# Analyse de données censurées : Méthodes non-paramétriques

# 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

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

- 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 :

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

- On a $0 = \tau_0 < \tau_1 < \ldots < \tau_L$ avec $L \leq n$.

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

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

$$S(\tau_i) = p_i \times p_{i-1} \times \ldots \times p_1 \times S(\tau_0)$$

# Groupe 6-MP, estimateur de Kaplan-Meier

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

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

où

  - $d_i$ est le nombre de rechutes observées au temps $\tau_i$ (en ne comptant pas les censures)
  - $R_i$ est le nombre d'individus à risque de rechute (individus toujours en rémission) juste avant $\tau_i$.

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

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

# Application sous R

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

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

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

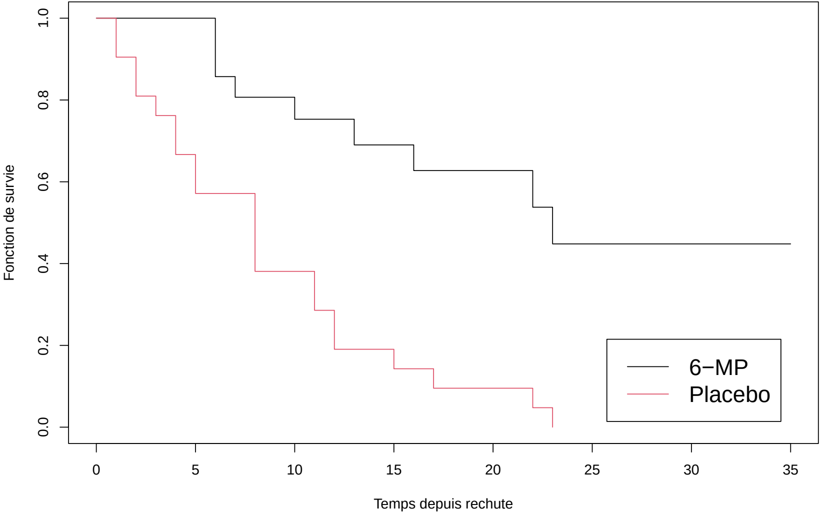
```
##  time n.risk n.event survival
##     6     21       3    0.857
##     7     17       1    0.807
##    10     15       1    0.753
##    13     12       1    0.690
##    16     11       1    0.627
##    22      7       1    0.538
##    23      6       1    0.448
```

# 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 !
- L'estimateur de Kaplan-Meier est consistant et asymptotiquement normal sauf dans les "queues de distribution".
- La variance asymptotique $\sigma^2$ est estimée par l'estimateur de Greenwood qui est un estimateur **consistant** (Greenwood, M. 1926 ; Breslow, N.E. et Crowley, J. J. 1974.)

$$\hat{\sigma}_{KM}(t) = (\hat{S}_{KM}(t))^2 \sum_{\tau_j \leq t} \frac{d_j}{R_j(R_j - d_j)}$$

# Intervalles de confiance de l'estimateur de Kaplan-Meier

- On peut donc construire des intervalles de confiance de $S(t)$ de la manière habituelle :

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

  où $c_\alpha$ est le quantile d'ordre $\alpha$ 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

