Analyse de survie : Méthodes non paramétriques

Olivier Bouaziz

olivier.bouaziz@parisdescartes.fr
http://www.math-info.univ-paris5.fr/~obouaziz

Prise en compte de la censure dans l'estimation du risque instantané

Rappels

- Le but est d'estimer la loi de $\tilde{T}$ à partir des observations :

$$\begin{cases} T_i = \min(\tilde{T}_i, C_i) \\ \Delta_i = I(\tilde{T}_i \leq C_i). \end{cases}$$

- On va montrer qu'il est possible d'estimer le risque instantané sans introduire de biais ! On rappelle :

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}[t \leq \tilde{T} < t + \Delta t | \tilde{T} \geq t]}{\Delta t} = \frac{f(t)}{S(t)}$$

Un peu de mathématiques...

- On a vu que :

$$\mathbb{P}[T \leq t, \Delta = 1] = \int_0^t (1 - G(u)) f(u) du \tag{1}$$

  où F est la f.d.r de $\tilde{T}$ et G la f.d.r de C.
- On note $H_1(t) = \mathbb{P}[T \leq t, \Delta = 1]$ la f.d.r des observations **non censurées** et f_1 sa densité.
- On note $H(t) = \mathbb{P}[T \leq t]$ la f.d.r des observations $T = \min(\tilde{T}, C)$.
- Dans toute la suite, on travaillera toujours sous l'hypothèse que $\tilde{T}$ est **indépendant** de C.

Un peu de mathématiques...

- Sous l'hypothèse de **censure indépendante**, on a

$$\begin{aligned} 1 - H(t) = \mathbb{P}[T \geq t] = \mathbb{P}[\min(\tilde{T}, C) \geq t] = \mathbb{P}[\tilde{T} \geq t, C \geq t] \\ = \mathbb{P}[\tilde{T} \geq t] \times \mathbb{P}[C \geq t] = S(t)(1 - G(t)) \end{aligned} \tag{2}$$

- On calcule la dérivée de chaque côté de l'équation (1), puis on divise chaque côté par $1 - H(t)$:

$$\begin{aligned} f_1(t) &= (1 - G(t))f(t) \\ \frac{f_1(t)}{1 - H(t)} &= \frac{1 - G(t)}{1 - H(t)} f(t) \\ \frac{f_1(t)}{1 - H(t)} &= \frac{f(t)}{S(t)}, \end{aligned}$$

d'après l'hypothèse de censure indépendante (équation (2)).

Un peu de mathématiques...

On a donc montré que le risque instantané de $\tilde{T}$, h est égal à $f_1(t)/(1-H(t))$. Or, par définition de la densité,

$$\frac{f_1(t)}{1-H(t)} = \lim_{\Delta t \to 0} \frac{\mathbb{P}[t \leq T < t+\Delta t, \Delta = 1]}{\Delta t} \times \frac{1}{\mathbb{P}[T \geq t]}$$
$$\frac{f_1(t)}{1-H(t)} = \lim_{\Delta t \to 0} \frac{\mathbb{P}[t \leq T < t+\Delta t, \Delta = 1 | T \geq t]}{\Delta t}.$$

En conclusion, on a montré que

$$\lim_{\Delta t \to 0} \frac{\mathbb{P}[t \leq T < t+\Delta t, \Delta = 1 | T \geq t]}{\Delta t} = \lim_{\Delta t \to 0} \frac{\mathbb{P}[t \leq \tilde{T} < t+\Delta t | \tilde{T} \geq t]}{\Delta t} \tag{3}$$

L'estimateur du risque instantané

- On ordonne les individus par temps observés (les T_i) croissants. On a $T_{(1)} < \cdots < T_{(l)}$ avec $l \leq n$.
- On estime le risque instantané au temps $T_{(i)}$ par d_i/R_i où
  - d_i représente le nombre d'évènements d'intérêts observés au temps $T_{(i)}$ (c'est à dire le nombre de $T_j = T_{(i)}$ pour lesquels $\Delta_j = 1$; on ne compte pas les censures !)
  - R_i représente le nombre d'individus à risque au temps $T_{(i)}$ (c'est à dire le nombre de T_j tels que $T_j \geq T_{(i)}$; **les censures sont incluses dans ce calcul !**).
- d_i/R_i est un estimateur de $\mathbb{P}[t \leq T < t + \Delta t, \Delta = 1 | T \geq t]$ dans l'équation (3), au temps $t = T_{(i)}$.

L'estimateur de Kaplan-Meier

Retour sur les données de Freireich

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

Retour sur les données de Freireich

- Dans le groupe placebo, il y a **21 patients** et **aucune donnée censurée**. On note $S_{placebo}$ la fonction de survie des patients traités par le placebo.
- Dans le groupe traité par le 6-MP, **21 patients** et **12 données censurées**. La fonction de survie va être estimée de façon différente dans les 2 groupes. On note S_{6-MP} la fonction de survie des patients traités par le 6-MP.

Groupe placebo

- Dans le groupe traité par un placebo, la fonction de survie $S_{placebo}(t)$ est simplement estimée par

$$\hat{S}_{placebo}(t) = \frac{1}{n}\sum_{i=1}^{n} I(T_i > t)$$
$$= \text{proportion d'individus tels que } T_i > t.$$

- Idée : on estime $\mathbb{P}(T > t) = \mathbb{P}(\text{ne pas rechuter avant } t)$ par la proportion de patients n'ayant pas rechutés avant t.

Groupe 6-MP, estimateur de Kaplan-Meier

- L'idée est d'écrire :

$$\begin{aligned}&\mathbb{P}(\text{être en rémission à la } i\text{ème semaine}) = \\ &\quad \mathbb{P}(\text{être en rémission à la ième semaine sachant} \\ &\quad \text{qu'il n'y a pas eu rechute à la (i-1)ème semaine}) \\ &\quad \times \mathbb{P}(\text{être en rémission à la } (i-1)\text{ème semaine})\end{aligned}$$

- On a $0 = T_{(0)} < T_{(1)} < \cdots < T_{(l)}$ avec $l \leq n$.

$$\mathbb{P}(\tilde{T} > t(i)) = \underbrace{\mathbb{P}(\tilde{T} > t_{(i)} | \tilde{T} > t_{(i-1)})}_{p_i} \times \mathbb{P}(\tilde{T} > t_{(i-1)})$$

$$S(t_{(i)}) = p_i \times S(t_{(i-1)})$$

$$S(t_{(i)}) = p_i \times p_{i-1} \times \ldots \times p_1 \times S(t_{(0)})$$

Groupe 6-MP, estimateur de Kaplan-Meier

- On estime $p_i = 1 - \mathbb{P}(\tilde{T} \leq t_{(i)} | \tilde{T} > t_{(i-1)})$ par

$$\hat{p}_i = \left(1 - \frac{d_i}{R_i}\right),$$

où

  - d_i est le nombre de rechutes observées au temps $t_{(i)}$.
  - R_i est le nombre d'individus à risque de rechute (individus toujours en rémission) juste avant $t_{(i)}$.

