GENERALIZED KAPLAN MEIER ESTIMATOR FOR FUZZY SURVIVAL TIMES

ŚLĄSKI
PRZEGLĄD
STATYSTYCZNY
Nr 13(19)

Muhammad Shafiq

Department of Economics
Kohat University of Science & Technology, Kohat, Pakistan

Reinhard Viertl

Institute of Statistics and Probability Theory
Vienna University of Technology, Vienna, Austria

ISSN 1644-6739
e-ISSN 2449-9765

DOI: 10.15611/sps.2015.13.01

Summary: Survival analysis can be defined as a set of methods where the response of interest is the time until a specified event occurred. The most common specified event is death and the related time is called survival time or life time in medical sciences. The Kaplan Meier estimator is one of the popular methods for precise survival times. It is natural that life time is of a continuous nature, therefore it is unrealistic to treat life time observations as precise numbers. In [Viertl 2009] it is shown that life time observations are not precise numbers, but more or less fuzzy. In this study a Generalized Kaplan Meier estimator for fuzzy survival time observations is proposed.

Keywords: characterizing function, fuzzy numbers, Kaplan Meier estimator, non-precise data, survival time.

1. Introduction

Statistical modeling for life time data started in the 20th century, and is now known as reliability analysis or survival analysis. Reliability analysis is mainly concerned with the models of life time data obtained from components and systems in engineering sciences, and survival analysis models are mainly concerned with life time data obtained in biological or life sciences.

Life time, survival time, failure time or event time can simply be defined as the waiting time till a specified event occurs. The event may be death in life science, failure in engineering sciences, divorce in sociology, change of residence in demography, and so on.

Survival analysis techniques are mainly concerned with predicting the probability of response, probability of survival, mean life time, and comparing survival functions [Deshpande and Purhit 2005].

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 13(19)

2. Survival Function

The survival function is conventionally denoted by $S(\cdot)$, which is defined as:

$$S(t) = Pr(T > t) \quad \forall\, t \geq 0.$$

Where t is some specified time, T is the stochastic quantity describing time of death, and "Pr" stands for probability. This function gives the probability that the unit will survive time t or we can say that the event will occur after time t.

For the survival function it is usually assumed that $S(0) = 1$, and $\lim_{t \to \infty} S(t) = 0$ [Lee and Wang 2003].

3. Kaplan Meier Estimator

Let $0 \leq t_1 \leq t_2 \leq t_3 \leq \ldots \leq t_n$ be n precise life times from a given population, and n_i be the number of observations "at risk" at time t_i, and d_i the number of deaths at time t_i. If d_i denotes the number of

Table 1. Kaplan Meier Survival probabilities

Time	d_i	n_i	$1 - \frac{d_i}{n_i}$	$S(t)$
0	0	5	1	1
10	1	5	0.8	0.8
20	1	4	0.75	0.6
30	1	3	0.67	0.402
40	1	2	0.5	0.201
50	1	1	0	0

Source: own elaboration.

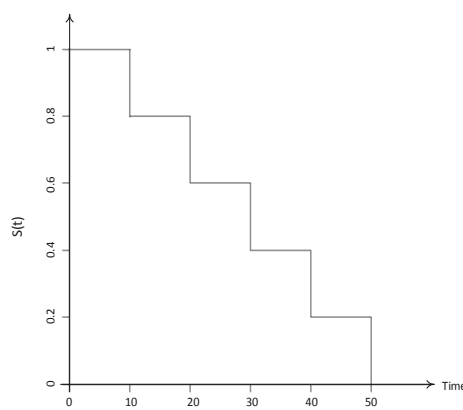

Figure 1. Kaplan Meier Survival Curve

Source: own elaboration.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 13(19)

deaths at time t_i, frequently it is either 0 or 1, but tied survival times are possible. In that case d_i may be greater than 1. The Kaplan Meier estimate can be expressed as:

$$S(t) = \prod_{t_i \le t}\left(1 - \frac{d_i}{n_i}\right) \; \forall\, t \ge 0$$ [Kaplan and Meier 1958].

For example, if we have five precise complete life time observations, i.e. 10, 20, 30, 40, 50, then the Kaplan Meier survival probabilities and survival curve are given in Table 1 and Figure 1 respectively.

4. Fuzzy Information

Standard statistical procedures like estimation of parameters and testing of hypotheses are based on precise numbers. It looks unrealistic to represent continuous real variables in the form of precise numbers or vectors because exact measurements of real continuous variables are not possible, they are more or less fuzzy. Some books and research papers have already been written dealing with fuzzy observations like [Klir and Yuan 1995; Lee 2005; Viertl and Hareter 2006; Huang et al. 2006; Wu 2009].

