STAT.Sara bahroz@MSc.2024.Non parametric Statistics

by Sara Bahroz

Submission date: 29-Dec-2023 08:40PM (UTC+0200)
Submission ID: 2265537813
File name: T.Sara_bahroz_MSc.2024.Non_parametric_Statistics_-_Sara_Math.pdf (208.15K)
Word count: 3445
Character count: 17004

Salahaddin University – Erbil

College of Administration and Economics

High Education: Master's

Department: Statistics

Subject: Non-Parametric

Semester: First semester

Kaplan – Meier Test to Analyze Survival Data

"Analytical Review"

Sara Bahrooz Ameen

Student Email: mathsara69@gmail.com

Under the Supervision of the Subject Professor

Asst. Prof. Dr. Nazeera Sedeek Kareem

Academic Year

2023 – 2024

Kaplan – Meier Test to Analyze Survival Data

Abstract:

To better understand the unanswered questions regarding the frequency rate of an event in a certain amount of time, the world needed an invention. That is when Edward L. Kaplan and Paul Meier came up with the brilliant idea of "time to event analysis" or "survival analysis" and published a paper on how to deal with incomplete observations in 1958. The two managed to analyze the data successfully and made it easier for us to understand how a particular event could take place in a certain amount of time. However, the beginning of any invention could have erroneous perspectives or findings. Thus, since then, many scientists and researchers have dug into the subject more deeply until it has become one of the main subjects of statistics nowadays.

The importance of survival analysis in Kurdistan is still taken for granted; as in general, data management is not seriously taken by institutions and establishments. For instance, there is a huge data from the health sector of the Kurdistan Regional Government. Nonetheless, according to the officials of the ministry, not much has been taken out of these data. Kaplan – Meier analysis can help us better analyze the data collected by such sectors and give us a pre-understanding method while dealing with different cases. To be more clear, this method of analysis could have proven benefit, especially in the health sector. In this Article Review, we will go over some of the scholarly articles that are written and published by different authors and try to find gaps, direct points to the critiques, and suggest changes.

Key Words:

Time–to–event: a clinical course duration variable for each subject having a beginning and an end anywhere along the timeline of the complete study.

Event of Interest: an event that we want to study. In other words, a case that we want to know how it happens in a certain duration.

Censored Data: data that is unobserved or misses. It does not count in the general observation.

Introduction:

The Kaplan – Meier estimator, which is also known as the product limit estimator, is a statistical way to guess and evaluate the probability of the taking place of an event in a certain amount of time. In other words, it is a non – non-parametric statistic used to estimate the survival function from already collected data. The biggest use of this estimator is in the field of health; albeit it is not the only field this estimator is used for. "The Kaplan–Meier method gives an unbiased estimate of survival only if censored cases are typical of the whole series. If patients are lost to follow-up for reasons related to the event being studied, e.g. because they appear to be cured and so are discharged from the clinic, or conversely because they are too ill to attend the clinic, then the Kaplan–Meier method will underestimate or overestimate the true survival. Consequently, Kaplan–Meier methods should be used only when follow-up is reasonably complete and when losses to follow-up are clearly due to unrelated events." (Damato, et al, 2007).

Like most other datasets, the data collected for analysis by Kaplan–Meier should have two main variables that are crucial for the analysis, though other variables can be added to the study according to the needs of the researcher. In such cases, the third variable can be "the study of different groups" in the same amount of time. For the main and most simple analysis by Kaplan–Meier, the first needed variable is "time," which is very important as it covers the lapse and the period in which a certain event takes place. The second variable is "event of interest," which is the exact case that we want to study in the analysis. For instance, one tries to understand how long people diagnosed with level 4 cancer can live after they are diagnosed. In this situation, the case or event we want to study is the "alive" or "survived" status of the patient. At this point, the event of interest in itself implies that the situation has an end: the patient is either "alive" or "deceased." Although we want to study how long someone has live, the "deceased" status is de facto end course of our study. Kaplan – Meier test either manually, or by using a computer statistics program, gives us the answer to the questions: how long do people live after they are diagnosed with level 4 cancer? How many people can live in a certain time? How many are recovered in that time? And from another perspective, how many have deceased?

