

The Kaplan Meier Estimate in Survival Analysis

Abstract

Kaplan-Meier is a statistical method used in the analysis of time to event data. Time to event means the time from entry into a study until a particular event, for example onset of illness. This method is very useful in survival analysis as it is used by the researchers to determine and/or analyze the patients or participants who lost to follow up or dropped out of the study, those who developed the disease of interest or survived it. It is also used to compare two groups of subjects such as a control group, the one that is given placebo and the other treatment group that is the one given the genuine drug. The method is not only applicable to the fields of public health, medicine and epidemiology, but it is also useful in other disciplines such as engineering, economics, among others. Most of the studies that use Kaplan Meier estimate are longitudinal in nature like a cohort study. Examples of studies that Kaplan-Meier estimate can be applicable include death times of kidney transplant patients, times to infection for burn patients and times to death for a breast-cancer trial. A fictive data was created concerning the treatment and control groups who were given Drug A and placebo respectively. The participants in each these two groups are ten and they were followed for 2 years (24 months). A survival table and Kaplan-Meier estimate curve were generated from the SPSS software using the fictive data and these were used to analyze the 24 month study.

Keywords: Survival analysis; Kaplan-Meier estimate

Research Article

Volume 5 Issue 2 - 2017

Ilker Etikan*, Sulaiman Abubakar and Rukayya Alkassim

Department of Biostatistics, Near East University Faculty of Medicine, Cyprus

***Corresponding author:** Ilker Etikan, Near East University Faculty of Medicine Department of Biostatistics, Nicosia-TRNC, Cyprus, Email: ietikan@gmail.com

Received: November 24, 2016 | **Published:** February 13, 2017

Introduction

The need for analyzing time to event data arises in a number of applied fields, such as epidemiology, public health and medicine [1]. 'Time to event' simply means the time from entry into a study until a subject has a particular outcome. The study that involves time to event can be a cohort study on a specific number of patients or participants which should be followed for a particular time period. In epidemiology, survival analysis is very important in the analysis of data involving patients/participants that should be followed to determine a particular event. Kaplan Meier estimate is best statistical method used in survival analysis to analyze the data and to make comparison between two groups of participants such as treatment group and control group using the log-rank test for hypothesis testing. In addition to medical disciplines, Kaplan-Meier analyses are also useful to other disciplines such as physics, engineering, economics, demography, among others. Example of Kaplan Meier estimate will be in the cohort study of lung cancer among smokers; here the selected number of smokers will be followed for 20 years. In this study, the Kaplan-Meier estimate will be used to determine or analyze the events, and censoring. Events here mean the development of the disease (the lung cancer) while censored are those who dropped out of the study or those who lost to follow-up. The fraction of smokers surviving the lung cancer will also be calculated using the survival table and Kaplan-Meier Estimate curve. Both the survival table and Kaplan-Meier estimate curve can be generated from the SPSS software or other statistical softwares such as Stata, SAS and R packages.

Material and Methods

A fictive data will be created regarding the two groups of participants. The first group will be the treatment group while the

second group will be the control group. The treatment group is a group that is given Drug A while placebo is given to control group. Each group will consist of ten participants. Tables and Kaplan-Meier estimate curves which will be generated from the SPSS software will be used to analyze the fictive data.

Survival Analysis

The time starting from a specified point to the occurrence of a given event, for example injury is called the survival time and hence, the analysis of group data is referred to the survival analysis [2]. Therefore survival analysis is a statistical technique for analyzing data on the occurrence of events especially in cohort study. Thus, it considers data from randomized clinical trials or cohort study. Clinical trials are controlled experiments which are conducted to compare efficacy and safety among human subjects [3]. Analysis and modeling of 'time-to-event' data is the primary objective of survival analysis. The event can be disappearance of a tumor, time to discharge from health facility/hospital, response to a treatment, death or the development of a disease. An injury, recovery from illness and onset of illness are also referred to events. Examples of an event include Ebola disease for people tested positive after been quarantined for three weeks in Serra Leone and Lassa fever or *Lassa hemorrhagic fever* (LHF) for those who showed its signs after been followed for one week in Maiduguri, Nigeria. The technique of survival analysis is used to estimate and interpret survival, to compare it between groups, and to assess the association or relationship of explanatory variables with survival time. Survival analysis considers time, the time until a particular event of interest occurs.

