Introduction aux analyses de survie

Yohann.Foucher@univ-nantes.fr

Master 2 - Modélisation en Pharmacologie Clinique et Epidémiologie

Plan

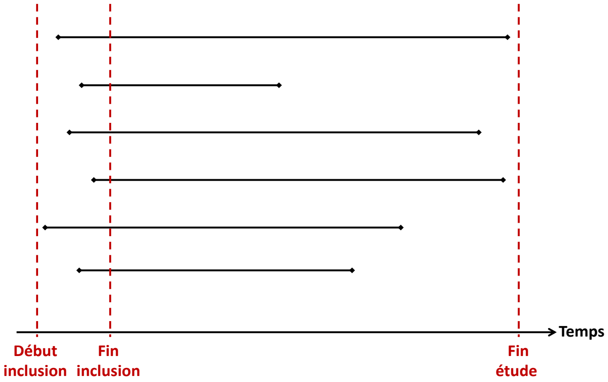
Plan

Contexte

- Etude des facteurs influençant l'aparition d'un événement.
- Variable à expliquer : Temps d'apparition d'un événement
- Soit T cette variable aléatoire (v.a.)
- Exemples :
  - Est ce que le tabagisme réduit le temps d'apparition d'un décès ?

  $\leftrightarrow$ Est ce que le tabagisme réduit la survie ?
- Le décès a souvent été l'événement étudié en médecine, d'où le terme survie.
- D'autres temps peuvent être étudié (*time-to-event analysis*).

Illustration



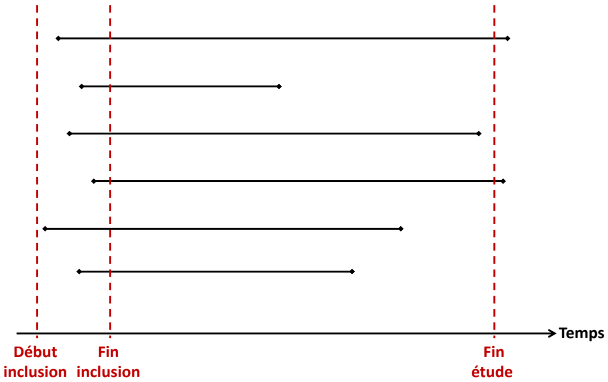
Une régression linéaire ?

$$E[T|x_1, x_2, ...] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...$$

- T : v.a représentant le temps de l'événement (ex : mois).
- x_1, x_2, ... les variables explicatives d'intérêt.
- β_1, β_2, ... les coefficients de régression.

$\rightarrow$ β_1 représente le nombre moyen de mois supplémentaires avant l'évènement dans le groupe $x_1 = 1$ par rapport au groupe $x_1 = 0$.

Problème de la censure à droite



On ne connait pas les temps d'apparition des événements pour les patients 1 et 4 $\rightarrow$ Régression linéaire impossible (valeurs manquantes informatives).

Définition de la censure à droite

- Rappel : T est la v.a du temps d'apparition de l'événement.
    - Soit t_i le temps observé d'apparition de l'événement pour le sujet i ($i = 1, ..., N$).
- Le temps de censure à droite, C, est une autre v.a. représentant le temps entre l'inclusion (origine, J0, *baseline*) et la fin du suivi qui aurait lieu indépendamment de l'événement.
    - Ex : Fin de l'étude, déménagement, perdu de vue, etc.
    - Soit c_i le temps de censure observé pour le sujet i ($i = 1, ..., N$).
- Si y_i est le temps de suivi du sujet i, deux situations peuvent être distinguées :
    - Si $t_i \leq c_i$: on observe exactement le temps d'événement.
      $\rightarrow$ $y_i = t_i$ et $\delta_i = 1$.
    - Si $t_i > c_i$: le temps d'événement est inconnu, mais on sait qu'il est supérieur à c_i.
      $\rightarrow$ $y_i = c_i$ et $\delta_i = 0$.

Présentation des données

Sujet (i)	Sexe (x_{1i})	Age (x_{2i})	Temps (y_i)	Evt (δ_i)
1	1	15	32	0
2	0	25	13	1
3	1	74	25	1
4	0	63	26	0
.	.	.	.	.
.	.	.	.	.
N	.	.	.	.

Comment modéliser T si ces observations sont incomplètes ?

- Régression linéaire impossible car données manquantes informatives (on observe Y mais pas T).
- Quid d'une régression logistique pour expliquer la variable binaire δ ?

Régression logistique

$$logit\Big(Pr(\delta = 1|x_1, x_2, ...)\Big) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...$$

- δ : v.a. à expliquer binaire
- x_1, x_2, ... les variables explicatives d'intérêt.
- β_1, β_2, ... les coefficients de régression.

$\rightarrow$ $\exp(\beta_1)$ représente l'odds ration du groupe $x_1 = 1$ par rapport au groupe $x_1 = 0$.

Problème

- Un patient avec un long suivi a plus de chance d'être $\delta = 1$.
- Biais majeur dans l'estimation des effets des variables explicatives qui seront aussi explicatives du temps de suivi...

$\Rightarrow$ Nécessité de modèles spécifiques $\Leftarrow$

Plan

Notations

- N : la taille de l'échantillon ($i = 1, ..., N$).
- T : la v.a. représentant le temps d'apparition de l'événement.
- C : la v.a. représentant le temps de censure à droite.
- On ne peut observer qu'une seule résultante des deux v.a. :

$$Y = min(T, C)$$

- y_i : l'observation pour le sujet i de Y.
- δ : l'indicatrice de l'événement avec :
  - $\delta_i = 1$ si y_i correspond à un temps d'événement.
  - $\delta_i = 0$ si y_i correspond à un temps de censure.

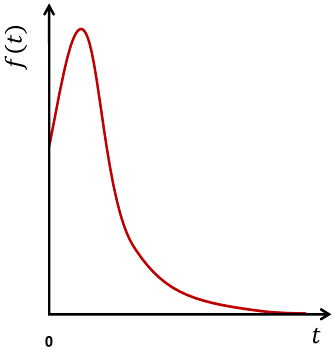
La fonction de densité $f(t)$

La fonction de densité de probabilité $f(t)$ représente la limite de probabilité que l'événement se produise au temps t.

$$f(t) = \lim_{\Delta t \to 0} Pr(t \leq T < t + \Delta t)/\Delta t$$

Même si il s'agit d'une limite, il est plus simple d'intuiter cette fonction comme la probabilité que l'événement se produise à peu près eu temps t.

La fonction de densité $f(t)$



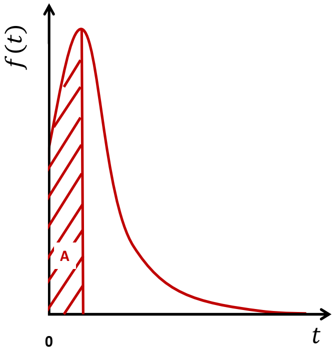
La fonction de répartition $F(t)$

La fonction de répartition $F(t)$ est la probabilité que l'événement se produise avant t.

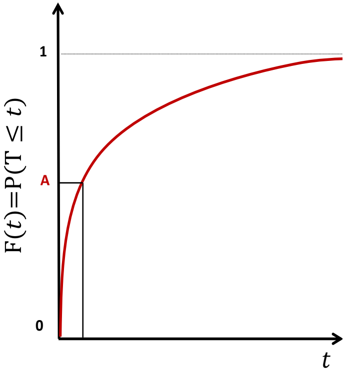
$$F(t) = Pr(T \leq t) = 1 - S(t) = \int_0^t f(u)du$$

$F(t) \in [0,1]$, c'est une probabilité. Elle est croissante de 0 en 1. La probabilité d'avoir eu l'événement avant l'origine est égale à nulle : $F(0) = 0$. La probabilité d'avoir eu l'événement avant en temps de suivi très long est certaine : $F(\infty) = 1$.