The Kaplan – Meier test of course works on a mathematical equation. The survival probability is computed using:

$$St+1 = St*((Nt+1-Dt+1)/Nt+1)$$

This article review is intended to look into some published scholarly articles about the Kaplan – Meier survival analysis. It particularly measures the content of the articles, and their methodologies and checks if they have succeeded in maintaining the aims of the articles themselves. This review will be a case–by–case study of the articles and will finally analyze the articles put together to see the conclusion obtained from thoroughly studying each of them.

Articles:

1. A PRACTICAL GUIDE TO UNDERSTANDING KAPLAN-MEIER CURVES, BY JASON T. RICH, MD, J. GAIL NEELY, MD, FACS, RANDAL C. PANIELLO, MD, FACS, COURTNEY C. J. VOELKER, MD, D. Phil (Oxon), BRIAN NUSSENBAUM, MD, FACS, and ERIC W. WANG, MD

Article Summary:

The article first briefly discusses what the Kaplan – Meier test is and then goes into the details of the survival analysis. It details the concepts around this approach of understanding the unobserved data and how it is analyzed. Then, using tables and figures, it illustrates the application of Kaplan–Meier to show the survival rate of cancer patients. This imagery explanation clarifies the mathematics and statistical approaches within the process. To get the required results from the given data, the research article has used the SigmaPlot computer program rather than doing it manually. The study uses hypothetical data that is not obtained from a real sample but rather made up. The data includes censored data, which according to the article is "the total survival time for that subject cannot be accurately determined. This can happen when something negative for the study occurs, such as the subject drops out, is lost in the follow-up, or required data is not available or, conversely, something good happens, such as the study ends before the subject had the event of interest occur, i.e., they survived at least until the end of the study, but there is no knowledge of what happened thereafter." (T. Rich, et al, 2010, p.p 303). The application of the equation is performed upon two different groups, or two different sets of data, which give different results and the research study analyzes the

difference obtained from these two cases. Nothing is very particular regarding the article other than further clarifying how the method works in itself. Nonetheless, the thesis of this article on the one hand claims that to get a precise result for a certain set of data, the researcher must use smaller intervals of time for collecting data. For example, instead of using trimesters or seasons, they better collect the data every month. On the other hand, it claims that "the Kaplan-Meier method's main focus is on the entire curve of mortality rather than on the traditional clinical concern with rates at fixed periodic intervals. Looking at the ends of the curves or points within them may easily miss the real message." (T. Rich, et al, 2010, p.p 306).

2. The Kaplan Meier Estimate in Survival Analysis, Ilker Etikan, Sulaiman Abubakar, Rukayya Al Kasim

Article Summary:

The researchers have clearly stated what Kaplan–Meier is and how it is used to conduct research and gain results from a given set of data. Besides mentioning the history of how this test came into being, they discuss how it works and how it benefits the fields of study. They also clearly mention that this test is not used only for health analysis, although it is the main field of interest. They claim that this can be used in other fields, such as demography, engineering, and agriculture among others. What makes this article unique is that the researchers have focused on comparison rather than one particular study. They claim that the Kaplan – Meier test is particularly beneficial for comparing two different data that take place within the same amount of time. Called log–rank, the test can be used to compare two or more groups. The methodology they used is that they have brought fictitious data of two different groups of smokers: a group that is on real treatment using medicine and another group that has been given a placebo instead of real treatment. They study the difference in the responses of these two groups to the given medications. They also claim that this approach of data analysis is different from other approaches since Kaplan–Meier considers censored data and hence no data is left out unanalyzed. According to the article, "Unlike other statistical methods such as logistic regression, among others, survival analysis considers censoring and time. Censoring can occur when the patients are lost to follow up to the end of the study. Censored data are data that arise when a person's

life length is known to happen only in a specified time. 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." (Etikan, et al, 2017). The conclusion states that the Kaplan – Meier method of analysis is highly applicable for comparing two different, treatment and control groups or even more, albeit its use in a simple analysis where a relation exists between the event of interest and time interval.