Survival time is a non-negative valued variable, and it is already shown in [Viertl 2009] that life time observations are not precise numbers but more or less fuzzy. Therefore dealing with time analysis instead of classical statistical tools, fuzzy numbers approaches are more suitable and realistic.

For fuzzy life time observations, a Generalized Kaplan Meier estimator is proposed in this paper.

5. Fuzzy Numbers

Let t^* be a fuzzy observation with a so-called *characterizing function* $\xi(\cdot)$, which is a function of one real variable obeying the following:

1. $\xi : \mathbb{R} \to [0;1]$.
2. For all $\delta \in (0;1]$ the so-called δ-cut $C_\delta(t^*) := \{t \in \mathbb{R} : \xi(t) \ge \delta\}$ is a finite union of compact intervals $[a_{\delta,j}\,;\, b_{\delta,j}]$, i.e $C_\delta(t^*) = \bigcup_{j=1}^{k_\delta}[a_{\delta,j}\,;\, b_{\delta,j}] \neq \phi$.
3. The support of $\xi(\cdot)$ is bounded, i.e. $\text{supp}[\xi(\cdot)] := \{t \in \mathbb{R} : \xi(t) > 0\} \subseteq [a\,;\, b]$.

The set of all fuzzy numbers is denoted by $\mathcal{F}(\mathbb{R})$.

If all δ-cuts of a fuzzy number are non-empty closed bounded intervals, the corresponding fuzzy number is called a fuzzy interval.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 13(19)

6. Fuzzy Vectors

A n-dimensional fuzzy vector $\underline{t}^*$ is determined by its so-called vector characterizing function $\zeta(.,\dots,.)$ which is a real function of n real variables $t_1, t_2, \dots, t_n$ obeying the following three conditions:

1. $\zeta : \mathbb{R}^n \to [0\,;\,1]$.

2. For all $\delta \in (0\,;\,1]$ the so-called δ-cut $C_\delta[\underline{t}^*] := \{\underline{t} \in \mathbb{R}^n : \zeta(\underline{t}) \geq \delta\}$ is non-empty, bounded, and a finite union of simply connected and closed sets.

3. The support of $\zeta(.,\dots,.)$ defined by supp $[\zeta(.,\dots,.)] := \{\underline{t} \in \mathbb{R} : \zeta(\underline{t}) > 0\}$ is a bounded set.

The set of all n-dimensional fuzzy vectors is denoted by $\mathcal{F}(\mathbb{R}^n)$.

Let T be a stochastic quantity with observation space $M_T \subseteq [0\,;\,\infty)$, and a sample of size n i.e $t_1, t_2, \dots, t_n$ is considered from it. Each t_i is an element of the observation space and $(t_1, t_2, \dots, t_n)$ is an element of the so-called sample space M_T^n which is the Cartesian product of n copies of M_T, i.e. $M_T^n := M_T \times M_T \times \dots \times M_T$.

While on the other hand in the case of fuzzy observations, each fuzzy observation $t_i^*, i = 1(1)n$ with characterizing function $\xi_i(\cdot)$ is a fuzzy element of M_T then $(t_1^*, t_2^*, \dots, t_n^*)$ is not a fuzzy element of M_T^n. In order to generalize the Kaplan Meier estimator, the aggregation of the fuzzy observations into a fuzzy element of the sample space is necessary.

To construct a fuzzy element (fuzzy vector) of the sample space M_T^n usually the so-called minimum t-norm is used.

For the vector-characterizing function of the combined fuzzy sample $\underline{t}^* := (t_1, t_2, \dots, t_n)^*$ applying the minimum t-norm, i.e. $\zeta(t_1, t_2, \dots, t_n) = \min\{\xi_1(t_1), \xi_2(t_2), \dots, \xi_n(t_n)\}$ $\forall\, (t_1, t_2, \dots, t_n) \in \mathbb{R}^n$, a fuzzy element of $M_T^n \subseteq \mathbb{R}^n$ is obtained, whose vector characterizing function is $\zeta(.,\dots,.)$.