La fonction de répartition $F(t)$



La fonction de répartition $F(t)$



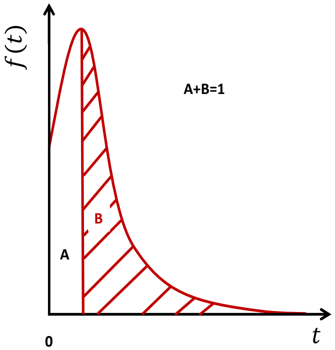
La fonction de survie $S(t)$

La fonction de survie $S(t)$ est la probabilité que l'événement se produise après t.

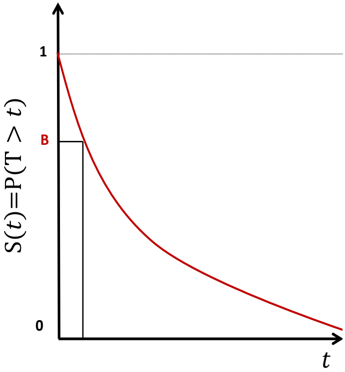
$$S(t) = Pr(T > t) = \int_t^{\infty} f(u)du$$

$S(t) \in [0, 1]$, c'est une probabilité. Elle est décroissante de 1 en 0. La probabilité d'être en vie à l'origine est égale à 1 : $S(0) = 1$. La probabilité d'être en vie après en temps infini est nulle : $S(\infty) = 0$.

La fonction de survie $S(t)$



La fonction de survie $S(t)$



La fonction de risque $h(t)$

La fonction de risque instantanée $h(t)$ représente la limite de probabilité que l'événement se produise au temps t sachant qu'il ne s'est pas produit avant.

$$h(t) = \lim_{\Delta t \to 0} Pr(t \leq T < t + \Delta t | T \geq t)/\Delta t$$

Elle peut être vue comme une vitesse. $h(t) > 0$, l'événement est unique et sans retour : on ne ressuscite pas. Tout processus vivant est vieillissant.

La fonction de risque cumulé $H(t)$

La fonction de risque cumulé $H(t)$ est la somme des risques instantanés jusqu'en t.

$$H(t) = \int_0^t h(u)du$$

Toutes ces fonctions sont équivalentes. Il suffit d'en définir une pour retrouver les autres. Liens importants :

$$S(t) = exp\Big(- H(t)\Big) = exp\Big(- \int_0^t h(u)du\Big) = f(t)/h(t)$$

$\Rightarrow$ Comment estimer $S(t)$? $\Leftarrow$

Plan

Plan

L'estimateur de Kaplan-Meier (1)

Etre en vie à deux ans après l'inclusion, c'était être en vie à un an et ne pas être décédé dans la seconde année. On écrire cette propriété pour tout temps t_1 et t_2 tels que $t_2 > t_1$:

$$S(t_2) = Pr(T > t_2)$$

$$\Downarrow$$

$$S(t_2) = Pr(T > t_2, T > t_1)$$

$$\Downarrow$$

$$S(t_2) = Pr(T > t_2 | T > t_1) Pr(T > t_1)$$

L'estimateur de Kaplan-Meier (2)

On observe tous les couples (y_i, δ_i) avec $i = 1, ..., N$. On peut ainsi généraliser la propriété précédente en ordonnant tous les sujets par y_i :

$$S(y_N) = Pr(T > y_N | T > y_{N-1}) \times ... \times Pr(T > y_1)$$

$$\Downarrow$$

$$S(y_N) = Pr(T > y_N | T > y_{N-1}) \times ... \times Pr(T > y_1 | T > t_0)$$

$$\Downarrow$$

$$S(y_N) = \prod_{i=1}^{N} Pr(T > y_i | T > y_{i-1})$$

avec $t_0 = 0$

L'estimateur de Kaplan-Meier (3)

Si on souhaite la survie à n'importe quel temps t tel que $t < y_N$, il suffit d'arrêter le produit précédent au dernier individu suivi juste avant t :

$$S(t) = \prod_{i:y_i<t} Pr(T > y_i | T > y_{i-1})$$

La question est maintenant de savoir comment estimer ces probabilités conditionnelles. La probabilité de connaitre l'événement au temps y_i sachant qu'on ne l'a pas subi avant y_{i-1} est estimée par le nombre d'événements entre y_{i-1} et y_i parmi tous les sujets à risque au temps y_i :

$$Pr(T \leq y_i | T > y_{i-1}) = \frac{\sum_j I(\delta_j = 1, y_{i-1} < y_j \leq y_i)}{\sum_j I(y_j > y_{i-1})} = \frac{d_i}{n_i}$$

L'estimateur de Kaplan-Meier (4)

On en déduit les probabilités de survie conditionnelles :

$$Pr(T > y_i | T > y_{i-1}) = 1 - Pr(T \leq y_i | T > y_{i-1}) = 1 - \frac{d_i}{n_i}$$

Remarquons que si le temps y_i correspond à une censure ($\delta_i = 0$), il n'y a aucun événement dans l'intervalle de y_{i-1} à y_i, alors cette probabilité vaut 1. La probabilité de survie $S(t)$ ne change que si y_i correspond à l'observation d'un événement dans l'échantillon ($\delta_i = 0$). Cette propriété simplifie les calculs.

On démontre ainsi que l'estimateur de Kaplan-Meier * de la survie est obtenu comme suit :

$$\hat{S}(t) = \prod_{i:y_i<t} \left(1 - \frac{d_i}{n_i}\right)$$

*. Kaplan E, Meier P. Non-parametric estimation from incomplete observations. Journal of the American Statistical Association 1958 ; 53 :457-481.

Un exemple simple

Soient deux groupes de patients atteints d'un cancer du poumon. On propose deux types de chimiothérapies (A/B) aléatoirement et on étudie le temps entre le traitement et le décès du patient. Voici les observations :

Chimio	Mois (t_i)	Evt (δ_i)
A	24	1
A	26	0
A	30	1
A	31	0
B	22	0
B	28	1
B	30	0
B	32	0

Un exemple simple

- Pour le groupe A.

t_i	d_i	c_i	n_i	proba	$\hat{S}(t)$
0	0	0	4	1	1
24	1	0	4	1-1/4 $=$ 0.75	1 $\times$ 0.75 $=$ 0.75
30	1	1	4-1-1 $=$ 2	1-1/2 $=$ 0.50	0.75 $\times$ 0.50 $=$ 0.375
31	0	0	1-0 $=$ 1	1-0/1=1	0.375 $\times$ 1 $=$ 0.375

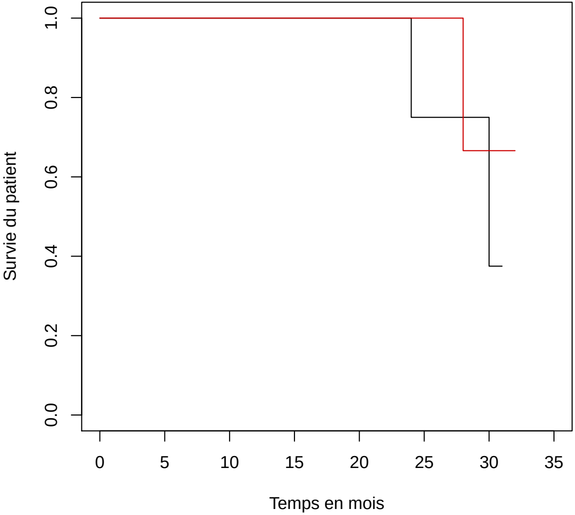
- Pour le groupe B.

