Vietnam Institute for Advanced Study in Mathematics

Survival analysis

Practical work 2: Non parametric estimations

(Lecturers: Agnès LAGNOUX & Jean-François DUPUY)

Exercise 1: Familiarization with the function `survfit`

The goal of this first exercise is to get familiarized with the function `survfit` of the software `R` that provides the Kaplan-Meier and Nelson-Aalen estimations of the survival function and of the cumulative hazard rate function. We work on the example `lung` already in `R`.
Write and comment the following commands:

```
library(survival)
help(lung)
kmfit<-survfit(Surv(time,status)~ 1,data=lung,conf.type="plain",type=’kaplan-meier’)
print(kmfit)
summary(kmfit)
plot(kmfit)
plot(kmfit,mark.time=F,xscale=365.25,xlab="Time (in years)",ylab="Survival S(t)")
legend(1,0.8, c("Kaplan-Meier function", "95\% pointwise CI"), lty=1:2)
fhfit<-survfit(Surv(time,status)~1,data=lung,conf.type="plain",type=’fh’)
plot(kmfit,mark.time=F,xscale=365.25,xlab="Time (in years)",ylab="Survival S(t)")
lines(fhfit,lty=3,mark.time=F,xscale=365.25,col="red")
plot(kmfit\$time,kmfit\$surv-fhfit\$surv)
naH =-log(fhfit\$surv)
time= fhfit\$time
plot(time,naH,type="s",ylab="Cumulative risk H(t)",xlab="Time (in months)")
```

Exercise 2

The following data come from a clinical trial led by Freireich, in 1963. The goal was to compare the remission durations (in weeks) of patients that suffer from leukemia. The patients are divided into two subgroups: some of them received a medicine (6-MP) and the others a placebo. The results are presented in the following tabular:

6-MP	6	6	6	6^+	7	9^+	10	10^+	11^+	13	16
	17^+	19^+	20^+	22	23	25^+	32^+	32^+	34^+	35^+	
Placebo	1	1	2	2	3	4	4	5	5	8	8
	8	8	11	11	12	12	15	17	22	23	

The patients with a + sign correspond to lost subjects at the considered time of observation: they are censored, ”excluded-alive” of the study and one only knows that their remission duration is greater than the observed delay.

1. Compute the Kaplan-Meier estimator of the survival function S. Estimate its variance.

   One may use the following tabular to lead the calculus:

Time of relapse	Number of relapses	Censoring in $[T_{(i)}, T_{(i-1)}[$	At risk numbers at $T_{(i)}$	Conditional probability	Survival probability without relapse
$T_{(i)}$	m_i	c_{i-1}	n_i	$(n_i - m_i)/n_i$	$\hat{S}_{n,KM}(T_{(i)})$
Placebo					
1					
2					
3					
4					
5					
8					
76					
11					
12					
15					
17					
22					
23					
6-MP					
6					
7					
10					
13					
16					
22					
23					

2. Recover the previous results using `R` and plot the graph of the estimated survival function with respect to the time.

3. In the group that has received the placebo, the remission times are:

   $$1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23.$$

   Estimate using `R` the survival functions in each group using the Kaplan-Meier estimator and plot them.

4. Determine the Breslow and the Nelson-Aalen estimation of the cumulative hazard rate function H.

   One may use the previous and following tabulars to lead the calculus:

Time of relapse	Number of relapses	At risk numbers at $T_{(i)}$	Nelson proportion h	Nelson estimation	Kaplan-Meier estimations	
$T_{(i)}$	m_i	n_i	m_i/n_i	$\hat{H}_{n,NA}(T_{(i)})$	$\hat{S}_{n,KM}(T_{(i)})$	$\hat{H}_{n,BR}(T_{(i)})$
Placebo						
1						
2						
3						
4						
5						
8						
76						
11						
12						
15						
17						
22						
23						
6-MP						
6						
7						
10						
13						
16						
22						
23						

Exercise 3

From February 1998 to February 2001, 29 patients that suffered from a severe viral hepatitis were admitted in a therapeutic trial of 16 weeks. The goal was to compare the effect of a therapy with steroids. The patients received randomly the treatment or the placebo. The survival times (in weeks) of the goups of the 14 patients treated are

$$1, 1, 1, 1^+, 4^+, 5, 7, 8, 10, 10^+, 12^+, 16^+, 16^+16^+.$$

1. No assumption has been done on the survival time distribution under treatment.
   (a) Estimate cumulative risk H with the Nelson-Aalen estimator.
   (b) Deduce the Harrington and Fleming estimator of S.
   (c) Determine the Kaplan-Meier estimator of S.
   (d) Represent these two estimators of S on a same figure using `R`.
2. Now we assume that the survival time is exponentially distributed with parameter λ.
   (a) Estimate λ by the maximum likelihood method.
   (b) Deduce the estimation of the probability to survive more than 16 weeks.
   (c) Estimate the median of the survival time.
3. Represent these three estimators of S on a same figure using `R`. Comment.

Exercise 4

1. Generate a sample of size 100 of a random variable X exponentially distributed with parameter $\lambda = 1.1$. Represent on the same figure the theoretical and empirical survival functions of X using `R`.

2. Generate a sample of size 100 of the pair $(T = \min(X, C), \delta)$, where $X \sim \mathcal{E}(1.1)$, $C \sim \mathcal{E}(1)$ and $\delta = 1\!\!1_{\{X \leqslant C\}}$.

   (a) Compute the Kaplan-Meier survival function estimation obtained considering the whole observations.

   (b) Determine the estimation of the survival function by the maximum likelihood method on the whole observations.

   (c) Represent on the same figure the theoretical survival function of X, its Kaplan-Meier estimation and the MLE estimation.

3. Select the uncensored observations.

   (a) Compute the Kaplan-Meier survival function estimation obtained considering the uncensored sample.

   (b) Estimate the survival function by the maximum likelihood method on the uncensored observations.

   (c) On the previous figure, represent these two functions.

4. Same question by making vary the sample size. Conclusion?

5. CI comparisons

   We work on the whole sample. Represent on the same figure the theoretical survival function of X and its Kaplan-Meier estimation.
   Add the confidence intervals of types “plain”, “log” and “log-log” for $S(t)$ on three different figures.
   To which formulas do these intervals correspond?
   Conclusion?