- L'estimateur de Kaplan-Meier (1958) est une fonction **en escalier** qui s'écrit :

$$\hat{S}_{KM}(t) = \prod_{j=1}^{i} \left(1 - \frac{d_j}{R_j}\right), \text{ où } T_{(i)} \leq t < T_{(i+1)}.$$

Application sous R

```
## Loading required package: survival
```

```
require(survival)
summary(survfit(Surv(Time,status)~groupe))
```

```
## groupe=6MP
```

```
##  time n.risk n.event survival
##     6     21       3    0.857
##     7     17       1    0.807
##    10     15       1    0.753
##    13     12       1    0.690
##    16     11       1    0.627
##    22      7       1    0.538
##    23      6       1    0.448
```

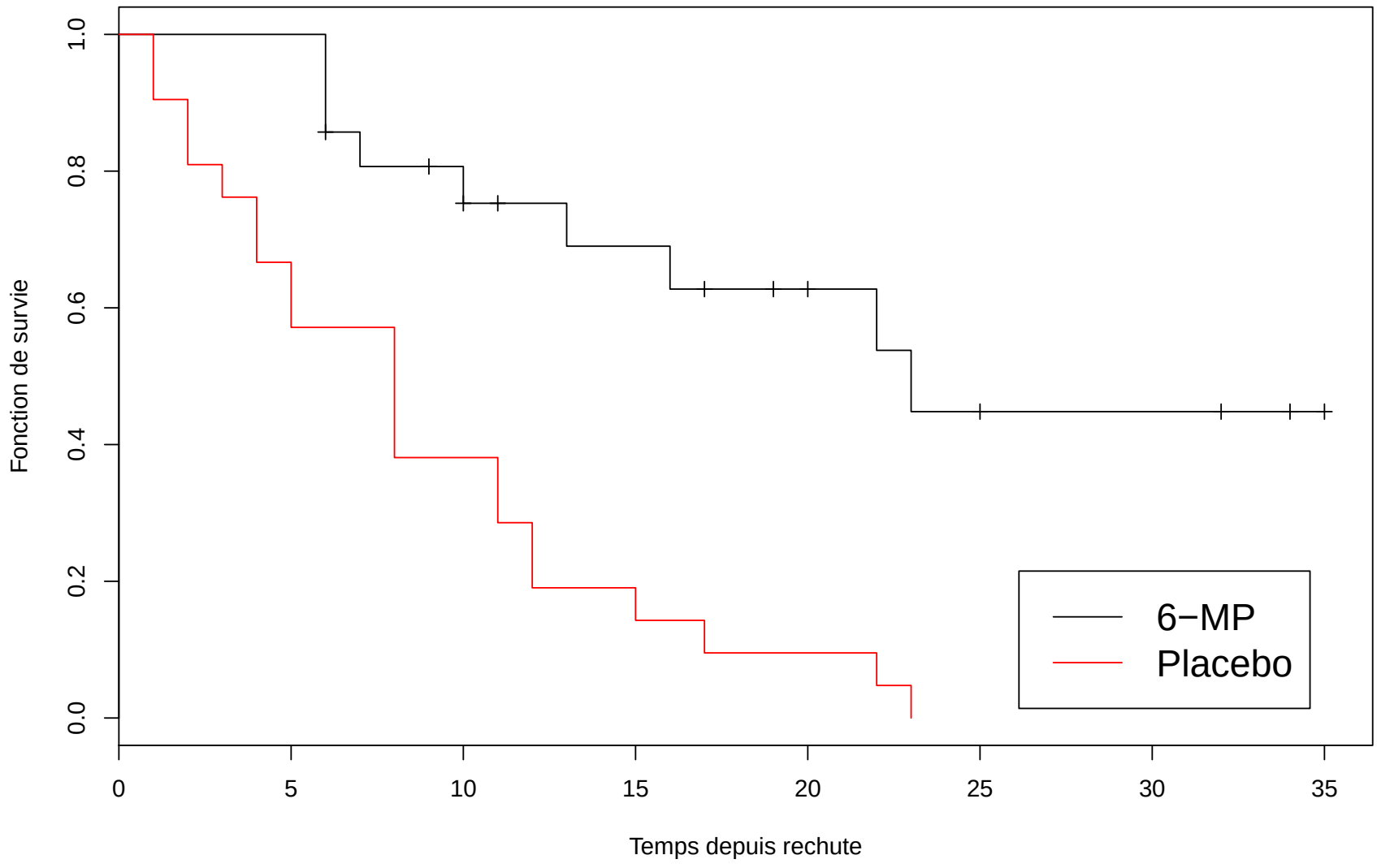
Application sous R

```
## groupe=Placebo

##  time n.risk n.event survival
##     1     21       2    0.905
##     2     19       2    0.810
##     3     17       1    0.762
##     4     16       2    0.667
##     5     14       2    0.571
##     8     12       4    0.381
##    11      8       2    0.286
##    12      6       2    0.190
##    15      4       1    0.143
##    17      3       1    0.095
##    22      2       1    0.048
##    23      1       1    0.000
```

Application sous R

```
plot(survfit(Surv(Time,status)~groupe))
```



Propriétés de l'estimateur de Kaplan-Meier

- En l'absence de censure, l'estimateur de Kaplan-Meier est équivalent à la fonction de survie empirique !
- Si $S(t) > 0$ alors,

$$0 \leq \mathbb{E}[\hat{S}_{KM}(t) - S(t)] \leq F(t)H(t)^n.$$

  L'estimateur de Kaplan-Meier est **biaisé**, mais **asymptotiquement sans biais** si $H(t) \neq 1$.
- Soit $\tau_H = \inf\{t \geq 0 : 1 - H(t) = 0\}$. On a la convergence en **probabilités** (Gill, R. 1980) :

$$\sup_{0 \leq t \leq \tau_H} |\hat{S}_{KM}(t) - S(t)| \xrightarrow[n \to \infty]{\mathbb{P}} 0.$$

Normalité asymptotique de l'estimateur de Kaplan-Meier

- Soit $\tau < \tau_H$, on a la convergence en **loi** suivante (Andersen, P. et Gill, R. 1983) :

$$\text{pour tout } t \leq \tau, \sqrt{n}(\hat{S}_{KM}(t) - S(t)) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2(t)),$$

avec

$$\sigma^2(t) = S^2(t) \int_0^t \frac{f(u)du}{S^2(u)(1-G(u))} = S^2(t) \int_0^t \frac{h(u)du}{(1-H(u))}.$$

- L'estimateur de Kaplan-Meier a des problèmes de convergence dans les queues de distribution causés par la censure.
  - Il est impossible qu'il soit consistant pour $t > \tau_H$ car il n'y a plus d'observations au delà de τ_H !
  - De plus, la normalité asymptotique n'est pas vérifiée pour $\tau < t \leq \tau_H$!!

