A Fuzzy Approach to Kaplan-Meier Estimator

Jaisankar R.[1], Parvatha Varshini K. S.[2], Siva M.[3]
[1] Professor, Department of Statistics, Bharathiar University, Coimbatore-641046.
[2, 3] Research Scholar, Department of Statistics, Bharathiar University, Coimbatore-641046.

Article Info
Page Number: 1107 – 1114
Publication Issue:
Vol. 71 No. 3s2 (2022)

Article History
Article Received: *28 April 2022*
Revised: *15 May 2022*
Accepted: *20 June 2022*
Publication: *21 July 2022*

Abstract

The application of the concept of fuzziness to statistical data is always meaningful because of the fact that data are imprecise in general. In survival analysis, the lifetimes which are continuous in nature or not precise numbers, and hence they are more or less fuzzy (Viertl (2009)). In this article, a procedure for the Kaplan-Meier estimator has been proposed by considering fuzzy survival times. The results obtained are substantiated through examples.

Key Words: Survival times, The Kaplan-Meier Estimator, Characteristic Functions, and Fuzzy Sets.

Introduction

The primary aim of survival analysis is to estimate the average survival time, the time starting from an initial epoch of a subject till the occurrence of a specified event. Though it is appeared to be simple, the main problem is 'Censoring' which is quite natural in survival data. Censored observations may give partial information but if they are true, including them in the analysis may be helpful to evolve better results. Apart from the estimation of average survival time, it is necessary to compare the survival experiences of two or more groups of subjects involved in the study by means of their corresponding hazard and survival curves. If the groups are significantly different it is possible to detect or identify the potential factor which is responsible for that difference. The estimator proposed by Kaplan-Meier (1958), is one of the most widely used familiar non-parametric methods in which the average survival time is estimated using conditional probabilities.

It is to be noted that, sample observations collected for statistical analyses of any nature, whether to estimate parameters or to test the hypotheses, are often taken as precise numbers. However, it is unreasonable to describe genuine continuous variables as precise integers or vectors since accurate measurements are not possible in many practical situations.

Hence, in survival analysis, life spans are continuous and may be inaccurate, and they are more or less ambiguous. A potential arena for dealing with such inaccuracies which may be prevailing in the data is the Fuzzy theory. Hence, involving fuzzy approaches to traditional statistical procedures may be more relevant and realistic, in particular, for analyzing survival durations. In this paper, a Fuzzy approach to Kaplan-Meier is proposed.

The Kaplan-Meier Estimator

Edward L. Kaplan and Paul Meier (1958) collaborated on a fundamental article on dealing with partial observations in lifetime data and observed the curves for survival and hazard. Since then, they become a common means of dealing with varying survival times with censored data. Apart from the usual independency, the Kaplan-Meier's procedure assumes random censoring and the independence of the censoring times and are independent of survival times with the same distribution, with the constant probability of survival. The censored are considered to have survived until after the time t at which they were last observed alive. The Kaplan-Meier estimator of the survivor function at times, for $t_{(k)} \leq t < t_{(k+1)}$ is given by,

$$\hat{S}(t) = \prod_{j=1}^{k} \frac{(n_j - d_j)}{n_j}$$

where, $t_1, t_2, t_3, \ldots$ denote the actual times of death of the n individuals in the cohort, $d_{1,} d_{2,} d_{3,}$ ...denote the number of deaths that occur at each of these times, and $n_1, n_2, n_{3,} \ldots$ be the corresponding number of patients remaining in the cohort.

Now a days the aspects of Fuzzy algebra/logic are applied to survival analysis, for dealing with ambiguous failure times. Grzegorzewski (2002) used fuzzy numbers, and computed the mean time to failure which is also observed as a fuzzy number and determined the corresponding confidence interval. For lifetime data collected as fuzzy numbers, lifetime data is usually more or less fuzzy; the statistical estimation of the reliability characterizing function was extended by Viertl (1995). Then Kian L Pokorny (2003) used the fuzzy logic to propose a fuzzy product limit estimator which could be used even when heavy censoring and has shown that they are more reliable than the classical fuzzy estimator.

As the survival times are continuous in nature there may certainly be a kind of fuzziness associated with them. Viertl (2009) shows that the lifetime observations are not precise numbers, but more or less fuzzy. Viertl (2015) have taken empirical probabilities for addressing fuzziness to the survival probabilities. However, as the frequency of occurrence of events to the lifetimes will be lesser, and hence taking such a measure for representing the fuzziness will no more be appropriate. The prime objective of survival analysis is to find a median or mean value which is measure of central tendency and which may not be thought of a crisp value. The survival times are mainly fuzzy and the problem is seeking an appropriate value which maybe expected based on the fuzzified observations of survival times.

