

# Analyse de Survie

Thibault ALLART

CNAM

2017

- 1 Généralités sur l'analyse de survie
  - Introduction
  - Définitions
  - Censure
  - Propriétés des fonctions de survie
- 2 Estimation et comparaison des courbes de survie
  - L'estimateur de Kaplan-Meier
  - L'estimateur de Nelson-Aalen
  - Tests de comparaison
- 3 Le modèle de Cox
  - Selection de variables

On s'intéresse au temps de réalisation d'un évènement

- Durée de vie d'un patient atteint d'un cancer.
- Durée d'un composant électronique avant une panne.
- Durée d'utilisation d'un service par des clients.

L'évènement d'intérêt peut être censuré.

- Mortalité : Quelle est l'espérance de vie d'une population ?
- Durée de vie d'un appareil électronique
- Peut on démontrer l'efficacité d'un traitement médical ?
- Comment améliorer la durée d'abonnement de mes clients ?

# Les deux méthodes Statistique les plus cités

## Nonparametric estimation from incomplete observations

EL Kaplan, P Meier - Journal of the American statistical association, 1958 - Taylor & Francis

Abstract In lifetesting, medical follow-up, and other fields the observation of the time of occurrence of the event of interest (called a death) may be prevented for some of the items of the sample by the previous occurrence of some other event (called a loss). Losses may be

Cité 48411 fois [Autres articles](#) [Web of Science: 42852](#) [Citer](#) [Enregistrer](#) [Plus](#)

## Regression Models and Life-Tables - JStor

<https://www.jstor.org/stable/2985181> - Traduire cette page

de DR Cox - 1972 - Cité 42495 fois [Autres articles](#)

1972] 187. Regression Models and Life-Tables. BY D. R. Cox. Imperial College, London. [Read before the ROYAL STATISTICAL SOCIETY, at a meeting ...

On distingue l'évènement d'intérêt

- Décès du patient après l'apparition du cancer
- Panne du composant
- Fin de contrat du client

de la variable à expliquer durée avant l'apparition de l'évènement

- temps écoulé avant le décès
- temps écoulé avant la panne
- temps écoulé avant la fin du contrat

## Définition : Durée de Vie ou Survie

La durée de survie désigne le temps qui s'écoule depuis un instant initial(début du traitement, diagnostic, ...), jusqu'à la survenue d'un évènement d'intérêt final (décès du patient, rechute, rémission, guérison, ...).

La variable étudiée est appelée **durée de vie  $T$** .

On dit que le patient survit au temps  $t$  si, à cet instant, l'évènement d'intérêt final n'a pas encore eu lieu.

$T$  est une variable aléatoire positive continue.

Supposons que l'étude soit un essai clinique portant sur deux groupes de patients, recevant 2 types de traitements. Deux questions importantes se posent aux médecins :

- ❶ L'un des deux traitements est-il plus efficace que l'autre en terme d'amélioration de la survie des patients ?
- ❷ Peut-on mettre en évidence des **facteurs pronostiques** qui améliorent/détériorient la survie ?

Exemple : âge, sexe, tabagisme, antécédents familiaux, ...



Pour répondre à ces questions, on peut :

- ❶ Mettre en place des méthodes statistiques qui vont permettre de comparer les deux groupes de patients qui reçoivent les deux types de traitement.
- ❷ Proposer un modèle qui relie la durée de survie des patients à des variables explicatives et mettre en évidence des facteurs pronostiques.

La durée de survie n'est pas toujours complètement observée. Pour certains individus, l'évènement d'intérêt n'est pas observé.

## Definition : Censure

La durée  $T$  est dite censurée si la durée n'est pas intégralement observée.

Exemples :

- Un patient peut être perdu de vue (déménagement, ...)
- un évènement peut survenir et entraîner la sortie de l'étude : le patient peut décéder d'une autre cause que de la maladie étudiées.
- L'étude s'arrête alors que des individus sont encore vivants

# Les différents types de censure

La censure est dite indépendante si elle n'apporte pas d'information sur la durée de survie.

Du fait de la censure, on ne peut pas utiliser les méthodes statistiques classiques (t-test, régression linéaire). On ne peut même pas calculer de moyenne !

Les différents types de censure :

- ① censure de type I : fixée
- ② censure de type II : attente
- ③ censure de type III : aléatoire

Pour chaque individu  $i \in 1, \dots, n$ , on note :

- $T_i^*$  le temps de survie (pas toujours observé)
- $C_i$  le temps de censure
- $\delta_i$  l'indicateur de censure (1 observé, 0 censuré)

En pratique on observe  $T_i = \min(T_i^*, C_i)$  et  $\delta_i$

En survie, les observations sont donc  $(T_1, \delta_1), \dots, (T_n, \delta_n)$

$$(T_i, \delta_i) = \begin{cases} (T_i, 1) & \text{si } C_i \geq T_i^* & \text{non censuré} \\ (T_i, 0) & \text{si } C_i < T_i^* & \text{censuré} \end{cases}$$

La censure de type I correspond à une censure à date fixe.

$$C_i = C \quad \forall i \in 1, \dots, n$$

Exemple :

- L'étude prend fin le 10 mars.
- On dispose des données clients jusqu'à aujourd'hui.