L'estimateur de Greenwood

Greenwood, M. 1926 ; Breslow, N.E. et Crowley, J. J. 1974.

- La variance asymptotique σ^2 est estimée par l'estimateur de Greenwood qui est un estimateur **consistant**.
- On peut donc construire des intervalles de confiance de $S(t)$ de la manière habituelle :

$$\mathbb{P}\left[\hat{S}_{KM}(t) - c_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}} \leq S(t) \leq \hat{S}_{KM}(t) + c_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right] \underset{n\to\infty}{\longrightarrow} 1-\alpha$$

  en probabilité, où c_α est le quantile d'ordre α de la loi $\mathcal{N}(0,1)$.
- sous R, la sortie "std.err" contient le terme $\hat{\sigma}/\sqrt{n}$.

Intervalles de confiance ponctuels sous R

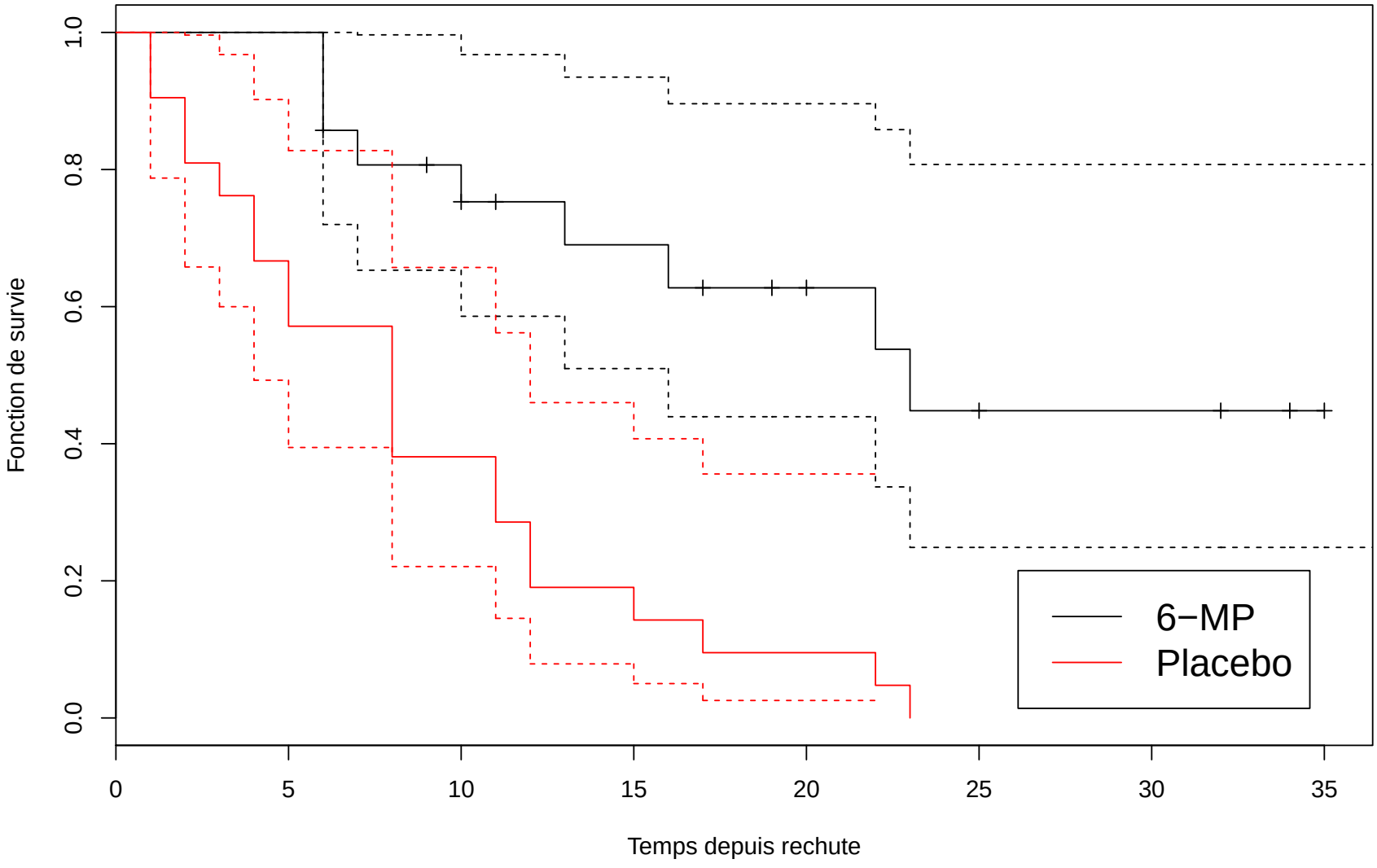
```
summary(survfit(Surv(Time,status)~groupe,conf.type="plain"))
```

```
## groupe=6MP

##  time std.err survival lower 95% CI upper 95% CI
##     6  0.0764    0.857        0.707        1.000
##     7  0.0869    0.807        0.636        0.977
##    10  0.0963    0.753        0.564        0.942
##    13  0.1068    0.690        0.481        0.900
##    16  0.1141    0.627        0.404        0.851
##    22  0.1282    0.538        0.286        0.789
##    23  0.1346    0.448        0.184        0.712
```

On a bien $0.807 - 0.0869 \times 1.96 = 0.636$; $0.807 + 0.0869 \times 1.96 = 0.977$ etc.

Intervalles de confiance ponctuels sous R



L'estimateur de Nelson-Aalen du risque cumulé

Nelson, W. 1969; Nelson, W. 1972; Aalen, O. O. 1978.

