International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.112

Statistical Investigation on Survival Analysis Using The Approach of Kaplan Meier

Smrithi Prakash
UG Student , Department of Data Science and Business Systems,
SRM Institute of Science And Technology
Kattankulathur, India
smrithireddy79@gmail.com

SnehaTheva
UG Student ,Department of Data Science and Business Systems,
SRM Institute of Science And Technology
Kattankulathur, India
snehatheva4@gmail.com

AnandMadasamy
Department of Data Science and Business Systems,
SRM Institute of Science And Technology
Kattankulathur, India
anandm4@srmist.edu.in

***Abstract*—Time-to-event details are hypothesised using the statistical means known as Kaplan-Meier. The phrase "time to event" indicates to the cycle of duration among a study's admission and a particular incident, like an illness's onset. Since it is used by researchers to classify and/or scrutinize sufferers who off-tracked to follow-up or at times left the study, as well as the ones that grew the condition of concern or pulled through it, this process is advantageous in survival analysis. Additionally, it is used to contrast clusters, like the placebo-treated clusters and the real drug-treated treatment clusters. In addition to being useful in epidemiology, public health, and medicine, the method can also be used in engineering, economics, and other fields. The vast majority of studies that make use of the Kaplan Meier estimate are, like cohort studies, longitudinal. The demise times of kidney transfer victim, the time to infection for smolder sufferers, and demise time for big C trial are all examples of studies for which the Kaplan-Meier estimate may be applicable. The analysis and control clusters, which received a particular medicine and a placebo, were the subject of fictitious data. There were 20 people equally divided in these two clusters, who were monitored for a period of 24 months over the course of two years. Using the fictitious data, the SPSS software fabricated a table and a estimation arc of Kaplan-Meier that be utilised in the analysis of the 24-month study.**

Keywords—Kaplan-Meier, epidemiology, survival analysis, log-rank-test

I. INTRODUCTION

Areas like epizootiological, communal well-being, and treatments, require the analysis of time-to-event data. This type of data measures the time between the start of a study and a specific outcome. To analyse such data, survival analysis is often used. This method involves following a group of participants for a predetermined length of time to measure time to event. In survival analysis, the Kaplan-Meier estimate is the preferred statistical method for comparing both clusters of participants, like a under prescription cluster and a cluster under control. This method is also useful in fields like physics, engineering, economics, and demography. To illustrate the Kaplan Meier estimate, consider a shared characteristic learning of the big C in lungs amid the ones who smoke. The study follows a group of smokers for 20 years to measure the occurrence of lung cancer. The events and censoring are analyzed using the Kaplan-Meier estimate, with events indicating the onset of lung cancer and censored data representing those who dropped out of the study or were unable to continue. Software like SPSS, Stata, SAS, or R can be used to create the survival table and KM measure curve, which are used to calculate the proportion of smokers who survive lung cancer.

II. LITERATURE SURVEY

First, when conducting cancer studies, survival time analysis is a common method used to answer research questions. These questions may involve determining the impact of clinical characteristics, such as blood sugar levels, on patient survival, or calculating the probability of an individual surviving a specific period of time after a cancer diagnosis. Additionally, researchers may compare survival times between groups of patients who received different treatments. In such cases, Kaplan-Meier plots are often used to visually represent endurance arcs, all while Log-rank compares endurance arcs among clusters.

A. Survival Analysis Using COX Proportional Hazard Model

Regression modelling is put to use to calculate the immediate risk of dying, and it is a little more challenging to visualise than the Kaplan-Meier estimate. It is made up of the hazard function h(t), which expresses the likelihood of an event or hazard h (such as survival) up until a specific period t. When comparing the survival of patient groups, the hazard function takes covariates (independent variables in regression) into account [1].

B. Survival Analysis Using Kaplan-Meir Method

When dealing with shortened or unpublished information, the Kaplan-Meier estimator is utilised in the survival distribution. We may estimate the survival function using this non-parametric statistic, which is independent of the underlying probability distribution. The Kaplan - Meier estimates are established on the total count of sufferers who survive for a specific amount of time following therapy (each patient is represented by a row of data) (which is the event) [2] .