Voir dessin au tableau

Dans le cas de la censure de type II, on observe les individus jusqu'à ce que  $r$  d'entre eux aient vu l'évènement d'intérêt se produire.

On observe  $T_1 < T_2 < \dots < T_r$

Les temps de censure  $C_i$  sont aléatoires et indépendants des temps d'évènement  $T_i$ .

Voir dessin au tableau

Dans la suite du cours, on considère que la censure est aléatoire.

De plus on suppose que la censure est indépendante du temps de survie (censure non informative). Mais si la censure est due à l'arrêt du traitement, l'hypothèse d'indépendance n'est pas valide.



Freireich (1963) à observé la durée de rémission (en semaines) de patients atteints de leucémie aiguë, traités soit par placebo soit par 6-mercaptopurine (6-MP)

6-MP	6	6	6	6 <sup>+</sup>	7	9 <sup>+</sup>	10	10 <sup>+</sup>	11 <sup>+</sup>	13		
	16	17 <sup>+</sup>	19 <sup>+</sup>	20 <sup>+</sup>	22	23	25 <sup>+</sup>	32 <sup>+</sup>				
	32 <sup>+</sup>	34 <sup>+</sup>	35 <sup>+</sup>									
Placebo	1	1	2	2	3	4	4	5	5	8	8	8
	8	11	11	12	12	15	17	22	23			

Le

signe + correspond à des patients qui ont quitté l'étude à la date considérée. Pour ces individus, les durées sont donc censurées.

Les données censurées demandent un traitement particulier. Si on enlève les données censurées → perte d'information.

Dans l'exemple précédent, si on enlève les données censurées, on ne tient pas compte des durées de rémission les plus longues et on sous-évalue l'effet du traitement 6-MP.

## 5 fonctions caractéristique

La loi du temps de survie  $T$  est décrite par 5 fonctions :

- La fonction de survie  $S(t)$
- La fonction de répartition  $F(t)$
- La densité  $f(t)$
- Le taux de risque instantané  $\lambda(t)$
- Le taux de risque cumulé  $H(t)$

A partir de l'une d'elles, on peut déduire toute les autres.

La fonction de survie  $S(t)$  représente la probabilité de survivre au moins jusqu'au temps  $t$ .

$$S(t) = \mathbb{P}(T \geq t)$$

## propriétés

- $S(t) = 1 - F(t)$

La fonction de répartition désigne la probabilité que l'évènement d'intérêt ait lieu avant  $t$

$$F(t) = \mathbb{P}(T \leq t)$$

## propriétés

- $F(t) = 1 - S(t)$

La densité désigne la probabilité que l'évènement d'intérêt ait lieu après  $t$ , dans un petit intervalle de temps.

$$f(t) = \lim_{dt \rightarrow 0} \frac{\mathbb{P}(t < T \leq t + dt)}{dt}$$

## propriétés

- $F(t) = \int_0^t f(s)ds$
- $f(t) = F'(t)$

Le taux de risque instantané  $h(t)$ , aussi noté  $\lambda(t)$ , est la probabilité qu'un évènement survienne dans un petit intervalle de temps après  $t$ , sachant qu'il n'a pas eu lieu avant  $t$ .

$$h(t) = \lim_{dt \rightarrow 0} \frac{\mathbb{P}(t \leq T < t + dt | T \geq t)}{dt}$$

Remarques :

$h(t)$  est lié à une unité de temps. Si  $t$  est en heure, alors  $h(t)$  mesure le risque qu'un évènement survienne dans l'heure. Ce n'est pas une densité donc son intégrale ne vaut pas nécessairement 1.

Le taux de risque cumulé est défini par :

$$H(t) = \int_0^t h(s) ds$$



## Propriétés

- $\lambda(t) = H'(t) = \frac{-S'(t)}{S(t)} = \frac{f(t)}{S(t)}$
- $H(t) = -\ln(S(t))$
- $S(t) = e^{-H(t)} = e^{-\int_0^t \lambda(s)ds}$

Ces 5 fonctions caractérisent la loi de T.

Elles sont inconnues.

On va chercher à les estimer à partir des observations  $(X_i, \delta_i)$ .

Pour répondre à la question 1 (**comparaison de traitements**) : on va travailler avec  $S(t)$ , que l'on cherchera à estimer.

Pour répondre à la question 2 (**facteurs pronostiques**) : on va travailler avec le taux de risque instantané  $\lambda(t)$

# Exercice : loi exponentielle

Calculez les 5 fonctions associés à la loi exponentielle, sachant que le taux de risque instantané est constant :  $\lambda(t) = \lambda$

## Loi exponentielle

- $S(t) = e^{-\lambda t}$
- $F(t) = 1 - e^{-\lambda t}$
- $f(t) = \lambda e^{-\lambda t}$
- $\lambda(t) = \lambda$
- $H(t) = \lambda t$

Nous verrons d'autres lois (Weibull, log-Normale) dans le chapitre sur les modèles paramétriques.