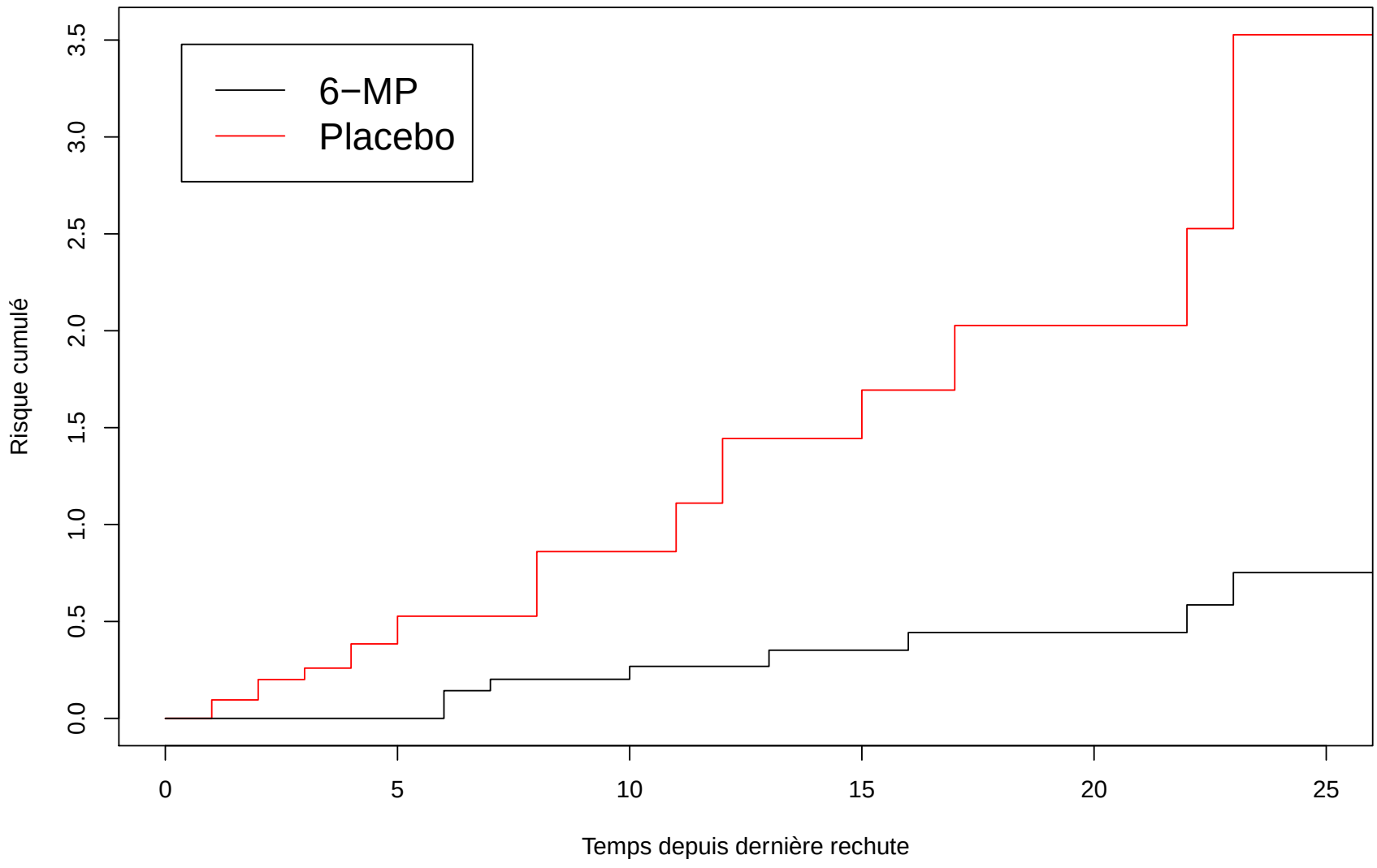
- On rappelle que le risque cumulé est défini par $H(t) = \int_0^t h(u)du$. C'est une version cumulée du risque instantané.
- On estime le risque cumulé par une fonction en escalier :

$$\hat{H}(t) = \sum_{j=1}^{i} \frac{d_j}{R_j}, \text{ où } T_{(i)} \leq t < T_{(i+1)}$$

L'estimateur de Nelson-Aalen sous R

```
result=survfit(Surv(Time,status)~groupe)
n1<-result$strata[1]; n2<-result$strata[2]
xval1=result$time[1:n1]
xval2=result$time[(n1+1):(n1+n2)]
yval1=cumsum(result$n.event[1:n1]/result$n.risk[1:n1])
yval2=cumsum(result$n.event[(n1+1):(n1+n2)]/
                result$n.risk[(n1+1):(n1+n2)])
plot(c(0,xval2,30),c(0,yval2,yval2[(n2)]),type="s",col="red",
     xlim=c(0,25),xlab="Temps depuis dernière rechute",
     ylab="Risque cumulé")
lines(c(0,xval1),c(0,yval1),type="s")
legend("topleft",c("6-MP","Placebo"),col=c(1,2),lty=c(1,1),
       cex=1.6,inset = 0.05 )
```

L'estimateur de Nelson-Aalen sous R



L'estimateur de Breslow du risque cumulé

- A partir de l'estimateur de Kaplan-Meier, on peut définir un estimateur alternatif du risque cumulé (Breslow, N. E. 1972).
- On utilise la formule

$$H(t) = -\log(S(t))$$

- L'estimateur de Breslow s'écrit :

$$\hat{H}(t) = -\log(\hat{S}_{KM}(t)).$$

- Les estimateurs de Nelson-Aalen et Breslow sont quasiment égaux en pratique !

