

Durées de vie – Approche non paramétrique

Exercice 1. (*Estimateur de Kaplan-Meier*)

Voici les durées de vie en jours de 21 patients atteints d'une infection virale : 6, 6, 6, 6+, 7, 9+, 10, 10+, 11+, 13, 16, 17+, 19+, 20+, 22, 23, 25+, 32+, 32+, 34+, 35+.

Représentez l'estimateur de Kaplan-Meier de la fonction de survie de cette population.

$t_{(i)}$: statistique d'ordre de rang i dans la série des observations expurgées des répétitions et rangées par ordre croissant ; d_i : nombre de décès observé en $t_{(i)}$; r_i : exposition au risque en $t_{(i)}$ (nombre de vivants juste avant cet instant). Ainsi : $t_{(1)} = 6$, $t_{(2)} = 7$, $t_{(3)} = 9$, $t_{(4)} = 10$, etc. Par ailleurs : $d_1 = 3$ puisque trois décès sont observés en $t_{(1)} = 6$ (6+ est une censure en 6, pas un décès). Enfin : $r_4 = 15$ puisque 15 sujets sont vivants juste avant l'instant $t_{(4)} = 10$.

Tableau TDR et estimateur $\hat{S}_{KM}$				
$t_{(i)}$	d_i	r_i	$1 - d_i/r_i$	$\prod_{j:t_{(j)} \leq t_{(i)}} (1 - d_j/r_j)$
6	3	21	0,8571	0,8571
7	1	17	0,9412	0,8067
10	1	15	0,9333	0,7529
13	1	12	0,9167	0,6902
16	1	11	0,9091	0,6275
22	1	7	0,8571	0,5378
23	1	6	0,8333	0,4482

$\hat{S}_{KM}(t) = 1$ si $t < 6$
 0,8571 si $6 \leq t < 7$
 0,8067 si $7 \leq t < 10$
 0,7529 si $10 \leq t < 13$
 0,6902 si $13 \leq t < 16$
 0,6275 si $16 \leq t < 22$
 0,5378 si $22 \leq t < 23$
 0,4482 si $t \geq 23$.

Remarque. La limite en $+\infty$ de la fonction $\hat{S}_{KM}(t)$ ne vaut pas 0 ; ainsi, $\hat{S}_{KM}$ n'est pas une fonction de survie à proprement parler. Plus généralement : pour que la limite de $\hat{S}_{KM}(t)$ en $+\infty$ soit 0, il faut et il suffit que le dernier événement ne soit pas une censure.

Un script R définissant et représentant $\hat{S}_{KM}$

```
par(mfrow=c(1,1))
t=c(6,7,10,13,16,22,23)
d=c(3,1,1,1,1,1,1)
r=c(21,17,15,12,11,7,6)
cq=1-d/r
s=cumprod(cq)
tab <- data.frame(t=t,d=round(d,4),r=round(r,4),cq=round(cq,4),s=round(s,4))
plot(stepfun(tab$t,c(1,tab$s)),right=FALSE,col='blue',
main='Fonction de survie estimee (Kaplan-Meier)',xlab='temps',ylab='probabilite')
print(tab)
```

Un autre script utilisant la fonction `survfit{survival}` de R :

```
library(survival)
sdat <- data.frame(
time=c(6, 6, 6, 6, 7, 9, 10, 10, 11, 13, 16, 17, 19, 20, 22, 23, 25, 32, 32, 34, 35) ,
status=c(1,1,1,0,1,0,1,0,0,1,1,0,0,0,1,1,0,0,0,0,0))
km <- survfit(Surv(time, status) ~ 1,data=sdat,type="kaplan-meier",conf.type='none')
KME <- stepfun(km$time,c(1,km$surv),right=FALSE)
print(summary(km))
plot(KME,col='red',main='Fonction de survie estimee (Kaplan-Meier)',xlab='temps',
ylab='probabilite')
```

Remarque. La courbe rouge obtenue avec `survfit` affiche les instants de censure.

Exercice 2. (*Estimateur de Nelson-Aalen*)

Déterminez et représentez le hasard cumulé estimé par Nelson-Aalen sur les données de Exercice 1.