C. Method

The two participant groups will be the subject of fictional data. The treatment cluster comes first, and the control cluster comes second. A treatment cluster gets a particular medicine while the control cluster gets a placebo. Ten people are present in each cluster. The fictitious data will be analyzed with the help of SPSS-generated data.III. Survival Analysis

Survival analysisa arithmetical approach, for analysing information on the basis of when the incident happen, particularly in associate study. The duration from a particular point to the instance of a given incident, such as trauma, is known as survival time.As a result, the investigation of cluster information is referred to as survival analysis. As a result, it takes into account data from cohort studies or randomised clinical trials. Main objective of survival analysis is the analysis plus modelling of "time-to-

DOI: rp-9788770040723.112

event" information, which is the subject of controlled experiments in clinical trials.The occurrence could be the removal of a tumor, the length of time it takes to be discharged from a medical facility or hospital, the reaction to medicine, or demise. Events can also refer to an injury, illness recovery, or illness onset. Lassa hemorrhagic fever (LHF) for the ones who displayed symptoms after being checked on for one week in Maiduguri, hemorrhagic fever disease for ones who were checked positive after being under quarantine for 21 days in Serra Leone are two examples of events. Estimating and interpreting survival, comparing it between groups, and determining whether or not explanatory variables have a connection to survival time are all accomplished through the use of survival analysis. Time is taken into account in survival analysis, the duration before an important event occurs[3].

The duration of remission, duration taken for a tumor to disappear, duration for a patient to die, and the duration taken for a response to develop are all examples of survival time data. Other examples include the duration taken for a patient to die and the time it takes for a patient to respond to treatment. There are two crucial aspects of survival time that needs to be precisely explained: a starting point and a finish point that is attained when the incident of importance happens or after the subsequent period has come to an end. Survival duration, consequence to a treatment provided, and sufferer symptoms in accordance to survival, reply, and illness progression are examples of survival data. Clinical and epidemiologic human studies with acute or chronic diseases can yield these data. Survival analysis takes into account censoring and time, in contrast to other statistical techniques like logistic regression, amidst rest[7].

Survival analysis is an essential tool for inspecting time-to-event information in areas like epizootiological, communal well-being, and treatments,. In studies where participants cannot be checked on until the learning ends, censorship can occur. Censored data refers to cases where the actual event time is not observed or is unknown. Censoring can occur in different ways, such as right censoring, where the event has not occurred by a certain point in time, left censoring, where the event has already occurred before the study started, and interval censoring, where only the information that the event occurred within a certain time interval is available[6]. In survival analysis, censored observations are included in the analysis, and the participants are considered to be at risk until they experience the event or are censored. One advantage of survival analysis is that it allows for the inclusion of participants with different follow-up times. Various statistical software, including SPSS, Stata, SAS, and R, can be put to use to conduct survival analysis and generate survival tables and Kaplan-Meier curves[8].

Survival analysis involves tracking participants from a defined point of start and recording the duration it takes for the incident of intrigue to occur. However, not all participants may experience the event before the study ends, and it is also uncertain what will happen to those who withdraw from the study. The duration of follow-up is noted for these cases, resulting in unpublished information. The Kaplan-Meier estimate is a useful tool for analyzing this data and describing the survival characteristics of the study population.

IV. KAPLAN-MEIR ESTIMATE

During survival analysis, participants are followed from a defined starting point, and duration until the happening of the incident of interest is recorded. However, not all participants experience the event before the study ends, and the outcomes for those who withdraw from the study are unknown. For these cases, the follow-up period is censored, and the estimate ,Kaplan-Meier is the most effective technique for demonstrating and describing characteristics of survival[5].

It is recommended to keep text and graphic files separate until after the text has been formatted and styled. Survival analysis using the Kaplan-Meier arc is commonly put to use in epizotiological to compare two groups and analyse time-to-event information. The survival curve calculates the percentage of survivors in a specific incident, such as demise, over a duration, as well as the statistical difference in survivals between the two groups. A Kaplan-Meier survival curve moves downward when the incident of intrigue happens, and tick marks indicate censoring. The Product Limit estimate (PLI), also known as Kaplan Meier's estimate, can estimate the fragment of beings or bodily instruments living past some age t, even when few of the objects are not concerned to pass away or malfunction, and the illustrative size is tiny. It presumes enumerates the likelihood of an incident happening at a specific duration, and dividing these consecutive likelihood by any previously calculated likelihoods to arrive at the end of the estimate. For instance, the likelihood of a not so fertile woman conceiving after hydrogenation and laparoscopy three months later can be calculated using conditional probability[6].