La moyenne :

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

Le moment d'ordre 2 :

$$\mathbb{E}(T^2) = 2 \int_0^{\infty} tS(t)dt$$

La variance :

$$\text{Var}(T) = 2 \int_0^{\infty} tS(t)dt - \left( \int_0^{\infty} S(t)dt \right)^2$$

En pratique, on peut rarement estimer la durée moyenne de survie.

$$\mathbb{E}(t) = \int_0^{\infty} S(u) du$$

On considère alors la moyenne restreinte, i.e le temps moyen de survie sur un intervalle de temps  $[0, t]$

$$\mu(t) = \int_0^t S(u) du$$

Que l'on peut estimer par

$$\hat{\mu}(t) = \int_0^t \hat{S}(u) du$$

La fonction quantile de la durée de survie est définie par

$$Q(p) = \inf\{t : 1 - S(t) \geq p\} = \inf\{t : F(t) \geq p\}, \quad 0 < p < 1$$

Si  $F$  est strictement croissante et continue, alors :

$$Q(p) = F^{-1}(p) = S^{-1}(1 - p), \quad 0 < p < 1$$

A partir des données de l'INSEE

<http://www.ined.fr/fr/tout-savoir-population/chiffres/france/mortalite-cause-deces/table-mortalite/>

- Affichez la courbe de survie
- Quelle est la probabilité de survivre au moins jusqu'à  $t$  ?
- Calculez le temps de vie moyen et médian.
- Quelle est la probabilité de survivre au moins jusqu'à  $t_2$ , sachant que l'on a survécu jusqu'à  $t_1$  ?
- Quelle est la probabilité de mourir aujourd'hui ?  
Calculer  $\lambda(t)$



## Correction :

Quelle est la probabilité de survivre au moins jusqu'à  $t_2$ , sachant que l'on a survécu jusqu'à  $t_1$  ?

Reponse :  $\frac{S(t_2)}{S(t_1)}$

Démonstration :

$$\begin{aligned} S(t_2 | T > t_1) &= \mathbb{P}(T \geq t_2 | T > t_1) \\ &= \mathbb{P}(T \geq t_2 | T > t_1) \cdot \mathbb{P}(T \geq t_1) \cdot \frac{1}{\mathbb{P}(T \geq t_1)} \\ &= \mathbb{P}(T \geq t_2 \cap T > t_1) \cdot \frac{1}{S(t_1)} \\ &= \mathbb{P}(T \geq t_2) \cdot \frac{1}{S(t_1)} \\ &= \frac{S(t_2)}{S(t_1)} \end{aligned}$$

- 1 Généralités sur l'analyse de survie
  - Introduction
  - Définitions
  - Censure
  - Propriétés des fonctions de survie
  
- 2 Estimation et comparaison des courbes de survie
  - L'estimateur de Kaplan-Meier
  - L'estimateur de Nelson-Aalen
  - Tests de comparaison
  
- 3 Le modèle de Cox
  - Selection de variables

Problématique : A partir des observations  $(T_i, \delta_i)_{i \in 1..n}$ , on cherche à estimer la fonction de survie  $S(t)$ .

Si les données ne sont pas censurées, la proportion d'individus encore en vie à l'instant  $t$  peut être estimée par :

$$\hat{S}_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{T_i > t}$$

Pour ce faire, il suffit de trier les temps d'évènement par ordre croissant et de tracer la courbe en escalier (voir au tableau), avec des sauts de  $\frac{1}{n}$ .

Exemple d'observations : 1, 3, 4, 5.

Dans le cas des données censurées, la hauteur des sauts n'est plus uniforme. Elle est donnée par l'estimateur de Kaplan-Meier.

## Estimateur de Kaplan-Meier

Soit  $(t_i)_{i \in 1..n}$  la suite ordonnée des temps d'évènement, alors

$$\hat{S}(t) = \prod_{t_i < t} \left(1 - \frac{d_i}{r_i}\right)$$

avec :

$d_i$  le nombre d'évènements à  $t_i$

$r_i$  le nombre d'individus à risque juste avant  $t_i$

Remarque : C'est l'article de Statistique le plus cité.

$$\begin{aligned} S(t_i) &= \mathbb{P}(T > t_i) \\ &= \mathbb{P}(T > t_i | T > t_{i-1}) \cdot \mathbb{P}(T > t_{i-1}) \\ &= (1 - \mathbb{P}(T \leq t_i | T > t_{i-1})) \cdot \mathbb{P}(T > t_{i-1}) \end{aligned}$$

Or  $\mathbb{P}(T \leq t_i | T > t_{i-1})$ , c'est la probabilité de mourir dans  $]t_{i-1}, t_i]$ . On peut l'estimer par le nombre de décès dans cet intervalle divisé par le nombre d'individus à risque, soit  $\frac{d_i}{r_i}$ . Et il ne reste plus qu'à dérouler le produit sur tous les temps d'évènements.