Hence, a new methodology is proposed in this article which takes in account of both survival times and survival probabilities. In this methodology, the survival times are adjusted for survival proportions in order to obtain the interval fuzzy survival times.

The present work assumes that the survival times are fuzzy and hence some modifications are carried out in Kaplan-Meier's estimators accordingly, which utilizes the following fuzzy concepts.

Preliminaries

Fuzzy Sets

Let X is a nonempty set. A fuzzy set B in X is characterized by its membership function $\mu_B: X \to [0,1]$ and is interpreted as the degree of membership of element x in the fuzzy set B for each $x \in X$. It is clear that B is completely determined by the set of tuples.

$$B = \{(x, \mu_B(x)) | x \in X\}$$

The family of all fuzzy (sub) sets in X is denoted by $F(X)$. Fuzzy subsets of the real line are called fuzzy quantities.

Fuzzy Numbers

Let $\tilde{t}$ be a fuzzy number with their characterizing function $\xi(.)$, which is a function of a real variable t obeying the following:

$$\xi(.): \mathbb{R} \to [0; 1].$$

1. Let $C_\delta(\tilde{t}) := \{t \in \mathbb{R}: \xi(t) \geq \delta\}\ \forall \delta \epsilon (0; 1]$ is a finite union of $\delta -$ cuts with the compact intervals. That is $C_\delta(\tilde{t}) = \bigcup_{j=1}^{k_\delta} [a_{\delta,j}; b_{\delta,j}] \neq \emptyset$.
2. The support $\sup[\xi(.)] := \{t \in \mathbb{R} : \xi(t) > 0\} \subseteq [a; b]$ of the characterizing the function $\xi(.)$ is bounded.

Let $\mathcal{F}(\mathbb{R})$ denotes the set of all fuzzy numbers. It is to be noted that the fuzzy numbers can be represented by the finite numbers of δ-cuts and $\xi(t) = \max\{\delta . I_{C_\delta(\tilde{t})}(t): \delta \epsilon [0; 1]\}\ \forall t \in \mathbb{R}$. If all δ -cuts of a fuzzy number are non-empty closed bounded intervals, then the corresponding fuzzy number is called a fuzzy interval.

Let T denotes a stochastic quantity with observational space $S_T \subseteq [0; \infty)$, and let $t_1, t_2, \dots, t_n$ be a sample of size n considered from it. Each t_i is an element of the observational space. Let $S_T^n := S_T \times S_T \times \dots \times S_T$ is the Cartesian product of n copies of S_T so that $(t_1, t_2, \dots, t_n)$ is its element. In the case of fuzzy observations, for each fuzzy observation $\tilde{t}_i, i = 1,2 \dots n$ with their characterizing function $\xi_i(.)$ is a fuzzy element of S_T and $(\tilde{t}_1, \tilde{t}_2, \tilde{t}_3, \dots, \tilde{t}_n)$ is not a fuzzy element of S_T^n. Here the aggregation of fuzzy observations into a fuzzy element of the sample space is taken up.

Characteristic Function – Triangular fuzzy Numbers

Let $(\tilde{t}_1, \tilde{t}_2, \tilde{t}_3, \dots, \tilde{t}_n)$ be n fuzzy observations with their corresponding characteristic functions $\xi_i(.)$ with the implications of a finite number of δ cuts. The Triangular Fuzzy Number (TFN) is the most popular of the numerous shapes of fuzzy numbers, which can be used them δ cuts in the present scenario as the objective can be viewed as the measure of central tendency. A Triangular fuzzy Number is represented by three points: $B = (b_1, b_2, b_3)$ where b_1, b_2, b_3 are real numbers with $b_1 < b_2 < b_3$ and the corresponding characteristic functions are given by,

$$\xi_{(B)}(x) = \begin{cases} 0, & x < b_1 \\ \dfrac{x - b_1}{b_2 - b_1}, & b_1 \le x \le b_2 \\ \dfrac{b_3 - x}{b_3 - b_2}, & b_2 \le x \le b_3 \\ 0, & x > b_3 \end{cases}$$

Methodology

The estimator proposed by Kaplan-Meier is based on the conditional probability, obtained on the basis of the number of deaths and people at risk, which are discrete values and hence it may not be possible to consider them as fuzzy. However, survival times are continuous in nature and hence treating them as precise may be unrealistic. In order to make them fuzzy, the individual survival times are made as fuzzy intervals. A fuzzification factor considered for each interval are proportional to the corresponding survival probability and the survival times are made as fuzzy intervals.

