A BRIEF INTRODUCTION TO SURVIVAL ANALYSIS

Research · February 2018

DOI: 10.13140/RG.2.2.22858.47046

Sucharith Thoutam
National Institute of Technology Rourkela

A BRIEF INTRODUCTION TO SURVIVAL ANALYSIS

Submitted by:

Sucharith Thoutam

Roll number: 412ma5082

[Integrated MSc. Mathematics]

Under the guidance of:

Dr. Sumathi Uma Maheshwari,

Assistant Professor,

Department of Mathematics,

Kakatiya University, Warangal, Telangana.

Duration: 10th May 2016 to 10th July, 2016

DECLARATION

I hereby declare that the work which is being presented in the report "A Brief Introduction to Survival Analysis" submitted in Department of Mathematics, National Institute of Technology, Rourkela is an authentic record of my own work carried out under the supervision of Dr. Sumathi Uma Maheshwari at Department of Mathematics, Kakatiya University.

The matter embodied in this has not been submitted by me for the award of any other degree.

Sucharith Thoutam

ACKNOWLEDGMENTS :

It is a great pleasure and proud privilege to express my deep sense of gratitude to my guide, Prof. T. Sumathi Uma Maheshwari. I am grateful to her for her continuous encouragement and guidance throughout the period of my project work. Without her active guidance it would not have been possible for me to complete the work.

I acknowledge my deep sense of gratitude to Prof. M. Tirumala Devi for her helpful guidance.

I thank all the staff members and the authorities of University College of Engineering and Technology for Women, Kakatiya University.

I express my sincere thanks to my parents and other family members, for their support, blessings and encouragement.

ABSTRACT:

Survival analysis has become a widely used methodology in diverse fields of research such as medicine, economics and political science. It has gained momentum particularly in medicine field in the past few decades. In this paper we provide a layman introduction to survival analysis and its terminology. Main emphasis is focused on understanding the basic concepts such as “censoring”. And clear explanation of a well-known technique called Kaplan- Meier product limit formula and its estimator curve is provided.

Keywords: Survival Analysis, Censoring, Kaplan-Meier Estimator.

INTRODUCTION:

Survival Analysis is a branch of statistics for analyzing the expected duration of time until one or more events happen such as death in biological organisms and failure in mechanical systems. This topic is called “reliability theory” or “reliability analysis” in engineering, “duration analysis” or “duration modelling” in economics, and “event history analysis” in sociology.

What is the proportion of the population which will survive past a certain time? Of those that survive, at what rate will they die or fail? Can multiple causes of death or failure be taken into account? How do particular circumstances or characteristics increase or decrease the probability of survival? Survival Analysis attempts to answer questions such as mentioned above.

THEORY :

Definitions of common terms in survival analysis:

Survival Analysis is a collection of statistical procedures for data analysis for which the outcome variable of interest is *time until an event occurs.*

By **time**, we mean years, months, weeks or days from the beginning of follow-up of an individual until an event occurs; alternatively, time can refer to the age of an individual when an event occurs.

By **event**, we mean death, disease incidence, relapse from remission, recovery or any designated experience of interest that may happen to an individual.

In a survival analysis, we usually refer to the time variable as **survival time,** because it gives the time that an individual has "survived" over some follow-up period. We also typically refer to the event as a **failure,** because the event of interest usually is death, disease incidence, or some other negative individual experience.

Examples:

Three examples of survival analysis problems are briefly mentioned here. The first is a study that follows leukaemia patients in remission over several weeks to see how long they stay in remission. The second example follows a disease-free cohort of individuals over several years to see who develops heart disease. A third example considers a 13-year follow-up of an elderly population (60+ years) to see how long subjects remain alive.

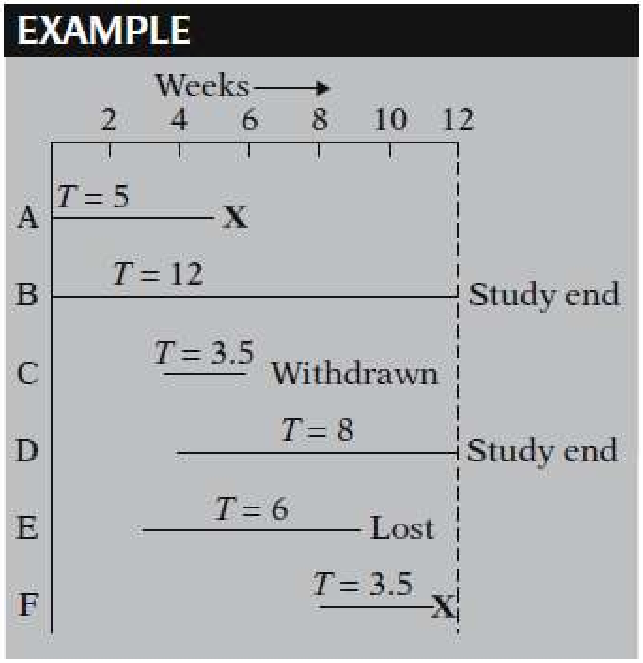
All of the above examples are survival analysis problems because the outcome variable is time until an event occurs. In the first example, involving leukaemia patients, the event of interest (i.e., failure) is "going out of remission," and the outcome is "time in weeks until a person goes out of remission." In the second example, the event is "developing heart disease," and the outcome is "time in years until a person develops heart disease." In the third example, the event is "death" and the outcome is "time in years until death."

Censored Data:

Most survival analyses must consider a key analytical problem called **censoring.** Censoring occurs when we have some information about individual survival time, but **we don't know the survival time exactly.**

There are generally three reasons why censoring may occur:
(1) a person does not experience the event before **the study ends;**
(2) a person is **lost to follow-up** during the study period;
(3) a person **withdraws from the study** because of some reason.

These situations are graphically illustrated here. The graph describes the experience of several persons followed over time. An **X** denotes a person who got the event.



Person A, for example, is followed from the start of the study until getting the event at week 5; his survival time is 5 weeks and is *not* censored.

Person B also is observed from the start of the study but is followed to the end of the 12-week study period without getting the event; the survival time here is censored because we can say only that it is *at least* 12 weeks.

Person C enters the study between the second and third week and is followed until he withdraws from the study at 6 weeks; this person's survival time is censored after 3.5 weeks.