Tableau TDR et estimateur $\hat{\Lambda}_{NA}$				
$t_{(i)}$	d_i	r_i	d_i/r_i	$\sum_{j:t_{(j)} \leq t_{(i)}} d_j/r_j$
6	3	21	0,143	0,143
7	1	17	0,059	0,202
10	1	15	0,067	0,269
13	1	12	0,083	0,352
16	1	11	0,091	0,443
22	1	7	0,143	0,586
23	1	6	0,167	0,753

$$\hat{\Lambda}_{NA}(t) = 0 \text{ si } t < 6$$

$$0,143 \text{ si } 6 \leq t < 7$$

$$0,202 \text{ si } 7 \leq t < 10$$

$$0,269 \text{ si } 10 \leq t < 13$$

$$0,352 \text{ si } 13 \leq t < 16$$

$$0,443 \text{ si } 16 \leq t < 22$$

$$0,586 \text{ si } 22 \leq t < 23$$

$$0,753 \text{ si } t \geq 23.$$

Un script R définissant et représentant $\hat{\Lambda}_{NA}$

```
t=c(6,7,10,13,16,22,23) # instants de deces
d=c(3,1,1,1,1,1,1) # nombre de deces
r=c(21,17,15,12,11,7,6) # exposition au risque
q=round(d/r,3) # ratios deces/exposition
h=cumsum(q) # hasard cumule
tab1 <- data.frame(t=t,d=d,r=r, q=q,h=h) # le tableau TDR + hasard cumule
#representation de l'estimateur NAE
step1 <- stepfun(tab1$t,c(0,tab1$h),right=FALSE)
plot(step1,lwd=2,col='blue',xlab='A la main',ylab='NAE',
main='Fonction de hasard cumule : Nelson-Aalen')
#le tableau TDR
print(tab1)
```

Un autre script utilisant la fonction `survfit{survival}` de R :

```
library(survival)
sdat <- data.frame( # donnees brutes
time=c(6, 6, 6, 6, 7, 9, 10, 10, 11, 13, 16, 17, 19, 20, 22, 23, 25, 32, 32, 34, 35) ,
status=c(1,1,1,0,1,0,1,0,0,1,1,0,0,0,1,1,0,0,0,0,0))

tab2 <- summary(survfit(Surv(time, status)~ 1, data=sdat, conf.type="none"))
print(tab2) # le tableau TDR
h <- cumsum(tab2$n.event/tab2$n.risk) # hasard cumule

# representation de l'estimateur NAE
NAE <- stepfun(tab2$time,c(0,h),right=FALSE)
plot(NAE,xlim=c(6,25),lwd=2,col='red',xlab='Avec survfit{survival}',ylab='NAE',main='')
```

Exercice 3. (Intervalle de confiance des valeurs de la fonction de survie)

Table 1 indique la durée de vie de dix patients observés pendant trente jours, ainsi que l'existence (statut=0) ou non (statut=1) d'une censure à droite.

patient	1	2	3	4	5	6	7	8	9	10
durée	6,4	11,0	13,4	13,4	15,8	16,8	16,8	16,8	20,6	30
statut	1	0	1	1	0	1	1	1	1	1

Table 1: durée de vie de dix patients observés pendant dix jours

Représentez l'estimateur de Kaplan-Meier de la fonction de survie ainsi que les intervalles de confiance à 95% des valeurs de cette fonction.

Notons S la fonction de survie de la durée de vie d'un sujet tiré de la population. Le script R suivant représente pour chaque $t > 0$:

(i) l'estimation $\hat{S}_{KM}(t)$ de $S(t)$

(ii) les bornes de l'intervalle de confiance à 95% de $S(t)$: $\hat{S}_{KM}(t) \pm z_{\alpha/2} \sqrt{\hat{V}(\hat{S}_{KM}(t))}$ où $\hat{V}(\hat{S}_{KM}(t))$ est l'estimation de la variance de $\hat{S}_{KM}(t)$ selon la formule de Greenwood.

```
library(survival)
sdat <- data.frame(
  time=c(6.4,11.0,13.4,13.4,15.8,16.8,16.8,16.8,20.6,30) ,
  status=c(1,0,1,1,0,1,1,1,1,1))
km <- survfit(Surv(time, status) ~ 1,data=sdat,type="kaplan-meier",conf.type='plain')
plot(km,main='Survival function estimate (Kaplan-Meier) & 95% confidence intervals')
```

La commande `print(summary(km))` donne (colonne `std.err`) les valeurs de $\sqrt{\hat{V}(\hat{S}_{KM}(t))}$ permettant de tracer les courbes de confiance.

Exercice 4. (*Estimateur de Kaplan-Meier : une application actuarielle*)

On vaccine cent patients contre la grippe, puis on les observe pendant 24 mois. Treize patients quittent l'étude à : 1, 3, 3, 4, 4, 5, 7, 7, 9, 10, 11, 13, 15 mois sans présenter de symptômes ; huit patients contractent la grippe au bout de : 2, 2, 2, 6, 6, 8, 8, 17 mois ; les soixante dix-neuf patients restants ne présentent toujours aucun symptôme à la fin de l'étude.

