Estimation of Survival Distribution Using R Software

M.Ramakrishnan*, R.Ravanan**

* Department of Mathematics, RKM Vivekananda College, Chennai 600004
** Department of Statistics, Presidency College, Chennai 600005

Abstract- Survival analysis is widely applied in many fields such as biology, medicine, and public health. A typical analysis of survival data involves the modeling of time-to-event data, such as the time until death. To find Survival probabilities of the given observations censored or not, use Kaplan-Meier methods of estimation. But this method gives survival probabilities at any specific time. So this method does not compare total survival experience of two survival group. The non-parametric test, log rank test takes the complete follow up period and testing the significance difference between two survival distributions. In this paper, R software is used for finding survival (remission) probabilities and testing survival (remission) distributions using log rank test for 30 Resected Melanoma Patients.

Index Terms- Kaplan-Meier estimation, log rank test, R Software, Resected Melanoma Patients

I. Introduction

R Package is used to do Survival analysis. A lot of functions for survival analysis are in the package **Survival**. First install R and load the package **survival**. A step function with jumps at the observed event times will be obtained by using Kaplan-Meier method to estimate the survival function. This estimation takes censored and uncensored observations information to find out survival probabilities. Survival up to any point of time is calculated as the product of the conditional probabilities of surviving each time interval. This method of survival distribution is also obtained using R package.

The problem of testing survival distributions arises often in medical research. Survival curve gives rough idea about the distributions. But researcher expects significant difference in two or more treatments to prolong life of maintain health. So statistical test like log rank is necessary to find out significance difference exists or not between two survival distributions.

II. Non-Parametric estimation and Non-Parametric test

2.1. Non Parametric estimation – Kaplan-Meier method of estimation.

Let n be the total number of individuals whose survival times, censored or not, are available. Relabeling the survival times in order of increasing magnitude such that $t_1 \leq t_2 \leq \cdots .. t_n$ and the values of r are consecutive integers 1,2,….,n if there are no censored observation . If there are censored observations, they are not. Then the survival probabilities are calculated using S (t) = $\prod_{t_r \leq t} \frac{(n-r)}{(n-r+1)}$, where r runs through those positive integers for which $t_r \leq t$ and t_r is uncensored.

2.2 Non-Parametric Test - The Log-rank Test

Let d_t be the number of deaths at time t and n_{1t} and n_{2t} be the numbers of patients still exposed to risk of dying at time up to t in the two treatment groups. The expected deaths for groups 1 and 2 at time t are

$$e_{1t} = \frac{n_{1t}}{n_{1t} + n_{2t}} * d_t, \quad e_{2t} = \frac{n_{2t}}{n_{1t} + n_{2t}} * d_t$$

Then the total numbers of expected deaths in the two groups $E_1 = \sum e_{1t}, E_2 = \sum e_{2t}$. Let O_1 and O_2 be the observed numbers and E_1 and E_2 the expected numbers of death in two treatment groups.

$$The\ Test\ statistic\ \chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$ has approximately the chi-square distribution with one degree of freedom. A large χ^2 value (e.g., $\geq \chi^2_{1,.05}$) would lead to the rejection of the null hypothesis in favor of the alternative that the two treatments are not equally effective at α= 0.05.

2.3 *R Software*

R software is used to find survival probabilities, survival curves and testing significant difference between two survival distributions using log rank test. First install package survival using

>install.packages('survival')

To load libraries, use

>library(survival)

III. Example

Thirty melanoma patients were studied to compare the immunotherapics BCG (Bacillus Calmette –Guerin) and Coryne bacterium parvum for their abilities to prolong remission time.

First create data set with remission time event (Censor = 0, uncensored = 1), sex (Male = 1, female = 2),treat(BCG = 1, C.Parvum = 2)

To find survival probabilities using Kaplan-Meier method of estimation use

> library(survival)
Loading required package: splines

>rtime=c(33.7,3.8,6.3,2.3,6.4,23.8,1.8,5.5,16.6,33.7,17.1,4.3,26.9,21.4,18.1,5.8,3.0,11.0,22.1,23.0,6.8,10.8,2.8,9.2,15.9,4.5,9.2,8.2,8.2,7.8)
>
revent=c(0,1,1,1,1,0,1,1,0,0,0,1,0,0,0,1,1,0,1,0,1,0,1,1,1,1,1,0,0,0)
> sex=c(2,2,1,2,1,2,2,1,1,2,2,1,1,1,1,2,1,2,2,1,1,2,2,1,1,2,2,1,1,1,1,2,2,2)
>
treat=c(1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2)
> data=data.frame (rtime, revent,sex,treat)
> fit1=survfit(Surv(rtime,revent)~1,data=data)
> summary(fit1)
Call: survfit(formula = Surv(rtime, revent) ~ 1, data = data)