Survival analysis is based on defining intervals by failures. The survival probabilities for different intervals are calculated by multiplying the probabilities for each preceding interval. This calculation leads to the Product Limit estimate (PLI) I.e equation (1) .

$$PLI(1) = \frac{P(SurvivalIntervalA)}{NumberofSubjectsatRiskuptofailureA} X \frac{P(SurvivingIntervalB)}{NumberofSubjectsatRiskuptofailureB} (1)$$

The survival probability is deliberated by dividing the count of beings at possibility by the count of beings who survived for each specified time period. Participants who have withdrawn, died, or relocated are not considered to be "at risk," which means that they will not be included in the denominator. This analysis relies on three presumptions. First, it is assumed that participants who withdraw or are censored will always have same chances of living like the ones, continue to be followed. Second, the survival probabilities for people part of recruited at the start and at the end in the study are assumed to be similar. Thirdly, the incident must happen at the specified duration. Because it only examines effects of one factor at once, the K-M measure cannot be applied to multivariate analysis[4].

III. THE LOG-RANK TEST

The above stated test is a statistical tool used to find the difference of the survival functions of multiple clusters. Its primary purpose is to test whether there is a change in the incident probability between populations at any given time. This test is widely used in clinical trials to compare the survival rates of under medical supervision and control clusters, or different treatment groups. Popular statistical software programs, such as SPSS, SAS, Stata, and R

DOI: rp-9788770040723.112

packages, can generate a log-rank test table. The null hypothesis is not accepted if the p-value is less than the predetermined significance level, usually 0.05. However, the test do not provide any estimate of the magnitude of the difference between groups or a confidence interval.

IV. BENCHMARK PROBLEM

The tables below contain fictitious data generated by SPSS software. Survival Table shows data for the treatment group, whereas Total distributions comparison table shows data for both the treatment and control groups. In Total distributions comparison tablethe starting cluster represents the under medical supervision cluster, and the next cluster represents the control cluster. Ten people in each group were followed for 24 months. Participants were labelled AA,BB, CC..., TT, and received various treatments. For both the treatment and control groups, these information will be put to use to calculate the Kaplan-Meier estimates, also known as the limit estimate of the product.

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	P	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	F	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

TABLE 2: TOTAL DISTRIBUTIONS COMPARISON

	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

The above curve shows six incidents (demise) in the treatment cluster (who received drug A) and seven events (demise) in the control group (who received placebo). The treatment cluster has four censored data, while the control group has three. When a participant dies, the curve decreases, and censoring is represented by tick marks on the curve indicating loss to check-up or withdrawal from the study.

The survival probabilities for the treatment group were estimated using the Kaplan-Meier method. Subject DD passed away at 60 days with a predicted survival probability of 0.9. Subject EE passed away at 120 days with a PLI of 0.8. Subject AA passed away at 180 days, resulting in a PLI of 0.7. Subjects BB, QQ, and HH were left out at 210, 240, and 425 days, respectively. Subject FF passed away at 570 days, and the PLI was 0.525. Subject L L passed away at 600 days, and the PLI was 0.35. Subject K K was censored at 660 days, and subject NN passed away at 720 days with a PLI of 0.00. The tick marks on the curve represent censoring, while a descending curve indicates the occurrence of events (deaths).

Subject CC passed away during the start month in the control group, with a survival rate of 0.90. In the third month, Subject II was hidden. Subject JJ died at the age of five months, and the PLI is 0.788. The survival rate for Subject PP's death at 9 months is 0.8571, and the PLI is 0.675. At 10 months, Subject MM died with a survival rate of 0.8333 and a PLI of 0.562. At 11 months, Subject OO was hidden. Subject GG demise after a year, and his PLI is 0.422. Subject TT was hidden at the age of 15 months. Subject RR died at the age of 17 months, with a PLI of 0.211. Subject SS died at the age of 18 months, and his PLI is 0.00. These calculations are visible on the curve.

The survival patterns of two participant groups, like the treatment and control cluster, can be compared by comparing the gaps in their curves, which can be linear or stable. A gap that is vertical indicates that one group had a higher probability of survival at a given time, whereas a horizontal gap indicates that one group died more slowly. Figure 1 : KM estimate curve compares the survival curves of the both clusters, with the null hypothesis that there is no much of a change among them. This hypothesis is tested using the SPSS-generated table below, and the fact that each of the three p-values in Total distributions comparison table is greater than 0.05 indicates that the null hypothesis was not rejected.