Survival time are data that measure the time to a certain event such as death, failure, response, relapse, divorce or the

development of a given disease [4]. Survival time can be length of remission, time to disappearance of a tumor, time to death and the time from the start of treatment to the response. Survival time has two important components that must be unambiguously defined: a starting point and an endpoint reached either when the event of interest occurs or when the follow-up time has ended. Survival data may include survival time, response to a given treatment, and patient characteristics related to survival, response and the development of disease. These data can be derived from clinical and epidemiologic studies of humans who have acute or chronic disease. Unlike other statistical methods such as logistic regression, among others, survival analysis considers censoring and time.

Censoring can occur when the patients lost to follow up to the end of the study. Censored data are data that arises when a person's life length is known to happen only in a specified period of time. Possible censoring schemes are said to be right censoring, when the participant is still alive at a specified period of time, left censoring when the participant has experienced the event of interest before the study begin, or where the only information is that the event of interest occurs within a given interval, that is interval censoring. In analysis of time to event data, censored observations contribute to the total number at risk till the time that the participant is no longer been followed. One advantage here is that the length of time that a participant is followed does not have to be the same for everyone. All observations could have different amounts of time of follow-up, and the analysis can take that into account.

The survival analysis can be conducted in such a way that the participants will be followed at a defined or specified starting-point, and the time needed for the event of interest to emerge will be recorded. Usually, the study ends before all participants have exhibited the event, and the outcome of the remaining participants or patients is unknown. Also the outcome of those participants who have dropped out of the study is unknown. The time of follow-up is recorded (censored data for all these cases). Hence, the data obtained from the study can be analyzed by means of Kaplan-Meier estimate, which is the most appropriate method to present and/or describe survival characteristics.

Kaplan Meier Estimate

Kaplan Meier is derived from the names of two statisticians; Edward L. Kaplan and Paul Meier, in 1958 when they made a collaborative effort and published a paper on how to deal with time to event data [5]. Therefore, they introduced the Kaplan-Meier estimator which serves as a tool for measuring the frequency or the number of patients surviving medical treatment. Later on, the Kaplan-Meier curves and estimates of survival data have become a better way of analyzing data in cohort study. Kaplan-Meier (KM) is non-parametric estimates of survival function that is commonly used to describe survivorship of a study population and to compare two study populations. KM estimate is one of the best statistical methods used to measure the survival probability of patients living for a certain period of time after treatment. It is an intuitive graphical presentation approach. In clinical trials or community trials, the intervention effect is assessed by measuring the number of participants saved or survived after that

intervention over a period of time. KM estimate is the simplest procedure of determining the survival over time in spite of all the difficulties associated with subjects or situations. Curves are used in Kaplan Meier estimate to determine the events, censoring and the survival probability.

Kaplan-Meier survival curve is used in epidemiology to analyze time to event data and to compare two groups of subjects. The survival curve is used to determine a fraction of patients surviving a specified event, like death during a given period of time. This can be calculated for two groups of patients or subjects and also their statistical difference in the survivals. Below is an example of Kaplan-Meier survival curve:

The tick marks on the curve indicate censoring and the curve moves down when the event of interest occurs.

Product Limit estimate (PLI) is another name of Kaplan Meier estimate. The product-limit formula estimates the fraction of organisms or physical devices surviving beyond any age t , even when some of the items are not observed to die or fail, and the sample is rather small [6]. It involves computing the probabilities of occurrence of event at a certain point of time. These successive probabilities will be multiplied by any earlier computed probabilities to determine the final estimate. For example, the probability of a sub-fertile woman surviving the pregnancy three months after laparoscopy and hydrotubation can be considered to be the probability of surviving the first month multiplied by the probabilities surviving the second and third months respectively given that the woman survived the first two months. The third probability is known as a conditional probability.