## Définition

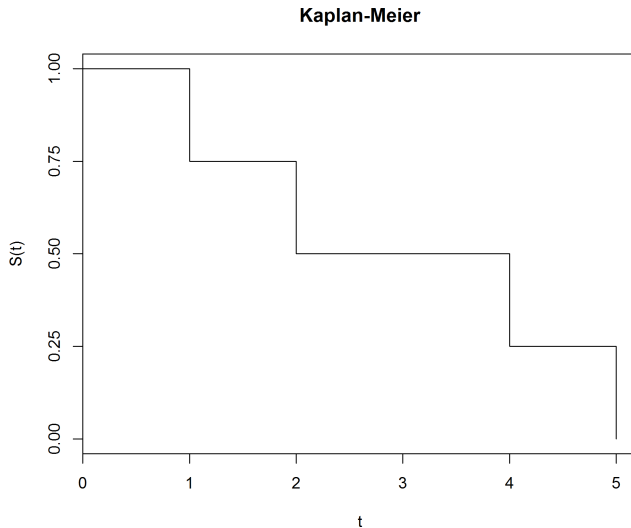
L'erreur standard de l'estimateur de Kaplan-Meier est approchée selon la formule de Greenwood, au temps de décès  $t_i$ , par :

$$\hat{\sigma} \left( \hat{S}(t_i) \right) = \hat{S}(t_i) \sqrt{\sum_{j=1}^n \frac{d_j}{r_j(r_j - d_j)}}$$

Calculez l'estimateur de Kaplan Meier pour les observations suivantes, où  $t^+$  indique un évènement censuré.

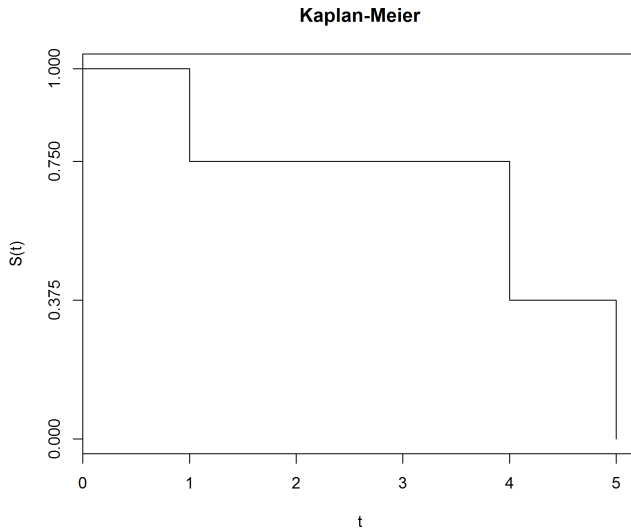
- ① 1, 2, 4, 5
- ② 1, 2<sup>+</sup>, 4, 5
- ③ 1, 2, 3<sup>+</sup>, 4, 5

# Correction 1 : 1, 2, 4, 5

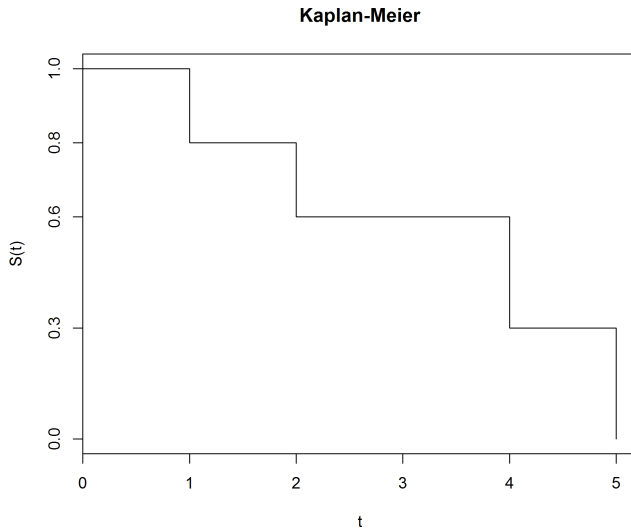




# Correction 2 : $1, 2^+, 4, 5$



# Correction 3 : 1, 2, 3<sup>+</sup>, 4, 5



Freireich (1963) à observé la durée de rémission (en semaines) de patients atteints de leucémie aiguë, traités soit par placebo soit par 6-mercaptopurine (6-MP)

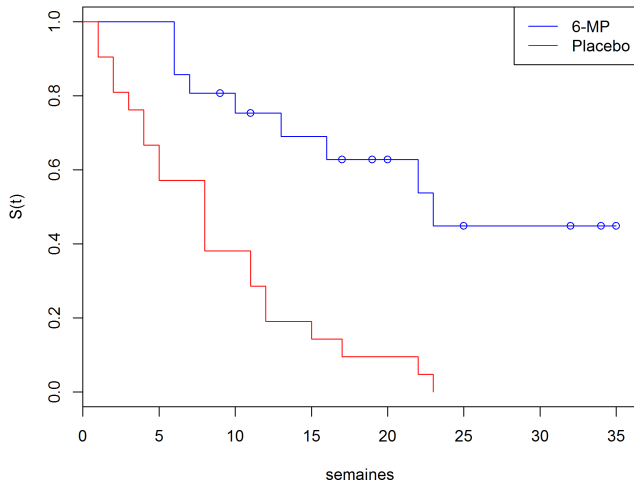
6-MP	6	6	6	6 <sup>+</sup>	7	9 <sup>+</sup>	10	10 <sup>+</sup>	11 <sup>+</sup>	13
	16	17 <sup>+</sup>	19 <sup>+</sup>	20 <sup>+</sup>	22	23	25 <sup>+</sup>	32 <sup>+</sup>		
	32 <sup>+</sup>	34 <sup>+</sup>	35 <sup>+</sup>							
Placebo	1	1	2	2	3	4	4	5	5	8
	8	11	11	12	12	15	17	22	23	

Calculez l'estimateur de Kaplan-Meier pour le groupe traité par 6-MP.

