Vietnam Institute for Advanced Study in Mathematics

Survival analysis

Practical work 1: Introduction to survival data analysis

(Lecturers: Agnès LAGNOUX & Jean-François DUPUY)

Exercise 1: Familiarization with the Weibull distribution

1. With the software R, plot on a same figure the hazard rate functions for the Weibull distribution for different sets of parameters α and λ.
2. Same question with the survival functions.
3. Generate a sample of size $n = 100$ of Weibull random variables.
4. Determine the empirical mean and the empirical variance. Compare these values to the theoretical ones.

Exercise 2: Construction of a sample of right censored data
Let X be a Weibull random variable $\mathcal{W}(\alpha, \lambda)$ and let C be censoring uniformly distributed over $[0, c]$. Assume that X and C are independent.
Define $T = \min(X, C)$ and δ as the indicator of the event $\{X \leqslant C\}$: $\delta = \mathbb{1}_{\{X \leqslant C\}}$.

1. Create a R function that generates a n-sample of independent realizations $(T_i, \delta_i)_{1 \leqslant i \leqslant n}$ of (T, δ) and determines the rate τ of censored data.
2. Make vary the point date c to observe it influence on the censoring rate τ.
3. Determine theoretically τ as a function of c and h in the case $\alpha = 1$ (that means that $X \sim \mathcal{E}(\lambda)$).
4. When $\lambda = 1$, which point date c one should to choose to get a theoretical censoring rate of 20%? of 50%? You may use the minimizing functions of R. Sample data and check that with these values of c, the censoring rates obtained by simulation are close to the theoretical ones.

Exercise 3: Maximum likelihood of a right censored model
Let X be a random variable exponentially distributed $\mathcal{E}(\lambda)$ and C a right random censoring also exponentially distributed $\mathcal{E}(\theta)$. Assume that X and C are independent.
Define $T = \min(X, C)$ and δ as the indicator of the event $\{X \leqslant C\}$: $\delta = \mathbb{1}_{\{X \leqslant C\}}$.

1. Determine the distribution of δ.
2. Determine the distribution of T.
3. Prove that T and δ are independent.

4. Now let $(T_i, \delta_i)_{1 \leqslant i \leqslant n}$ be n independent replicas of (T, δ) and $(X_i)_{1 \leqslant i \leqslant n}$ n independent replicas of X.

   (a) Determine the Fisher information provided on λ by the sample $(T_i, \delta_i)_{1 \leqslant i \leqslant n}$ and then by the sample $(T_i)_{1 \leqslant i \leqslant n}$. Comment.

   (b) Determine the maximum likelihood estimator $\hat{\lambda}_n$ of λ using the sample $(T_i, \delta_i)_{1 \leqslant i \leqslant n}$.

   (c) Determine the maximum likelihood estimator $\hat{\lambda}_n^*$ using the sample $(T_i)_{1 \leqslant i \leqslant n}$.

   (d) We want to compare $\hat{\lambda}_n$ and $\hat{\lambda}_n^*$. Using the results of the previous questions, determine the expectation of $\hat{\lambda}_n$ and $\hat{\lambda}_n^*$ and deduce $\mathrm{Var}(\hat{\lambda}_n)$ and $\mathrm{Var}(\hat{\lambda}_n^*)$. Then compute the ration $\mathrm{Var}(\hat{\lambda}_n)/\mathrm{Var}(\hat{\lambda}_n^*)$. Conclude.

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 proba-bility	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 $\delta = \mathbb{1}_{\{X \leqslant C\}}$.

   (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?

VIASM INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES INSTITUT de MATHEMATIQUES de TOULOUSE UNIVERSITÉ TOULOUSE Jean Jaurès

Vietnam Institute for Advanced Study in Mathematics

Survival analysis

Practical work 3: Logrank tests

(Lecturers: Agnès LAGNOUX & Jean-François DUPUY)

Exercise 1: Logrank test

We want to realize the logrank test on the data of Freireich (presented in the course). Remind that Freireich, in 1963, realized a therapeutic trial in order 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.
One may use the following tabular to lead the different tests

	6-MP		Placebo			Weights w_i		
						Logrank	Gehan	Peto-Prentice
$T_{(i)}$	m_{Ai}	n_{Ai}	m_{Bi}	n_{Bi}	e_{Bi}	1	n_i	S_i^*
1								
2								
⋮								
23								

and present the results in

Test	Test stat.	p	RR_1
Logrank LR^2			
Approx. Logrank LRA^2			
Gehan			
Peto-Prentice			

Exercise 2: Logrank test

The following tabular presents the survival times after mastectomy of 45 women that suffers from a breast cancer. The patients are divided into two groups according to the presence or not of metastases. A + indicates a censored data. You can find the data in `breast.txt`.