Estimation de quantités d'intérêt : quantiles et moyenne

Estimation des quantiles

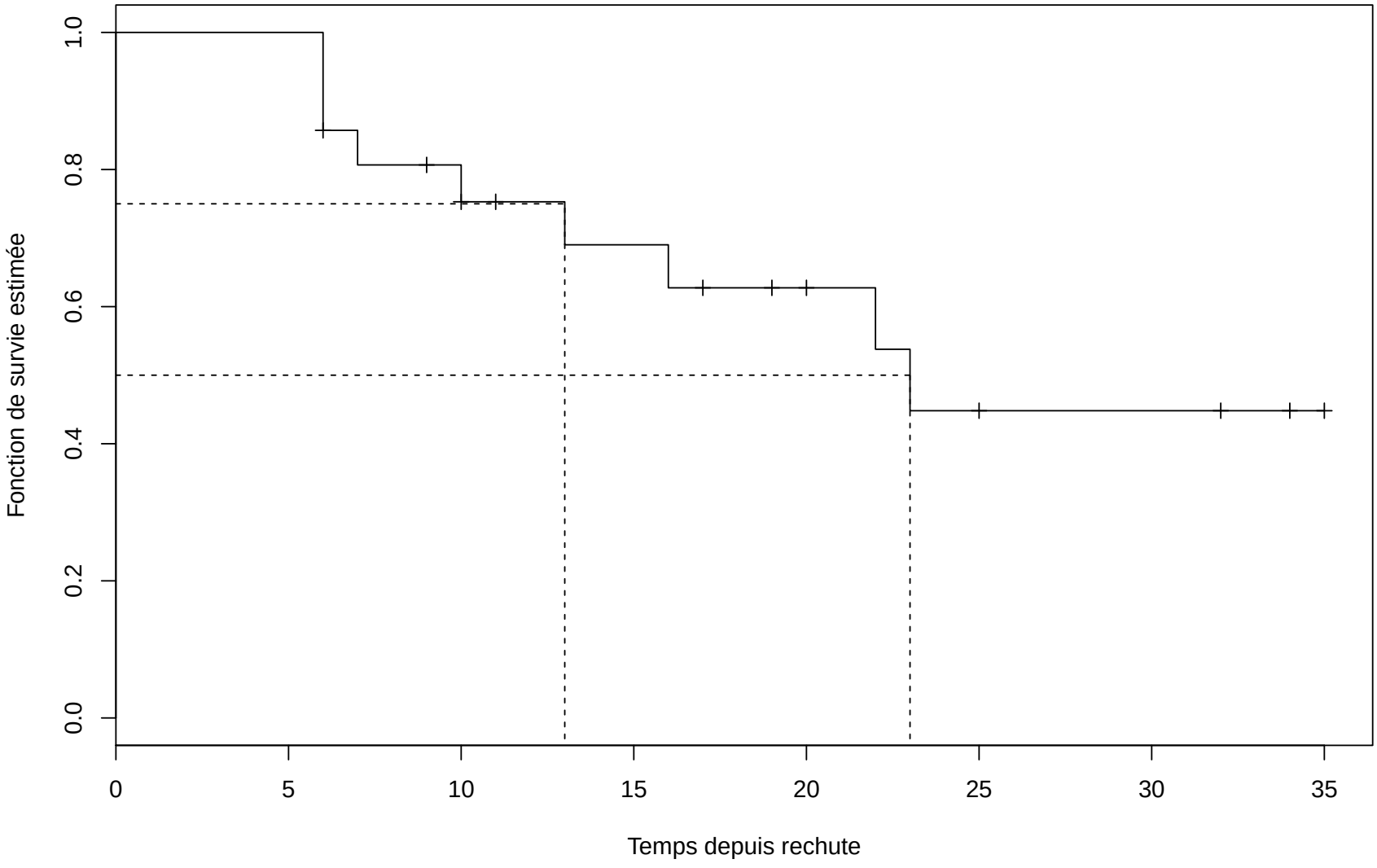
```
summary(survfit(Surv(Time,status)~groupe))
```

```
## groupe=6MP

##  time n.risk n.event survival
##     6     21       3    0.857
##     7     17       1    0.807
##    10     15       1    0.753
##    13     12       1    0.690
##    16     11       1    0.627
##    22      7       1    0.538
##    23      6       1    0.448
```

Donner une estimation du premier quartile et de la médiane dans le groupe 6-MP. Que peut-on dire concernant le troisième quartile ?

Estimation des quantiles



Estimation de l'espérance

- On a vu en cours la formule :

$$\mathbb{E}[\tilde{T}] = \int_0^\infty S(t)dt.$$

- On peut donc estimer l'espérance en calculant l'aire sur la courbe de $\hat{S}_{KM}$, ce qui est facile puisque $\hat{S}_{KM}$ est une fonction en escalier et il suffit donc d'additionner des aires de rectangles.
- Mais on a un problème si la dernière observation est censurée ! Le dernier rectangle a une aire infinie. Selon où on "coupe", on obtient une moyenne différente.
- A cause des problèmes d'estimation dans les **queues de distribution**, on ne peut pas proposer d'estimateur sans biais de l'espérance.
- On préférera estimer les quantiles : ces estimateurs sont très robustes et asymptotiquement sans biais !
- Même problème pour estimer la variance !

Estimation de l'espérance sous R

```
result<-survfit(Surv(Time,status)~groupe)
print(result, print.rmean=TRUE,rmean=23)
```

```
## Call: survfit(formula = Surv(Time, status) ~ groupe)
##
##                records n.max n.start events *rmean *se(rmean)
## groupe=6MP          21    21       21      9  17.91       1.55
## groupe=Placebo      21    21       21     21   8.67       1.38
##                0.95LCL 0.95UCL
## groupe=6MP          16      NA
## groupe=Placebo       4      12
##     * restricted mean with upper limit =  23
```

Estimation de l'espérance sous R

```
print(result, print.rmean=TRUE,rmean=30)
```

```
## Call: survfit(formula = Surv(Time, status) ~ groupe)
##
##                records n.max n.start events *rmean *se(rmean)
## groupe=6MP          21    21       21      9  21.05       2.24
## groupe=Placebo      21    21       21     21   8.67       1.38
##                0.95LCL 0.95UCL
## groupe=6MP          16      NA
## groupe=Placebo       4      12
##     * restricted mean with upper limit =  30
```

Estimation de l'espérance sous R

```
print(result, print.rmean=TRUE,rmean=35)
```

```
## Call: survfit(formula = Surv(Time, status) ~ groupe)
##
##                records n.max n.start events *rmean *se(rmean)
## groupe=6MP          21    21       21      9  23.29       2.83
## groupe=Placebo      21    21       21     21   8.67       1.38
##                0.95LCL 0.95UCL
## groupe=6MP          16      NA
## groupe=Placebo       4      12
##     * restricted mean with upper limit =  35
```

Tests de comparaison des courbes de survie

But du test

Notons S_A et S_B les fonctions de survie dans deux groupes A et B. Par exemple, A est le groupe Placebo et B le groupe $6-MP$ dans les données de Freireich.

On souhaite tester :

$$(H_0) : S_A = S_B \text{ contre } (H_1) : S_A \neq S_B.$$

Dans la suite, on va proposer un **test non-paramétrique** asymptotique qui marche en présence de données censurées.

Rappels en l'absence de données censurées

Si il n'y avait pas de données censurées, pour comparer la loi de $\tilde{T}$ entre les groupes A et B on peut proposer des tests paramétriques comme :

- Test de comparaison d'espérance : le test de Student.
- Test de comparaison de variance : le test Levene (ou Bartlett ou Fisher dans le cas Gaussien).

On peut également utiliser des tests non-paramétriques pour tester

$$(H_0) : S_A = S_B \text{ contre } (H_1) : S_A \neq S_B.$$

- Test de Kolomogorov Smirnov de comparaison des f.d.r.
- Test de la somme des rangs ou test de Mann-Whitney.

En présence de données censurées

On généralise les tests non-paramétriques usuels aux tests du log-rang (log-rank en anglais) et ses extensions.

- le test du log-rang ; Gehan, E. A. 1965 et Mantel, N. 1966.
- le test de Gehan-Wilcoxon; Gehan, E. A. 1965.
- le test de Prentice-Wilcoxon ou Peto-Wilcoxon; Prentice, R. L. 1978 et Peto R., Peto, J. 1972.

Principe du test du log-rang

On ordonne par ordre croissant les individus par les temps observés T_i dans les deux groupes A et B réunis. On a $T_{(1)} < \cdots < T_{(l)}$ avec $l \leq n$. On note :

- $d_{B,i}$: nombre de décès observés au temps $T_{(i)}$ dans le groupe B.
- $R_{B,i}$: nbre de sujets exposés au risque de décès juste avant $T_{(i)}$, dans le groupes B.