Person D enters at week 4 and is followed for the remainder of the study without getting the event; this person's censored time is 8 weeks.

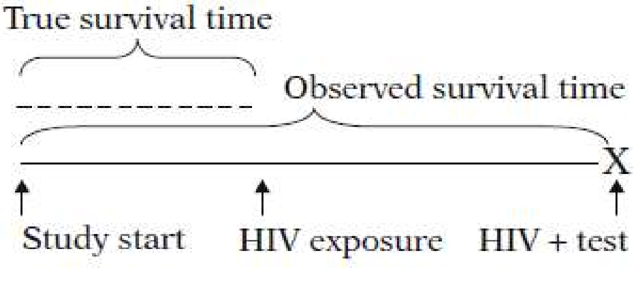
Person E enters the study at week 3 and is followed until week 9, when he is lost to follow-up; his censored time is 6 weeks.

Person F enters at week 8 and is followed until getting the event at week 11.5. As with person A, there is no censoring here; the survival time is 3.5 weeks.

In summary, of the six persons observed, two get the event (persons A and F) and four are censored (B, C, D, and E).

Notice in our example that for each of the four persons censored, we know that the person's exact survival time becomes incomplete at the **right** side of the follow-up period, occurring when the study ends or when the person is lost to follow-up or is withdrawn. We generally refer to this kind of data as **right-censored.** For these data, the complete survival time interval, which we don't really know, has been cut off (i.e., censored) at the right side of the observed survival time interval. Although data can also be left-censored**,** most survival data is right-censored.

Left-censored data can occur when a person's true survival time is less than or equal to that person's observed survival time. For example, if we are following persons until they become HIV positive, we may record a failure when a subject first tests positive for the virus. However, we may not know exactly the time of first exposure to the virus, and therefore do not know exactly when the failure occurred. Thus, the survival time is censored on the left side since the true



survival time, which ends at exposure, is shorter than the follow-up time, which ends when the subject tests positive.

Interval censoring occurs when failure is only known to have occurred during an interval.

Here we are mainly concerned with right censoring, which is most common in survival analysis.

Terminology and Notation:

First, we denote a person's survival time by the random variable a **capital *T***. Since *T* denotes time, its possible values include all nonnegative numbers; that is, *T* can be any number equal to or greater than zero. Next, we denote by a small letter *t* any specific value of interest for the random variable capital *T*.

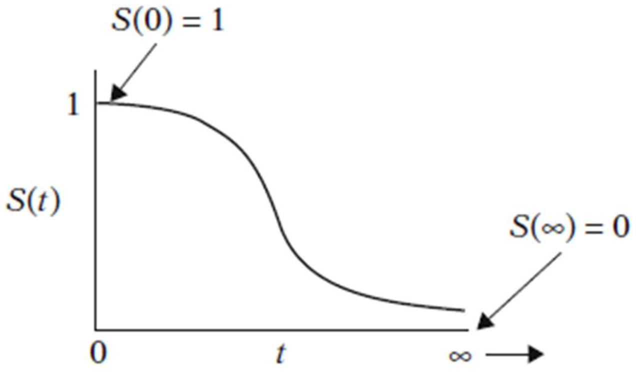
Finally, we let the Greek letter delta (δ) denote a (0,1) random variable indicating either failure or censorship. That is, δ = 1 for failure if the event occurs during the study period, or δ = 0 if the survival time is censored by the end of the study period.

Survival Function:

The survival function *S*(*t*) gives the probability that a person survives longer than some specified time *t:* that is, *S*(*t*) gives the probability that the random variable *T* exceeds the specified time *t*.

$$S(t) = P\ (T > t)$$

Theoretical $S(t)$:

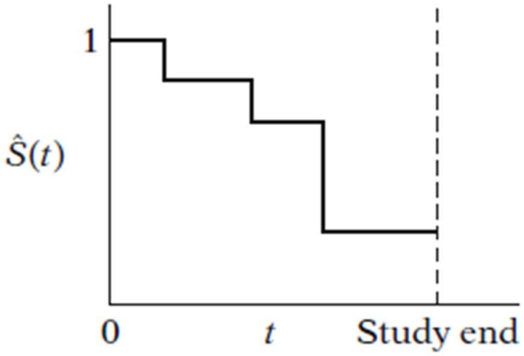


Theoretically, as t ranges from 0 up to infinity, the survival function can be graphed as a smooth curve. As illustrated by the graph, where t identifies the X-axis, all survivor functions have the following characteristics:

- they are non-increasing; that is, they head downward as t increases;
- at time t = 0, S(t) = S(0) = 1; that is, at the start of the study, since no one has gotten the event yet, the probability of surviving past time 0 is one;
- at time t=∞, S(t) = S(∞) = 0; that is, theoretically, if the study period increased without limit, eventually nobody would survive, so the survivor curve must eventually fall to zero.

Note that these are theoretical properties of survival curves.

$\hat{S}(t)$ in practice:



In practice, when using actual data, we usually obtain graphs that are step functions, as illustrated here, rather than smooth curves.

Moreover, because the study period is never infinite in length and there may be competing risks for failure, it is possible that not everyone studied gets the event. The estimated survival function, denoted by a caret over the S in the graph, thus may not go all the way down to zero at the end of the study.

Hazard Function:

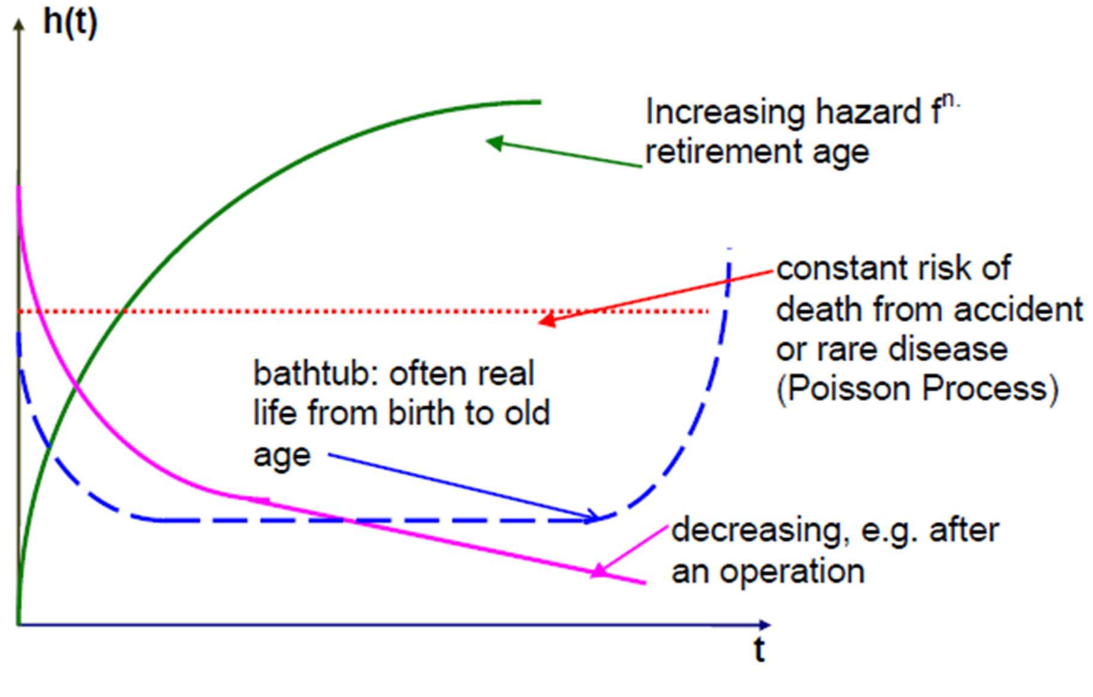
The hazard function h(t) gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t.

$$h(t) = \lim_{\delta t \to 0} \left[\frac{P[t \le T < t + \delta t \mid T \ge t]}{\delta t} \right]$$

This is also known as “hazard rate” or “failure rate”.
As with a survival function, the hazard function h(t) can be graphed as t ranges over various values. In contrast to a survival function, the graph of h(t) does not have to start at 1 and go down to zero, but rather can start anywhere and go up and down in any direction over time. In particular, for a specified value of t, the hazard function h(t) has the following characteristics:

- it is always nonnegative, that is, equal to or greater than zero;
- it has no upper bound.

Some examples of the hazard function are given below.



Regardless of which function S(t) or h(t) one prefers, there is a clearly defined relationship between the two. In fact, if one knows the form of S(t), one can derive the corresponding h(t), and vice versa.
More generally, the relationship between S(t) and h(t) can be expressed equivalently in either of two calculus formulae shown here.

$$S(t) = \exp\left[-\int_0^t h(u)du\right]$$

$$h(t) = -\left[\frac{dS(t)/dt}{S(t)}\right]$$

The first of these formulae describes how the survival function S(t) can be written in terms of an integral involving the hazard function. The formula says that S(t) equals the exponential of the negative integral of the hazard function between integration limits of 0 and t.
The second formula describes how the hazard function h(t) can be written in terms of a derivative involving the survival function. This formula says that h(t) equals minus the derivative of S(t) with respect to t divided by S(t).
In any actual data analysis, a computer program can make the numerical transformation from S(t) to h(t), or vice versa, without the user ever having to use either formula. The point here is simply that if you know either S(t) or h(t), you can get the other directly.

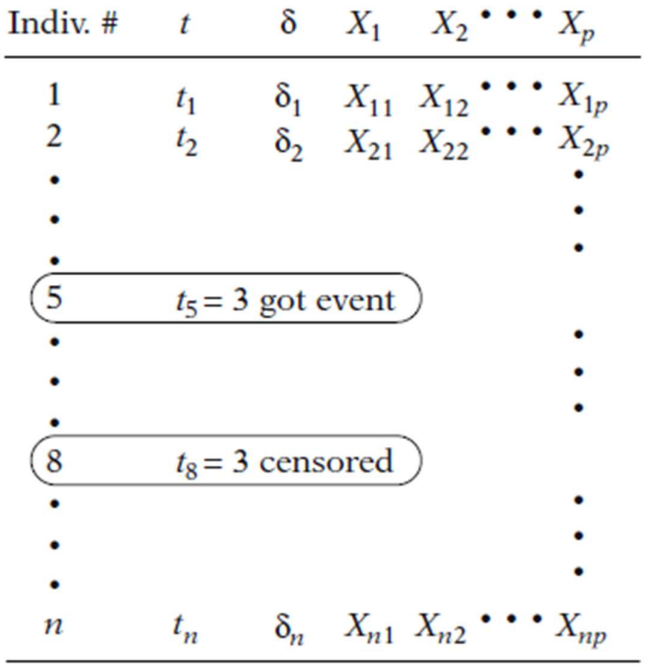
The main goal of survival analysis is to estimate and interpret survivor and/or hazard functions from survival data.

ANALYSIS:

Basic Data Layout for Computer:

Assume that we have a data set consisting of n persons. The first column of the table identifies each person from 1, starting at the top, to n, at the bottom. The remaining columns after the first one provides survival time and other information for each person. The second column gives the survival time information, which is denoted t_1 for

individual 1, t_2 for individual 2, and so on, up to t_n for individual n. Each of these t's gives the observed survival time regardless of whether the person got the event or is censored. For example, if person 5 got the event at 3 weeks of follow-up, then $t_5 = 3$; on the other hand, if person 8 was censored at 3 weeks, without getting the event, then $t_8 = 3$ also.
To distinguish persons who get the event from those who are censored, we turn to the third column, which gives the information for status (i.e. δ) the dichotomous variable that indicates censorship status.
Thus, δ_1 is 1 if person 1 gets the event or is 0 if person 1 is censored; δ_2 is 1 or 0 similarly, and so on, up through δ_n. In the example just considered, person 5, who failed at 3 weeks, has a δ of 1; that is, δ_5 equals 1. In contrast, person 8, who was censored at 3 weeks, has a δ of 0; that is, δ_8 equals 0.