In survival analysis, intervals are defined by failures. For example, the probability of surviving intervals A and B is equal to the probability of surviving interval A multiplied by the probability of surviving interval B. thus, the PLI be:

$$\frac{P(\text{Surviving interval A})}{\text{Number of subjects at risk upto failure A}} \times \frac{P(\text{Surviving interval B})}{\text{Number of subjects at risk upto failure B}}$$

For each specified interval of time, survival probability is calculated as the number of participants surviving divided by the number of persons at risk. Participants who have dropped out, died, or move out are not counted as "at risk" that is, those who are lost (censored) will not be included in the denominator.

There are three assumptions used in this analysis [7]. *Firstly*, it is assumed that at any time participants who are dropped out or censored have the same survival prospects as those who continue to be followed. *Secondly*, it is assumed that the survival probabilities are the same for participants recruited early and late in the study. *Thirdly*, it is assumed that the event occurs at the time specified.

The limitation of Kaplan Meier estimate is that it cannot be used for multivariate analysis as it only studies the effect of one factor at the time.

The Log-Rank Test

Log-rank test is used to compare two or more groups by testing the null hypothesis. The null hypothesis states that the

populations do not differ in the probability of an event at any time point. Thus, log-rank test is the most commonly-used statistical test to compare the survival functions of two or more groups. These groups can be treatment and control groups or different treatment groups in a clinical trial. The log rank test can be generated in form of table from the statistical softwares such as SPSS, SAS, Stata and R packages. The null hypothesis will be rejected when the p value is less than α value (α can be 0.05, etc.) or fail to be rejected when the p value is large. The log-rank test cannot provide an estimate of the size of the difference between a related confidence interval and groups as it is purely a significance test.

Benchmark Problem

The tables below are the tables of fictive data generated from the SPSS software. (Table 1) contains the data of treatment group only while table 2 contains the data for both the two groups. The first group in the second table is the treatment group while the second group is the control group. Each group comprises ten participants who have been followed for the period of 24 months. The participants in the treatment and control groups were given Drug A and placebo respectively and they were given alphabetical names like A, B, C..., T. The data will be used to determine the Kaplan-Meier estimates (the product limit estimate) of the both the control and the treatment groups.

Table 1: Survival Table.

Treat	ID	Time	Status	Cumulative Proportion Surviving at the Time		No of Cumulative Events	No of Remaining Cases	
				Estimate	Std. Error			
Drug A	1	D	2	Dead	0.9	0.095	1	9
	2	E	4	Dead	0.8	0.126	2	8
	3	A	6	Dead	0.7	0.145	3	7
	4	B	7	Censored	.	.	3	6
	5	Q	8	Censored	.	.	3	5
	6	H	14	Censored	.	.	3	4
	7	F	19	Dead	0.525	0.186	4	3
	8	L	20	Dead	0.35	0.189	5	2
	9	K	22	Censored	.	.	5	1
	10	N	24	Dead	0	0	6	0
Placebo	1	C	1	Dead	0.9	0.095	1	9
	2	I	3	Censored	.	.	1	8
	3	J	5	Dead	0.788	0.134	2	7
	4	P	9	Dead	0.675	0.155	3	6
	5	M	10	Dead	0.563	0.165	4	5
	6	O	11	Censored	.	.	4	4
	7	G	12	Dead	0.422	0.174	5	3
	8	T	15	Censored	.	.	5	2
	9	R	17	Dead	0.211	0.173	6	1
	10	S	18	Dead	0	0	7	0

The product limit estimate is:

$$\frac{P(\text{Surviving interval } A)}{\text{Number of subjects at risk upto failure } A} \times \frac{P(\text{Surviving interval } B)}{\text{Number of subjects at risk upto failure } B}$$

From the curve above, the number of events (deaths) in the treatment group (those given drug A) is 6 while that of the control group (those given placebo) is 7. The number of censored for treatment and control groups are 4 and 3 respectively. The curve

takes a step down when a participant dies and the tick marks on the curve indicate censoring, that is when they lost to follow-up or dropped out of the study.