Mêmes notations pour le groupe A ($d_{A,i}$ et $R_{A,i}$).

- $e_{B,i}$: nombre de décès **attendus** (i.e sous (H_0)) au temps $T_{(i)}$ dans le groupe B,

$$e_{B,i} = \frac{d_{A,i} + d_{B,i}}{R_{A,i} + R_{B,i}} \times R_{B,i}$$

- w_i : poids associé au temps $T_{(i)}$.

Principe du test du log-rang

La statistique de test compare les décès **observés** dans le groupe B aux décès **attendus sous** (H_0) dans le groupe B :

$$U = \sum_{i=1}^{I} w_i \left(d_{B,i} - e_{B,i}\right).$$

On peut montrer que **sous** (H_0) : $\mathbb{E}[U] = 0$ et

$$\frac{U}{\sqrt{\hat{V}}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

avec $\hat{V} = \sum_{i=1}^{I} w_i^2 \, v_i$ et les v_i qui s'écrivent en fonction de $R_{A,i}$, $R_{B,i}$, $d_{A,i}$ et $d_{B,i}$.

Statistique de test et zone de rejet

La statistique de test usuel est :

$$T_n = \frac{U^2}{\hat{V}}.$$

- On a, sous (H_0), $T_n \sim \chi^2(1)$.
- Pour un test **asymptotique** de niveau α, la zone de rejet est telle que $R_\alpha = \{T_n \geq c_\alpha\}$ où c_α est le quantile d'ordre $1-\alpha$ de la loi $\chi^2(1)$.
- La p-valeur du test est égale (quand n est *grand*) à :

$$\mathbb{P}_{H_0}[T_n \geq t_n] \approx \mathbb{P}[\chi^2(1) \geq t_n] = 1 - \phi(t_n),$$

où ϕ est la f.d.r de la loi $\chi^2(1)$.

Choix du poids attribué à chaque individu

Le choix des w_i donne un test différent.

- $w_i = 1, \forall i = 1, \ldots, n$ donne le test du **log-rang**.
- $w_i = R_{A,i} + R_{B,i}, \forall i = 1, \ldots, n$ donne le test de **Gehan-Wilcoxon**. Il donne plus de poids aux évènements (les T_i pour lesquels $\Delta_i = 1$) qui se produisent à des temps précoces.
- $w_i = \hat{S}_{KM}(T_{(i)}), \forall i = 1, \ldots, n$ donne le test de **Peto/Prentice**. On l'appelle également le **test du log-rang généralisé**. Il donne également plus de poids aux évènements (les T_i pour lesquels $\Delta_i = 1$) qui se produisent à des temps précoces.

Remarques

- Le test fait intervenir uniquement le **rang** des observations.
- Le test s'étend facilement à plus de deux groupes. La statistique de test suit asymptotiquement une loi du χ^2 dont le nombre de degrés de liberté est égal aux nombres de groupes moins 1.
- Quand il n'y a que deux groupes à comparer, on a :

$$\sum_{i=1}^{I} w_i \left(d_{B,i} - e_{B,i}\right) = -\sum_{i=1}^{I} w_i \left(d_{A,i} - e_{A,i}\right)$$

- Le choix des poids w_i influence la puissance des tests.
- On peut facilement montrer quand il n'y a que deux groupes que la statistique de test peut s'écrire :

$$U = \sum_{i=1}^{I} w_i \frac{R_{A,i} R_{B,i}}{R_{A,i} + R_{B,i}} \left(\frac{d_{B,i}}{R_{B,i}} - \frac{d_{A,i}}{R_{A,i}}\right).$$

Application sur les données de Freireich (le test du log-rang)

```
survdiff(Surv(Time,status)~groupe)
```

```
## Call:
## survdiff(formula = Surv(Time, status) ~ groupe)
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## groupe=6MP      21        9     19.3      5.46      16.8
## groupe=Placebo 21       21     10.7      9.77      16.8
##
##  Chisq= 16.8  on 1 degrees of freedom, p= 4.17e-05
```

Application sur les données de Freireich (le test du log-rang généralisé)

```
survdiff(Surv(Time,status)~groupe,rho=1)
```

```
## Call:
## survdiff(formula = Surv(Time, status) ~ groupe, rho = 1)
##
##                 N Observed Expected (O-E)^2/E (O-E)^2/V
## groupe=6MP     21     5.12    12.00      3.94      14.5
## groupe=Placebo 21    14.55     7.68      6.16      14.5
##
##  Chisq= 14.5  on 1 degrees of freedom, p= 0.000143
```

Le test du log-rang stratifié

- Le test du log-rang ne compare que deux groupes d'individus, sans prendre en compte d'autres variables.
- Le test du log-rang stratifié permet d'ajuster sur d'autres variables, pour comparer des individus comparables entre eux.
- On considère que les données sont divisées en S strates et que l'on veut comparer deux groupes A et B.

Les données s'écrivent : $T_{(1s)} < T_{(2s)} < \cdots < T_{(l_s s)}$ pour $s = 1, \ldots, S$.

- On calcule comme précédemment, $d_{B,s}$ et $e_{B,s}$ où le calcul ne s'effectue que dans la strate s pour le groupe B.
- La statistique de test est :

$$\frac{\sum_{s=1}^{S}(d_{B,s} - e_{B,s})}{\sqrt{\sum_s \hat{V}_s}} \xrightarrow[(H_0)]{\mathcal{L}} \chi^2(1)$$

Application sur les données de mélanome

```
library(ISwR)
```

```
##
## Attaching package: 'ISwR'
```