t_i	d_i	c_i	n_i	proba	$\hat{S}(t)$

Un exemple simple


```
> ta <- c(0, 24, 30, 31)
> sa <- c(1, 0.75, 0.375, 0.375)
> plot(ta, sa, type="s", ylab="Survie du patient",
+  xlab="Temps en mois", ylim=c(0,1), xlim=c(0, 35))
> tb <- c(0, 28, 32)
> sb <- c(1, 0.666, 0.666)
> lines(tb, sb, type="s", col="red3")
```

Un exemple simple



Un exemple plus complet

- Etude du cancer du poumon.
- Essai clinique sur 137 hommes (traitement par chimiothérapie vs. standard).
- Temps étudié = Délai entre le début du traitement et le décès du patient.
- 6 variables explicatives mesurées à l'inclusion :
  - Type de la cellule (sqamous cell, large cell, small cell and adenocarcinoma)
  - Score de Karnofsky
  - Délai entre diagnostic et traitement (en mois)
  - Age (en années)
  - Thérapie antérieure (yes/no)
  - Traitement (chemotherapy vs. sandard)

Objectif : Evaluer l'effet du médicament

Un exemple plus complet

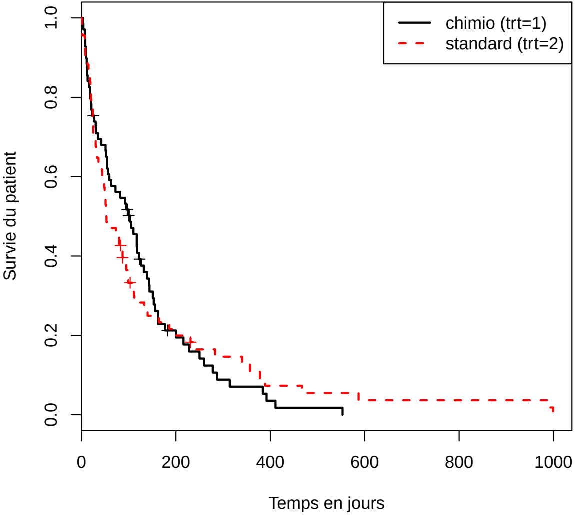

```
> library(survival)
> veteran[1:10,]

   trt celltype time status karno diagtime age prior
1    1 squamous   72      1    60        7  69     0
2    1 squamous  411      1    70        5  64    10
3    1 squamous  228      1    60        3  38     0
4    1 squamous  126      1    60        9  63    10
5    1 squamous  118      1    70       11  65    10
6    1 squamous   10      1    20        5  49     0
7    1 squamous   82      1    40       10  69    10
8    1 squamous  110      1    80       29  68     0
9    1 squamous  314      1    50       18  43     0
10   1 squamous  100      0    70        6  70     0
```

Un exemple plus complet


```
> f <- survfit(Surv(time, status) ~ trt,
+  data = veteran)
> plot(f, ylab="Survie du patient",
+  col=1:2, lwd=2, lty=1:2, xlab="Temps en jours")
> legend("topright", c("chimio (trt=1)", "standard (trt=2)"),
+  col=1:2, lty=1:2, lwd=2)
```

Un exemple plus complet



Bonnes pratiques de représentation †

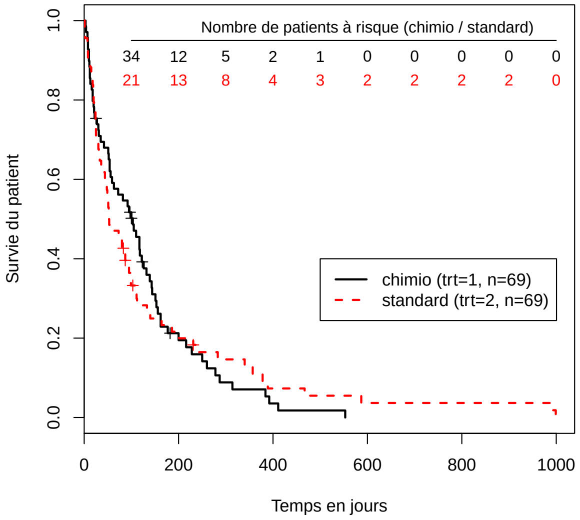
- Légende.
- Libellé des axes.
- Unité du temps.
- Fonction en escalier.
- Axe des ordonnées entre 0 et 1, sinon symbole "//".
- Effectifs à risque au cours du temps.

†. Pocock SJ, Clayton TC, Altman DG. Survival plots of time-to-event outcomes in clinical trials : good practice and pitfalls. Lancet. 2002 ;359 :1686-9.

Bonnes pratiques de représentation


```
> plot(f, ylab="Survie du patient",
+  col=1:4, lwd=2, lty=1:2, xlab="Temps en jours")
> nb.risk1<-function(x) {
+  dim(veteran[veteran$trt==1,])[1]-sum(veteran$time[veteran$trt==1]<x) }
> nb.risk2<-function(x) {
+  dim(veteran[veteran$trt==2,])[1]-sum(veteran$time[veteran$trt==2]<x) }
> .time <- seq(100, 1000, by=100)
> text(x=500, y=0.98, "Nombre de patients à risque (chimio / standard)", cex=0.9)
>  segments(x0=0, y0=0.95, x1=1000, y1=0.95)
> text(x=.time, y=rep(0.91, length(.time)),
+  as.character(sapply(.time, FUN="nb.risk1")), cex=0.9)
> text(x=.time, y=rep(0.85, length(.time)),
+  as.character(sapply(.time, FUN="nb.risk2")), col=2, cex=0.9)
> legend(600, 0.4, c("chimio (trt=1)", "standard (trt=2)"),
+  col=1:2, lty=1:2, lwd=2)
```

Bonnes pratiques de représentation



L'hypothèse principale : censure non-informative

Cet estimateur est non-biaisé si les temps de censures à droite sont indépendants des temps d'événements.

Ex1 : Etude de la mortalité liée au VIH en Afrique. Si tous les patients quittent le centre de soins pour mourir dans leur village (sous-estimation de la mortalité).

Ex2 : Etude de la mortalité des patients sur liste de greffe : les patients greffés vont en moyenne mieux que les autres (sur estimation de la mortalité).

Plan

Introduction
Définitions
Kaplan-Meier
Estimation
Variance
LogRank
Distributions paramétriques
Les principales lois
L'estimation des paramètres
Modèle de Cox
Définitions
Inférence
Hypothèses
Perspectives

Formule de Greenwood et intervalle de confiance

Cet estimateur est assez facilement démontrable en remarquant que l'estimateur de Kaplan-Meier est un estimateur obtenu par maximisation et en utilisant la delta-méthode :

$$Var[\hat{S}(t)] = \hat{S}(t)^2 \times \sum_{i:y_i \leq t} \frac{d_i}{(n_i - d_i)n_i}$$