Dans un échantillon de référence composé de patients non vaccinés, la fonction de survie est définie par : $S_0(t) = 1$ si $0 \leq t < 5$; $S_0(t) = 0,92$ si $5 \leq t < 8$; $S_0(t) = 0,895$ si $8 \leq t < 12$; $S_0(t) = 0,87$ si $12 \leq t < 24$; $S_0(t) = 0$ si $t \geq 24$.

Chaque vaccin coûte 0,05 € et le traitement d'un malade coûte 1 € de façon forfaitaire au moment du diagnostic.

1. Déterminez l'estimation de Kaplan-Meier de la fonction de survie chez les patients vaccinés.

Notons T la durée avant qu'un patient vacciné n'attrape la grippe et $S(t)$ la fonction de survie de T .

Les observations sont : 1+, 2, 2, 2, 3+, 3+, 4+, 4+, 5+, 6, 6, 7+, 7+, 8, 8, 9+, 10+, 11+, 13+, 15+, 17, 24+, ..., 24+ (avec soixante-dix-neuf répétitions de la censure 24+).

L'estimation de Kaplan-Meier de $S(t)$ est : $\hat{S}_{KM}(t) = 1$ si $t < 2$; 0,970 si $2 \leq t < 6$; 0,948 si $6 \leq t < 8$; 0,927 si $8 \leq t < 17$; 0,915 si $t \geq 17$.

2. Comparez la valeur actuelle moyenne du coût de traitement d'un sujet grippé dans chacune des populations : vaccinée/non vaccinée, pour un taux d'intérêt annuel de 4%.

Le taux d'intérêt vaut 4% par an : 1 euro après t années vaut $1/1,04^t$ euro à l'instant 0.

La loi de probabilité du coût de traitement d'un patient vacciné :

k	1	2	3	4	5
Valeur de T : t_k	2	6	8	17	24+
Probabilité associée : p_k	0,03	0,022	0,021	0,012	0,915
Coût du traitement en euros	1	1	1	1	0
Valeur actuelle du coût du traitement : c_k	$1/1,04^{2/12}$	$1/1,04^{6/12}$	$1/1,04^{8/12}$	$1/1,04^{17/12}$	0

Remarque. Estimer $S(t)$ par la fonction en escalier $\hat{S}_{KM}(t)$ revient à assimiler T à une variable discrète. Or, pour une variable discrète, la valeur de la fonction de probabilité en un point est toujours la différence entre la limite à gauche de la fonction de survie en ce point et la valeur de la fonction de survie en ce point. Ainsi : $\forall t \in \mathbb{R}, \mathbb{P}(T = t) = \lim_{\substack{x \rightarrow t \\ x < t}} \hat{S}_{KM}(x) - \hat{S}_{KM}(t)$. Par exemple : $\mathbb{P}(T = 8) = \lim_{\substack{x \rightarrow 8 \\ x < 8}} \hat{S}_{KM}(x) - \hat{S}_{KM}(8) = 0,948 - 0,927 = 0,021$.

Ainsi, la valeur actuelle attendue du coût de traitement d'un patient vacciné est : $\sum_{k=1}^5 c_k p_k \approx 0,067$ euros auquel s'ajoute le coût fixe du vaccin : 0,05 euros. Ainsi, la valeur actuelle du coût global (vaccin et traitement) d'un patient vacciné est 0,117 euros environ.

La variable T_0 dont S_0 est la fonction de survie peut être interprétée comme la durée avant qu'un patient non vacciné n'attrape la grippe. Mais alors : $\mathbb{P}(T_0 = 24) = \lim_{\substack{x \rightarrow 24 \\ x < 24}} S_0(x) - S_0(24) = 0,87$ indique : 87% des non vaccinés attrapent la grippe au bout de 24 mois exactement. Et la loi de probabilité du coût de traitement d'un patient non vacciné est :

k	1	2	3	4
Valeur de T_0 : τ_k	5	8	12	24
Probabilité associée : π_k	0,08	0,025	0,025	0,87
Coût du traitement en euros	1	1	1	1
Valeur actuelle du coût du traitement : γ_k	$1/1,04^{5/12}$	$1/1,04^{8/12}$	$1/1,04^{12/12}$	$1/1,04^{24/12}$

La valeur actuelle attendue du coût de traitement d'un patient non vacciné est alors : $\sum_{k=1}^4 \gamma_k \pi_k \approx 0,931$ euros.