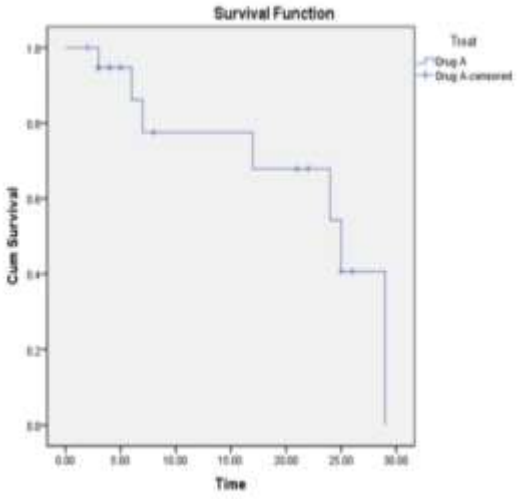

Fig.1. Kaplan-Meir estimate curve

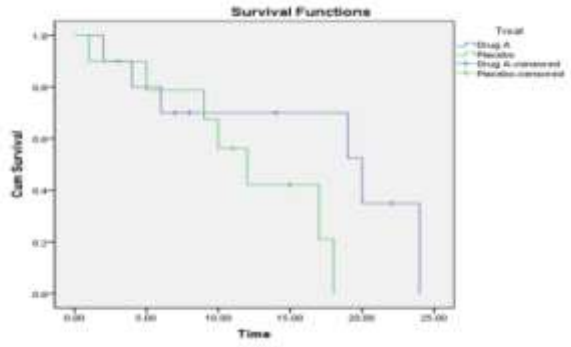

Fig.2.The SPSS software - generated Kaplan-Meir estimate curve

Therefore, statistically speaking, there is no dissimilarity among the treatment and control clusters survival arc. In this context, survival curves refer to the population or actual survival arcs. The events that take place later are more heavily emphasised by the Low Rank in the table, while the events that take place earlier are more heavily emphasised by the Generalized Wilcoxon, with Taron-ware in between the two.

V. CONCLUSION

The Kaplan-Meier arithmetic technique is extremely helpful in the area ofepizootiological, particularly when analyzing data pertaining to time to events. In survival analysis, the method is used to look at patients who reached a certain point and those who were censored for a certain amount of time. It also works well when comparing participant groups like the control group and the treatment group. The generation of survival tables, Kaplan-Meier estimate curves, and other significant and pertinent tables,such as the overall comparisons table, can be accomplished using statistical software packages like SPSS, Stata, SAS, and R. Engineering, economics, physics, and other fields use the KM estimate as well.

REFERENCES

[1] Rajesh, M., &Sitharthan, R. (2022). Image fusion and enhancement based on energy of the pixel using Deep Convolutional Neural Network. Multimedia Tools and Applications, 81(1), 873-885

[2] S.Rai, P. Mishra, and U.C. Ghoshal, "Survival analysis: A primer for the clinician scientists," Indian J Gastroenterol, vol. 40, pp. 541–549, 2021, https://doi.org/10.1007/s12664-021-01232-1.

[3] Schober, M.D. Patrick, PhD, MMedStat*; Vetter, Thomas R. MD, MPH†, "Survival Analysis and Interpretation of Time-to-Event Data: The Tortoise and the Hare," Anesthesia & Analgesia, vol. 127, no. 3, p 792-798, September 2018. | DOI: 10.1213/ANE.0000000000003653

[4] Moshika, A., Thirumaran, M., Natarajan, B., Andal, K., Sambasivam, G., &Manoharan, R. (2021).Vulnerability assessment in heterogeneous web environment using probabilistic arithmetic automata. IEEE Access, 9, 74659-74673

[5] P.Guyot, A.Ades,and M.J. Ouwens,et al., "Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves," BMC Med Res Methodol., vol. 12, no. 9, 2012, https://doi.org/10.1186/1471-2288-12-9.

[6] T.P. Morris, C.I. Jarvis, and W. Cragg, et al., "Proposals on Kaplan–Meier plots in medical research and a survey of stakeholder views,"KMunicateBMJOpen, vol. 9, p. e030215, 2019, doi: 10.1136/bmjopen-2019-030215

[7] J.T. Rich, J.G. Neely, and R.C. Paniello, et al., "A practical guide to understanding kaplan-meier curves,"Otolaryngol Head Neck Surg. vol. 143, no. 3, pp. 331–336, 2010.

[8] D.G. Altman, Chapman, and Hall, "Analysis of Survival times," In:Practical Statistics for Medical Research, pp. 365–393, 1992.