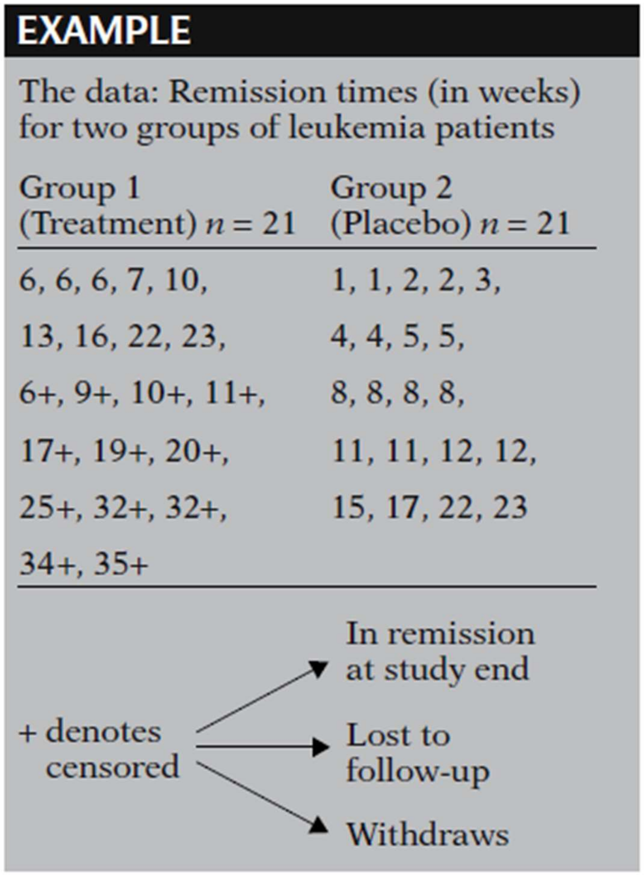


Indiv. #	t	δ	X_1	X_2	• • •	X_p
1	t_1	δ_1	X_{11}	X_{12}	• • •	X_{1p}
2	t_2	δ_2	X_{21}	X_{22}	• • •	X_{2p}
•						•
•						•
•						•
5	$t_5 = 3$ got event					
•						•
•						•
•						•
8	$t_8 = 3$ censored					
•						•
•						•
•						•
n	t_n	δ_n	X_{n1}	X_{n2}	• • •	X_{np}

The remainder of the information in the table gives values for explanatory variables of interest. An explanatory variable, X_i, is any variable like age or exposure status, E, or a product term like age×race that the investigator wishes to consider to predict survival time. These variables are listed at the top of the table as X_1, X_2, and so on, up to X_p.

Example:

As an example of this data layout, consider the following set of data for two groups of leukemia patients: one group of 21 persons has received a certain treatment; the other group of 21 persons has received a placebo.



EXAMPLE

The data: Remission times (in weeks) for two groups of leukemia patients

Group 1 (Treatment) $n = 21$	Group 2 (Placebo) $n = 21$
6, 6, 6, 7, 10,	1, 1, 2, 2, 3,
13, 16, 22, 23,	4, 4, 5, 5,
6+, 9+, 10+, 11+,	8, 8, 8, 8,
17+, 19+, 20+,	11, 11, 12, 12,
25+, 32+, 32+,	15, 17, 22, 23
34+, 35+	