3. Understanding survival analysis: Kaplan-Meier estimate, by Manish Kumar Goel, Pardeep Khanna, and Jugal Kishore

Article Summary:

As the previous studies, this article gives a clear definition of what Kaplan – Meier test is. According to the article, "Kaplan-Meier estimate is one of the best options to be used to measure the fraction of subjects living for a certain amount of time after treatment. In clinical trials or community trials, the effect of an intervention is assessed by measuring the number of subjects survived or saved after that intervention over some time." (Goel, et al, 2010). In statistical language, the article mentions the uses of this estimate and where it could be applied to get the best results out of it. This study also suggests that the shorter the time interval is, the better the results will be. It suggests that the most effective method of statistics used in clinical trials is the Kaplan – Meier estimate. As well as other research studies, it states that in data collection, one may lose too much data that could be crucial for the results of the test. Those lost data are called censored data. If the researcher simply ignores the missing data, they might end up having much smaller data, which will negatively affect the result of the research. Unlike other methods, Kaplan–Meier found a way to include the censored data and give it a special value within the application of the test. These censored data not only are protected, but they also add to the final results. After having applied fictional data on Kaplan–Meier by using a computer program, it states that another set of data can be added to the original results. Thus, having used fictional data regarding some cancer patients, the researchers then added another set of data of those who have gotten a different therapy in their treatments of cancer. As a result, a comparison is created on the curve that compares two different groups of cancer patients: those who have gotten a standard treatment

method and those who have been under a different therapy. The article finally states that the importance and smartness of Kaplan–Meier is because no data is lost; rather, the lost data can be added to the test and will serve as data that has been recorded for a certain amount of time. This is very particular to Kaplan–Meier since other analyses cannot include this sort of data. It also claims that Kaplan – Meier method can generate evidence–based information on a survival analysis.

4. Survival Analysis I: The Kaplan – Meier Method, by Vianda S. Stel, Friedo W. Dekker, Giovanni Tripeppi, Carmine Zoccali, Kitty J. Jager

Article Summary:

The article very briefly states what Kaplan–Meier is and does not go into the details of its application. It does not show a statistical way of how the test is conducted. Rather, it focuses on the importance of the estimate and very clearly shows the difference between Kaplan – Meier and a simple statistical calculation; it shows that simple statistical equations may give wrong results in estimates that require this method to be used. Different from the other articles, this article uses actual data to clarify the subject. The data is collected from a research study conducted by Tsakiris et al. The data was collected from 159,637 patients who had gone under renal replacement between 1986 and 2005. Among these patients, some people suffer from a disease called Multiple Myeloma and are referred to as MM patients. Others are merely patients who needed renal replacement. Thus, there are two groups of people from the same category. The data used for Kaplan – Meier test from this big dataset, is from the data collected within 2005. Kaplan – Meier test is applied to 20 patients who have Multiple Myeloma and 20 patients who do not have this disease.

The study shows that out of the taken examples of 20 patients, one of them died on Day 37, one of them recovered on Day 50, and another died on Day 105. The deceased ones are marked as “dead,” and thus are not considered in the survival list. However, the recovered one is marked in the censored data as we do not know what happened to him after he got out of the hospital, meaning that the data is lost on Day 50. The researchers argue that if we do a simple statistical calculation, we would conclude that the survival rate is 17/19, which is 89.4% while one member variable from our data set is lost. Nonetheless, using the Kaplan – Meier method, they conclude that the actual survival rate for the MM patients is 32.9% and 84.1% for the non-

MM patients. In other words, the mortality rate of the MM patients is 68.1% and 16.9% for the non-MM patients. In conclusion, the researchers argue that the importance of Kaplan – Meier method in this sort of statistical calculation is very high and much more precise than simple statistical calculations since it includes censored data as an important part of the analysis. According to the article, "The results of the KM analyses suggested that the unadjusted 1-year mortality probability was higher in MM patients (68.1%) than in non-MM patients (16.9%) and that according to the log-rank test, this difference was highly statistically significant (p = 0.002). However, to answer a research question, additional analyses may be needed. A limitation of the KM method is that the log-rank test is purely a significance test and cannot provide an estimate of the size of the difference between the groups and its related confidence interval. Another limitation of the KM method is that it only provides unadjusted mortality (and survival) probabilities. However, to make a fair comparison between the MM and non-MM groups, it may be needed to adjust for potential confounders, like age and sex." (Stel S, et al, 2011).