Remark: The δ-cuts of the combined fuzzy sample will be obtained as the Cartesian products of the δ-cuts of respective fuzzy observations, i.e.

$$C_\delta[\zeta(.,\dots,.)] = \times_{i=1}^n C_\delta[\xi_i(\cdot)] \quad \forall\, \delta \in (0\,;\,1] \text{ [Viertl 2011].}$$

Extension Principle:

This is the generalization of an arbitrary function $g: M \to N$ for fuzzy argument value a^* in M. Let a^* be a fuzzy element of M with membership function $\mu: M \to [0\,;1]$, then the fuzzy value $y^* = g(a^*)$ is the fuzzy element y^* in N whose membership function $\vartheta(\cdot)$ is defined by

ŚLĄSKI PRZEGLĄD STATYSTYCZNY
Nr 13(19)

$$\vartheta(y) := \left\{ \begin{array}{l} sup\{\, \mu(a) : a \in M, g(a) = y\} \;\; \text{if } \exists a: g(a) = y \\ 0 \text{ if } \nexists a: g(a) = y \end{array} \right\} \forall\, y \in N$$

[Klir and Yuan 1995].

Theorem: For a continuous function f: $\mathbb{R} \rightarrow \mathbb{R}$ and for a fuzzy interval t^* the following holds true:

$$C_\delta[f(t^*)] = \left[\min f(t)_{t \in C_\delta(t^*)} \; ; \; \max f(t)_{t \in C_\delta(t^*)} \right] \forall\, \delta \in (0\,;1]$$

where $\min\limits_{t \in C_\delta(t^*)} f(t)$, $t \geq 0$ determines the lower end of the δ-cut, and $\max f(t)_{t \in C_\delta(t^*)}$ determines the upper end of the δ-cut of the fuzzy value $f(t^*)$ [Viertl 2011].

Examples of characterizing functions of fuzzy life times are depicted in Figure 2.

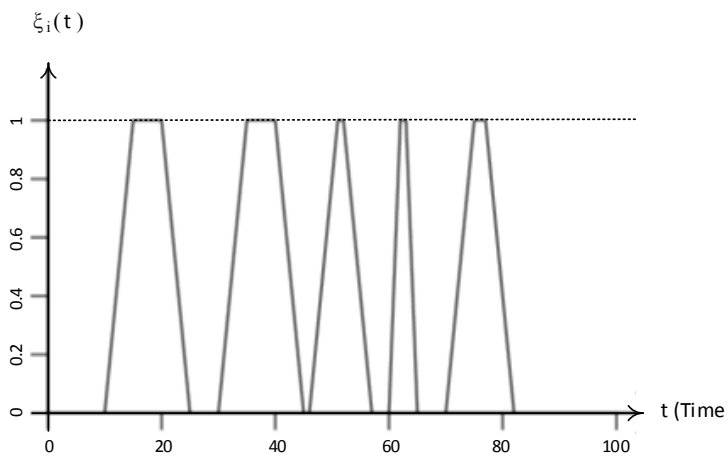

Figure 2. Fuzzy sample

Source: own elaboration.

For the generalized Kaplan Meier estimator $\hat{S}^*(t)$, upper and lower δ-level curves are obtained with the help of δ-cuts from the above mentioned theorem in the following way:

$$C_\delta(S^*(t)) = \left[\min_{\underline{t} \in \times_{i=1}^{n} C_\delta(t_i^*)} S(\underline{t}) \;\; ; \;\; \max_{\underline{t} \in \times_{i=1}^{n} C_\delta(t_i^*)} S(\underline{t}) \right]$$

with $\underline{t} = (t_1, t_2, \ldots, t_n) \in [\,0\,;\,\infty)^n \;\forall\, \delta \in (0\,;1]$.

Where $\min_{\underline{t}\in\times_{i=1}^{n} C_\delta(t_i^*)} S(\underline{t})$ is the lower end of the δ-cut which defines the lower δ-level curve and $\max_{\underline{t}\in\times_{i=1}^{n} C_\delta(t_i^*)} S(\underline{t})$ is the upper end of the δ-cut which defines the upper δ-level curve.

The above mathematical calculations are made through the following algorithm:

1. The values for δ are taken from 0 to 1 with an increment $\Delta \in (0\,;1)$.

2. For a given value of δ calculate the δ-cut of the fuzzy combined sample $\underline{t}^*$.

3. Taking minimum and maximum from the δ-cuts to generate hypothetical classical samples.

4. The Kaplan Meier survival probabilities are calculated and the Kaplan Meier survival curves are drawn for fixed δ-level.

5. Steps 2-4 are performed for each $\delta = 0\ (\Delta)\ 1$.

Example: For the fuzzy life time data given in Figure 2 the lower δ-level curves and upper δ-level curves of the generalized Kaplan Meier estimator are calculated for $\delta = 0, 0.2, 0.4, 0.6, 0.8, 1$. They are depicted in Figure 3.