+ denotes censored

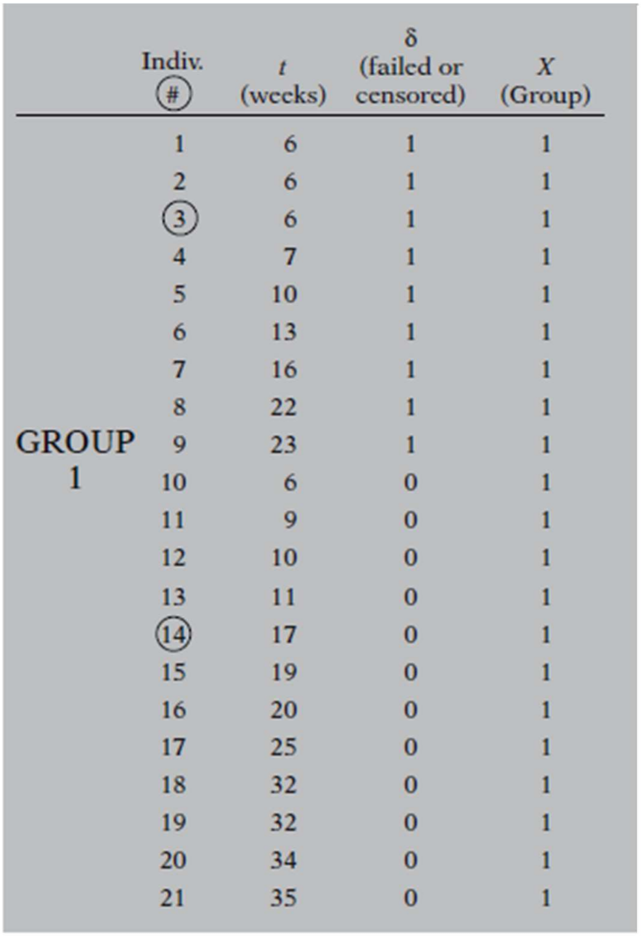
In remission at study end

Lost to follow-up

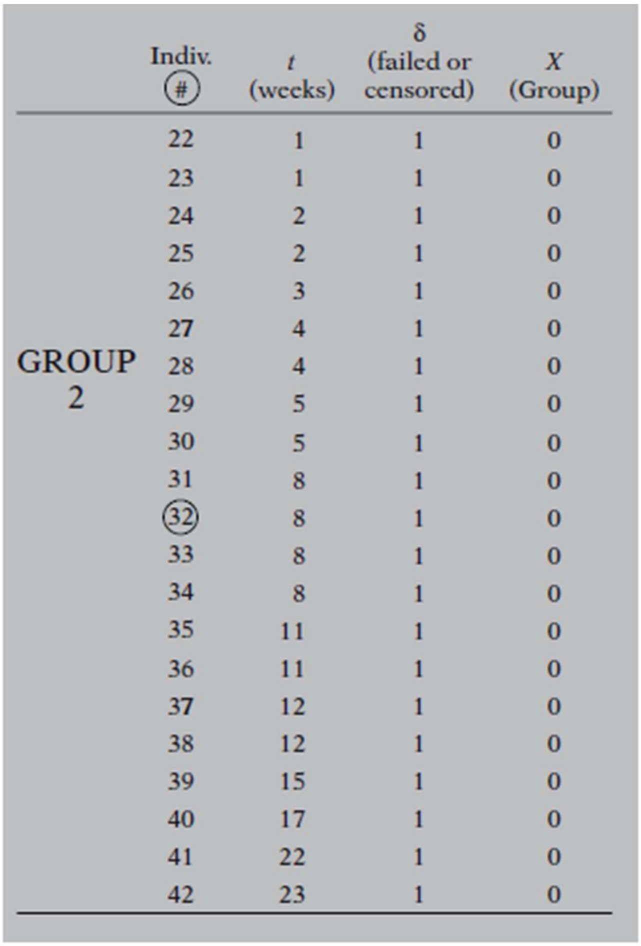
Withdraws

The values given for each group consist of time in weeks a patient is in remission, up to the point of the patient's either going out of remission or being censored. Here, going out of remission is a failure. A person is censored if he or she remains in remission until the end of the study, is lost to follow-up, or withdraws before the end of the study. The censored data here are denoted by a plus sign next to the survival time.

Notice that the first three persons in group 1 went out of remission at 6 weeks; the next six persons also went out of remission, but at failure times ranging from 7 to 23. All of the remaining persons in group 1 with pluses next to their survival times are censored. For example, on line three the first person who has a plus sign next to a 6 is censored at six weeks. The remaining persons in group one are also censored, but at times ranging from 9 to 35 weeks.



	Indiv. #	t (weeks)	δ (failed or censored)	X (Group)
	1	6	1	1
	2	6	1	1
	③	6	1	1
	4	7	1	1
	5	10	1	1
	6	13	1	1
	7	16	1	1
	8	22	1	1
GROUP 1	9	23	1	1
	10	6	0	1
	11	9	0	1
	12	10	0	1
	13	11	0	1
	⑭	17	0	1
	15	19	0	1
	16	20	0	1
	17	25	0	1
	18	32	0	1
	19	32	0	1
	20	34	0	1
	21	35	0	1



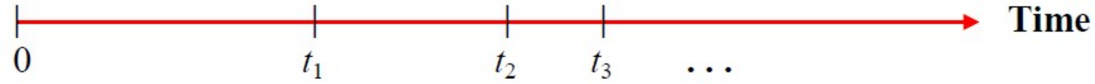
	Indiv. #	t (weeks)	δ (failed or censored)	X (Group)
	22	1	1	0
	23	1	1	0
	24	2	1	0
	25	2	1	0
	26	3	1	0
	27	4	1	0
GROUP 2	28	4	1	0
	29	5	1	0
	30	5	1	0
	31	8	1	0
	㉜	8	1	0
	33	8	1	0
	34	8	1	0
	35	11	1	0
	36	11	1	0
	37	12	1	0
	38	12	1	0
	39	15	1	0
	40	17	1	0
	41	22	1	0
	42	23	1	0

We now put this data in tabular form for the computer, as shown at the top. The list starts with the 21 persons in group 1 (listed 1–21) and follows with the 21 persons in group 2 (listed 22–42). Our n for the composite group is 42. The second column of the table gives the survival times in weeks for all 42 persons. The third column indicates failure or censorship for each person. Finally, the fourth column lists the values of the only explanatory variable we have considered so far, namely, group status, with 1 denoting treatment and 0 denoting placebo.

There are parametric and non-parametric methods to estimate and analyse the survivor function. We here discuss a very famous non-parametric method known as Kaplan-Meier Estimator.

Kaplan-Meier Estimator:

The Kaplan–Meier estimator, also known as the product limit estimator, is a non-parametric statistic used to estimate the survival function from lifetime data. The estimator is named after Edward L. Kaplan and Paul Meier, who each submitted similar manuscripts to the Journal of the American Statistical Association. The journal editor, John Tukey, convinced them to combine their work into one paper.



Let t_1, t_2, t_3, … denote the actual times of death of the n individuals in the cohort. Also let d_1, d_2, d_3, … denote the number of deaths that occur at each of these times, and let n_1, n_2, n_3, … be the corresponding number of patients remaining in the cohort. Note that $n_2 = n_1 – d_1$, $n_3 = n_2– d_2$, etc. Then, loosely speaking, $S(t_2) = P(T > t_2)$ = "Probability of surviving beyond time t_2" depends conditionally on $S(t_1) = P(T > t_1)$ = "Probability of surviving beyond time t_1." Likewise, $S(t_3) = P(T > t_3)$ = "Probability of surviving beyond time t_3" depends conditionally on $S(t_2) = P(T > t_2)$ = "Probability of surviving beyond time t_2," etc. By using this recursive idea, we can iteratively build a numerical estimate Sˆ(t) of the true survival function S(t).

- For any time $t \in [0, t_1)$, we have $S(t) = P(T > t)$ = "Probability of surviving beyond time t" = 1, because no deaths have as yet occurred. Therefore, for all t in this interval, let $S\hat{}(t) = 1$.

(For any two events A and B, $\boldsymbol{P(A \textbf{ and } B) = P(A) \times P(B \mid A)}$.)
Let A = "survive to time t_1" and B = "survive from time t_1 to beyond some time t before t_2." Having *both* events occur is therefore

equivalent to the event "*A* and *B*" = "survive to beyond time *t* before t_2," i.e., "$T > t$." Hence, the following holds.

- For any time $t \in [t_1, t_2)$, we have…

$$S(t) \;=\; P(T > t) \;=\; \underbrace{P(\text{survive in } [0, t_1))} \;\times\; \underbrace{P(\text{survive in } [t_1, t] \mid \text{survive in } [0, t_1))},$$

i.e,

$$\hat{S}(t) \;=\; 1 \;\times\; \frac{n_1 - d_1}{n_1}, \qquad \text{or}$$

$$\hat{S}(t) \;=\; 1 - \frac{d_1}{n_1}.$$

Similarly for any time $t \in [t_2, t_3)$, we have…

$$S(t) \;=\; P(T > t) \;=\; \underbrace{P(\text{survive in } [t_1, t_2))} \;\times\; \underbrace{P(\text{survive in } [t_2, t] \mid \text{survive in } [t_1, t_2))},$$