Analysis and Comparison:

Articles written around such a statistical subject can be very similar or even identical. Since science and especially mathematics is considered fact, it does not allow deep argumentation as other subjects. So, the articles about Kaplan–Meier are very much alike in terms of content. They all make a similar definition of what Kaplan–Meier is and how it can be applied to a set of data. Also, their approaches to clarifying the subject in examples are very much alike because they all must use data and apply the estimate method to the data so that they can explain how it works and what results it gives. However, what makes them different is the examples they use and what claims they make to define and introduce the method. Out of the four articles summarized above, three of them use fictional data and one uses actual data. One article is written around the idea that to get more precise data, shorter intervals of time should be used in data collection and that the Kaplan-Meier method's main emphasis is on the mortality curve rather than on the traditional clinical concern with rates at fixed periodic intervals. Another one suggests that the importance of Kaplan–Meier is that no data is lost compared to other methods of analyzing the same sets of data while one other argues that the

method is very effective for comparing two different sets of data that have occurred at the same period. The final one claims that although Kaplan–Meier's method is very useful and effective in the area, it fails to answer some other questions, like how age, sex, and other factors affect the mortality rate of the cases. The researchers in the final article suggest that to fully analyze the obtained data, other methods of statistics should be used alongside the Kaplan – Meier estimate.

It is well known for all the authors that the Kaplan-Meier estimate is a successful way of analyzing data. Their research articles are written with pre-judgment, which is quite normal because the method has already been tested thousands of times way before writing these journal articles. Different fields can be highly controversial as people may view them in different ways. Nonetheless, statistics, as a part of mathematical sciences, mostly includes facts, not assumptions. The Kaplan-Meier estimate is based on mathematical equations that are proven to be correct and accurate. Consequently, the results obtained from the estimate may not be fallacious. All the scholarly articles argue that the Kaplan-Meier estimate is a crucial tool for analyzing data that has missing or lost particles, which makes the estimate highly unique among others.

Conclusion:

Since "necessity is the mother of invention," the two scientists Edward L. Kaplan and Paul Meier felt the need to find a statistical equation that would solve the then unsolved problem of how to calculate and analyze data that has lost values or data, which is called censored data nowadays. They eventually managed to find an equation that would consider the censored data and is calculated within the normal existing data as an important part of the analysis. Although this was an invention in statistics, it helped researchers and scientists in many fields. Now, this method is particularly used by health scientists to analyze their data. It is obvious that statistics is a tool used for almost all aspects of this life and this invention certainly added to the value of statistics. Nonetheless, despite the importance of this estimate, there is still a need for other statistical equations if the data sets that need to be analyzed contain different factors other than the two main variables that are considered in Kaplan – Meier's method: Time and Event of Interest.

Reference

- Damato, B & Taktak, A (2007). Outcome Prediction in Cancer, Ch. 2, Pp. 38
- Etikan, I, Abubakar, S & Alkassim, R (2017). The Kaplan Meier Estimate in Survival Analysis, *Biom Biostat Int J*. Vol 5, No 2.
- Goel, M. K, Khanna, P & Kishore, J (2010). Understanding Survival Analysis: Kaplan Meier Estimate. *National Library of Medicine*. Vol 1, No 4.
- Stel, V. S, Dekker, F. W, Tripepi, G, Zoccali, C, Jagger, K. J (2011). *Karger*. Vol 2, No 2, Pp. 83 – 88.
- Rich J. T, Neely, J. G, Paniello, R. C, Voelker, C. C. J, Nussenbaum, B, Wang E. W (2014). A Practical Guide to Understanding Kaplan Meier Curves, *National Library of Medicine*. Vol 4, No 5, Pp. 331 – 336.

STAT.Sara bahroz@MSc.2024.Non parametric Statistics

ORIGINALITY REPORT

22%	21%	20%	18%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.ncbi.nlm.nih.gov Internet Source	5%
2	karger.com Internet Source	4%
3	medcraveonline.com Internet Source	3%
4	coek.info Internet Source	3%
5	Submitted to Queen Margaret University College, Edinburgh Student Paper	2%
6	c.coek.info Internet Source	1%
7	Submitted to University of Florida Student Paper	1%
8	Submitted to Harrisburg University of Science and Technology Student Paper	1%

9	Submitted to Conroe Independent School District Student Paper	<1%
10	Submitted to Queensland University of Technology Student Paper	<1%
11	sphweb.bumc.bu.edu Internet Source	<1%
12	Submitted to University of Witwatersrand Student Paper	<1%
13	Jason T. Rich, J. Gail Neely, Randal C. Paniello, Courtney C. J. Voelker, Brian Nussenbaum, Eric W. Wang. "A practical guide to understanding Kaplan-Meier curves", Otolaryngology–Head and Neck Surgery, 2010 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		