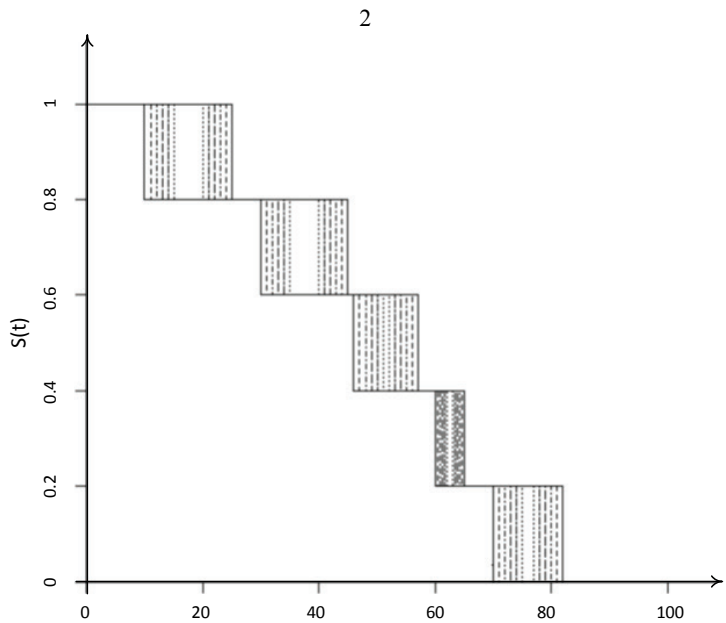

Figure 3. Generalized Kaplan Meier estimator for the fuzzy sample from Figure

Source: own elaboration.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 13(19)

The generalized estimated survival curve (generalized Kaplan Meier estimator) is depicted in Figure 3.

The functions are the lower and upper δ-level curves defined by the considered δ-levels.

7. Conclusion

The precise measurement of a continuous variable is impossible. Survival time observations are usually assumed as precise numbers. However, these observations are of a continuous nature and therefore survival time observations are more or less fuzzy. Consequently, fuzzy numbers are more suitable and realistic to describe real survival times. In the given study, the classical Kaplan Meier estimator based on precise observations is generalized for fuzzy life time observations.

References

Deshpande J.V., Purhit S.G., *Life Time Data: Statistical Models and Methods*, World Scientific Publishing, Singapore 2005.

Huang H.-Z., Zuo M.J., Sun, Z.-Q., *Bayesian reliability analysis for fuzzy lifetime data*, "Fuzzy Sets and Systems" 2006, Vol. 157(12), pp. 1674–1686.

Kaplan E.L., Meier P., *Nonparametric estimation from incomplete observations*, "Journal of the American Statistical Association" 1958, Vol. 53(282), pp. 457–481.

Klir G., Yuan B., *Fuzzy Sets and Fuzzy Logic – Theory and Applications*, Upper Saddle River: Prentice Hall, 1995.

Lee E.T., Wang J.W, *Statistical Methods for Survival Data Analysis*, Wiley, New Yersey 2003.

Lee K.H., *First Course on Fuzzy Theory and Applications*, Springer, Heidelberg 2005.

Viertl R., *On reliability estimation based on fuzzy lifetime data*, "Journal of Statistical Planning and Inference" 2009, Vol. 139(5), pp. 1750–1755.

Viertl R., *Statistical Methods for Fuzzy Data*, Wiley, Chichester 2011.

Viertl R., Hareter D., *Beschreibung und Analyse unscharfer Information – Statistische Methoden für unscharfe Daten*, Springer, Wien 2006.

Wu H.-C., Statistical confidence intervals for fuzzy data, "Expert Systems with Applications" 2009, Vol. 36(2), pp. 2670–2676.

ŚLĄSKI
PRZEGLĄD
STATYSTYCZNY

Nr 13(19)

UOGÓLNIONY ESTYMATOR KAPLANA MEIERA DLA ROZMYTEGO CZASU PRZEŻYCIA

Streszczenie: Analiza przeżycia definiowana jest jako zestaw metod badawczych służących do określenia czasu zajścia pewnego wyspecyfikowanego zdarzenia (losowego). W szczególności zdarzeniem takim jest śmierć człowieka. Do estymacji czasu przeżycia stosowana jest metoda Kaplana-Mayera. W 2009 r. Viertl wykazał, że czasu życia nie można określić precyzyjnie i zaproponował, by stosować liczby rozmyte. W niniejszym artykule zaproponowano uogólniony estymator Kaplana-Mayera wykorzystujący obserwacje rozmyte.

Słowa kluczowe: liczby rozmyte, estymatory Kaplana Meiera, dane nieprecyzyjne, czas przeżycia.