No metastases	23	47	69	70+	71+	100+	101+	148	181	198+	208+	212+	224+
	5	8	10	13	18	24	26	26	31	35	40	41	48
Metastases	50	59	61	68	71	76+	105+	107+	109+	113	116+	118	143
	154+	162+	188+	212+	217+	225+							

1. Determine using `R` the Kaplan-Meier estimations of the survival functions in each group. Plot these estimations on the same figure adding the confidence intervals.

2. Compare the survival of the two groups using the the classical logrank test (function `survdiff`).

3. Use now the weighted logrank test with $w_i = \hat{S}_{KM}(t_i)$ (obtained for $\rho = 1$) to realize another comparison. Conclusion?

Exercise 3: Stratified logrank test

We consider a clinical trial conducted by Peto (1979) on comparison of the survival functions of two groups. We have an extra information: the kidney function that is known to influence the survival:

Participation time	Group	Kidney function	Participation time	Group	Kidney function
8	A	A	220	A	N
8	A	N	365^+	A	N
13	B	A	632	B	N
18	B	A	700	B	N
23	B	A	852^+	A	N
52	A	A	1296	B	N
63	A	A	1296^+	A	N
63	A	A	1328^+	A	N
70	B	N	1460^+	A	N
76	B	N	1976^+	A	N
180	B	N	1990^+	B	N
195	B	N	2240^+	B	N
210	B	N			

The letter A (respectively N) means an abnormal (resp. N) kidney function. The censored data are indicated by a +.

1. Check by the logrank test that the kidney function influences the survival. You can also plot the survival Kaplan-Meier function according to the kidney function.

2. Compare using a logrank test the survival functions of the two groups. Validate your results using the argument `subset` of the function `survdiff`. Conclusion?

3. Compare using a logrank test the survival functions of the two groups, separately for the patients with a normal kidney function and the patients with an abnormal one. Validate your results using the argument `subset` of the function `survdiff`. Conclusion?

4. Compare using a logrank test the survival functions of the two groups, stratifying on the kidney function. Validate your results using the option `+strata` of the function `survdiff`. Conclusion?

One may use the following tabular to lead the results analytically:

Death times	Treatment								Kidney function					
	A				B									
	N		AN		N		AN		N		AN		Total	
t_i	m_{A_i}	n_{A_i}	m_{A_i}	n_{A_i}	m_{B_i}	n_{B_i}	m_{B_i}	n_{B_i}	e_{B_i}	v_{B_i}	e_{B_i}	v_{B_i}	e_{B_i}	v_{B_i}
8														
13														
18														
23														
52														
63														
70														
76														
180														
195														
210														
220														
632														
700														
1296														
Total														

Exercise 4: Comparison of three subgroups

The data analyzed in this example concern three small (fictive) samples corresponding to the three different treatment doses (Thomas 1977). The survival and censoring are represented in the following tabular.

Group	N_j	Dose x_j	Survival and censoring									
A	9	0	73^+	74^+	75^+	76	76	76^+	99	166	246^+	
B	10	1.5	43^+	44^+	45^+	67	68^+	136	136	150	150	150
C	10	2	41^+	41^+	47	47^+	47^+	58	58	58	100^+	117

The censored data are indicated by a +.

One may use the following tabular to compute the heterogeneity and trend statistics.

	Group										
	A		B		C						
t_i	m_{A_i}	n_{A_i}	m_{B_i}	n_{B_i}	m_{C_i}	n_{C_i}	e_{B_i}	v_{B_i}	e_{C_i}	v_{C_i}	c_{B_i,C_i}
47											
58											
67											
76											
99											
117											
136											
150											
166											
Total											

The expectations $e.$, variances $v.$ and covariances $c_{.,.}$ of the death numbers under H_0 are indicated only for groups B and C, since these quantities are not necessary to the computation of the statistics. The reader can compute E_A and check that

$$E_A + E_B + E_C = O_A + O_B + O_C = 15.$$