Let $t_1, t_2, \dots, t_n$ denotes the times at which the event or events occurred as observed. Let η_i represent fuzzification factor t_i, which is proportional to the quantity ξ_i with proportionality constant $(c, 0 \le c < 1)$, determined on the basis of triangular fuzzy membership functions developed based on classical Kaplan-Meier estimator. The survival times are then fuzzified through the fuzzification factor η_i in such a way that $t_i^F = [(t_i - \eta_i), (t_i + \eta_i)]$ where $0 \le \eta_i < 1$. Hence, $\{t_i^F, i = 1,2, \dots n\}$ fuzzy intervals are obtained corresponding to each survival time.

Since, most of the survival data are skewed, the measure of Median is preferred. For finding the median, it is necessary to order the data and even when the values are fuzzy. Hence, a fuzzy-number ranking procedure is required. There are numerous ranking methods suggested by various researchers in the literature, each of which can produce distinct ranking outcomes.

The survival times are ranked by using Lee and Li's (1998) ranking technique for triangular fuzzy numbers. Lee and Li's technique is based on finding the quantity defined by the following equation.

$$\bar{y}_q(\tilde{y}_j) = \frac{y_{jl} + 2y_{jm} + y_{ju}}{4} \quad (1)$$

Where, y_{jl} be the lower triangular fuzzy number, y_{jm} is the middle triangular fuzzy number and y_{ju} is the upper triangular fuzzy number. The fuzzy sample median is found by the method suggested by Nguyen and Wu (2006). Let V be the universe set, and $\{Fy_j = [a_j, b_j], a_j, b_j \in R, j = 1,2, \dots, n\}$ be a sequence of random fuzzy sample on V. Let d_j be the epicenter of the interval value of $[a_j, b_j]$ and l_j be its distance. Then the fuzzy sample median is defined by,

$$F_{(Median)} = (d; q), d = \text{median}\{d_j\}, q = \frac{\text{median } (l_j)}{2} \quad (2)$$

If one needs a crisp value for the fuzzy median, then this quantity may be defuzzified in such a way that the Median of the triangular $\tilde{B} = (a, b, c)$ equals to the centroid $\left(C_{\tilde{B}} = \frac{a+b+c}{3}\right)$ of the triangular $\tilde{B}$ when the triangular is isosceles. The survival curve may be drawn using the strong α −cut of the triangular fuzzy numbers corresponding to each fuzzy survival time intervals.

Illustration

The following example has been taken for the purpose of illustration of the proposed methodology. The survival times (t_i), the number of deaths (d_i) with 12 subjects are given in the following Table 1. Using the classical Kaplan Meier procedure, the survival probabilities $S(t)$ are calculated and the median survival time is 9 Months.

Table 1: Classical Kaplan Meier Estimator

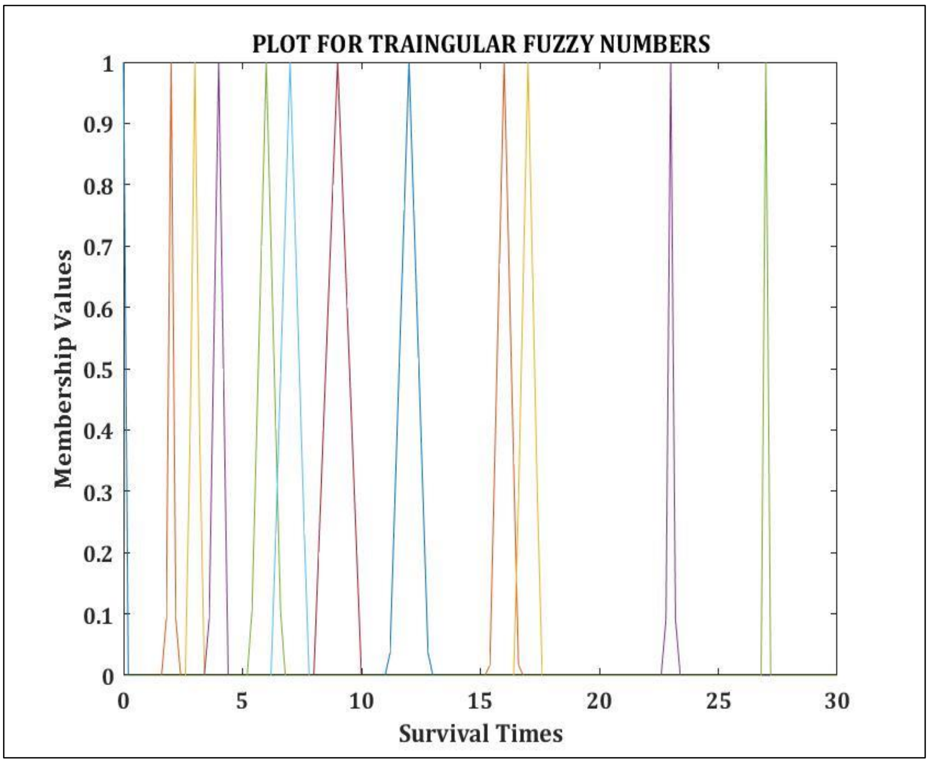
Time (t_i)	**Number of Deaths (d_i)**	**Survival Probabilities $S(t)$**
0	12	0.93
2	11	0.85
3	10	0.79
4	09	0.71
6	08	0.64
7	07	0.57
9	06	0.43
12	05	0.36
16	04	0.29
17	03	0.21
23	02	0.14
27	01	0.07