Table 1: Remission Probabilities of melanoma patients obtained through K-M estimate

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1.8	30	1	0.967	0.0328	0.905	1.000
2.3	29	1	0.933	0.0455	0.848	1.000
2.8	28	1	0.900	0.0548	0.799	1.000
3.0	27	1	0.867	0.0621	0.753	0.997
3.8	26	1	0.833	0.0680	0.710	0.978
4.3	25	1	0.800	0.0730	0.669	0.957
4.5	24	1	0.767	0.0772	0.629	0.934
5.5	23	1	0.733	0.0807	0.591	0.910
5.8	22	1	0.700	0.0837	0.554	0.885
6.3	21	1	0.667	0.0861	0.518	0.859
6.4	20	1	0.633	0.0880	0.482	0.832
6.8	19	1	0.600	0.0894	0.448	0.804
9.2	15	2	0.520	0.0937	0.365	0.740
15.9	11	1	0.473	0.0964	0.317	0.705
22.1	6	1	0.394	0.1078	0.230	0.674

This table shows corresponding survival (remission) probability, Standard error, lower and upper limits for 95% condifidence interval
For Survival curves use
> plot(fit1,xlab="time in months",ylab="Survival Probability")

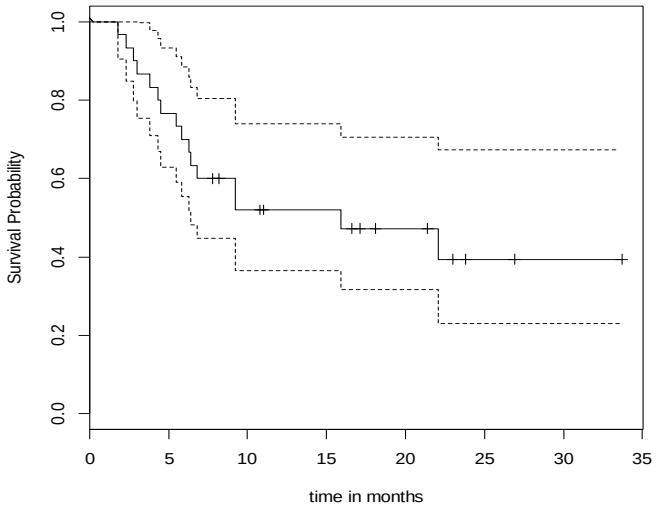

Figure 1. KM survival curves with 95% Confidence interval for melanoma patients

> fit3=survfit(Surv(rtime,revent==1)~sex,data=data)
summary(fit3)
Call: survfit(formula = Surv(rtime, revent == 1) ~ sex, data = data)

Table 2: Remission Probabilities of melanoma patients obtained through K-M estimate related with sex

sex=1

time	*n.risk*	*n.event*	*survival*	*std.err*	*lower 95% CI*	*upper 95% CI*
3.0	*15*	*1*	*0.933*	*0.0644*	*0.815*	*1.000*
4.3	*14*	*1*	*0.867*	*0.0878*	*0.711*	*1.000*
4.5	*13*	*1*	*0.800*	*0.1033*	*0.621*	*1.000*
5.5	*12*	*1*	*0.733*	*0.1142*	*0.540*	*0.995*
6.3	*11*	*1*	*0.667*	*0.1217*	*0.466*	*0.953*
6.4	*10*	*1*	*0.600*	*0.1265*	*0.397*	*0.907*
6.8	*9*	*1*	*0.533*	*0.1288*	*0.332*	*0.856*
9.2	*8*	*2*	*0.400*	*0.1265*	*0.215*	*0.743*
15.9	*6*	*1*	*0.333*	*0.1217*	*0.163*	*0.682*

sex=2

time	*n.risk*	*n.event*	*survival*	*std.err*	*lower 95% CI*	*upper 95% CI*
1.8	*15*	*1*	*0.933*	*0.0644*	*0.815*	*1.000*
2.3	*14*	*1*	*0.867*	*0.0878*	*0.711*	*1.000*
2.8	*13*	*1*	*0.800*	*0.1033*	*0.621*	*1.000*
3.8	*12*	*1*	*0.733*	*0.1142*	*0.540*	*0.995*
5.8	*11*	*1*	*0.667*	*0.1217*	*0.466*	*0.953*
22.1	*4*	*1*	*0.500*	*0.1708*	*0.256*	*0.977*