On peut en déduire l'intervalle de confiance à (1-α)% :

$$IC95\%[\hat{S}(t)] = \hat{S}(t) \pm z_{\alpha/2}\sqrt{Var[\hat{S}(t)]}$$

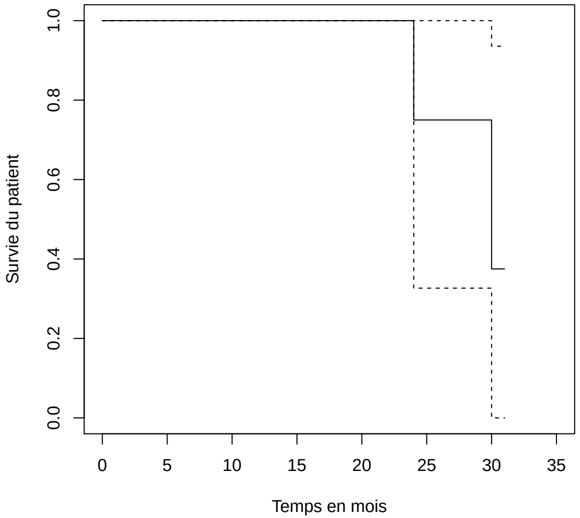
Reprenons l'exemple du groupe A

t_i	d_i	c_i	n_i	$\hat{S}(t)$	$g_i = d_i/(n_i - d_i)n_i$	$\sum_i g_i$	$\sqrt{Var[\hat{S}(t)]}$
0	0	0	4	1.000	0.000	0.000	0.000
24	1	0	4	0.750	1/(4 (4-1)) = 0.083	0.083	0.216
30	1	1	2	0.375	1/(2 (2-1)) = 0.500	0.583	0.286
31	0	0	1	0.375	0.000	0.583	0.286

Reprenons l'exemple du groupe A


```
> ta <- c(0, 24, 30, 31)
> sa <- c(1, 0.75, 0.375, 0.375)
> sda <- c(0,0.216, 0.286, 0.286)
> plot(ta, sa, type="s", ylab="Survie du patient",
+  xlab="Temps en mois", ylim=c(0,1), xlim=c(0, 35))
> lines(ta, pmax(sa-qnorm(0.975)*sda, 0), lty=2, type="s")
> lines(ta, pmin(sa+qnorm(0.975)*sda, 1), lty=2, type="s")
```

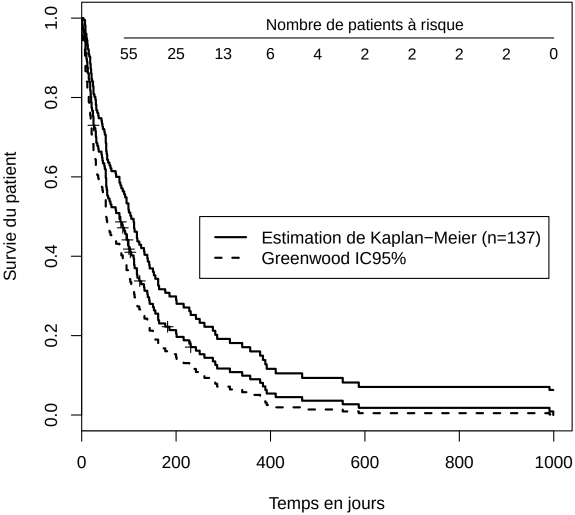
Reprenons l'exemple du groupe A



Les données de cancer de poumon


```
> f <- survfit(Surv(time, status) ~ 1,
+  error="greenwood", data = veteran)
> plot(f, ylab="Survie du patient",
+  col=1, lwd=2, lty=1:2, xlab="Temps en jours",
+  conf.int=TRUE)
> nb.risk<-function(x) {
+  dim(veteran)[1]-sum(veteran$time<x) }
> .time <- seq(100, 1000, by=100)
> text(x=500, y=0.98, "Nombre de patients à risque", cex=0.9)
>  segments(x0=0, y0=0.95, x1=1000, y1=0.95)
> text(x=.time, y=rep(0.91, length(.time)),
+  as.character(sapply(.time, FUN="nb.risk")), cex=0.9)
> legend(250, 0.5, c("Estimation de Kaplan-Meier (n=137)", "Greenwood IC95%"),
+  col=1, lty=1:2, lwd=2)
```

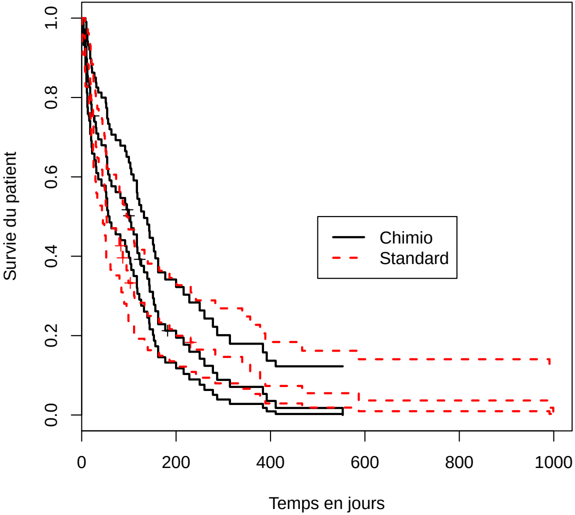
Les données de cancer de poumon



Il y a-t-il une différence de survie selon le traitement ?


```
> f <- survfit(Surv(time, status) ~ trt,
+  error="greenwood", data = veteran)
> plot(f, ylab="Survie du patient",
+  col=1:4, lwd=2, lty=1:2, xlab="Temps en jours",
+  conf.int=TRUE)
> legend(500, 0.5, c("Chimio", "Standard"),
+  col=1:2, lty=1:2, lwd=2)
```

Comme d'habitude, un chevauchement d'IC95% ne prouve rien...



Plan

Principe du test du LogRank

Il s'agit d'une statistique construite sous la forme d'une statistique du χ^2. L'idée est donc de calculer des effectifs dits *attendus* d'événements (e) et de les comparer aux effectifs observés (o). Ces effectifs attendus sont calculés en considérant l'hypothèse nulle vraie :

H_0 : Les deux coubes de survie sont similaires : $S_A(t) = S_B(t)$.

H_1 : Les deux coubes de survie sont différentes : $S_A(t) \neq S_B(t)$.

Pour cela, il suffit de considérer qu'à chaque temps d'événement observé $(\delta_i = 1)$, la probabilité d'apparition de l'événement est équivalente dans les deux groupes. Elle est simplement estimée par le nombre d'événement divisé par l'effectif à risque quel que soit le groupe (d_i/n_i).

Si n_{Ai} et n_{Bi} sont les effectifs à risque dans chacun des groupes (tels que $n_i = n_{Ai} + n_{Bi}$), alors le nombre de décès attendus est $e_{Ai} = \frac{d_i}{n_i} n_{Ai}$ dans le groupe A et $e_{Bi} = \frac{d_i}{n_i} n_{Bi} = d_i - e_{Ai}$ dans le groupe B.

Reprenons notre exemple simple

t_i	d_{Ai}	n_{Ai}	d_{Bi}	n_{Bi}	d_i	n_i	e_{Ai}	e_{Bi}
24	1	4	0	3	1	7	4*1/7=0.57	3*1/7=0.43
28	0	2	1	3	1	5	2*1/5=0.40	3*1/5=0.60
30	1	2	0	2	1	4	2*1/4=0.50	2*1/4=0.50

$\rightarrow o_A = 2,\ o_B = 1,\ e_A = 1.47,\ e_B = 1.53$

La statistique de test

$$\chi^2_{LR} = \frac{(o_A - e_A)^2}{e_A} + \frac{(o_B - e_B)^2}{e_B} \sim \chi^2_{1\ ddl}$$

Pour notre exemple, on a :

```
> stat <- ((2-1.47)^2)/1.47 + ((1-1.53)^2)/1.53
> 1-pchisq(3.84, df = 1) # pour comprendre...
[1] 0.05004352
> 1-pchisq(stat, df = 1)
[1] 0.5404625
```

L'étude ne permet pas de montrer une relation significative entre le type de chimiothérapie et la survie des patients ($p > 0.05$).