In the treatment group, Subject D died at 2 months. The estimated survival probability $[P(T>t)]$ will be: $9/10 = 0.9$. Subject E died at 4 months, the estimated survival probability or fraction surviving this death is $8/9$, and thus the product limit estimate (PLI) is: $0.9 \times 8/9 = 0.8$. Subject A also died at 6 months, therefore the PLI is: $0.8 \times 7/8 = 0.7$. Subjects B, Q and H were censored at

7, 8 and 14 months respectively. Subject F died at 19 months, the estimate will be: $0.7 \times \frac{3}{4} = 0.525$. Subject L died at 20 months, the PLI will be $0.525 \times \frac{2}{3} = 0.35$. The next subject in the group, which is subject K, was censored at 22 months while subject N, the last subject in the group died at 24 months and that is the last month of the study. The product limit estimate will be $0.35 \times 0 = 0.00$.

In the control group, subject C died at the first month, the fraction surviving this death will be $\frac{9}{10} = 0.90$ while subject I was censored at the third month. Subject J died at 5 months, the estimated survival probability is $\frac{7}{8}$ and thus, the product limit estimate will be $0.9 \times \frac{7}{8} = 0.788$. Subject P also died at 9 month, the estimated survival probability or fraction surviving this death is $\frac{6}{7} = 0.8571$, therefore the PLI will be $0.788 \times 0.8571 = 0.675$. The next subject in the group, subject M died at 10 months, the fraction surviving this death is $\frac{5}{6} = 0.8333$ and the PLI will be $0.675 \times 0.8333 = 0.562$. Subject O was censored at 11 months. Subject G died at 12 months, the product limit estimate will be $\frac{3}{4} \times 0.562 = 0.422$. Subject T was censored at 15 months. The next subject, which is R died at 17 months, the product limit estimate will be $\frac{1}{2} \times 0.422 = 0.211$. S is the subject that died last in the group, the subject died at 18 months, therefore the product limit estimate will be $0 \times 0.211 = 0.00$.

Note: censored are assumed to be the participants who lost to followed-up or dropped out during the 24 month study.

Table 2: Overall Comparisons.

	Chi-Square	Df	Sig.
Log Rank (Mantel-Cox)	2.603	1	0.107
Breslow (Generalized Wilcoxon)	0.603	1	0.437
Tarone-Ware	1.318	1	0.251

Test of equality of survival distributions for the different levels of Treat.

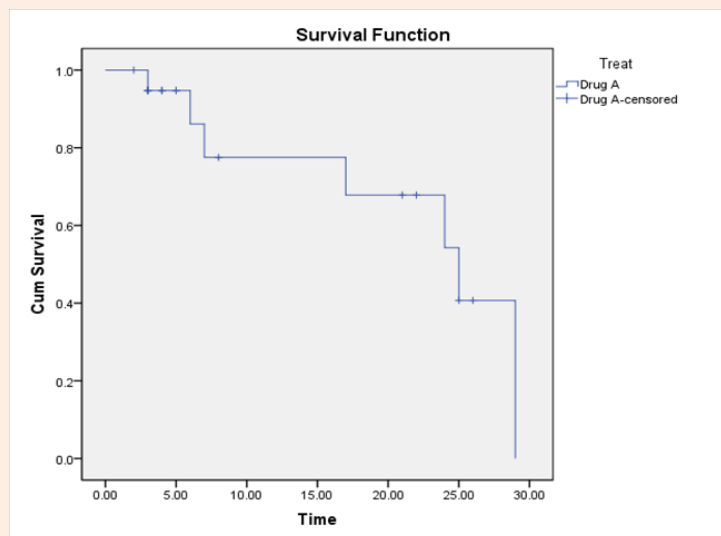


Figure 1: Kaplan Meier estimate curve.

It is seen from the curve

The curves for two different groups of participants can be compared. For example, compare the survival pattern for participants on a treatment with a control. We can identify the gaps in these curves in a vertical or horizontal direction. A vertical gap signifies that at a specific period of time, one group had a greater probability of participants surviving while a horizontal gap signifies that it took longer for one group to experience a certain fraction of deaths.