On peut aussi voir en T_0 la durée avant qu'un patient non vacciné n'attrape la grippe ou n'atteigne la fin de l'étude sans avoir attrapé la grippe ; et supposer que les 87% de sujets pour lesquels $T_0 = 24$ sont des patients

sortis de l'étude sans attraper la grippe. La loi de probabilité du coût de traitement d'un patient non vacciné est alors :

k	1	2	3	4
Valeur de $T_0 : \pi_k$	5	8	12	24
Probabilité associée : π_k	0,08	0,025	0,025	0,87
Coût du traitement en euros	1	1	1	0
Valeur actuelle du coût du traitement : γ_k	$1/1,04^{5/12}$	$1/1,04^{8/12}$	$1/1,04^{12/12}$	0

La valeur actuelle attendue du coût de traitement d'un patient non vacciné est alors : $\sum_{k=1}^4 \gamma_k \pi_k \approx 0,127$ euros.

Quelle que soit l'interprétation de T_0 , un patient non vacciné coûte en moyenne plus qu'un patient vacciné.

Exercice 5. (*Comparaison de trois estimateurs non paramétriques de la fonction de survie*)

Une durée de vie $X \sim \mathcal{E}(1)$ est censurée à droite par une variable Y uniforme entre 0 et $b = 2$.

- (a) Représentez la fonction de survie S de X entre les instants 0 et $b = 2$.

La fonction de survie de X est définie en $t > 0$ par : $S(t) = e^{-t}$; on la représente entre 0 et 2 sous R par :
`bb=2 ; S <- function(t){exp(-t)} ; plot(S,xlim=c(0,bb),ylim=c(0,1)).`

- (b) Générez un échantillon de taille 20 de la loi de $T = \min\{X, Y\}$.

```
nn=20 # taille de l'échantillon
x=rexp(nn,rate=1) # les durees de vie issues d'une loi exponentielle
c=runif(nn,0,bb) # les censures
complete <- data.frame(x=x,c=c) # les donnees completes
time<- apply(complete,1,min) # les observations
status <- (time==x)*1 # les indicatrices de non censure
observed <- data.frame(time=time,status=status) # observations et censures
```

- (c) Représentez l'estimateur (i) empirique (ii) de Kaplan-Meier (iii) de Harrington-Fleming de la fonction S .

n désigne la taille de l'échantillon et $t_1, \dots, t_n$ les observations. L'estimateur empirique de la survie donne en $t > 0$ la proportion de sujets ayant survécu à l'instant t : $\hat{S}_E(t) = (\sum_{i=1}^n \mathbb{1}_{\{t_i > t\}}) / n$.

$t_{(i)}$ désigne la statistique d'ordre de rang i dans la série des observations expurgées des répétitions et rangées par ordre croissant ; d_i , le nombre de décès observé en $t_{(i)}$; r_i , l'exposition au risque en $t_{(i)}$ (nombre de vivants juste avant cet instant). Les estimateurs de Kaplan-Meier et de Harrington-Fleming de $S(t)$ sont respectivement définis en $t > 0$ par : $\hat{S}_{KM}(t) = \prod_{i:t_{(i)} \leq t} (1 - d_i/r_i)$; $\hat{S}_{HF}(t) = \exp\left(-\sum_{i:t_{(i)} \leq t} d_i/r_i\right)$.

Représentation de $\hat{S}_E(t)$, $\hat{S}_{KM}(t)$, $\hat{S}_{HF}(t)$ sous R.

```
# ---- estimateur empirique ----

esurv <- function(t){1-ecdf(time)(t)} # proportion de sujets qui survivent à t
xx=seq(from=0,to=10,by=0.1)
points(xx,esurv(xx),'l',col='red',xlim=c(0,5),
       xlab='trois estimateurs non parametriques de la fonction de survie',ylab='')

# ---- estimateur Kaplan-Meier ----

library(survival)
sdat <- data.frame(time=time,status=status)
km <- summary(survfit(Surv(time, status) ~ 1,data=sdat,type="kaplan-meier",
conf.type='none'))
print(km)
KME <- stepfun(km$time,c(1,km$surv),right=TRUE)
points(xx,KME(xx),'l',col='blue')

# ---- estimateur de Harrington-Fleming ----

hf <- summary(survfit(Surv(time, status)~ 1, data=sdat, type='fleming',
conf.type="none"))
HFE <- stepfun(hf$time,c(1,hf$surv),right=TRUE)
points(xx,HFE(xx),'l',col='magenta')
```

```

legend('topright',
c('Fonction de survie', 'Empirique', 'Kaplan Meier', 'Harrington Fleming'),
lwd=c(1,1,1,1), lty=c(1,1,1,1), col=c('black', 'red', 'blue', 'magenta'))

```

2. Comparez entre eux les trois estimateurs de la fonction de survie.

On observe : $\forall t > 0 : \hat{S}_{HF}(t) \geq \hat{S}_{KM}(t) \geq \hat{S}_E(t)$.