Reprenons l'étude du cancer du poumon


```
> survdiff(Surv(time, status) ~ trt, data = veteran)

Call:
survdiff(formula = Surv(time, status) ~ trt, data = veteran)

       N Observed Expected (O-E)^2/E (O-E)^2/V
trt=1 69       64     64.5   0.00388   0.00823
trt=2 68       64     63.5   0.00394   0.00823

 Chisq= 0  on 1 degrees of freedom, p= 0.928
```

L'étude ne permet pas de montrer une relation significative entre le traitement et la survie des patients atteints d'un cancer du poumon ($p > 0.05$).

Plan

Plan

Le choix d'une distribution des temps de survie

- Hypothèse : la distribution des temps de survie répond à une loi paramétrique prédéfinie
- Le livre de Bagdonavicius et Nikulin ‡ détaille beaucoup de possibilités
- Les deux lois simples les plus utilisées :
  - Loi Exponentielle
  - Loi de Weibull

‡. Vilijandas Bagdonavicius et Mikhail Nikulin. Accelerated Life Models. Modeling and Statistical Analysis. 2002, Chapman and Hall/CRC : Boca Raton, London.

La loi Exponentielle

- Hypothèse : la fonction de risque est constante au cours du temps

$$h(t) = \alpha \quad (1 \text{ paramètre})$$

d'où

$$H(t) = \int_0^t \alpha du = \alpha \int_0^t du = \alpha t,$$

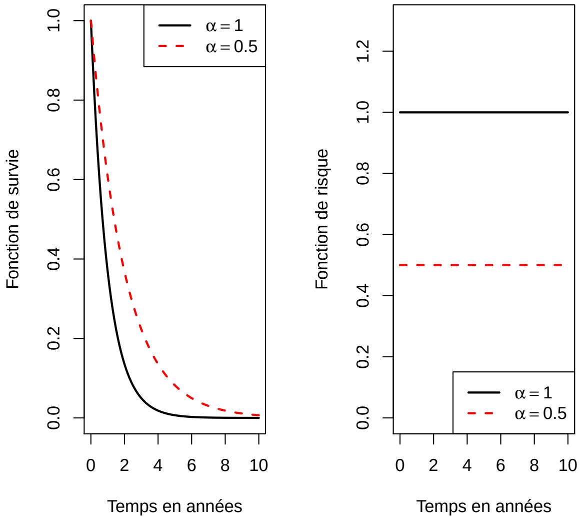
$$S(t) = exp(-\alpha t),$$

$$f(t) = \alpha exp(-\alpha t).$$

La loi Exponentielle


```
> h.exp <- function(a, t) { a * rep(1, length(tps)) }
> s.exp <- function(a, t) { exp(-a*t) }
> tps <- seq(0, 10, by=0.1)
> par(mfrow=c(1,2))
> plot(tps, h.exp(a=1, t=tps), ylab="Fonction de risque",
+  xlab="Temps en années", ylim=c(0, 1.3), type="l", lwd=2)
> lines(tps, h.exp(a=0.5, t=tps), col=2, type="l", lwd=2, lty=2)
> legend("bottomright", c(expression(alpha==1.0),
+  expression(alpha==0.5)), lwd=2, col=1:2, , lty=1:2)
> plot(tps, s.exp(a=1, t=tps), ylab="Fonction de survie",
+  xlab="Temps en années", ylim=c(0, 1), type="l", lwd=2)
> lines(tps, s.exp(a=0.5, t=tps), col=2, type="l", lwd=2, lty=2)
> legend("bottomright", c(expression(alpha==1.0),
+  expression(alpha==0.5)), lwd=2, col=1:2, , lty=1:2)
```

La loi Exponentielle



La loi de Weibull

- Hypothèse : la fonction de risque est monotone (croissante ou décroissante au cours du temps)

$$h(t) = \gamma\alpha^\gamma t^{\gamma-1} \quad \text{(2 paramètres)}$$

d'où

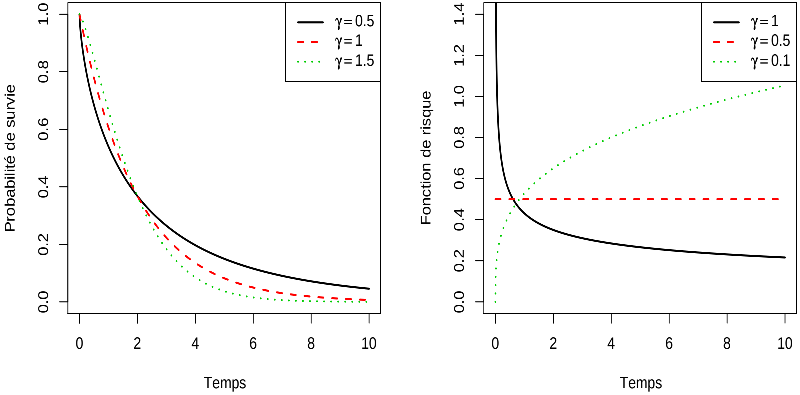
$$H(t) = (\alpha t)^\gamma \quad \text{(exo : calculer l'intégrale)}$$

$$S(t) = exp(-(\alpha t)^\gamma),$$

$$f(t) = \gamma\alpha^\gamma t^{\gamma-1} exp(-\alpha t).$$

- Si $\gamma = 1 \Rightarrow$ Loi Exponentielle

La loi de Weibull (pour $\alpha = 0.5$ quelque soit la courbe)



Plan

L'estimateur du maximum de Vraisemblance (1)

La Vraisemblance, ℓ, représente la probabilité d'observer l'échantillon.
Rappels de notations :

- Soit N individus inclus dans l'étude ($i = 1, ..., N$).
- Soit y_i le temps de suivi de l'individu i.
- Soit δ_i l'indicatrice de censure : $\delta_i = 1$ si y_i correspond à un événement et $\delta_i = 0$ si y_i correspond à une censure à droite.

L'estimateur du maximum de Vraisemblance (2)

- Si tous les individus sont indépendants :

$$\ell = \prod_{i}^{N} \ell_i$$

$\rightarrow$ ℓ_i représente la probabilité d'observer l'individu i.
$\rightarrow$ Elle est aussi appelée contribution individuelle.

- $\ell_i = f(y_i)$ si l'événement est observé au temps y_i $(\delta_i = 1)$
- $\ell_i = S(y_i)$ si le temps de survie est censuré au temps y_i $(\delta_i = 0)$

$$\ell = \prod_{i}^{N} \ell_i = \prod_{i}^{N} f(y_i)^{\delta_i} S(y_i)^{1-\delta_i} = \prod_{i}^{N} h(y_i)^{\delta_i} S(y_i)$$

L'estimateur du maximum de Vraisemblance (3)

Il est plus simple de travailler à partir de la logVraisemblance :

$$\log(\ell) = \sum_i^N \delta_i log(h(y_i)) + log(S(y_i))$$

$$\log(\ell) = \sum_i^N \delta_i log(h(y_i)) - H(y_i)$$

Exemple de la loi Exponentielle

$$\log(\ell) = \sum_i^N \delta_i log(h(y_i)) - H(y_i) \Longleftrightarrow \log(\ell) = \sum_i^N \delta_i log(\alpha) - \alpha y_i$$

$$\partial \log(\ell)/\partial\alpha = \sum_i^N \delta_i/\alpha - y_i$$

$$\hat{\alpha} \text{ satisfait } \partial \log(\ell)/\partial\alpha = 0 \Longleftrightarrow \sum_i^N \delta_i/\hat{\alpha} - y_i = 0$$

$$\sum_i^N \delta_i/\hat{\alpha} = \sum_i^N y_i \Longleftrightarrow \hat{\alpha} = \sum_i^N \delta_i / \sum_i^N y_i$$

$$\hat{\alpha} = \text{Nb Evt/Suivi cumulé total}$$

Exemple de la loi Exponentielle

- Exo : Retrouver l'estimateur de la variance de $\hat{\alpha}$. Rappel : $\hat{V}(\hat{\alpha})$ est obtenue par :

$$\hat{V}(\hat{\alpha}) = \left[\frac{-\partial^2 \log(\ell)}{\partial^2 \alpha}\right]^{-1}$$

- En pratique, le modèle est souvent complexe et est basé sur de nombreux paramètres. Il n'est alors plus possible d'estimer à la main leur valeur. Les logiciels utilisent des algorithmes qui permettent d'obtenir numériquement ces résultats (estimations et variances).

