Econométrie des modèles de durée

Guillaume Horny*

*Banque de France et UCLouvain

Master 2 MOSEF

Chapitre 2 : Temps discret et estimation non-paramétrique

Plan

Modèles en temps discret

Chapitre 2

Durées discrètes

- le processus générant les données peut-être fondamentalement discret (ex: durée séparant deux livraisons...)
- En pratique, on mesure souvent des durées continues à des dates prédéterminés (semaine, mois...)
  $\Rightarrow$ on parle alors de **durées groupées**, car les transitions sont regroupées sur des intervalles de temps

Exemple de durées groupées

TABLE II

FAILURES, CENSORINGS, AND THE KAPLAN-MEIER EMPIRICAL HAZARD[a]

Week t	Risk set R_t	Failures D_t	Censorings C_t	Hazard H_t	Standard error
1	3365	277	0	.08232	.0047
2	3062	203	26	.06630	.0045
3	2832	159	27	.05614	.0043
4	2657	161	16	.06059	.0046
5	2458	123	38	.05004	.0044
6	2271	112	64	.04932	.0045
7	2112	88	47	.04167	.0043
8	1984	82	40	.04133	.0045
9	1850	86	52	.04649	.0049
10	1722	63	42	.03659	.0045
11	1621	68	38	.04195	.0050
12	1520	91	33	.05987	.0061
13	1402	71	27	.05064	.0059
14	1300	58	31	.04462	.0057
15	1210	55	32	.04545	.0060
16	1134	46	21	.04056	.0059
17	1077	60	11	.05571	.0070
18	999	58	18	.05806	.0074
19	936	44	5	.04701	.0069
20	880	41	12	.04659	.0071
21	829	49	10	.05911	.0082
22	773	45	7	.05821	.0084
23	721	44	7	.06103	.0089
24	662	34	15	.05136	.0086
25	610	48	18	.07869	.0109
26	430	45	132	.10465	.0148
27	378	26	7	.06878	.0130
28	317	30	35	.09464	.0164
29	279	21	8	.07527	.0158
30	245	13	13	.05306	.0143
31	226	9	6	.03982	.0130
32	212	17	5	.08019	.0187
33	190	5	5	.02632	.0116
34	178	8	7	.04494	.0155
35	165	13	5	.07879	.0210
36	121	12	31	.09917	.0272
37	105	6	4	.05714	.0227
38	91	9	8	.09890	.0313

[a] 2380 failures were observed, and 985 censorings. 201 of the censorings occurred at exhaustion of benefits.

Hasard en temps discret

Concepts équivalents au temps continu (et plus intuitifs!)
Soient les dates de sorties t_j $(j = 1, \ldots)$
Hasard en temps discret:

$$\lambda^d(t_j) = \Pr(T = t_j | T \geq t_j) \tag{1}$$

Le hasard est ici une probabilité, dans $[0, 1]$

Survie en temps discret

Survie en temps discret

$$S^d(t) = \Pr(T > t). \tag{2}$$

Intuition

$$\Pr(T > t_2) = \Pr(T > t_1) \Pr(T > t_2 | T > t_1).$$

Or:

- $\Pr(T > t_1) = 1 - \lambda(t_1)$
- $\Pr(T > t_2 | T > t_1) = 1 - \Pr(T = t_2 | T > t_1) = 1 - \lambda(t_2)$

D'où:

$$S^d(t) = \prod_{j | t_j \leq t} [1 - \lambda(t_j)]. \tag{3}$$

Réécriture du hasard en temps discret

$$\begin{aligned}\lambda^d(t_j) &= \Pr(T = t_j | T \geq t_j) \\ &= \frac{f^d(t_j)}{S^d(t_{j_-})}.\end{aligned}$$

On utilise $S^d(t_{j_-}) = \lim_{t \to t_{j_-}} S^d(t)$, car $S^d(t_j) = \Pr(T > t_j)$ et non pas $\Pr(T \geq t_j)$.

Hasard intégré en temps discret

$$\Lambda^d(t) = \sum_{j|t_j \leq t} \lambda(t_j). \tag{4}$$

Estimation non-paramétrique de la survie

Chapitre 2

Estimateur empirique

Soit un échantillon de durées $(t_i)_{i=1,\ldots,n}$

Estimateur empirique:

$$S_n(t) = \frac{1}{n} \sum_{i=1,\ldots,n} \mathbb{1}[t_i > t] \tag{5}$$

Estimateur empirique avec données censurées

On observe $(T_i, \delta_i)_{i=1,\dots,n}$ au lieu de $(T_i^*)_{i=1,\dots,n}$.

- estimer S par la survie empirique évaluée sur $(T_i)_{i=1,\dots,n}$ fournit une estimation biaisée
  Intuition: $T_i \leq T_i^*$, et S_n sous-estime S
- estimer S par la survie empirique évaluée sur $(T_i)_{i=1,\dots,n|\delta_i=1}$ fournit une estimation biaisée
  Intuition: en cas de censure à droite, on évalue S_n en utilisant les durées les plus courtes et S_n sous-estime S

Conclusion: N'utilisez pas l'estimateur empirique de S!