i.e,

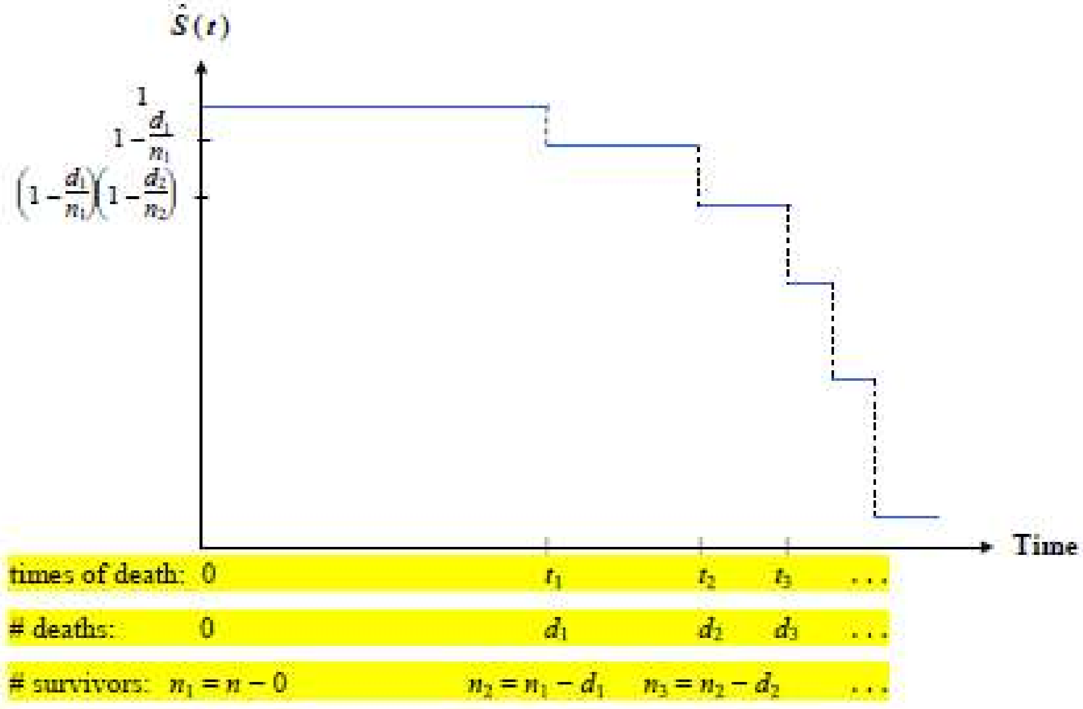
$$\hat{S}(t) \;=\; \left(1 - \frac{d_1}{n_1}\right) \;\times\; \frac{n_2 - d_2}{n_2}, \qquad \text{or}$$

$$\hat{S}(t) \;=\; \left(1 - \frac{d_1}{n_1}\right)\left(1 - \frac{d_2}{n_2}\right)$$

- In general, for $t \in [t_j, t_{j+1})$, $j = 1, 2, 3, \ldots,$ we have…

$$\boxed{\hat{S}(t) \;=\; \left(1 - \frac{d_1}{n_1}\right)\left(1 - \frac{d_2}{n_2}\right) \ldots \left(1 - \frac{d_j}{n_j}\right) \;=\; \prod_{i=1}^{j} \left(1 - \frac{d_i}{n_i}\right).}$$

This is known as the **Kaplan-Meier estimator** of the survival function $S(t)$. (Theory developed in 1950s, but first implemented with computers in 1970s.) Note that it is *not continuous*, but only *piecewise-continuous* (actually, *piecewise-constant*, or “step function”).


$\hat{S}(t)$
1
$1-\frac{d_1}{n_1}$
$\left(1-\frac{d_1}{n_1}\right)\left(1-\frac{d_2}{n_2}\right)$
Time
times of death: 0 t_1 t_2 t_3 . . .
# deaths: 0 d_1 d_2 d_3 . . .
# survivors: $n_1 = n - 0$ $n_2 = n_1 - d_1$ $n_3 = n_2 - d_2$ . . .

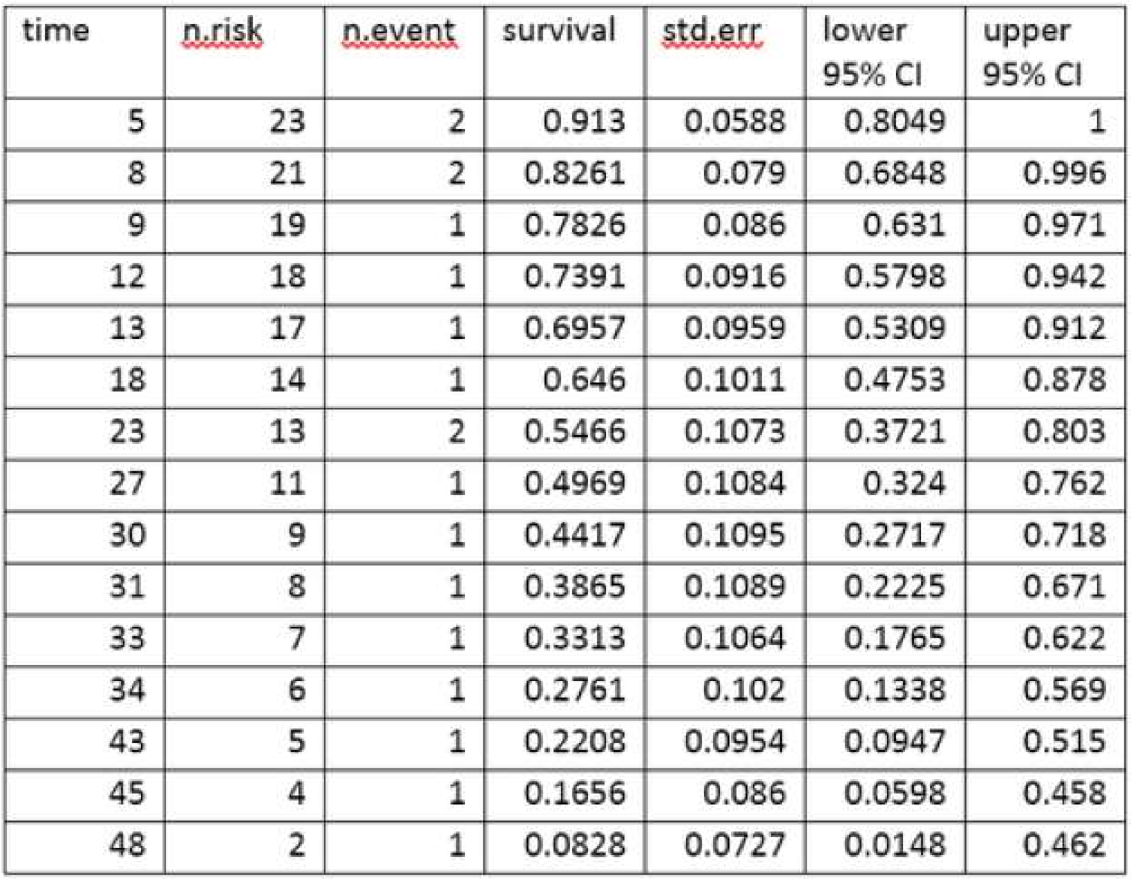
In this paper, we perform Kaplan Meier Analysis of two datasets using R software. The datasets are obtained from following site: `https://www.umass.edu/statdata/statdata/data/`