3. Conjecturez le comportement de chaque estimateur lorsque la taille d'échantillon augmente et/ou lorsque b augmente.

Reprenons les représentations précédentes en combinant deux hypothèses sur b et n : $\{b \text{ petit (bb=2) ou grand (bb=10)}\}$ et $\{n \text{ petit (nn=20) ou grand (nn=2000)}\}$

Ce qu'on observe :

	$b \text{ petit (bb=2)}$	$b \text{ grand (bb=10)}$
$n \text{ petit (nn=20)}$	$\forall t > 0, \hat{S}_{HF}(t) \geq \hat{S}_{KM}(t) \geq \hat{S}_E(t)$ Les trois estimateurs sont éloignés de $S(t)$	$\forall t > 0, \hat{S}_{HF}(t) \geq \hat{S}_{KM}(t) \geq \hat{S}_E(t)$ Les trois estimateurs sont proches de $S(t)$
$n \text{ grand (nn=2000)}$	$\forall t > 0, \hat{S}_{HF}(t) \geq \hat{S}_{KM}(t) \geq \hat{S}_E(t)$ $\hat{S}_{HF}(t), \hat{S}_{KM}(t)$ sont proches de $S(t)$ $\hat{S}_E(t)$ sous estime $S(t)$	$\forall t > 0, \hat{S}_{HF}(t) \geq \hat{S}_{KM}(t) \geq \hat{S}_E(t)$ Les trois estimateurs sont proches de $S(t)$

o Puisque les estimateurs $\hat{S}_{HF}(t)$, $\hat{S}_{KM}(t)$, $\hat{S}_E(t)$ sont systématiquement (quel que soit l'échantillon) rangés dans le même ordre, il doit exister une raison algébrique à cet ordre.

Justifions : $\forall t > 0, \hat{S}_{HF}(t) \geq \hat{S}_{KM}(t)$.

* Si $0 < t < t_{(1)}$ alors : $\hat{S}_{KM}(t) = \hat{S}_{HF}(t) = 1$.

* Si t est compris entre deux observations il existe i_0 tel que : $t_{(i_0)} \leq t < t_{(i_0+1)}$. En $t_{(i)} \leq t_{(i_0)}$ le nombre de décès d_i est strictement inférieur à l'exposition r_i et $-d_i/r_i > -1$. Or, pour tout $x > -1$: $\ln(1+x) \leq x$. Ainsi : $\ln(1 - d_i/r_i) \leq -d_i/r_i$ et $\sum_{i:t_{(i)} \leq t} \ln(1 - d_i/r_i) \leq \sum_{i:t_{(i)} \leq t} -d_i/r_i$. En passant à l'exponentielle dans chaque membre : $\prod_{i:t_{(i)} \leq t} (1 - d_i/r_i) \leq \exp\left(-\sum_{i:t_{(i)} \leq t} d_i/r_i\right)$ et $\hat{S}_{KM}(t) \leq \hat{S}_{HF}(t)$.

* Sinon t est supérieur ou égal à la dernière observation $t_{(i_{\max})}$ et deux cas peuvent se présenter :

- aucune censure n'est observée en $t_{(i_{\max})}$ et alors : $d_{i_{\max}} = r_{i_{\max}}$ et $S_{KM}(t) = 0 \leq \hat{S}_{HF}(t)$.
- une censure au moins est observée en $t_{(i_{\max})}$ et l'inégalité découle d'un raisonnement similaire à celui du second cas.

Justifions : $\forall t > 0, \hat{S}_{KM}(t) \geq \hat{S}_E(t)$.

* Si $0 < t < t_{(1)}$ alors : $\hat{S}_E(t) = \hat{S}_{KM}(t) = 1$.

* Si t est supérieur à la dernière observation alors : $\hat{S}_E(t) = 0$ et $\hat{S}_{KM}(t) \geq \hat{S}_E(t)$.