Hypothèses et notations

On dispose d'un échantillon de n durées groupées, censurées à droite.
Les durées sont observées à dates fixes
$0 = t_0 < t_1 < t_2 < \ldots < t_{K-1} < t_K < +\infty$. Toutes les durées sont inférieures à t_K.
Pour chaque intervalle, on observe:

- d_j, le nombre de transitions dans l'intervalle $]t_{j-1}, t_j]$
- c_j, le nombre de durées censurées dans l'intervalle $]t_{j-1}, t_j]$
- r_j, le nombre de durées "au risque" de se terminer par une censure ou une transition dans $]t_{j-1}, t_j]$, cad les durées supérieurs à t_{j-1}

Estimateur de Kaplan-Meier

Un estimateur immédiat de $\lambda(t_j)$ est:

$$\widehat{\lambda}(t_j) = \frac{d_j}{r_j} \tag{6}$$

estimateur de **Kaplan-Meier** de la fonction de survie:

$$\begin{aligned}\widehat{S}_{KM}(t) &= \prod_{j|t_j \leq t} [1 - \widehat{\lambda}(t_j)] \\ &= \prod_{j|t_j \leq t} \frac{r_j - d_j}{r_j}.\end{aligned}$$

Exemple de calcul

On observe 21 durées, où $*$ indique une censure à droite
1,1,1,1*, 2, 3*, 4, 4*, 5*, 6,7, 8*, 9*, 10*, 11, 12, 13*, 14*, 14*, 15*, 16*.

Exemple de calcul (suite)

durées	r_j	d_j	$(r_j - d_j)/r_j$	$\widehat{S}_{KM}(t)$
1	21	3	0.86	0.86
2	17	1	0.94	0.81
4	15	1	0.93	0.75
6	12	1	0.92	0.69
7	11	1	0.91	0.63
11	7	1	0.86	0.54
12	6	1	0.83	0.45

Propriétés de l'estimateur de Kaplan-Meier

- la fonction est constante sur chaque intervalle
- elle est *cadlag*
- $\widehat{S}_{KM}(t) = 0$ pour $t > t_K$
- lorsqu'il n'y a pas de censure, $c_j = 0, \forall j$, et $r_j - d_j = r_{j+1}$. On a:

$$\widehat{S}_{KM}(t) = \prod_{j|t_j \leq t} \frac{r_{j+1}}{r_j} = \frac{r_1}{r_0}\frac{r_2}{r_1} \ldots \frac{r_{j+1}}{r_j}$$
$$= \frac{r_{j+1}}{n}.$$

Propriétés de l'estimateur de Kaplan-Meier (suite)

- plusieurs transitions peuvent se produire dans le même intervalle de temps, en contradiction avec l'hypothèse de temps continu. Ceci n'affecte pas $\widehat{S}_{KM}$.
  En effet, supposons dans un échantillon sans censure que l'on ait 2 transitions dans l'intervalle $]t_j, t_{j+1}]$:

$$\underbrace{\left(\frac{n-1}{n}\right)}_{]t_0,t_1]} \cdots \underbrace{\left(\frac{n-j}{n-j+1}\right)}_{]t_{j-1},t_j]} \underbrace{\left(\frac{n-j-2}{n-j}\right)}_{]t_j,t_{j+1}]} \underbrace{\left(\frac{n-j-3}{n-j-2}\right)}_{]t_{j+1},t_{j+2}]} \cdots \quad (7)$$

Conclusion: Après simplification, on retrouve l'estimateur de Kaplan-Meier

Propriétés de l'estimateur de Kaplan-Meier (suite)

- on peut montrer que $\widehat{S}_{KM}$ est un estimateur non-paramétrique du maximum de vraisemblance cad qu'il maximise la vraisemblance approchée:

$$L_{app}(S) = \prod_{i=1,\ldots,n} \left([S(t_{i-1}) - S(t_i)]^{d_i} \, S(t_{i-1})^{c_i} \right). \tag{8}$$

- il est donc sans biais et de variance:

$$\text{Var}[\widehat{S}_{KM}(t)] = -\widehat{S}^2_{KM}(t) \sum_{j|t_j \leq t} \frac{d_j}{r_j(r_j - d_j)}. \tag{9}$$

Exemple: survie de patients atteints d'une leucémie aiguë

Calcul de l'estimateur de Kaplan-Meier avec le logiciel R, package survival.
Source des données : R. Miller (1997), *Survival Analysis*, John Wiley & Sons.
Programme:

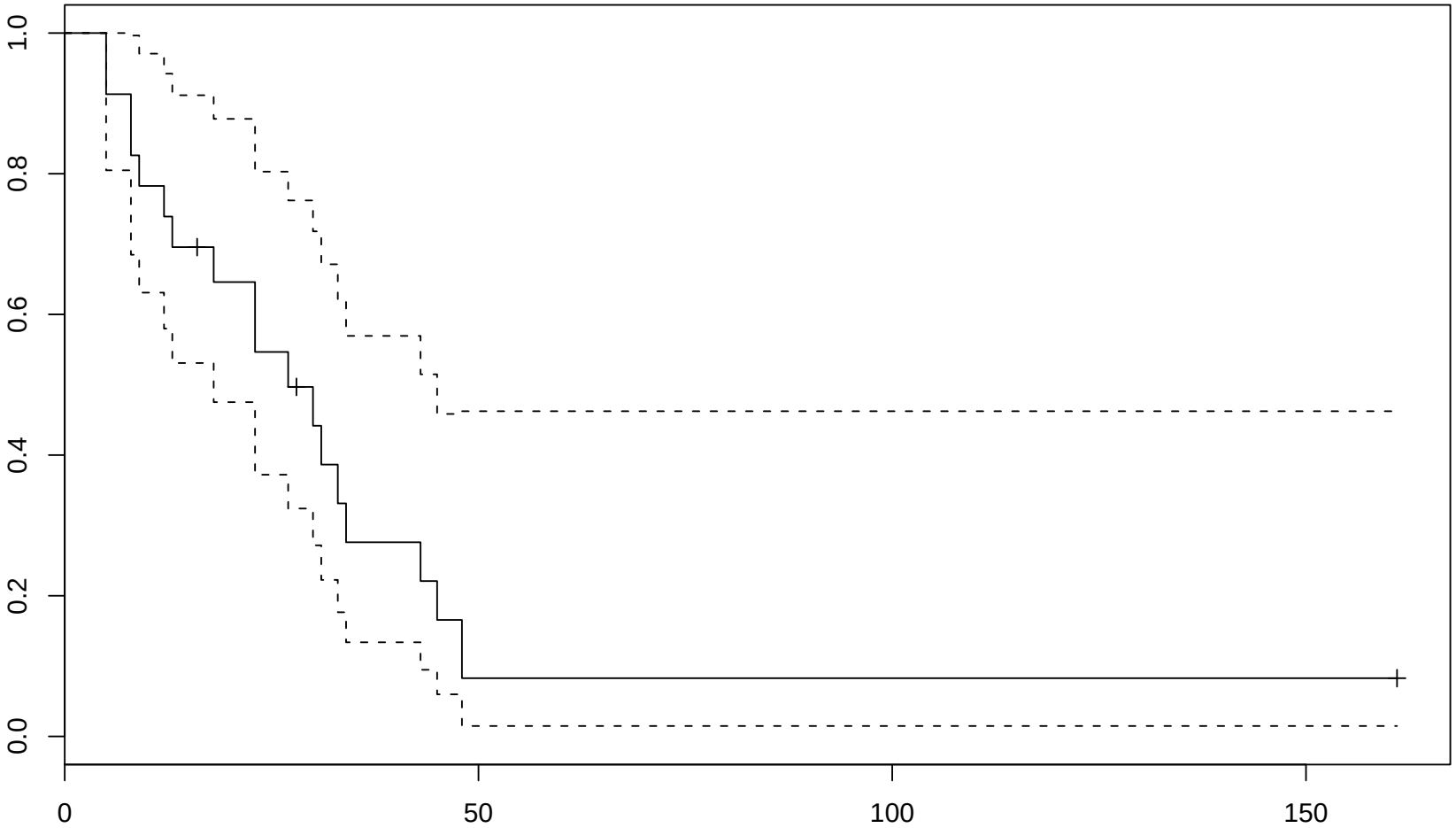
```
library(survival)
plot(survfit(Surv(time,status), data=aml))
```

Exemple: données

	time	status	x
1	9	1	Maintained
2	13	1	Maintained
3	13	0	Maintained
4	18	1	Maintained
5	23	1	Maintained
6	28	0	Maintained
7	31	1	Maintained
8	34	1	Maintained
9	45	0	Maintained
10	48	1	Maintained
11	161	0	Maintained
12	5	1	Nonmaintained
13	5	1	Nonmaintained
14	8	1	Nonmaintained
15	8	1	Nonmaintained
16	12	1	Nonmaintained
17	16	0	Nonmaintained
18	23	1	Nonmaintained
19	27	1	Nonmaintained
20	30	1	Nonmaintained
21	33	1	Nonmaintained
22	43	1	Nonmaintained
23	45	1	Nonmaintained

