Biosciences*, vol 6, 1970, pp 1-11.

[10] D.W. Newton, "Some pitfalls in reliability data analysis", *Reliability Engineering and System Safety*, vol 34, 1991, pp 7-21.

[11] J.I. Ansell, M.J. Phillips, "Practical problems in the statistical analysis of reliability data", *Applied Statistics*, vol 38, num 2, 1989, pp 205-247.

AUTHOR

Dr. G. A. Bohoris; School of Manufacturing & Mechanical Engineering; Univ. of Birmingham; POBox 363; Birmingham B15 2TT GREAT BRITAIN. Internet (e-mail): bohorisg@ibm3090.bham.ac.uk

George A. Bohoris is a lecturer in Operational Research & Statistics and the Director of the MSc course in Quality & Reliability Engineering in the School of Manufacturing & Mechanical Engineering, University of Birmingham. He holds a BSc in Mechanical Engineering, a MSc in Quality & Reliability Engineering, a PhD in Engineering Production, a Masters in Business Administration, and a Diploma in Marketing. He is a member of Operations Research Society, ASQC, and CIM.

Manuscript TR92-067 received 1992 April 15; revised 1993 June 7.

IEEE Log Number 92-12077 ◄TR►

CORRECTION 1993 DECEMBER ISSUE CORRECTION 1993 DECEMBER ISSUE CORRECTION 1993 DECEMBER ISSUE CORRECTION

Condition-Based Predictive Maintenance by Multiple Logistic Function

The following changes/corrections are needed in [1].

p 556, col 2, *Notation*, line 9 ↓

μ_{0i}, μ_{1i} means of x_i from a [normal, failing] system

p 557, col 1, (3b)

$$\beta_0 = \ldots (\mu_{0i} + \mu_{1i})\beta_i. \quad (3b)$$

REFERENCE

[1] K.S. Park, "Condition-based predictive maintenance by multiple logistic function", *IEEE Trans. Reliability*, vol 42, 1993 Dec, pp 556-560. (Original IEEE Log Number 92-07401)

Corrections received 1994 February 25 ◄TR►