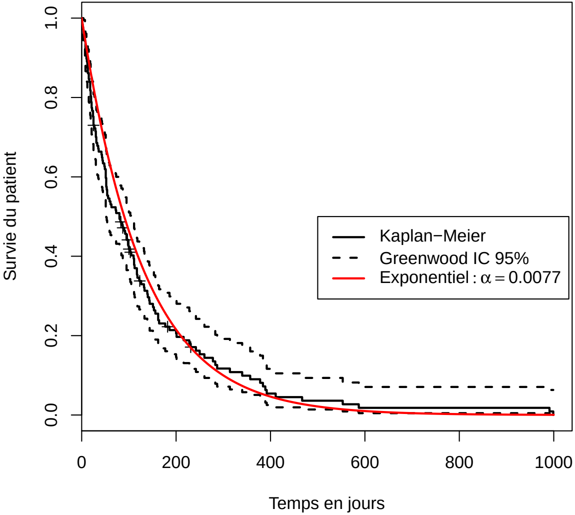
Application au cancer du poumon


```
> f <- survfit(Surv(time, status) ~ 1,
+  error="greenwood", data = veteran)
> plot(f, ylab="Survie du patient",
+  col=1, lwd=2, xlab="Temps en jours",
+  conf.int=TRUE)
>  alpha <- sum(veteran$status)/sum(veteran$time)
>  alpha

[1] 0.00768169

>  tps <- 0:1000
>  lines(tps, s.exp(a=alpha, tps), col=2, lwd=2)
> legend(500, 0.5, c("Kaplan-Meier", "Greenwood IC 95%", expression(Exponentiel:alp
```

Application au cancer du poumon



Plan

Introduction
Définitions
Kaplan-Meier
Estimation
Variance
LogRank
Distributions paramétriques
Les principales lois
L'estimation des paramètres
Modèle de Cox
Définitions
Inférence
Hypothèses
Perspectives

Plan

Introduction
Définitions
Kaplan-Meier
Estimation
Variance
LogRank
Distributions paramétriques
Les principales lois
L'estimation des paramètres
Modèle de Cox
Définitions
Inférence
Hypothèses
Perspectives

Le modèle de Cox [§]

$$h(t|X_1 = x_1, X_2 = x_2, ...) = h_0(t)exp(\beta_1 x_1 + \beta_2 x_2 + ...)$$

- $h_0(t)$ est la fonction de risque de base : fonction de risque dans la population de référence (toutes les variables explicatives sont nulles).
- $\beta = (\beta_1, \beta_2, ...)$ sont les coefficients de régression.
- Principe fondamental : la fonction de risque n'est pas estimée : Modèle semi-paramétrique.
  - $\rightarrow$ Partie "non paramétrique" (pas d'estimation de $h_0(t)$)
  - $\rightarrow$ Partie paramétrique (fonction exponentielle d'un régresseur linéaire)

§. Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society : Series B, 34 :187-220, 1972.

Exemple d'interprétation pour β_1 (H/F)

$$h(t|X_1 = x_1, X_2 = x_2, ...) = h_0(t)exp(\beta_1 x_1 + \beta_2 x_2 + ...)$$

- Chez les hommes ($x_1 = 1$) et toutes les autres covariables fixées ($X_2 = x_2$, ...) :

$$h(t|x_1 = 1, X_2 = x_2, ...) = h_0(t)exp(\beta_1)exp(\beta_2 x_2 + ...)$$

- Chez les femmes ($x_1 = 0$) et toutes les autres covariables fixées ($X_2 = x_2$, ...) :

$$h(t|x_1 = 0, X_2 = x_2, ...) = h_0(t)exp(\beta_2 x_2 + ...)$$

- $HR_{H/F} = exp(\beta_1)$ représente le rapport des risques (hazard ratio : HR) entre hommes et femmes, toutes les autres variables étant constantes.

Généralisation

- Si $\beta_1 > 0$ alors $exp(\beta_1) > 1$: x_1 est un facteur de risque.
- Si $\beta_1 < 0$ alors $exp(\beta_1) < 1$: x_1 est un facteur protecteur.
- Si $\beta_1 = 0$ alors $exp(\beta_1) = 1$: x_1 n'est pas un facteur influençant le temps d'événement.

La notion d'interaction (1)

$$h(t|X_1 = x_1, X_2 = x_2) = h_0(t)exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + ...)$$

- Dans le groupe $\{X_1 = 0\}$, c'est à dire les femmes :
  $\rightarrow$ $h_0(t)exp(\beta_2 X_2)$
  $\rightarrow$ $HR_{X_2=1/X_2=0} = exp(\beta_2)$
- Dans le groupe $\{X_1 = 1\}$, c'est à dire les hommes :
  $\rightarrow$ $h_0(t)exp(\beta_1 + \beta_2 X_2 + \beta_3 X_2) = h_0(t)exp(\beta_1 + (\beta_2 + \beta_3)X_2)$
  $\rightarrow$ $HR_{X_2=1/X_2=0} = exp(\beta_2 + \beta_3)$
- Le HR associé X_2 est différent selon les modalités de la variable X_1 (et inversement)
- $HR_{x_2} = exp(\beta_2)exp(\beta_3 X_1)$: $exp(\beta_3)$ représente le facteur entre les HR des groupes $\{X_1 = 0\}$ et $\{X_1 = 1\}$.

La notion d'interaction (2)

$$h(t|X_1 = x_1, X_2 = x_2) = h_0(t) exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2)$$

- Si $\beta_3 > 0$ alors $exp(\beta_3) > 1$: le risque est supérieur dans le groupe $\{x_1 = 1\}$.
- Si $\beta_3 < 0$ alors $exp(\beta_3) < 1$: le risque est inférieur dans le groupe $\{x_1 = 1\}$.
- Si $\beta_3 = 0$ alors $exp(\beta_3) = 1$: le risque est inchangé entre les deux groupes.

Plan

La Vraisemblance partielle (1)

- Soit $y_1 < y_2 < ... < y_N$ les différents temps observés pour les N individus.
- On démontre que la probabilité conditionnelle que le sujet i subisse l'événement en y_i sachant qu'il est à risque au temps y_i parmi tous les individus à risque au même temps est égale à :

$$\ell_i = \frac{h(y_i|x_{1i}, ...)}{\sum_{j:y_j \geq y_i} h(y_i|x_{1j}, ...)}$$

$$\ell_i = \frac{h_0(y_i)exp(\beta_1 x_{1i} + ...)}{\sum_{j:y_j \geq y_i} h_0(y_i)exp(\beta_1 x_{1j} + ...)}$$

$$\ell_i = \frac{exp(\beta_1 x_{1i} + ...)}{\sum_{j:y_j \geq y_i} exp(\beta_1 x_{1j} + ...,)}$$

- Il s'agit de la contribution individuelle à la vraisemblance partielle. Elle ne dépend pas de $h_0(t)$.