Exemple: estimation

Figure: Estimation par Kaplan-Meier



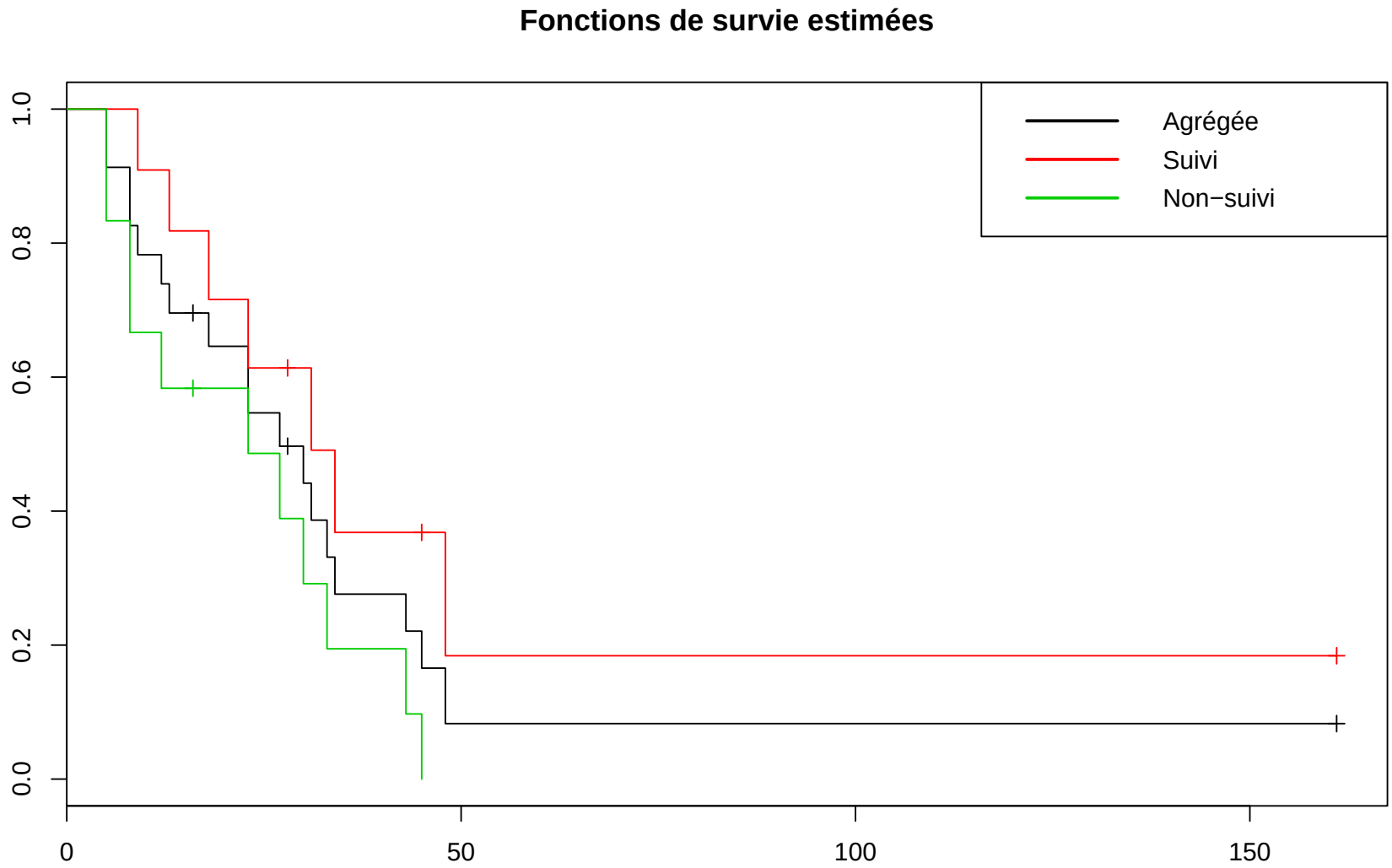
Exemple: stratification par traitement

Programme:

```
# Estimation par Kaplan-Meier
sagr <- survfit(Surv(time, status), data = aml)
s  <- survfit(Surv(time, status) ~x, data = aml)

plot(sagr, conf.int=FALSE, main="Fonctions de survie estimées")
lines(s, col=2:3)
legend("topright", legend = c("Agrégée","Suivi", "Non-suivi"),
 col=1:3, lty = 1, lwd = 2)
```

Exemple: estimation



Test de l'égalité des fonctions de survie

Intuition: A partir du tableau de contingence, on peut calculer le nombre espéré de transition, sa variance et faire un test du Chi-2.

	Groupe 1	Groupe 2	Total
Transition	d_1	d_2	$d1 + d_2 = d$
Non-transitions	$n_1 - d_1$	$n_2 - d_2$	$n - d$
Total	n_1	n_2	n

- nombre espéré de transition dans le groupe 1: $e_1 = dn_1/n$.
- sa variance (loi hypergéométrique): $V_1 = n_1 n_2 d(n-d))/[n^2(n-1)]$.

Test: $[\sum_i (d_{1i} - e_{1i})]^2 / \sum_i V_{1i} \sim \chi^2_1$

Exemple: test égalité des fonctions de survie

Programme:

```
# test d'égalité des fonctions de survie
survdiff(Surv(time,status)~x, data=aml, rho=0)
```

Exemple: résultat test égalité

```
> survdiff(Surv(time,status)~x, data=aml, rho=0)
Call:
survdiff(formula = Surv(time, status) ~ x, data = aml, rho = 0)

                 N Observed Expected (O-E)^2/E (O-E)^2/V
x=Maintained    11        7    10.69      1.27      3.40
x=Nonmaintained 12       11     7.31      1.86      3.40

 Chisq= 3.4  on 1 degrees of freedom, p= 0.0653
```

Exemple: code Stata

```
# déclaration des variables de durées
stset time, failure(status)
# Estimateur de Kaplan-Meier
sts list
# Estimateur de Kaplan-Meier stratifié
sts list, by(x)
# Représentation graphique
sts graph, by(x)
# test ‘‘log-rank’’
sts test x
```

Estimation non-paramétrique du hasard intégré

Chapitre 2

Estimateur déduit de Kaplan-Meier

Comme:

$$S(t) = \exp\left[-\Lambda(t)\right],$$
$$\Rightarrow \widehat{\Lambda}(t) = -\ln \widehat{S}_{KM}(t).$$

D'où:

$$\widehat{\Lambda}(t) = \sum_{j|t_j \leq t} \ln\left(\frac{r_j - d_j}{r_j}\right), \tag{10}$$

pour tout t tel que $\widehat{S}_{KM}(t) > 0$.