>plot(fit3, xlab="time in months",ylab="Survival Probability")

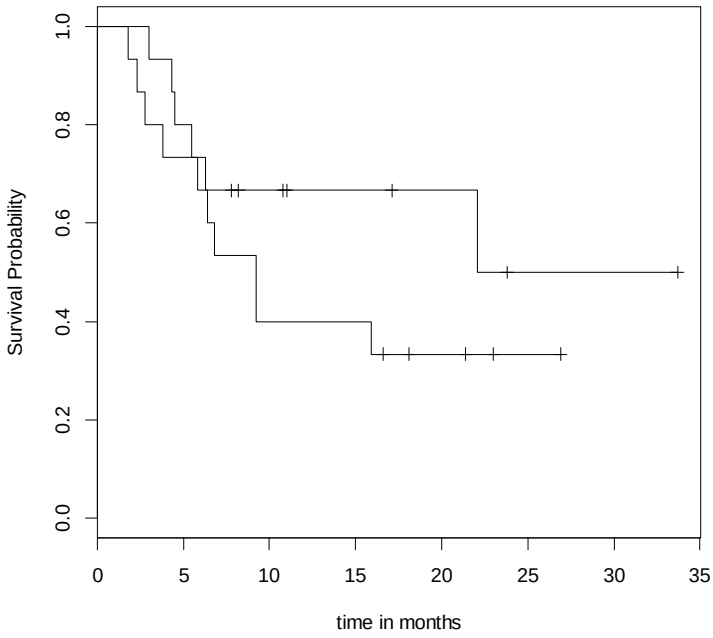

Figure 2. KM survival curves for remission times of patients compared with sex

Testing survival distributions related with sex using Log-rank test

> fit5=survdiff(Surv(rtime,revent)~sex,data=data,rho=0)
> fit5
Call:
survdiff(formula = Surv(rtime, revent) ~ sex, data = data, rho = 0)

	N	Observed	Expected	(O-E)^2/E
sex=1	15	10	8.21	0.389
sex=2	15	6	7.79	0.410

Chisq= 0.8 on 1 degrees of freedom, p= 0.367
> fit7=survfit(Surv(rtime,revent==1)~treat,data=data)

> summary(fit7)
Call: survfit(formula = Surv(rtime, revent == 1) ~ treat, data = data)

Table 3: Remission Probabilities of melanoma patients obtained through K-M estimate related with treatment type.

treat=1

time	*n.risk*	*n.event*	*survival*	*std.err*	*lower 95% CI*	*upper 95% CI*
1.8	*11*	*1*	*0.909*	*0.0867*	*0.754*	*1.000*
2.3	*10*	*1*	*0.818*	*0.1163*	*0.619*	*1.000*
3.8	*9*	*1*	*0.727*	*0.1343*	*0.506*	*1.000*
5.5	*8*	*1*	*0.636*	*0.1450*	*0.407*	*0.995*
6.3	*7*	*1*	*0.545*	*0.1501*	*0.318*	*0.936*
6.4	*6*	*1*	*0.455*	*0.1501*	*0.238*	*0.868*

treat=2

time	*n.risk*	*n.event*	*survival*	*std.err*	*lower 95% CI*	*upper 95% CI*
2.8	*19*	*1*	*0.947*	*0.0512*	*0.852*	*1.000*
3.0	*18*	*1*	*0.895*	*0.0704*	*0.767*	*1.000*
4.3	*17*	*1*	*0.842*	*0.0837*	*0.693*	*1.000*
4.5	*16*	*1*	*0.789*	*0.0935*	*0.626*	*0.996*
5.8	*15*	*1*	*0.737*	*0.1010*	*0.563*	*0.964*
6.8	*14*	*1*	*0.684*	*0.1066*	*0.504*	*0.929*
9.2	*10*	*2*	*0.547*	*0.1215*	*0.354*	*0.846*
15.9	*6*	*1*	*0.456*	*0.1311*	*0.260*	*0.801*
22.1	*3*	*1*	*0.304*	*0.1518*	*0.114*	*0.809*