La Vraisemblance partielle (2)

- En supposant les individus indépendants, le produit des contributions individuelles précédentes pour les sujets ayant subi l'événement permet d'obtenir la vraisemblance partielle :

$$\wp\ell = \prod_{i=1}^{N} \left\{ \frac{exp(\beta_1 x_{1i} + \ldots)}{\sum_{j:y_j \geq y_i} exp(\beta_1 x_{1j} + \ldots)} \right\}^{\delta_i}$$

- La maximisation de cette fonction permet d'estimer les coefficients de régression, $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots)$, sans avoir à estimer la distribution de ce dernier.
- La matrice Hessienne correspondante permet d'obtenir la matrice de variance-covariance des coefficients de régression : $Var(\hat{\beta})$.
- On peut en déduire l'intervalle de confiance à (1-α)% toujours grâce au théorème central limite, par exemple pour β_1 :

$$IC95\%[\hat{\beta}_1] = \hat{\beta}_1 \pm z_{\alpha/2}\sqrt{Var(\hat{\beta}_1)}$$

Application au cancer du poumon


```
> dim(veteran)

[1] 137   8

> names(veteran)

[1] "trt"      "celltype" "time"     "status"   "karno"    "diagtime" "age"
[8] "prior"

> fit.cox1<-coxph(Surv(time, status)~ trt,
+  data = veteran)
> fit.cox1

Call:
coxph(formula = Surv(time, status) ~ trt, data = veteran)


      coef exp(coef) se(coef)      z    p
trt 0.0177      1.02    0.181 0.0982 0.92

Likelihood ratio test=0.01  on 1 df, p=0.922  n= 137, number of events= 128
```

Les patients avec un traitement standard (trt=2) ont 1.02 fois plus de risque de décéder que les patients avec la nouvelle chimiothérapie (trt=1).

Application au cancer du poumon


```
> summary(fit.cox1)

Call:
coxph(formula = Surv(time, status) ~ trt, data = veteran)

  n= 137, number of events= 128

       coef exp(coef) se(coef)     z Pr(>|z|)
trt 0.01774   1.01790  0.18066 0.098    0.922

    exp(coef) exp(-coef) lower .95 upper .95
trt     1.018     0.9824    0.7144      1.45

Concordance= 0.525  (se = 0.026 )
Rsquare= 0   (max possible= 0.999 )
Likelihood ratio test= 0.01  on 1 df,   p=0.9218
Wald test            = 0.01  on 1 df,   p=0.9218
Score (logrank) test = 0.01  on 1 df,   p=0.9218
```

L'intervalle de confiance à 95% du HR estimé est compris entre 0,71 et 1,45.

```
> names(fit.cox1)
 [1] "coefficients"      "var"               "loglik"
 [4] "score"             "iter"              "linear.predictors"
 [7] "residuals"         "means"             "concordance"
[10] "method"            "n"                 "nevent"
[13] "terms"             "assign"            "wald.test"
[16] "y"                 "formula"           "call"
> fit.cox1$coefficients
       trt
0.01774257
> fit.cox1$loglik
[1] -505.4491 -505.4442
```

La logVraisemblance de ce modèle est de -505,444. Une belle jambe (pour l'instant...).

Tests d'hypothèses

H_0 : $HR_{x_1} = 1 \Leftrightarrow \beta_1 = 0$

H_1 : $HR_{x_1} \neq 1 \Leftrightarrow \beta_1 \neq 0$

- Test de Wald (bilatéral ou unilatéral) :

$$T = \frac{\hat{\beta}_1}{\sqrt{Var(\hat{\beta}_1)}} \sim N(0,1)$$

- Test du rapport des vraisemblances (bilatérale) :

$$\chi^2_{LRS} = -2 \times \{\log(\wp\ell_0) - \log(\wp\ell_1)\} \sim \chi^2_{1ddl}$$

où ℓ_0 est la logVraisemblance sans β_1 (i.e. $\beta = 0$) et ℓ_1 celle du modèle complet.

Tester plusieurs paramètres simultanément

H_0 : $\beta_1 = \beta_2 = ... = \beta_p = 0$

H_1 : Au moins un coefficient différent de 0.

Test du rapport des vraisemblances :

$$\chi^2_{LRS} = -2 \times \{\log(\wp\ell_0) - \log(\wp\ell_1)\} \sim \chi^2_{pddl}$$

où ℓ_0 est la logVraisemblance sans $\{x_1, ..., x_p\}$ et ℓ_1 celle avec du modèle complet.

Application au cancer du poumon


```
> dim(veteran)

[1] 137   8

> names(veteran)

[1] "trt"      "celltype" "time"     "status"   "karno"    "diagtime" "age"
[8] "prior"

> fit.cox2<-coxph(Surv(time, status)~ trt + age,
+  data = veteran)
> fit.cox2

Call:
coxph(formula = Surv(time, status) ~ trt + age, data = veteran)


        coef exp(coef) se(coef)      z    p
trt -0.00365     0.996  0.18251 -0.020 0.98
age  0.00753     1.008  0.00966  0.779 0.44

Likelihood ratio test=0.63  on 2 df, p=0.731  n= 137, number of events= 128
```

Pour un âge fixé, les patients avec un traitement standard (trt=2) ont 1.00 fois plus de risque de décéder que les patients avec la nouvelle chimiothérapie (trt=1). On ne peut pas montrer que ce HR entre les deux groupe soit statistiquement significatif ($p > 0.05$).

Application au cancer du poumon


```
> fit.cox1$loglik
[1] -505.4491 -505.4442
> fit.cox2$loglik
[1] -505.4491 -505.1361
> -2*(fit.cox1$loglik[2]-fit.cox2$loglik[2])
[1] 0.6161708
> 1-pchisq(-2*(fit.cox1$loglik[2]-fit.cox2$loglik[2]), df=1)
[1] 0.4324738
> summary(fit.cox2)$coefficients
            coef exp(coef)    se(coef)           z  Pr(>|z|)
trt -0.003654234 0.9963524 0.182513535 -0.02002171 0.9840261
age  0.007527262 1.0075557 0.009661283  0.77911618 0.4359113
> summary(fit.cox2)$coefficients[2,5]
[1] 0.4359113
```

Les tests du rapport de vraisemblance et de Wald renvoient des probabilités critiques équivalentes.

Application au cancer du poumon


```
> summary(veteran$age[veteran$trt==1])

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  34.00   49.00   62.00   57.51   66.00   81.00

> summary(veteran$age[veteran$trt==2])

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  35.00   52.00   62.00   59.12   66.00   81.00

> t.test(veteran$age[veteran$trt==1], veteran$age[veteran$trt==2])

Welch Two Sample t-test

data:  veteran$age[veteran$trt == 1] and veteran$age[veteran$trt == 2]
t = -0.8937, df = 134.827, p-value = 0.3731
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.174206  1.953405
sample estimates:
mean of x mean of y
 57.50725  59.11765
```

Pas étonnant que les résultats liés au traitement soient inchangés, les moyennes d'âge semblent proches entre les deux groupes. On ne devrait pas faire un test de différence de moyennes d'âge entre les deux groupes car l'hypothèse nulle est vraie à coup sûr (tirage au sort).

Application au cancer du poumon


```
> summary(coxph(Surv(time, status)~ trt * age,
+   data = veteran))$coefficients
                 coef exp(coef)   se(coef)           z  Pr(>|z|)
trt     -0.3113942213 0.7324251 1.20722624 -0.25794189 0.7964518
age      0.0004951811 1.0004953 0.02884639  0.01716614 0.9863041
trt:age  0.0052054323 1.0052190 0.02016938  0.25808585 0.7963406
```

Nous ne pouvons pas mettre en évidence un effet différent du traitement selon l'âge ($p = 0.7960$).