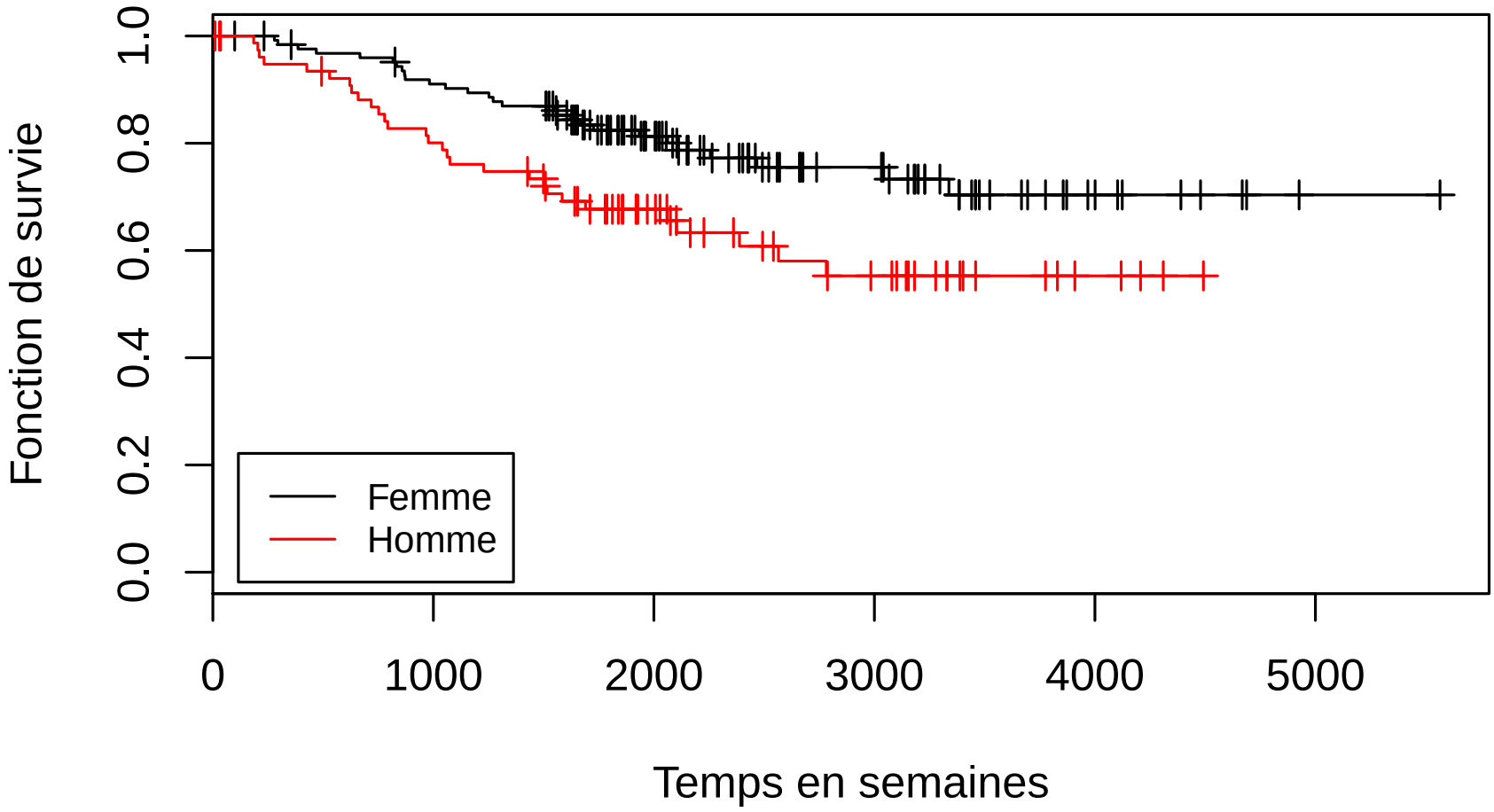
```
## The following object is masked from 'package:survival':
##
##     lung
```

```
head(melanom)
```

```
##    no status days ulc thick sex
## 1 789      3   10   1   676   2
## 2  13      3   30   2    65   2
## 3  97      2   35   2   134   2
## 4  16      3   99   2   290   1
## 5  21      1  185   1  1208   2
## 6 469      1  204   1   484   2
```

La survie en fonction du sexe

```
plot(survfit(Surv(days,status==1)~sex,data=melanom),
col = c(1,2),xlab="Temps en semaines",ylab="Fonction de survie")
legend("bottomleft", c("Femme","Homme"), cex=0.8,
inset=0.02,col=c("black","red"),lty=1)
```



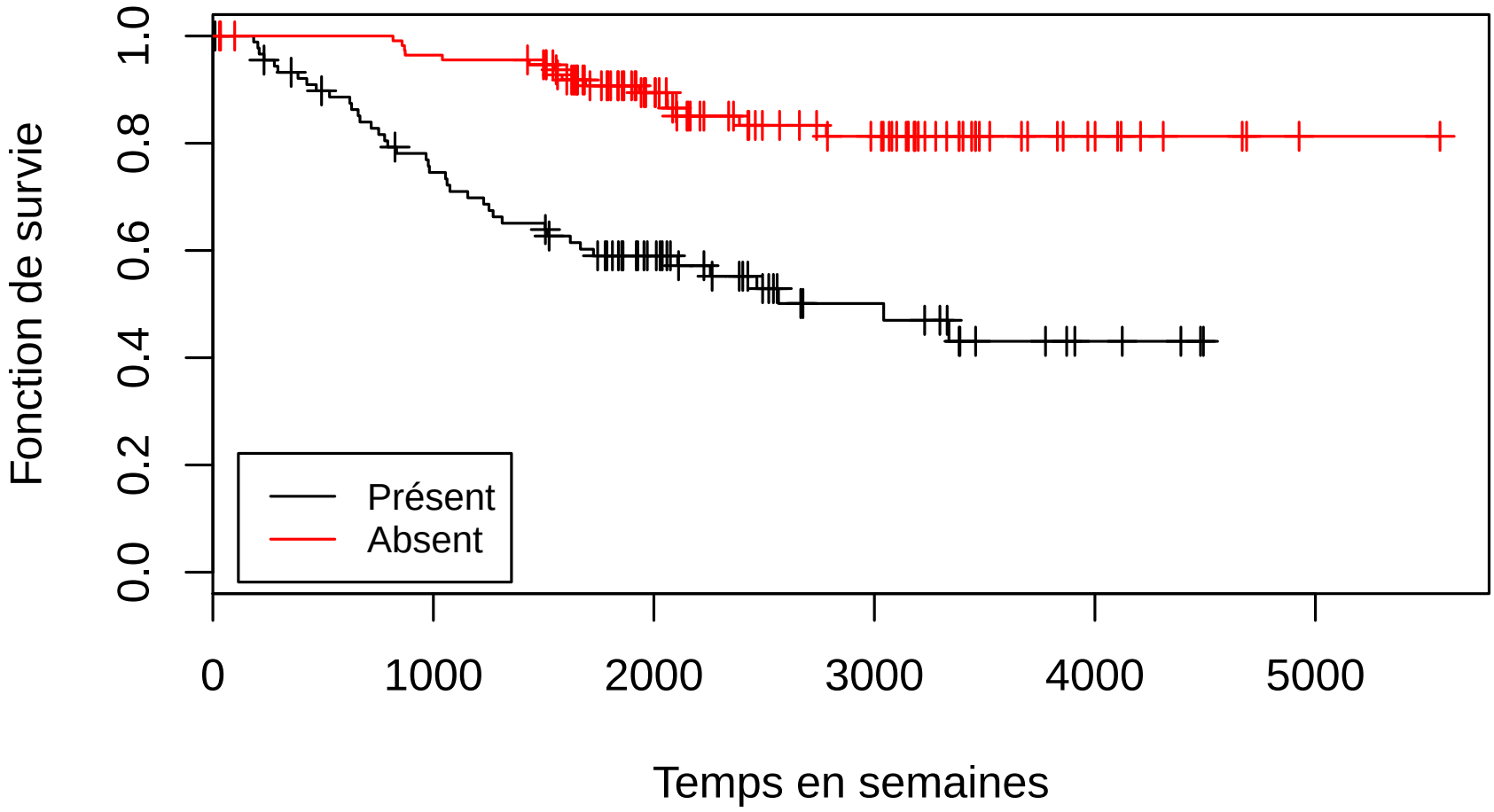
La survie en fonction du sexe

```
survdiff(Surv(days,status==1)~sex,data=melanom)
```

```
## Call:
## survdiff(formula = Surv(days, status == 1) ~ sex, data = mela
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=1 126       28     37.1      2.25      6.47
## sex=2  79       29     19.9      4.21      6.47
##
##  Chisq= 6.5  on 1 degrees of freedom, p= 0.011
```