RESULTS :

We first analyse a data comprising the heart attack study of 100 patients in the city of Worcester. The following is the code for Kaplan Meier estimator in “R”:

```
library(survival)
attach(whas100)
whas100.sv<-Surv(days,censor,type="right")
whas100Surv<-survfit(whas100.sv~1,data=whas100)
summary(whas100Surv)
plot(whas100Surv,main="Kaplan-Meier  Estimator  Curve  for
Worcester  Heart  Attack  Study",xlab="Time",ylab="Survival
function")
```

The 5th line which says summary produces a table known as “Life Table”. A sample of life table is given here to understand the basic data layout of a life table.



time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
5	23	2	0.913	0.0588	0.8049	1
8	21	2	0.8261	0.079	0.6848	0.996
9	19	1	0.7826	0.086	0.631	0.971
12	18	1	0.7391	0.0916	0.5798	0.942
13	17	1	0.6957	0.0959	0.5309	0.912
18	14	1	0.646	0.1011	0.4753	0.878
23	13	2	0.5466	0.1073	0.3721	0.803
27	11	1	0.4969	0.1084	0.324	0.762
30	9	1	0.4417	0.1095	0.2717	0.718
31	8	1	0.3865	0.1089	0.2225	0.671
33	7	1	0.3313	0.1064	0.1765	0.622
34	6	1	0.2761	0.102	0.1338	0.569
43	5	1	0.2208	0.0954	0.0947	0.515
45	4	1	0.1656	0.086	0.0598	0.458
48	2	1	0.0828	0.0727	0.0148	0.462

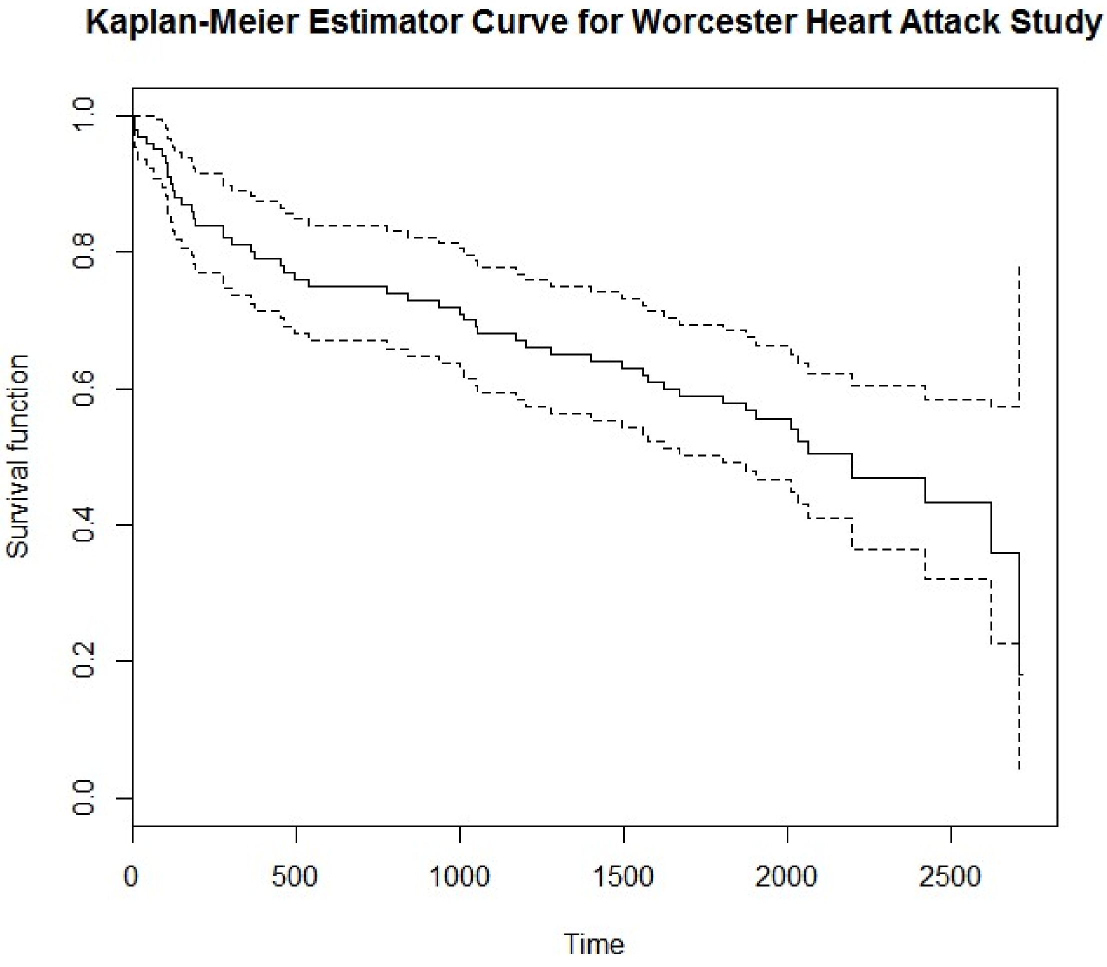
A Life table summarizes survival data in terms of the number of events and the proportion surviving at each event time point.