Estimateur de Nelson-Aalen

Alternative: estimateur de Nelson-Aalen, plus simple et avec de meilleures propriétés de convergence que l'estimateur déduit de Kaplan-Meier.
On peut estimer le hasard intégré:

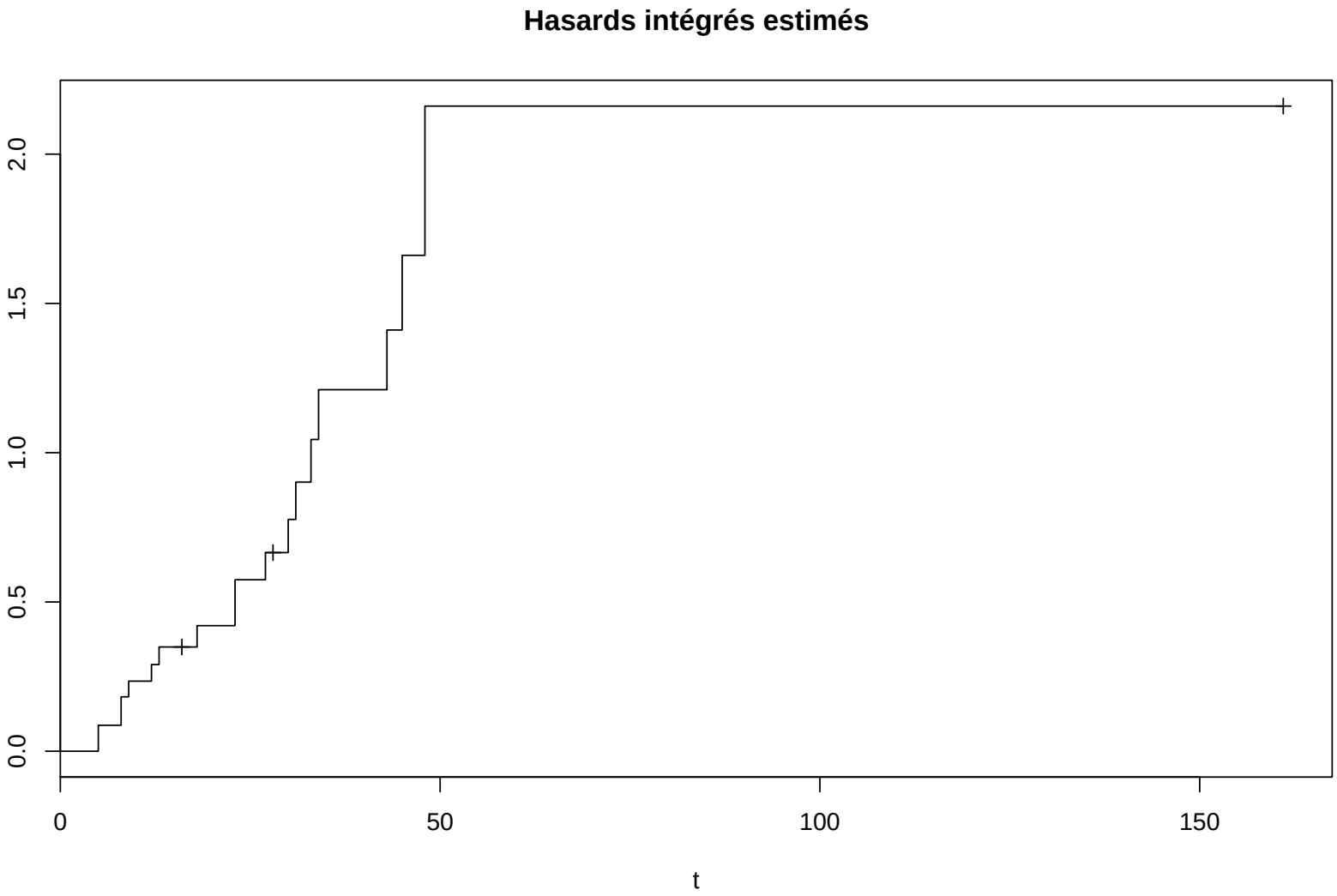
$$\widehat{\Lambda}_{NA}(t_j) = \sum_{j|t_j \leq t} \widehat{\lambda}(t_j) = \sum_{j|t_j \leq t} \frac{d_j}{r_j}. \tag{11}$$

Exemple: hasard intégré de Nelson-Aalen

Programme:

```
# hasard intégré de Nelson-Aalen
sagr.b <- survfit(Surv(time, status),\\
 type="fleming-harrington", data = aml)
plot(sagr.b, conf.int=FALSE, main="Hasards intégrés estimés",\\
 fun="cumhaz", xlab="t")
```

Exemple: hasard intégré de Nelson-Aalen



Estimateur de S déduit de Nelson-Aalen

De même, on peut déduire l'estimateur de Breslow de S:

$$\widehat{S}_B(t) = \exp\left(-\sum_{j|t_j \leq t} \frac{d_j}{r_j}\right). \tag{12}$$

Comparaison des estimateurs de Kaplan Meier et de Breslow

Chapitre 2

Comparaison KM-B de S

$$\widehat{S}_{KM}(t) = \prod_{j|t_j \leq t} \left[1 - d\widehat{\Lambda}(t_j)\right],$$

$$\widehat{S}_{B}(t) = \prod_{j|t_j \leq t} \exp\left[-d\widehat{\Lambda}(t_j)\right].$$

- $1 - x \leq \exp(-x), \forall x \geq 0$, $\widehat{S}_{KM} \leq \widehat{S}_B$
- si la durée la plus longue est une sortie: $\widehat{S}_{KM}(t_{\max}) = 0$ et $\widehat{S}_B(t_{\max}) > 0$
  $\Rightarrow$ comme $\widehat{S}_{KM}$ est un estimateur du maximum de vraisemblance, il est sans biais et $\widehat{S}_B$ surestime légèrement S

Comparaison KM-B de S (suite)

- les deux estimateurs sont toutefois asymptotiquement équivalents
  Intuition: $1 - x \approx \exp(-x)$ pour x petit.
- les écarts sont donc faibles tant que les individus au risque ne sont pas trop peu nombreux, ce qui est le cas avec les échantillons rencontrés en pratique ($\approx$ tant qu'il y a une dizaine d'individus au risque)