# Détail des calculs

durée de rémission observées	rechutes observées en la semaine $i$	sujets en rémission au début de la semaine $i$	prob. de ne pas rechuter à la semaine $i$ sachant qu'on est en rémission à la semaine $(i - 1)$	prob. d'être en rémission à la semaine $i$
0	0	21	$21/21 = 1$	1
6	3	21	$18/21 = 0,857$	$1 * 18 / 21 = 0,857$
7	1	17	$16/17 = 0,941$	$0,857 * 16/17 = 0,807$
10	1	15	$14/15 = 0,933$	$0,807 * 14/15 = 0,753$
13	1	12	$11/12 = 0,917$	$0,753 * 11/12 = 0,690$
16	1	11	$10/11 = 0,909$	$0,690 * 10/11 = 0,627$
22	1	7	$6/7 = 0,857$	$0,627 * 6/7 = 0,538$
23	1	6	$5/6 = 0,833$	$0,538 * 5/6 = 0,448$

Kaplan Meier par traitement



On cherche à estimer le risque cumulé  $H$ , défini par :

$$H(t) = \int_0^t \lambda(s) ds$$

Première approche : On utilise la relation  $H(t) = -\log(S(t))$ .

Définition : Estimateur de Breslow

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

Seconde approche : On estime le taux de hasard (ou risque instantané de décès) en  $t$  par

$$\hat{\lambda}(t) = \frac{d_i}{r_i}$$

Définition : Estimateur de Nelson-Aalen

$$\hat{H}_{Nelson\_Aalen}(t_i) = \sum_{t_j \leq t_i} \frac{d_j}{r_j}$$

## Estimateur de la moyenne

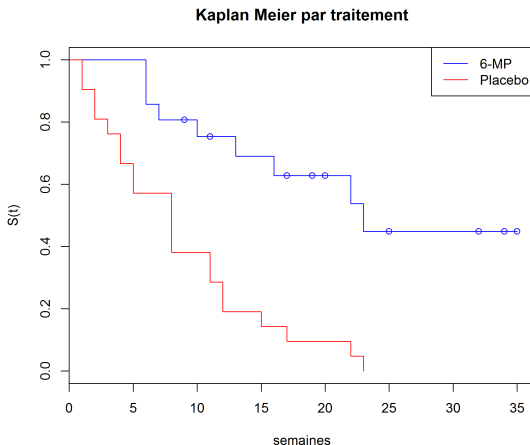
$$\hat{\mathbb{E}}(T) = \int_0^\infty \hat{S}_{KM}(t) dt = \sum_{i=1}^k \hat{S}_{KM}(t_{i-1})(t_i - t_{i-1})$$

## Estimateur d'un quantile

$$\hat{q}_\alpha = \min\{t : 1 - \hat{S}_{KM}(t) \geq \alpha\}$$



On souhaite comparer la survie de deux groupes ayant reçu des traitements différents



Notons  $S_A$  et  $S_B$  les fonctions de survie des groupes A et B. On souhaite tester

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

S'il n'y avait pas de censure, on pourrait utiliser

- Test de Kolmogorov Smirnov (comparaison de lois)
- Test de la somme des rangs
- Test de Mann-Whitney

Notons  $S_A$  et  $S_B$  les fonctions de survie des groupes A et B. On souhaite tester

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

En présence de censure, on peut utiliser

- Test de Wilcoxon généralisé
- Test du log-rank

Soient  $t_1, \dots, t_k$  les temps de décès ordonnés des deux groupes A et B réunis. On calcule

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

avec :

- $w_i$  une pondération dépendant du choix du test
- $d_{B,i}$  le nombre d'évènements observés au temps  $t_i$  dans le groupe  $B$
- $e_{B,i}$  le nombre d'évènements attendus au temps  $t_i$  dans le groupe  $B$  sous  $H_0$ , voir slide suivante.

- le nombre d'évènements attendus au temps  $t_i$  dans le groupe  $B$  sous  $H_0$  est :

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

- $d_{A,i}, d_{B,i}$  le nombre de décès observés en  $t_i$  dans les groupes  $A$  et  $B$  avec  $d_i = d_{A,i} + d_{B,i}$
- $R_{A,i}, R_{B,i}$  le nombre d'individus à risque juste avant  $t_i$  dans les groupes  $A$  et  $B$  avec  $R_i = R_{A,i} + R_{B,i}$

Sous  $H_0$ ,  $\mathbb{E}(U) = 0$  et on a :

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

avec

$$V(u) = \sum_{i=1}^k w_i^2 d_i \frac{R_i - d_i}{R_i - 1} \frac{R_{A,i} R_{B,i}}{R_i^2}$$

La Statistique de Test est donc

$$T_n = \frac{U^2}{V(u)}$$

avec sous  $H_0$

$$T_n \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \chi^2(1)$$

Notons  $t_{n,obs}$  la valeur observée de  $T_n$ .