- time gives the time points at which events occur.
- n.risk is the number of subjects at risk immediately before the time point, t. Being "at risk" means that the subject has not had an event before time t, and is not censored before or at time t.
- n.event is the number of subjects who have events at time t.
- survival is the proportion surviving, as determined using the Kaplan- Meier product-limit estimate.
- std.err is the standard error of the estimated survival. The standard error of the Kaplan-Meier product limit estimate at time t_i is calculated using Greenwood’s formula, and depends on the

number at risk (n.risk in the table), the number of deaths (n.event in the table) and the proportion surviving (survival in the table).

- lower 95% CI and upper 95% CI are the lower and upper 95% confidence bounds for the proportion surviving.

Taking into account the above said points a plot of Kaplan Meir Estimator Curve is given below.

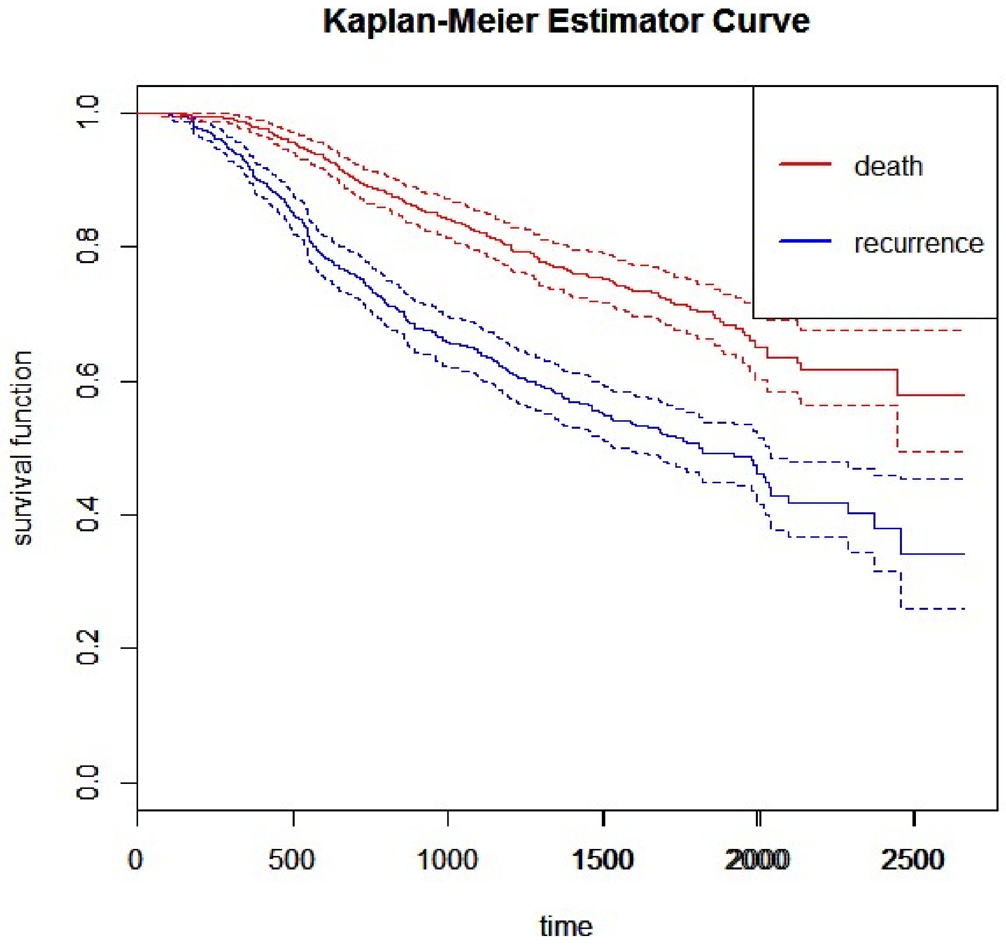


The solid line represents the survival probability whereas the dotted lines above and below the solid line represents upper 95% Confidence Interval and lower 95% Confidence Interval.

Another dataset we got to analyse is “German Breast Cancer Study”. The main difference of this analysis with the previous one is this contains two curves in a single plot. Unlike the previous dataset which contain only death information, it contains both disease recurrence and death information. The code in “R” is given below:

```
1. library(survival)
2. attach(gbcs)
3. gbcs.sv<-Surv(rectime,censrec,type="right")
4. gbcsSurv<-survfit(gbcs.sv~1,data=gbcs)
5. summary(gbcsSurv)
6. x<-plot(gbcsSurv,main="Kaplan-Meier Estimator
   Curve",col="blue",xlab="time")
7. gbcs.sv<-Surv(survtime,censdead,type="right")
8. gbcsSurv<-survfit(gbcs.sv~1,data=gbcs)
9. par(new=TRUE)

10.y<-plot(gbcsSurv,main="Kaplan-Meier Estimator
Curve",col="red",xlab="time",ylab="survival
function")11.legend('topright',c("recurrence","death"),lty=c
(1,1),lwd=c(2.5,2.5),col=c("blue","red"))
```



CONCLUSION:

Survival Data Analysis, as we have seen, provides a clear view of survival chance of a patient at a particular time taking into account his previous health status. This is a great help for doctors, health researchers and for humanity in general. So this study is a small step towards understanding the fundamentals of Survival Analysis. Censoring which is common in day to day life is discussed thoroughly in this paper. All the basic terminology and notations regarding survival analysis are given emphasis.

To conclude, Kaplan-Meier method is a clever method of statistical treatment of survival times which not only makes proper allowances for those observations that are censored, but also makes use of the information from these subjects up to the time when they are censored. There are many more software like SAS, STATA etc., to model the survival data and predict the survival function. As this is an introductory paper, more advanced topics are not discussed.

REFERENCES :

- https://en.wikipedia.org/wiki/Survival_analysis
- https://en.wikipedia.org/wiki/Kaplan%E2%80%93Meier_estimator
- http://pages.stat.wisc.edu/~ifischer/Intro_Stat/Lecture_Notes/8_-_Survival_Analysis/8.2_-_Kaplan Meier_Formula.pdf
- https://www.umass.edu/statdata/statdata/data/
- “Medical Statistics: Survival Data” by Nick Fieller.
- “Survival Analysis – A Self Learning Text” by David G. Kleinbaum and Mitchel Klein.