Plan

Modèle à risque proportionnel

$$h(t|x_1, x_2, ...) = h_0(t) exp(\beta_1 x_1 + \beta_2 x_2 + ...)$$

$\rightarrow$ Fonction de risque dépendante du temps

$\rightarrow$ Facteurs de risque indépendants du temps

Modèle à risque proportionnel

$$h(t|x_1, x_2, ...) = h_0(t) exp(\beta_1 x_1 + \beta_2 x_2 + ...)$$

$\rightarrow$ Fonction de risque dépendante du temps

$\rightarrow$ Facteurs de risque indépendants du temps

Modèle à risque proportionnel

$$h(t|x_1, x_2, ...) = h_0(t)exp(\beta_1 x_1 + \beta_2 x_2 + ...)$$

→ Fonction de risque dépendante du temps.

→ Facteurs de risque indépendants du temps. Les HR sont constants au cours du temps.

Cette dernière hypothèse (*proportional hazard*, PH) est régulièrement non respectée.

Modèle à risque proportionnel

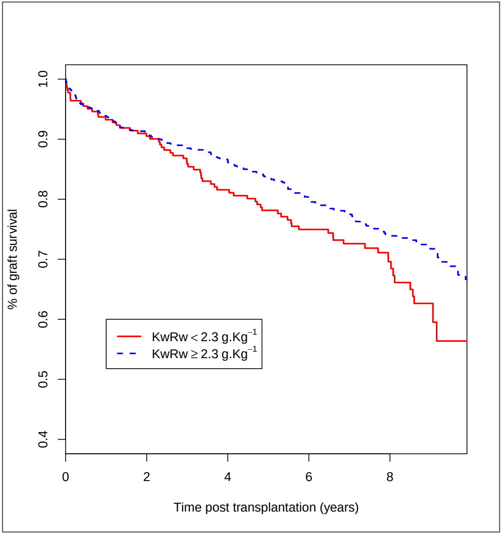
$$h(t|x_1, x_2, ...) = h_0(t)exp(\beta_1 x_1 + \beta_2 x_2 + ...)$$

$\rightarrow$ Fonction de risque dépendante du temps.

$\rightarrow$ Facteurs de risque indépendants du temps. Les HR sont constants au cours du temps.

Cette dernière hypothèse (*proportional hazard*, PH) est régulièrement non respectée.

$$KwRw = \frac{\text{Poids du rein}}{\text{Poids du receveur}}$$



Hypothèse PH

- Tracer les courbes $log(-log(S(t)))$ dans chacun des groupes en fonction du temps t.
- Si courbes parallèles : hypothèse respectée.
- Démonstration :

$$h(t|x_1) = h_0(t)exp(\beta_1 x_1)$$

$$\Longleftrightarrow$$

$$H(t|x_1) = \int_0^t h_0(u)exp(\beta_1 x_1)du$$

$$\Longleftrightarrow$$

$$H(t|x_1) = exp(\beta_1 x_1)\int_0^t h_0(u)du = H_0(t)exp(\beta_1 x_1)$$

$$\Longleftrightarrow$$

$$S(t|x_1) = exp(-H(t|x_1)) = exp(-H_0(t)exp(\beta_1 x_1)) = S_0(t)^{exp(\beta_1 x_1)}$$

Hypothèse PH

Donc si l'hypothèse PH est respectée :

$$log(S(t|x_1)) = exp(\beta_1 x_1) log(S_0(t))$$

$$\Longleftrightarrow$$

$$-log(S(t|x_1)) = exp(\beta_1 x_1) H_0(t)$$

$$\Longleftrightarrow$$

$$log(-log(S(t|x_1))) = \beta_1 x_1 + log(H_0(t))$$

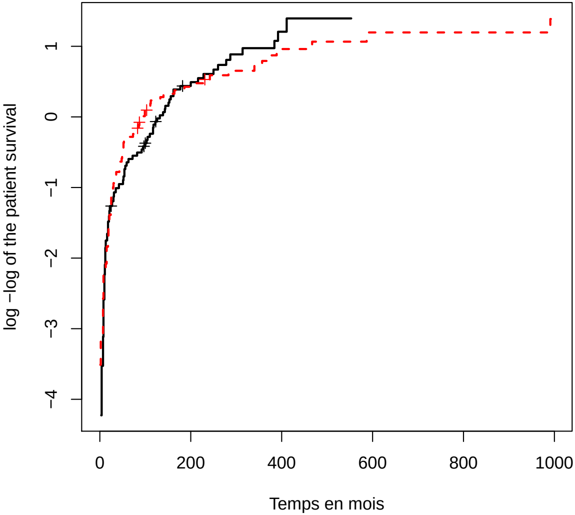
Ex : si $x_1 = 1/0$:

$$log(-log(S(t|x_1 = 1))) - log(-log(S(t|x_1 = 0))) = \beta_1$$

Application au cancer du poumon


```
> f <- survfit(Surv(time, status) ~ trt, data = veteran)
> log.minus.log<-function(y) { log(-log(y)) }
> plot(f , fun=log.minus.log, ylab="log -log of the patient survival",
+  col=1:2, lwd=c(2,2), lty=1:2, xlab="Temps en mois")
```

Application au cancer du poumon



La Log linéarité

$$h(t|x_1, x_2, ...) = h_0(t) exp(\beta_1 x_1 + \beta_2 x_2 + ...)$$

→ Si x_2 est une variable continue :

$$HR_{u+1/u} = exp(\beta_2) \ \forall u$$

→ Cette relation dose-effet constante peut souvent être remise en question, en particulier quand il s'agit de la variable d'intérêt (et pas d'une "simple" variable d'ajustement).

→ Une solution simple et pragmatique :

1. catégoriser la variable en p classes équilibrées,
2. choisir la première classe comme référence ($p - 1$ variables indicatrices),
3. tracer le log des HR en fonction du centre de chaque classe,
4. vérifier la linéarité.

La Log linéarité

$$h(t|x_1, x_2, ...) = h_0(t)exp(\beta_1 x_1 + \beta_2 x_2 + ...)$$

→ Si x_2 est une variable continue :

$$HR_{u+1/u} = exp(\beta_2)\ \forall u$$

→ Cette relation dose-effet constante peut souvent être remise en question, en particulier quand il s'agit de la variable d'intérêt (et pas d'une "simple" variable d'ajustement).

→ Une solution simple et pragmatique :

1. catégoriser la variable en p classes équilibrées,
2. choisir la première classe comme référence ($p - 1$ variables indicatrices),
3. tracer le log des HR en fonction du centre de chaque classe,
4. vérifier la linéarité.

Plan

A venir...

- Le modèle de Cox stratifié : une fonction de risque base différente par modalité d'une variable (ajuster sur cette variable sans estimer les effets fixes).
- Analyses des résidus.
- Le modèle de Cox étendu aux variables dont l'effet ou la valeur dépendent du temps.
- Les modèles à risques compétitifs permettent de modéliser des événements en concurrence et de prendre en compte une éventuelle censure informative.
- Les modèles de survie relative permettent d'éliminer la mortalité attendue d'une population de référence.
- Les modèles multi-états permettent de modéliser l'évolution d'un patient à travers différents états santé.
- Les modèles de fragilités permettent d'inclure des effets aléatoires et donc une structure de dépendance des données.

Quelques mots d'anglais

- LogVraisemblance $\rightarrow$ LogLikelihood
- Hazard Ratio (HR)
- Modèle à Risque Proportionnel $\rightarrow$ Proportional Hazard (PH) model
- Test du rapport de Vraisemblances $\rightarrow$ Likelihood Ratio Statistic (LRS)
- Vraisemblance partielle $\rightarrow$ Partial Likelihood
- Fonction de répartition : Cumulative Incidence Function (CIF).