Pour un test de niveau  $\alpha$ , on rejette  $H_0$  si

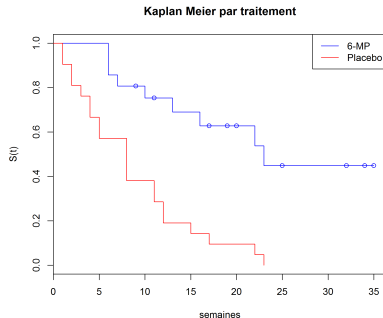
la p-value  $= \mathbb{P}(T_n > t_{n,obs}) \leq \alpha$

Choix des poids  $w_i$  :

- Test du log-rank :  $w_i = 1$
- Test de Wilcoxon généralisé (ou Breslow)  $w_i = R_i$  (nombre de sujets exposés à  $t_i$ . Les décès précoces ont un poids plus important. Utile pour montrer une différence sur les courtes durées de survie.



# Application aux données sur la Leucémie



test 1	Khi2	DF	p-value
Log-Rank	16.8	1	0.0000417
Wilcoxon	14.5	1	0.000143

On rejette l'hypothèse  $H_0$ . Les deux courbes de survie sont significativement différentes au seuil  $\alpha = 5\%$ . Le traitement 6-MP a un effet positif sur la fonction de survie

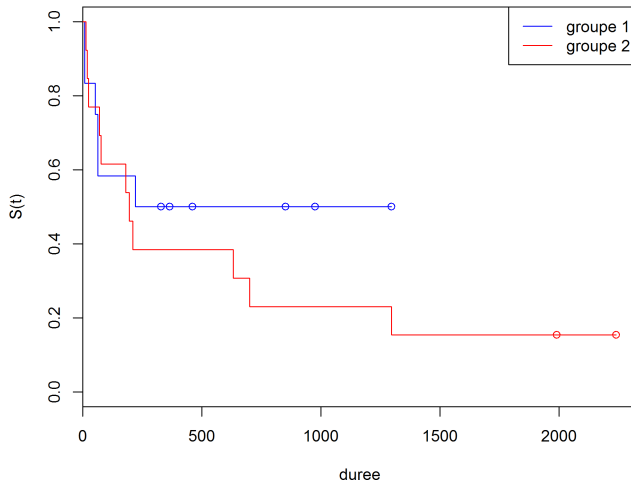
## Autre exemple : données de Peto

Durée de survie de deux groupes de patients à qui l'on a administré deux types de traitements. On dispose en plus de la fonction rénale, connue pour influencer la survie.

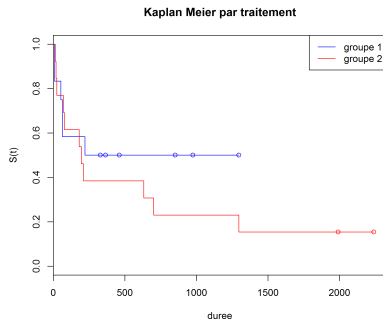
durée de survie	groupe de traitement	fonction rénale	durée de survie	groupe de traitement	fonction rénale
8	1	A	220	1	N
8	1	N	365*	1	N
13	2	A	632	2	N
18	2	A	700	2	N
23	2	A	852*	1	N
52	1	A	1296	2	N
63	1	A	1296*	1	N
63	1	A	1328*	1	N
70	2	N	1460*	1	N
76	2	N	1976*	1	N
180	2	N	1990*	2	N
195	2	N	2240*	2	N
210	2	N			

Fonction rénale N=normale, A=anormale.

Kaplan Meier par traitement



# Application aux données de Peto



test 1	Khi2	DF	p-value
Log-Rank	0.8	1	0.383
Wilcoxon	0.2	1	0.692

On ne rejette pas l'hypothèse  $H_0$ . Les deux courbes de survie ne sont pas significativement différent au seuil  $\alpha = 5\%$ . On ne peut pas conclure à un effet du traitement.

L'effet du traitement peut être caché par la fonction rénale.  
Problème : comment relier la durée de survie d'un patient à plusieurs facteurs pronostiques ? ex : traitement et fonction rénale.

## 1 Généralités sur l'analyse de survie

- Introduction
- Définitions
- Censure
- Propriétés des fonctions de survie

## 2 Estimation et comparaison des courbes de survie

- L'estimateur de Kaplan-Meier
- L'estimateur de Nelson-Aalen
- Tests de comparaison

## 3 Le modèle de Cox

- Selection de variables

## Définition : Modèle de Cox

$$\alpha(t|X) = \alpha_0(t)e^{(\beta_1 X_1 + \dots + \beta_p X_p)} = e^{\beta^T X}$$

- $\alpha_0(t)$  le risque de base (baseline) inconnu, dépend du temps mais pas des variables  $X$ . C'est donc le même pour tous les individus.
- $\beta$  les paramètres de régression.