* Sinon, il existe i_0 tel que : $t_{(i_0)} \leq t < t_{(i_0+1)}$. Les sujets qui dépassent l'instant t sont toujours vivants juste avant $t_{(i_0+1)}$; donc $\sum_{i=1}^n \mathbb{1}_{\{t_i > t\}} = r_{i_0+1}$. Par ailleurs, personne ne disparaît avant $t_{(1)}$; donc $n = r_1$. Ainsi : $\hat{S}_E(t) = (\sum_{i=1}^n \mathbb{1}_{\{t_i > t\}}) / n = r_{i_0+1}/r_1 = \prod_{i=1}^{i_0} r_{i+1}/r_i$. Or, les sujets vivants juste avant $t_{(i)}$ se divisent entre : ceux qui disparaissent ou sont censurés dans l'intervalle $[t_{(i)}, t_{(i+1)})$ et ceux qui vivent toujours juste avant $t_{(i+1)}$; ainsi : $r_i = \tilde{d}_i + r_{i+1}$ où $\tilde{d}_i$ désigne le nombre de censures et de décès réunis en $t_{(i)}$. Donc : $\hat{S}_E(t) = \prod_{i=1}^{i_0} (r_i - \tilde{d}_i)/r_i = \prod_{i=1}^{i_0} (1 - \tilde{d}_i/r_i)$. Mais, pour tout i : $\tilde{d}_i \geq d_i$ (il y a au moins autant de décès et de censures réunis en $t_{(i)}$ que de décès). Donc : $\prod_{i=1}^{i_0} (1 - \tilde{d}_i/r_i) \leq \prod_{i=1}^{i_0} (1 - d_i/r_i)$ et $\hat{S}_E(t) \leq \hat{S}_{KM}(t)$.

- o Quand b croît, la censure disparaît et pour tout $t > 0$: $\hat{S}_{HF}(t) \approx \hat{S}_{KM}(t) \approx \hat{S}_E(t)$.

Montrons : $\lim_{b \rightarrow +\infty} \mathbb{P}(X > Y) = 0$.

La densité de X est $f_X(x) = e^{-x} \mathbb{1}_{\{x > 0\}}$ et celle de Y : $f_Y(y) = \frac{1}{b} \mathbb{1}_{\{0 < y < b\}}$. La variable $Z = X - Y$ a pour support $] -b, +\infty[$ et puisque X et Y sont indépendantes, la densité de Z est définie en $z > -b$ par : $f_Z(z) = \int_{\mathbb{R}} f_X(u) f_{-Y}(z-u) du = \int_{\mathbb{R}} f_X(u) f_Y(u-z) du = \int_{\mathbb{R}} e^{-u} \mathbb{1}_{\{u > 0\}} \times \frac{1}{b} \mathbb{1}_{\{0 < u-z < b\}} du = \int_{\mathbb{R}} (e^{-u}/b) \mathbb{1}_{\{\max\{0, z\} < u < z+b\}} du = (e^{-\max\{0, z\}} - e^{-(z+b)})/b$. Ainsi, $\mathbb{P}(X > Y) = \mathbb{P}(Z > 0) = \int_0^{+\infty} f_Z(z) dz = (1 - e^{-b})/b \xrightarrow{b \rightarrow +\infty} 0$.

Exercice 6. (Equivalence en l'absence de censure des estimateurs empirique et Kaplan-Meier de la survie)

T est une durée de vie **non censurée** de fonction de survie $S(t)$; $t_1, \dots, t_n$ sont des observations de T .

1. Montrez que l'estimateur $\hat{S}_{KM}(t)$ coïncide avec la fonction de survie empirique $\hat{S}_E(t) = \sum_{i=1}^n \mathbb{1}_{\{t_i > t\}}/n$.

Justifions : $\forall t > 0 : \hat{S}_E(t) = \hat{S}_{KM}(t)$.

Supposons t compris entre deux événements : $\exists j, t_{(j)} \leq t < t_{(j+1)}$. Alors : $\hat{S}_{KM}(t) = \prod_{i:t_{(i)} \leq t} (1 - d_i/r_i) = \prod_{i=1}^j (r_i - d_i)/r_i$. Or, en l'absence de censure : $r_i - d_i = r_{i+1}$ (les sujets exposés au risque en $t_{(i)}$ sont à nouveau exposés au risque en $t_{(i+1)}$ s'il ne disparaissent pas en $t_{(i)}$). Ainsi, $\hat{S}_{KM}(t) = \prod_{i=1}^j r_{i+1}/r_i = r_{j+1}/r_1$. Mais $r_{j+1} = \sum_{i=1}^n \mathbb{1}_{\{t_i > t\}}$ et $r_1 = n$. Donc $\hat{S}_{KM}(t) = (\sum_{i=1}^n \mathbb{1}_{\{t_i > t\}})/n = \hat{S}_E(t)$.