Now the two groups in figure 3 will be compared in terms of their survival curves. The null hypothesis is that "there is no difference between the groups' survival curves". The table below generated from the SPSS software will be used to test the hypothesis.

Table 2 indicates that all the three p-values are greater than 0.05, and this means that the null hypothesis is failed to be rejected. Therefore, statistically, the survival curves of the treatment and control groups do not differ. Survival curves here mean the population or the true survival curves. The Low Rank in the table place more emphasis on the events happening later in time, Generalized Wilcoxon place more emphasis on the events happening earlier in time while Taron-ware in between the two.

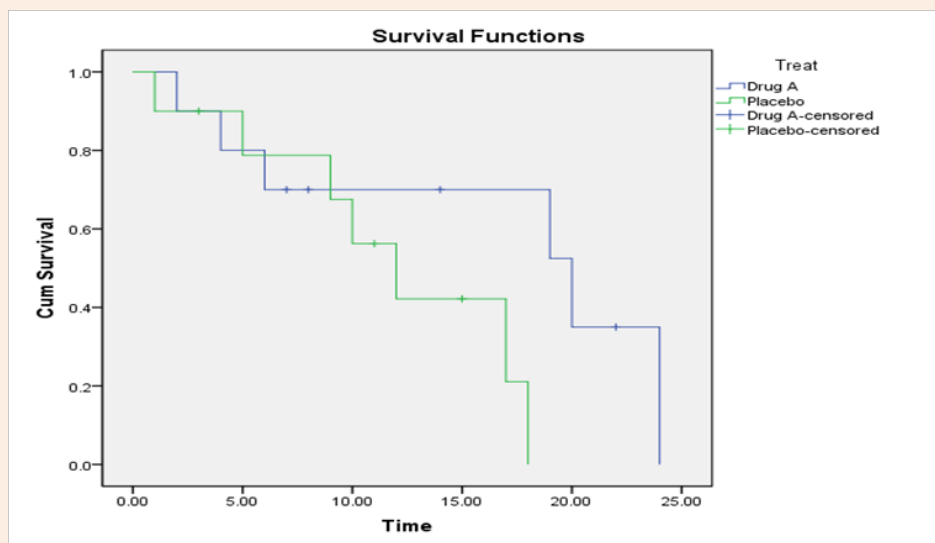


Figure 2: The Kaplan-Meier estimate curve generated by SPSS software from the data used as an example in Table 1.

Conclusion

Kaplan-Meier statistical method is very useful in the field of epidemiology especially in the analysis of time to event data. The method is used in survival analysis to analyze the patients that reached a certain event and those that are censored during a given period of time. It is also very applicable in making comparison between groups of participants such as control group and treatment group. Statistical softwares such as SPSS, Stata, SAS and R packages can be used to generate survival table and Kaplan-Meier estimate curve as well as other important and relevant tables like overall comparisons table. The KM estimate is also applied in other disciplines such as engineering, economics, physics etc.

Acknowledgement

None.

Conflict of Interest

None.

References

1. Gail M, Samet JM, Singer B, Tsiatis A (2002) Statistics for Biology and Health. Survival Analysis, Edition Springer.
2. Goel MK, Khanna P, Kishore J (2010) Understanding survival analysis: Kaplan-Meier estimate. *Int J Ayurveda Res* 1(4): 274–278.
3. Armitage P, Berry G, Matthews JN (2002) Clinical trials. Statistical methods in medical research, pp. 591.
4. Lee ET, Wang J (2003) Statistical methods for survival data analysis 476.
5. Rich JT, Neely JG, Paniello RC, Voelker CC, Nussenbaum B, et al. (2010) A practical guide to understanding kaplan-meier curves. *Otolaryngol Head Neck Surg* 143(3): 331–336.
6. Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53: 457–81.
7. Altman DG, Chapman, Hall (1992) Analysis of Survival times. In: Practical statistics for Medical research, pp. 365–93.