Le risque relatif  $e^{\beta^T X}$  est indépendant du temps.

## Hypothèses du modèle de Cox

- ① Les risques sont proportionnels : le rapport des risque instantané d'évènement (hazard rate) de deux individus est proportionnel.
- ② Log-linéarité : le logarithme de  $\alpha(t|X)$  est une fonction linéaire en  $X$ .

$$\log(\alpha(t|X)) = \log(\alpha_0(t)) + \beta^T X$$



Interprétation et avantages du modèle de Cox :

- $\alpha_0(t|X=0) = \alpha_0(t)$ . Le risque de base correspond au risque quand toutes les covariables sont nulles.
- Les paramètres du modèle  $\alpha_0(t)$  et  $\beta$  s'estiment de manière indépendante.
- La forme de  $\alpha_0(t)$  (et donc de la courbe de survie) n'as pas besoin d'être spécifiée(vs exponentielle ou Weibull). On parle de modèle semi-paramétrique.

Application aux données sur la Leucémie.

$$X = \begin{cases} 0 & \text{si groupe} = \text{"placebo"} \\ 1 & \text{si groupe} = \text{"6MP"} \end{cases} \quad \alpha(t|x) = \alpha_0(t)e^{\beta X}$$

Hazard Ratio (rapport des risques) :

$$HR = \frac{\alpha(t|X=1)}{\alpha(t|X=0)} = \frac{\alpha_0(t)e^{\beta}}{\alpha_0(t)} = e^{\beta}$$

Interprétation :

$HR = 1 \Leftrightarrow \beta = 0 \Leftrightarrow$  risque identique pour les deux groupes

$HR > 1 \Leftrightarrow \beta > 0 \Leftrightarrow$  risque plus élevé pour le groupe 6MP

$HR < 1 \Leftrightarrow \beta < 0 \Leftrightarrow$  risque plus faible pour le groupe 6MP

## Vraisemblance partielle de Cox

$$\hat{\beta} = \operatorname{argmax} \frac{1}{n} \prod_{i \in D} \frac{\exp(x_i^T \theta)}{\sum_{j \in R_i} \exp(x_j^T \theta)}$$

$D$  ensemble de tous les évènements.

$R_i$  individus à risque à  $t_i$  (temps où l'individu  $i$  réalise son évènement).

Freireich (1963) à observé la durée de rémission (en semaines) de patients atteints de leucémie aiguë, traités soit par placebo soit par 6-mercaptopurine (6-MP)

6-MP	6	6	6	6 <sup>+</sup>	7	9 <sup>+</sup>	10	10 <sup>+</sup>	11 <sup>+</sup>	13		
	16	17 <sup>+</sup>	19 <sup>+</sup>	20 <sup>+</sup>	22	23	25 <sup>+</sup>	32 <sup>+</sup>				
	32 <sup>+</sup>	34 <sup>+</sup>	35 <sup>+</sup>									
Placebo	1	1	2	2	3	4	4	5	5	8	8	8
	8	11	11	12	12	15	17	22	23			

Quel est le gain apporté par le médicament ?

```
library(survival)
library(bpcp)
data("leuk2")
leuk.surv=Surv(leuk2$time,leuk2$status)
fit<-coxph(leuk.surv~ treatment,data=leuk2, method="breslow")
summary(fit)
```

# Modèle de Cox avec une variable binaire

Call:

```
coxph(formula = leuk.surv ~ treatment, data = leuk2, method = "breslow")
```

n= 42, number of events= 30

	coef	exp(coef)	se(coef)	z	Pr(> z )	
treatmentplacebo	1.5092	4.5231	0.4096	3.685	0.000229	***

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

	exp(coef)	exp(-coef)	lower .95	upper .95
treatmentplacebo	4.523	0.2211	2.027	10.09

Concordance= 0.69 (se = 0.053 )

Rsquare= 0.304 (max possible= 0.989 )

Likelihood ratio test= 15.21 on 1 df, p=9.615e-05

Wald test = 13.58 on 1 df, p=0.0002288

Score (logrank) test = 15.93 on 1 df, p=6.571e-05

Le taux de risque instantané des patients traités au 6-MP est 4.5 fois plus petit que celui des patients sous placebo.

## Kidney catheter data

Data on the recurrence times to infection, at the point of insertion of the catheter, for kidney patients using portable dialysis equipment. Catheters may be removed for reasons other than infection, in which case the observation is censored.

patient: id

time: time

status: event status

age: in years

sex: 1=male, 2=female

disease: disease type (0=GN, 1=AN, 2=PKD, 3=Other)



$X$  est une variable catégorielle à 4 modalités (GN, AN, PKD and Other).