On montre par ailleurs : $\hat{S}_E(t) = \hat{S}_{KM}(t) = 1$ si t est strictement inférieur au premier événement et $\hat{S}_E(t) = \hat{S}_{KM}(t) = 0$ si t est supérieur au dernier événement.

2. Déterminez la distribution asymptotique de $\hat{S}_E(t)$.

Les variables aléatoires : $\mathbb{1}_{\{t_i > t\}}$; $i = 1, \dots, n$, constituent un échantillon aléatoire d'une loi de Bernoulli de paramètre $S(t)$. Selon le *Theorem Central Limit*, leur moyenne empirique : $\hat{S}_E(t)$ est approximativement distribuée quand n est grand selon une loi normale d'espérance $S(t)$ et de variance $S(t)(1 - S(t))/n$.

3. Déduisez-en un intervalle de confiance à 95% de $S(t)$.

La statistique $(\hat{S}_E(t) - S(t))/\sqrt{S(t)(1 - S(t))/n}$ est approximativement distribuée selon $\mathcal{N}(0; 1)$. Donc, l'intervalle $\hat{S}_E(t) \pm z_{0,975} \sqrt{S(t)(1 - S(t))/n}$ contient $S(t)$ dans 95% des cas. Puisqu'on ne connaît pas la variance $S(t)(1 - S(t))/n$, on la remplace par son estimation : $\hat{S}_E(t)(1 - \hat{S}_E(t))/n$. Ainsi, un intervalle de confiance à 95% de $S(t)$ est : $IC_{0,95}(S(t)) = \hat{S}_E(t) \pm z_{0,975} \sqrt{\hat{S}_E(t)(1 - \hat{S}_E(t))/n}$.

Exercice 7. (Une comparaison des estimateurs de Breslow et de Nelson-Aalen)

Une durée de vie X dont on note Λ la fonction de hasard cumulé est distribuée selon *Gompertz*(1,2). Une variable C indépendante de X et distribuée selon $\mathcal{U}_{[0,3]}$ censure X à droite.

1. Générez un échantillon de taille 100 de la loi de $T = \min(X, C)$.

```
library(flexsurv)
x=rgeomertz(100,rate=1,shape=2) # un echantillon de 100 durees de vies issues de Gompertz
c=runif(100,min=0,max=3) # les censures uniformes entre 0 et 3
complete <- data.frame(x=x,c=c) # les donnees completes
time <- apply(complete,1,min) # les observations
status <- (time==x)*1 # les indicatrices de non censure
observed <- data.frame(time=time,status=status) # observations et censures
```

2. Représentez dans la même fenêtre graphique :

- (a) l'estimation de Nelson-Aalen de Λ ,

```
tab <- summary(survfit(Surv(time, status)~1, data=observed, conf.type="none")) # TDR
ratios <- tab$n.event/tab$n.risk # ratios : deces/exposition
NAE <- function(t){sum(ratios[tab$time<=t])} # estimateur de Nelson-Aalen
# representation de NAE
abs=seq(from=0,to=max(time),by=0.01) ; ord=sapply(abs,NAE)
plot(abs,ord,col='red','l', xlab='time',ylab='')
```

- (b) les intervalles de confiance à 95% de Λ ,

```
ratios <- tab$n.event/tab$n.risk^2 # ratios : deces/exposition au carre
VARNAE <- function(t){sum(ratios[tab$time<=t])} # variance de l'estimateur de Nelson-Aalen
ee=sqrt(sapply(abs,VARNAE)) # ecart-type de l'estimateur de Nelson-Aalen
lines(abs,ord-qnorm(0.975)*ee,col='orange','l',lty=2) # borne inferieure de IC(95%)
lines(abs,ord+qnorm(0.975)*ee,col='orange','l',lty=2) # borne superieure de IC(95%)
```

- (c) l'estimation de Breslow de Λ ,

```
km <- survfit(Surv(time, status) ~ 1, type="kaplan-meier")
BRE <- stepfun(km$time,c(0,-log(km$surv)))
ord=sapply(abs,BRE)
points(abs,ord,col='blue','l')
```

(d) la fonction Λ .

```
ord=1/2*(exp(2*abs)-1) # hasard cumule de Gompertz(1,2)
points(abs,ord,col='black','l')

legend('topleft',
c('Nelson-Aalen','95% NAE confidence boundaries','Breslow','Cumulated Hazard'),
lwd=c(1,1,1,1),lty=c(1,2,1,1),col=c('red','orange','blue','black'))
```

Exercice 8. (Tests de comparaison de deux fonctions de survie)

On considère le jeu de données `colon{survival}` de R.