Comparaison KM-B de S (suite)

- toutefois, les estimateurs peuvent être assez différents en présence de durées groupées, où lorsqu'il y a des *ex-aequo* dans des durées discrètes
- on a vu que l'estimateur de Kaplan-Meier n'est pas sensible à ce problème
- l'estimateur de Breslow y est sensible, d'où un certain nombre de corrections qui ont été proposées. Toutefois, les résultats sont moins satisfaisant qu'avec KM

Conclusion: Utiliser les estimateurs pour ce pourquoi ils ont été conçus:

- pour estimer $S \Rightarrow \widehat{S}_{KM}$,
- pour estimer $\Lambda \Rightarrow \widehat{\Lambda}_{NA}$.

Estimation non-paramétrique du hasard

Chapitre 2

Estimation par KM et NA

Les estimateurs $\widehat{S}_{KM}$ et $\widehat{\Lambda}_{NA}$ approchent $\lambda(t)$ par:

$$\widehat{\lambda(t)} = \frac{d_j}{r_j}. \tag{13}$$

Note: il est préférable d'interpréter plutôt le taux de transition par unité de temps:

$$\widehat{\lambda^u(t)} = \frac{d_j}{r_j(t_j - t_{j-1})}. \tag{14}$$

Lissage

L'estimateur (13) est toutefois une fonction en escalier, que l'on préfère lisser par la technique du **noyau de convolution**.

Note: C'est ce que fait Stata par défaut à l'issue de la commande

```
sts graph, hazard
```

Noyau de convolution

Hypothèses: Soit K un **noyau réel**, c'est-à-dire une fonction intégrable et intégrant à 1. On suppose K continue, symétrique, à support compact et à variations bornées.

Exemples:

- Noyau gaussien

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2). \tag{15}$$

- Noyau uniforme

$$K(x) = 0.5\mathbb{1}(|x| \leq 1). \tag{16}$$

Soit une suite de paramètres strictement positifs $(h_n)_{n\geq 1}$ appelés **fenêtres** vérifiant $h_n \longrightarrow 0$.

Estimation de f par noyau de convolution, durées non-censurées

On suppose qu'on a un échantillon, trié par ordre croissant, de durées non-censurées $(t_i)_{i=1,\ldots,n}$ et de densité f inconnue. On estime f au point t par:

$$\widehat{f}_n(t) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{t - t_i}{h_n}\right). \tag{17}$$

Estimation du hasard par noyau de convolution, durées non-censurées

Un estimateur à noyau du hasard est donné par:

$$\begin{aligned}\widehat{\lambda}_n(t) &= \int \frac{1}{h_n} K\left(\frac{t-u}{h_n}\right) d\Lambda_n(u) \\ &= \int \frac{1}{h_n} K\left(\frac{t-u}{h_n}\right) \lambda_{NA}(u) du \\ &= \frac{1}{h_n} \sum_{i=1}^{n} K\left(\frac{t-t_i}{h_n}\right) \frac{1}{n-i+1}.\end{aligned}$$

où le dernier terme est le hasard estimé pour l'individu i.
$\Rightarrow$ cette méthode lisse l'estimateur de Nelson-Aalen par le noyau K

Estimation par noyau de convolution, durées censurées

L'estimateur à noyau de f doit être adapté pour prendre en compte la censure

$\Rightarrow$ on utilise l'expression de l'estimateur de Nelson-Aalen en présence de données censurées

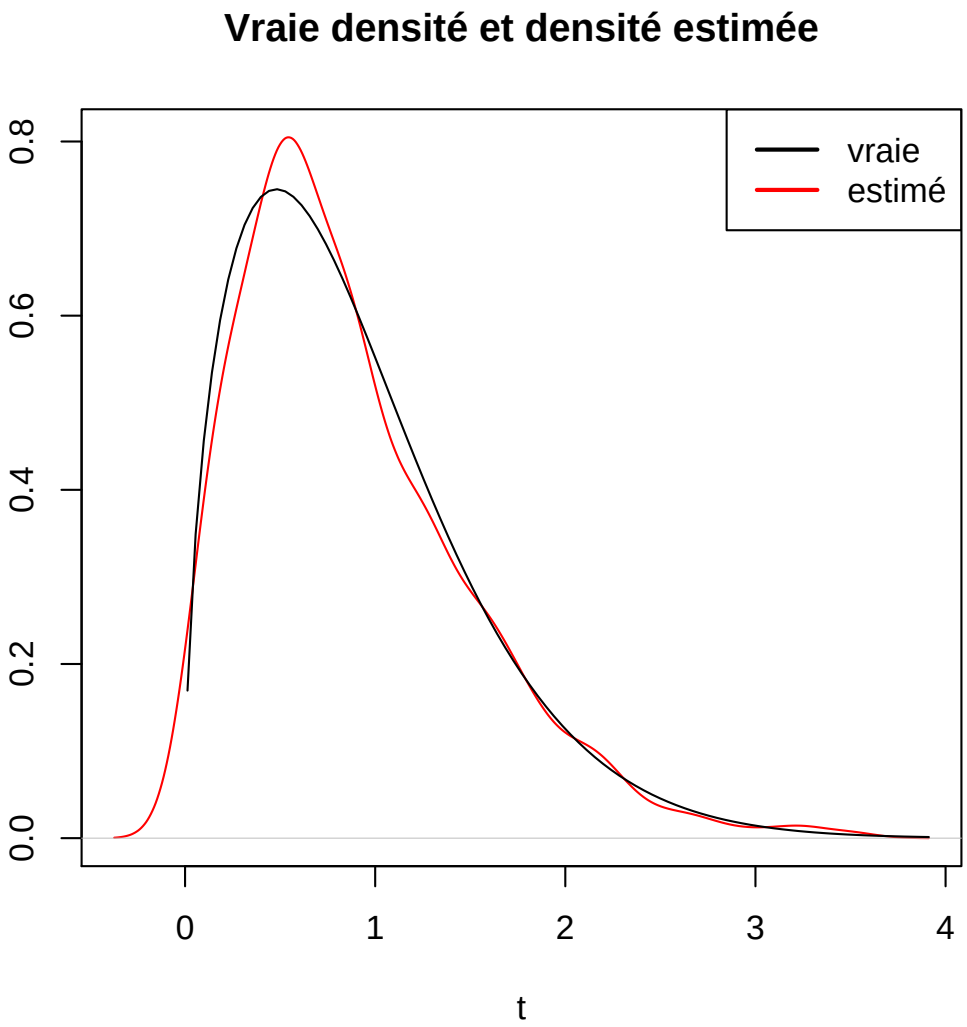
$$\widehat{\lambda}_n(t) = \frac{1}{h_n} \sum_{i=1}^{n} K\left(\frac{t - t_i}{h_n}\right) \frac{\delta_i}{n - i + 1}. \tag{18}$$