Now for the fuzzy approach, corresponding to each survival time the triangular fuzzy numbers are obtained as below in Table 2 and the corresponding membership functions are depicted in the Figure 1.

Table 2: Triangular Fuzzy Numbers

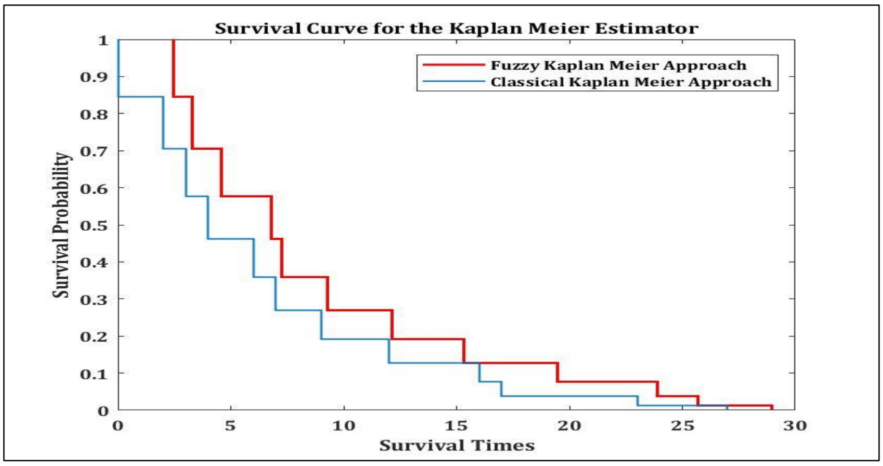
(0,0,0)	(1.78,2,2.22)
(2.67,3,3.33)	(3.56,4,4.44)
(5.33,6,6.67)	(6.22,7,7.78)
(8,9,10)	(11.17,12,12.83)
(15.39,16,16.61)	(16.44,17,17.56)
(22.78,23,23.22)	(27,27,27)

Figure 1: Plot for Triangular Fuzzy Numbers



The following Figure 2 shows the survival curves obtained by the classical Kaplan-Meier method and the proposed Fuzzy Kaplan-Meier procedure. The later one has been drawn based on the alpha cuts of the fuzzified survival times. Note that the pattern of the survival curve depends on the fuzzification factor chosen and the Median Survival time for the Fuzzy Kaplan Meier is 6 Months.

Figure 2: Survival Curve



Conclusion

Statistics as a data science effectively reveal the hidden facts exhibited in the data observed. The non-parametric procedure gives a method preamble to understand the behavior of any disease, which is important in this era of emergence of many epidemics. However, in reality every observation may have certain amount of impreciseness which cannot be defined. This is true in particular, when the data is of continuous in nature. The theory of fuzzy provides a way to incorporate a measure of impreciseness. In this article the survival times are fuzzified and the median survival times and the survival times are obtained accordingly by which maybe expected to be more applicable them the classical approach as it involves fuzzy features. The fuzzification procedure applied here a novel methodology which could be applied further both for theoretical development and practical applications