1. Au seuil de 5% doit-on rejeter l'hypothèse : la proportion de survivants au delà de 2500 jours (variable `time`) est la même chez les hommes et chez les femmes ?

T_H désigne la durée de vie d'un homme choisi au hasard et T_F celle d'une femme. S_H est la fonction de survie de T_H et S_F celle de T_F . On considère l'instant $t_0 = 2500$. On teste au seuil $\alpha = 0,05$ l'hypothèse $\mathcal{H}_0 : S_H(t_0) = S_F(t_0)$ contre $\mathcal{H}_1 : S_H(t_0) \neq S_F(t_0)$.

Notons $\hat{S}_{F,KM}(t_0)$ l'estimateur de Kaplan-Meier de $S_F(t_0)$ et $\hat{V}(\hat{S}_{F,KM}(t_0))$ l'estimateur de sa variance selon la formule de Greenwood. De même : $\hat{S}_{H,KM}(t_0)$ est l'estimateur de Kaplan-Meier de $S_H(t_0)$ et $\hat{V}(\hat{S}_{H,KM}(t_0))$ l'estimateur de sa variance selon la formule de Greenwood.

$(\hat{S}_{F,KM}(t_0) - S_F(t_0)) / \sqrt{\hat{V}(\hat{S}_{F,KM}(t_0))}$ et $(\hat{S}_{H,KM}(t_0) - S_H(t_0)) / \sqrt{\hat{V}(\hat{S}_{H,KM}(t_0))}$ sont approximativement normales centrées réduites et supposées indépendantes. Si $\mathcal{H}_0$ est vraie, la statistique :

$$z = (\hat{S}_{F,KM}(t_0) - \hat{S}_{H,KM}(t_0)) / \sqrt{\hat{V}(\hat{S}_{F,KM}(t_0)) + \hat{V}(\hat{S}_{H,KM}(t_0))}$$

est donc approximativement distribuée selon $\mathcal{N}(0;1)$.

Au seuil 5%, on rejette $\mathcal{H}_0$ si la valeur observée de z est en dehors de l'intervalle $\pm z_{0,975}$; on ne rejette pas $\mathcal{H}_0$ sinon.

Le script R suivant donne la valeur observée de la statistique de décision : $z_{\text{obs}} = 0,327$.

```
library(survival)
# Statistiques hommes
colonh <- colon[colon$sex==1,c(15,10)] # donnees completes (deces+censure)
kmh <- summary(survfit(Surv(time, status) ~ 1,data=colonh,type="kaplan-meier")) # KM
KMEH <- stepfun(kmh$time,c(1,kmh$surv),right=FALSE) # fonction SKM(t)
vsh <- kmh$std.err[kmh$time==max(kmh$time[kmh$time<=2500])]^2 # Greenwood en 2500
# Statistiques femmes
colonf <- colon[colon$sex==0,c(15,10)] # donnees completes (deces+censure)
kmf <- summary(survfit(Surv(time, status) ~ 1,data=colonf,type="kaplan-meier")) # KM
KMEF <- stepfun(kmf$time,c(1,kmf$surv),right=FALSE) # fonction SKM(t)
vsf <- kmf$std.err[kmf$time==max(kmf$time[kmf$time<=2500])]^2 # Greenwood en 2500
# Statistique de test
(KMEH(2500)-KMEF(2500))/sqrt(vsh+vsf)
```

Puisque la valeur observée de z est comprise entre $\pm z_{0,975} = \pm 1,96$ on ne rejette pas $\mathcal{H}_0$ au seuil 5% : il est possible que la proportion de sujets survivant à 2500 jours soit la même chez les hommes et chez les femmes.

2. Au seuil de 5% doit-on rejeter l'hypothèse : en tout point la fonction de survie est la même chez les hommes et chez les femmes ?

Dans un test de Mantel-Haenszel, on teste au seuil 5% l'hypothèse $\mathcal{H}_0$: en tout point $t > 0$ les fonctions de survie S_F et S_H coïncident contre $\mathcal{H}_1$: il existe au moins un instant en lequel les fonctions de survie S_F et S_H ne coïncident pas.

```
attach(colon)
MHAENSZEL <- (survdif(Surv(time, status) ~ sex, data=colon, rho=0))
MHAENSZEL$pvalue
```

La p -valeur du test vaut 0,61 ; ainsi, au seuil 5% on ne rejette pas $\mathcal{H}_0$.