> plot(fit7, xlab="time in months",ylab="Survival Probability")

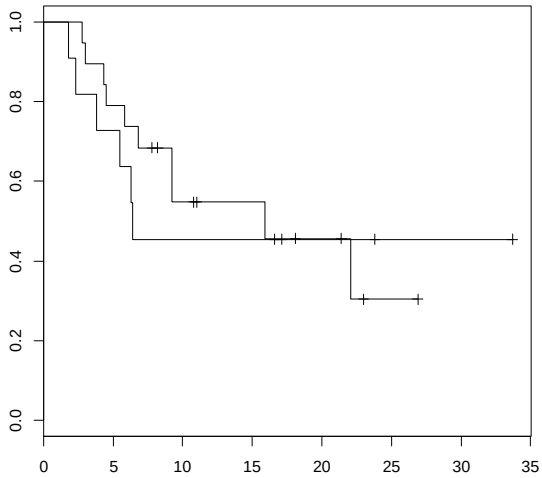

Figure 3. KM survival curves for remission times related with treatment type

```
>  fit7=survdiff(Surv(rtime,revent)~treat,data=data,rho=0)
> fit7
Call:
survdiff(formula = Surv(rtime, revent) ~ treat, data = data,
    rho = 0)

             N  Observed  Expected  (O-E)^2/E
treat=1     11         6      5.55     0.0366
treat=2     19        10     10.45     0.0194

 Chisq= 0.1  on 1 degrees of freedom, p= 0.811
```

IV. DISCUSSION

Figure 1 Shows survival probabilities for remission time of all patients in survival curve form. It shows the relation between survival (remission) probabilities and time for each observation. Figure 2 shows the survival curves of male and female patients. This figure also gives the idea about survival probabilities of male and female patients. Survival curve for male is differing from the survival curve for female. In early period two curves are overlapped.

Figure 3 shows the survival curve of patient receiving BCG and C. Parvum treatment. It also shows some difference in survival curves. But these significantly difference may be check using log-rank test.

When using log-rank test regarding sex of the patients, the value of p is 0.367>0.05.So accept null hypothesis is that the survival distribution of male patient is same as that of female patients.

When using log-rank test regarding treatment type of the patients, the value of p is 0.811>0.05.So accept null hypothesis is that the survival distribution of BCG treatment receiving patients is same as that of C.Parvum treatment receiving patients.

V. CONCLUSION

We obtained Survival Probabilities, Survival curve and test value using log-rank test are obtained by the use of R Statistical software which is freely available. Kaplan-Meier method of estimating survival curves only gives pictorial representation of the survival distributions but it does not take whole follow-up period into account. Survival curves related to sex and related to type of treatment shows slight difference in the survival curves of corresponding survival distributions but log-rank test determines there is no significant difference between these curves for both related to sex and related to treatment type. So Log-rank test is used to test the null hypothesis that there is no significant difference between the two survival distributions in the probability of an event at any time point.

REFERENCES

[1] Altman DG, Bland JM.(1998),” Time to event (survival) data”. British Medical Journal (BMJ); vol. 317: pp. 468-469.

[2] Altman DG, Bland JM. .(1998), “Survival probabilities (The Kaplan-Meier method)”. BMJ; vol. 317; pp. 1572 - 1580.

[3] Altman DG, Bland JM. (2004),” The log rank Test”, BMJ; vol. 328 ; pp. 1073.

[4] Collett D. (2003), “Modeling of Survival Data in Medical Research”. Chapman Hall, London, U.K.

[5] Elisa T. Lee. (1992), “Statistical methods for Survival Data Analysis”. Second Edition. A Wiley-Inter science publication, United States of America.

[6] Kaplan, E. L., and Paul Meier (Jun. 1958),”Non parametric Estimation from Incomplete Observations”, Journal of the American Statistical Association. Vol. 53, pp457 – 481.

[7] Rupert G. Miller J R (1981), “Survival Analysis”, John Wiley & Sons, United States of America

AUTHORS

First Author – M. Ramakrishnan ,M.Sc., M.Phil., M.B.A., P.G.D.C.A., RKM Vivekananda College, Chennai, email-mramkey69@gmail.com.

Second Author – Dr. R. Ravanan, M.Sc., M.Phil., Ph.D., Presidency College, Chennai. Email-ravananstat@gmail.com.