Exemple: estimation de f par noyau (1/1)

```
# simulations de durées weibull (alpha = 1)
theta <- 1.5
time <- (- log (runif(1000)))^(1/theta)
status <- seq(1,1, length=1000)
skm  <- survfit(Surv(time, status))

# estimation de la densité f
plot(density(time, kernel = "gaussian"), col=2, main="Vraie \\
densité et densité estimée", xlab="t", ylab=" ")
f <- function(t) {theta * t^(theta - 1) * exp(- t^theta)}
curve(f, add=TRUE, col = 1)
legend("topright", legend = c("vraie","estimé"), col=1:2, \\
lty = 1, lwd = 2)
```

Exemple: estimation de f par noyau (2/2)

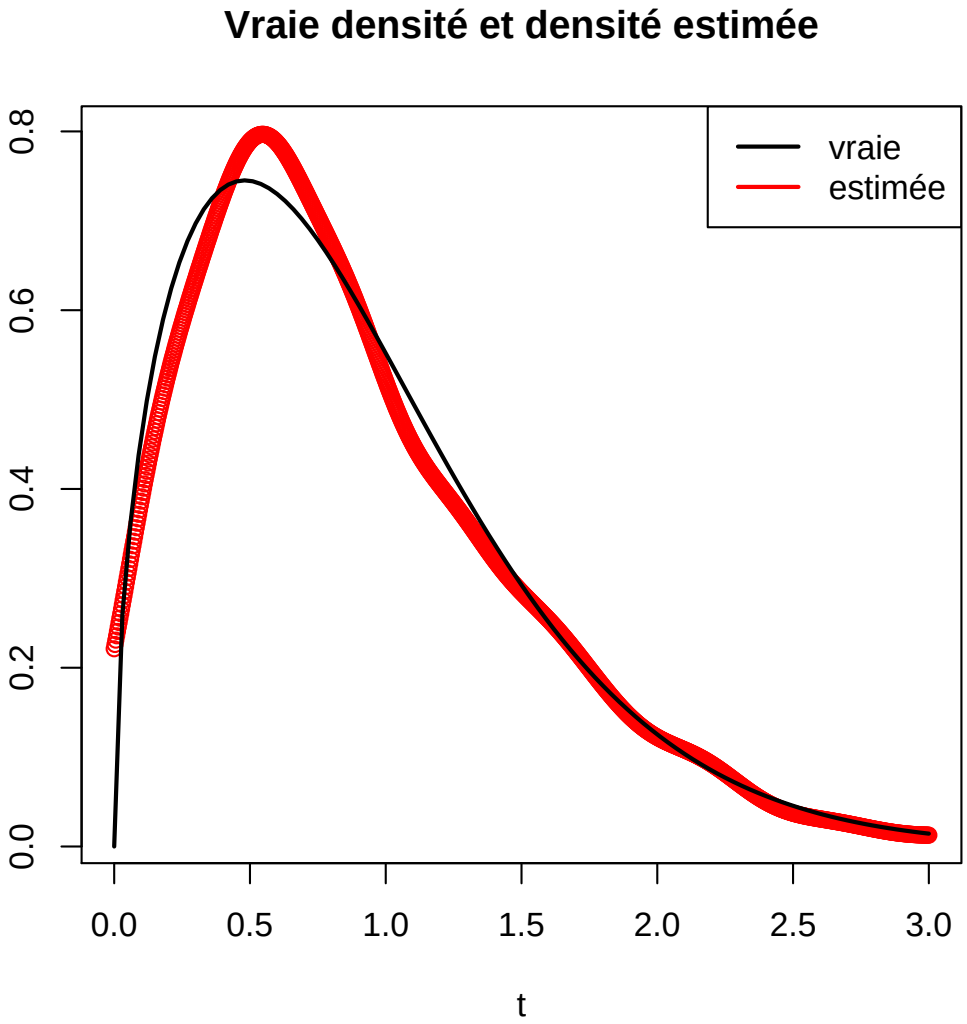


Exemple: estimation de f par noyau (3/3)

```
# estimation "manuelle" de f avec le noyau gaussien
t <- seq(0,3, length=1000)
h <- 0.138
ksmooth <- function(x, xpts, h) {
n <- length(x)
D <- outer(x, xpts, "-")
K <- dnorm(D / h)
dens <- colSums(K) / (n*h)
}

kest <- ksmooth(time, t, h)
plot(t, kest, col=2, main="Vraie densité et densité estimée")
 curve(f, add=TRUE, col = 1, lwd=2)
legend("topright", legend = c("vraie","estimée"), col=1:2,\\
 lty = 1, lwd = 2)
```