La survie en fonction de l'ulcération

```
plot(survfit(Surv(days,status==1)~ulc,data=melanom),
col = c(1,2),xlab="Temps en semaines",ylab="Fonction de survie")
legend("bottomleft", c("Présent","Absent"), cex=0.8,
inset = 0.02, col=c("black","red"),lty=1)
```



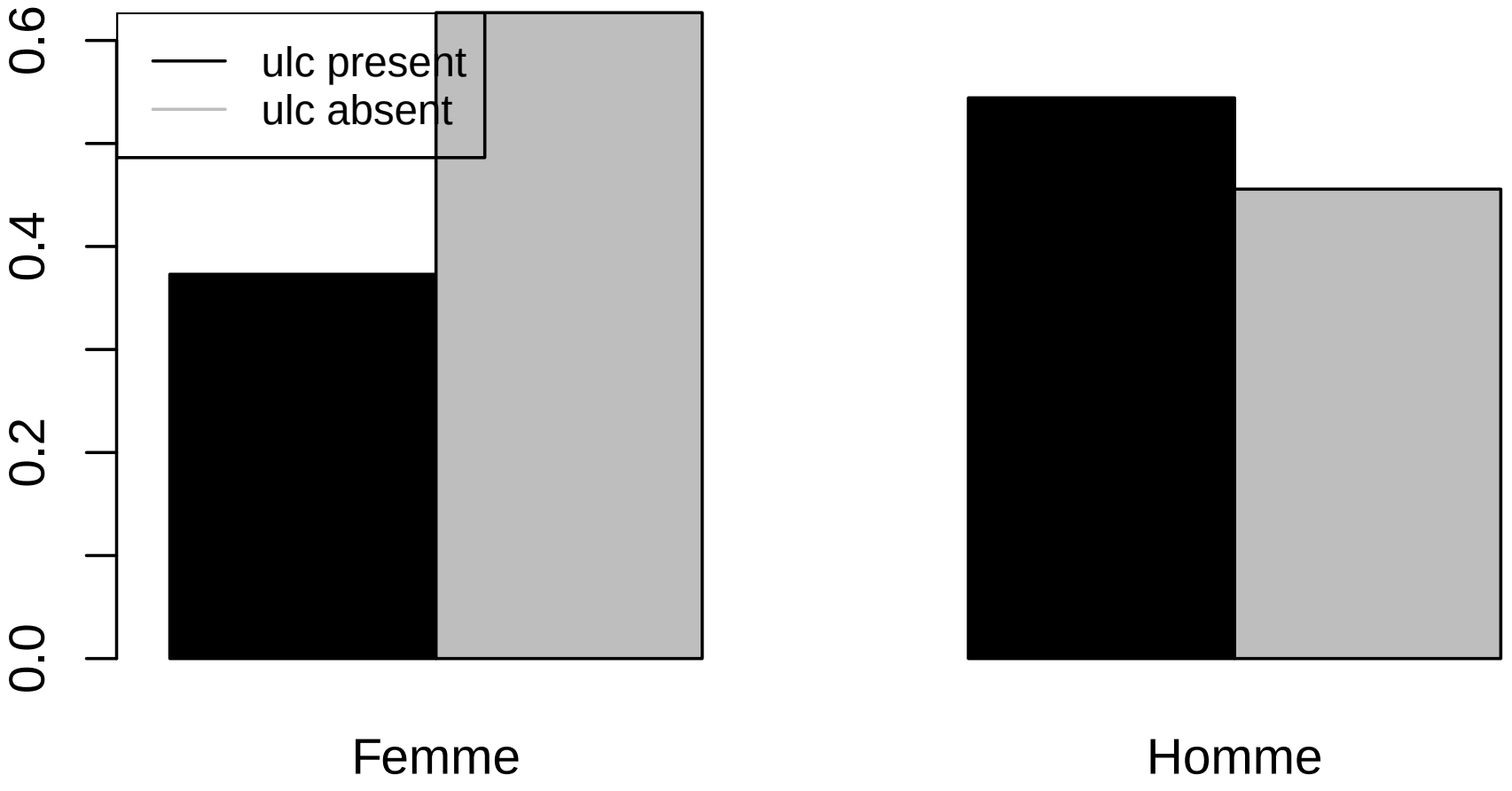
La survie en fonction de l'ulcération

```
survdiff(Surv(days,status==1)~ulc,data=melanom)
```

```
## Call:
## survdiff(formula = Surv(days, status == 1) ~ ulc, data = mela
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## ulc=1  90       41     21.2      18.5      29.6
## ulc=2 115       16     35.8      10.9      29.6
##
##  Chisq= 29.6  on 1 degrees of freedom, p= 5.41e-08
```

Lien entre ulcération et sexe

```
TabProp=with(melanom,prop.table(table(sex,ulc),margin=1))
rownames(TabProp)=c("Femme","Homme")
barplot(t(TabProp),beside=TRUE,col=c(1,8))
legend("topleft", c("ulc present","ulc absent"), cex=0.8, col=c(
```



Lien entre ulcération et sexe

```
with(melanom,chisq.test(sex,ulc))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  sex and ulc
## X-squared = 5.1099, df = 1, p-value = 0.02379
```

Sexe et ulcération sont très fortement liés !! Les hommes ont plus tendance à avoir de l'ulcération que les femmes !

Test du log-rang pour le sexe, stratifié sur l'ulcération

```
survdiff(Surv(days,status==1)~sex+strata(ulc),data=melanom)
```

```
## Call:
## survdiff(formula = Surv(days, status == 1) ~ sex + strata(ulc
##     data = melanom)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=1 126       28     34.7      1.28      3.31
## sex=2  79       29     22.3      1.99      3.31
##
##  Chisq= 3.3  on 1 degrees of freedom, p= 0.0687
```

Après stratification sur l'ulcération, l'effet sexe est **beaucoup moins significatif** (il passe d'une pvaleur de 0.011 à une p-valeur de 0.0687).