References

1. Colosimo, E. A., Ferreira, F. F., Oliveira, M. D. & Sousa, C. B. Empirical comparisons between Kaplan-Meier and Nelson-Aalen survival function estimators. Journal of Statistical Computation and Simulation **72**, 299–308 (2002).
2. Denoeux, T., Héì, M., Masson, E. & Hébert, P.-A. Nonparametric Rank-based Statistics and Significance Tests for Fuzzy Data.
3. Dubois, D. & Fu, K. S. Ranking Fuzzy Numbers in the Setting of Possibility Theory. INFORMATION SCIENCES vol. 30 (1983).
4. Dijkman, J. G., van Haeringen, H. & de Lange, S. J. Fuzzy Numbers. JOURNAL OF MATHEMATICAL ANALYSIS AND APPLICATIONS vol. 92 (1983).
5. George J.Klir, Bo Yuan 1995 Fuzzy Sets And Fuzzy Logic Theory and Applications, Prentice Hall PTR.
6. Hung T.Nguyen, Berlin Wu 2006, Fundamentals of Statistics with Fuzzy Data Springer.
7. Kahraman, C. & Sarı, İ. U. Fuzzy central tendency measures. Studies in Fuzziness and Soft Computing **343**, 65–83 (2016).
8. Kaplan, E. L. & Meier, P. Nonparametric Estimation from Incomplete Observations NONPARAMETRIC ESTIMATION FROM INCOMPLETE OBSERVATIONS*. Source: Journal of the American Statistical Association vol. 53 (1958).
9. Musavi, S., Pokorny, K. L., Poorolajal, J. & Mahjub, H. Fuzzy survival analysis of AIDS patients under ten years old in Hamadan-Iran. Journal of Intelligent and Fuzzy Systems **28**, 1385–1392 (2015).
10. Saranya & Karthikeyan. A Comparison study of Kaplan Meier and Nelson-Aalen Methods in Survival Analysis. INTERNATIONAL JOURNAL FOR RESEARCH IN EMERGING SCIENCE AND TECHNOLOGY (2015).
11. Nouby M. Ghazaly, M. M. A. . (2022). A Review on Engine Fault Diagnosis through Vibration Analysis . International Journal on Recent Technologies in Mechanical and Electrical Engineering, 9(2), 01–06. https://doi.org/10.17762/ijrmee.v9i2.364
12. Shafiq, M. & Viertl, R. Generalized Kaplan Meier Estimator for Fuzzy Survival Times. Śląski Przegląd Statystyczny (2015) doi:10.15611/sps.2015.13.01.
13. Shah, S. H., Shafiq, M. & Zaman, Q. Generalized Estimation for Two-Parameter Life Time Distributions Based on Fuzzy Life Times. Mathematical Problems in Engineering **2022**, 1–11 (2022).
14. Agarwal, D. A. . (2022). Advancing Privacy and Security of Internet of Things to Find Integrated Solutions. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 8(2), 05–08. https://doi.org/10.17762/ijfrcsce.v8i2.2067

15. Taheri, S. M. Trends in Fuzzy Statistics. AUSTRIAN JOURNAL OF STATISTICS vol. 32 (2003).
16. Pokorny, K. & Sule, D. EMPIRICAL FUZZY ESTIMATE OF THE SURVIVAL CURVE. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems vol. 12 www.worldscientific.com (2004).
17. Alaria, S. K., A. . Raj, V. Sharma, and V. Kumar. "Simulation and Analysis of Hand Gesture Recognition for Indian Sign Language Using CNN". International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 4, Apr. 2022, pp. 10-14, doi:10.17762/ijritcc.v10i4.5556.
18. A. Chawla, "Phishing website analysis and detection using Machine Learning", Int J Intell Syst Appl Eng, vol. 10, no. 1, pp. 10–16, Mar. 2022.
19. Viertl, R. Univariate statistical analysis with fuzzy data. Computational Statistics and Data Analysis **51**, 133–147 (2006).
20. Ghazaly, N. M. . (2022). Data Catalogue Approaches, Implementation and Adoption: A Study of Purpose of Data Catalogue. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 8(1), 01–04. https://doi.org/10.17762/ijfrcsce.v8i1.2063
21. Viertl, R. On reliability estimation based on fuzzy lifetime data. Journal of Statistical Planning and Inference **139**, 1750–1755 (2009).
22. Viertl, R. Fuzzy Numbers and Non-Precise Data Encyclopedia of Environmetrics, John Wiley & Sons, Ltd (2014)
23. M. . Parhi, A. . Roul, B. Ghosh, and A. Pati, "IOATS: an Intelligent Online Attendance Tracking System based on Facial Recognition and Edge Computing", Int J Intell Syst Appl Eng, vol. 10, no. 2, pp. 252–259, May 2022.
24. Viertl, R. Statistical Methods for Fuzzy Data John Wiley& Sons. (2011)