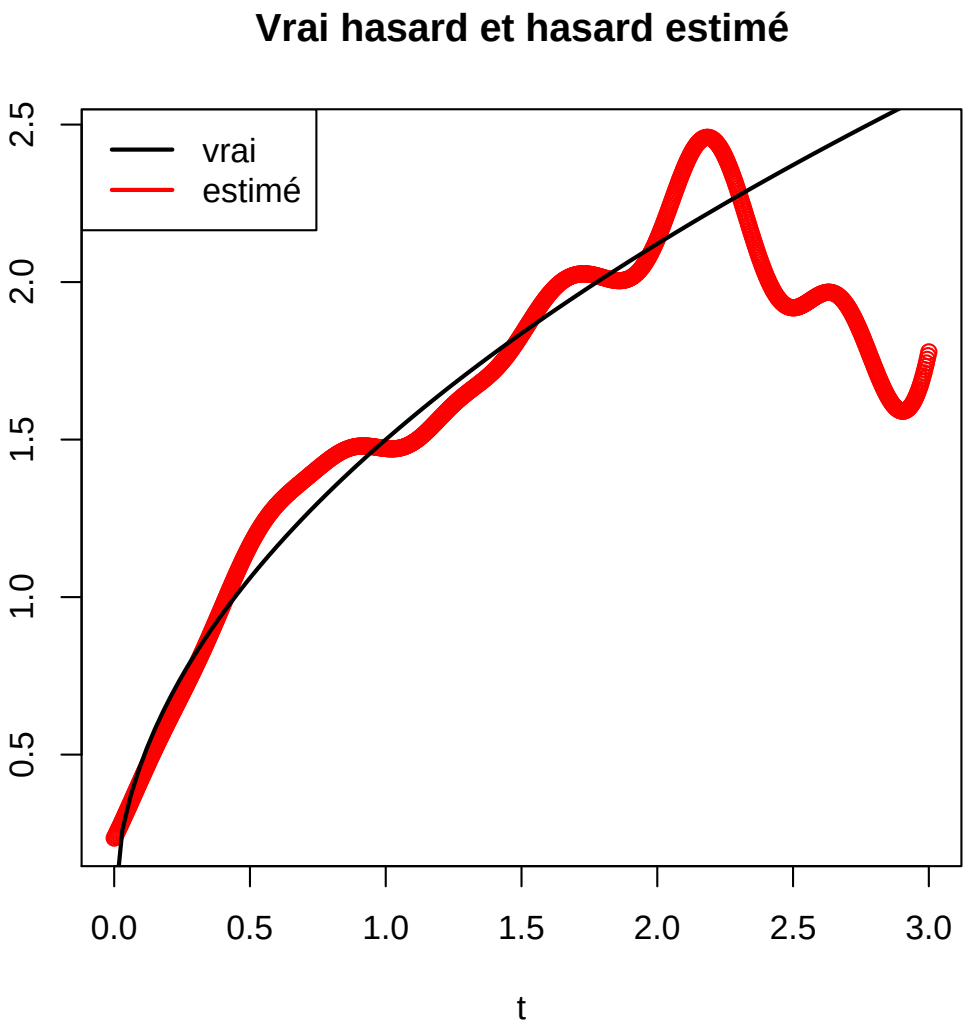
Exemple: estimation de f par noyau (4/4)



Exemple: estimation de λ par noyau (1/1)

```
# durées non-censurées
h <- 0.138
ksmooth2 <- function(x, xpts, h) {
x <- sort(x)
n <- length(x)
D <- outer(x, xpts, "-")
K <- dnorm(D / h) / (n- (1:n) + 1)
dens <- colSums(K) / (h)
}
kest2 <- ksmooth2(time, t, h)
has_true <- function(t) {theta * t^(theta - 1)}
plot(t, kest2, col=2, main="Vrai hasard et hasard estimé")
curve(has_true, add=TRUE, col = 1, lwd=2)
legend("topleft", legend = c("vrai","estimé"), col=1:2)
```

Exemple: estimation de λ par noyau (2/2)



Exemple: estimation de λ par noyau (3/3)

```
# durées censurées
status <- ifelse(time < quantile(time)[4], 1, 0)
time <- ifelse(time > quantile(time)[4], quantile(time)[4],\\
 time)
t <- seq(0,quantile(time)[4], length=1000)

ksmooth3 <- function(x, xpts, delta, h) {
x <- sort(x)
delta <- sort(delta, decreasing = TRUE)
n <- length(x)
D <- outer(x, xpts, "-")
K <- delta * dnorm(D / h) / (n - (1:n) + 1)
dens <- colSums(K) / (h)
}
kest3 <- ksmooth3(time, t, status, h)
```

Exemple: estimation de λ par noyau (4/4)