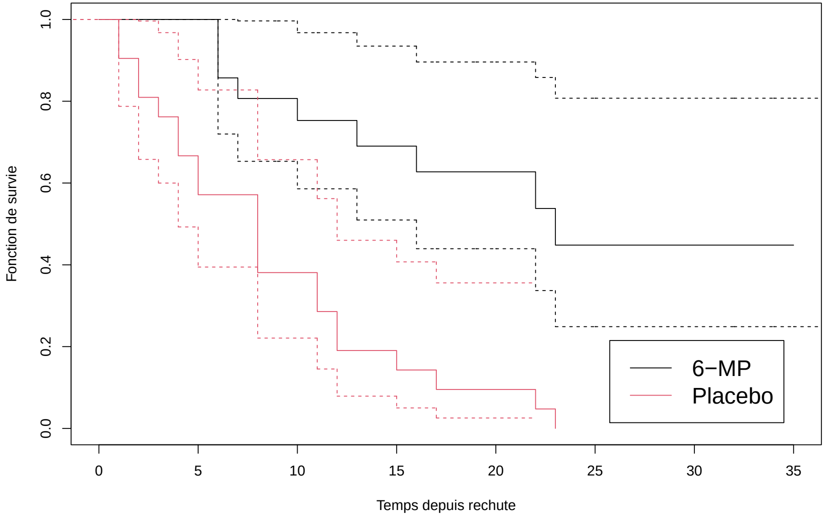
```
summary(survfit(Surv(Time,status)~groupe,conf.type="plain"))
```

```
## 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



# Estimation de quantités d'intérêt : quantiles et moyenne

# Estimation des quantiles

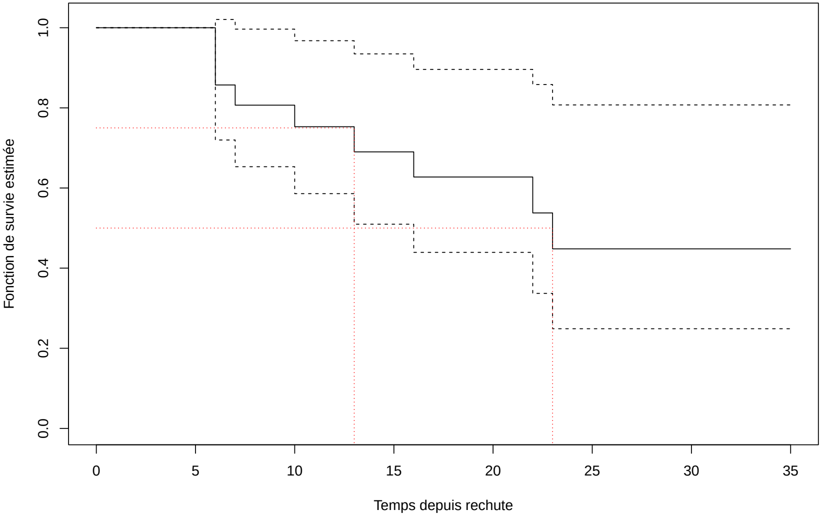
```
summary(survfit(Surv(Time,status)~groupe))
```

```
## 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

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

```
## Call: survfit(formula = Surv(Time, status) ~ groupe)
##
##                   n events rmean* se(rmean) median 0.95LCL 0.95
## groupe=6MP       21      9  17.91      1.55     23      16
## groupe=Placebo 21     21   8.67      1.38      8       4
##     * restricted mean with upper limit =  23
```

# Estimation de l'espérance sous R

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

```
## Call: survfit(formula = Surv(Time, status) ~ groupe)
##
##                   n events rmean* se(rmean) median 0.95LCL 0.95
## groupe=6MP      21      9  21.05      2.24     23      16
## groupe=Placebo 21     21   8.67      1.38      8       4
##     * restricted mean with upper limit =  30
```

# Estimation de l'espérance sous R

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

```
## Call: survfit(formula = Surv(Time, status) ~ groupe)
##
##                    n events rmean* se(rmean) median 0.95LCL 0.95
## groupe=6MP        21      9  23.29      2.83     23      16
## groupe=Placebo 21     21   8.67      1.38      8       4
##     * 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.

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 $\tau_i$ dans les deux groupes $A$ et $B$ réunis. On a $\tau_1 < \cdots < \tau_L$ avec $L \leq n$. On note :

- $d_{B,i}$: nombre de décès observés au temps $\tau_i$ dans le groupe $B$.
- $R_{B,i}$ : nombre de sujets exposés au risque de décès juste avant $\tau_i$, 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 $\tau_i$ 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 $\tau_i$.

# 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 $B$ :

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

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

$$v_i = \frac{R_{A,i} R_{B,i}}{(R_{A,i} + R_{B,i})^2} \frac{\left(d_{A,i} + d_{B,i}\right)\left((R_{A,i} + R_{B,i}) - (d_{A,i} + d_{B,i})\right)}{R_{A,i} + R_{B,i} - 1}.$$

# 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 $\alpha$, la zone de rejet est telle que $R_\alpha = \{T_n \geq c_\alpha\}$ où $c_\alpha$ 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 $\phi$ 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 écarts entre $d_{B,i}$ et $e_{B,i}$ qui se produisent à des temps précoces.
- $w_i = \hat{S}_{KM}(\tau_i), \forall i = 1, \ldots, n$ donne le test de **Peto/Prentice**. On l'appelle également le **test du log-rang généralisé**. Il donne également plus de poids aux écarts entre $d_{B,i}$ et $e_{B,i}$ 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 $\chi^2$ 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)

```r
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= 4e-05
```