- Si l'on considère que les données sont ordinales :

$$\lambda(t) = \lambda_0(t)e^{X^t\beta}$$

avec  $X \in \{0=\text{GN}, 1=\text{AN}, 2=\text{PKD and } 3=\text{Other}\}$

- Si les données ne sont pas ordinales, on choisit une référence (ex : GN)

$$\lambda(t) = \lambda_0(t)e^{\beta_1\text{AN}+\beta_2\text{PKD}+\beta_3\text{Other}}$$

# Modèle de Cox avec une variable catégorielle

```
library(survival)
data(kidney)

# catégorielle ordinale
kidney$X[kidney$disease=="GN"] <- 0
kidney$X[kidney$disease=="AN"] <- 1
kidney$X[kidney$disease=="PKD"] <- 2
kidney$X[kidney$disease=="Other"] <- 3

# catégorielle non ordinale
kidney$GN <- (kidney$disease=="GN")
kidney$AN <- (kidney$disease=="AN")
kidney$PKD <- (kidney$disease=="PKD")
kidney$Other <- (kidney$disease=="Other")
```

# Modèle de Cox avec une variable ordinale

```
> coxph(Surv(time,status)~X,data=kidney,method="breslow")
Call:
coxph(formula = Surv(time, status) ~ X, data = kidney, method = "breslow")

      coef exp(coef) se(coef)      z    p
X -0.144      0.866    0.113 -1.27 0.2

Likelihood ratio test=1.62 on 1 df, p=0.203
n= 76, number of events= 58
```

Écrire l'équation associée et donner une prédiction.

# Modèle de Cox avec une variable catégorielle

```
> coxph(Surv(time,status)~AN+PKD+Other,data=kidney,method="breslow")  
Call:  
coxph(formula = Surv(time, status) ~ AN + PKD + Other, data = kidney,  
      method = "breslow")
```

	coef	exp(coef)	se(coef)	z	p
ANTRUE	0.031	1.032	0.364	0.09	0.93
PKDTRUE	-0.620	0.538	0.531	-1.17	0.24
OtherTRUE	-0.351	0.704	0.354	-0.99	0.32

Likelihood ratio test=2.71 on 3 df, p=0.439  
n= 76, number of events= 58

Écrire l'équation associée et donner une prédiction.

## Données sur la récidive (Rossi 1980)

- week : week of first arrest after release or censoring; all censored observations are censored at 52 weeks.
- arrest : 1 if arrested, 0 if not arrested.
- fin : financial aid: no yes.
- age : in years at time of release.
- race : black or other.
- wexp : full-time work experience before incarceration: no or yes.
- mar : marital status at time of release: married or not married.
- paro : released on parole? no or yes.
- prio : number of convictions prior to current incarceration.
- educ : level of education: 2 = 6th grade or less; 3 = 7th to 9th grade; 4 = 10th to 11th grade; 5 = 12th grade; 6 = some college.

```
library(RcmdrPlugin.survival)
data(Rossi)
model <- coxph(Surv(week, arrest) ~
               fin + age + race + wexp + mar + paro + prio,
               data=Rossi)
```

# Modèle de Cox avec une variable catégorielle

Call:

```
coxph(formula = Surv(week, arrest) ~ fin + age + race + wexp +  
      mar + paro + prio, data = Rossi)
```

	coef	exp(coef)	se(coef)	z	p
finyes	-0.3794	0.6843	0.1914	-1.98	0.0474
age	-0.0574	0.9442	0.0220	-2.61	0.0090
raceother	-0.3139	0.7306	0.3080	-1.02	0.3081
wexpyes	-0.1498	0.8609	0.2122	-0.71	0.4803
marnot married	0.4337	1.5430	0.3819	1.14	0.2561
paroyes	-0.0849	0.9186	0.1958	-0.43	0.6646
prio	0.0915	1.0958	0.0286	3.19	0.0014

Likelihood ratio test=33.3 on 7 df, p=2.36e-05  
n= 432, number of events= 114

Le risque de base  $\alpha_0(t)$  peut être estimé par une fonction constante par morceaux (estimateur de Breslow)

$$\hat{\alpha}_0(t) = \prod_{i=1}^k \frac{d_i}{(t_{i+1} - t_i) \sum_{j \in R_{t_i}} \exp(X_j^T \hat{\beta})}$$

avec

- $t_i$  les temps ordonnées d'évènement
- $d_i$  le nombre d'évènements en  $t_i$
- $R_{t_i}$  le nombre d'individus à risque juste avant  $t_i$



Soit  $\mathcal{M}_k$  un modèle à  $k$  variables.

## Akaike information criterion

$$AIC(\mathcal{M}_k) = -2 \log(\hat{L}) + 2k$$

Si les modèles sont imbriqués :

- + : minimise l'erreur de prédiction
- : Tendance à choisir un modèle plus grand (ne pénalise pas assez)

## Bayesian information criterion

$$BIC(\mathcal{M}_k) = -2 \log(\hat{L}) + k \log(n)$$

Si les modèles sont imbriqués :

- + : Critère consistant : trouve le bon modèle
- : ne minimise pas l'erreur quadratique (mauvais en prédiction)

Problème si on veut tester toutes les combinaisons de variables parmi  $p$ , il y a  $2^p$  combinaisons.