Introduction à l'analyse de survie

Michaël Genin

Université de Lille 2

EA 2694 - Santé Publique : Epidémiologie et Qualité des soins

michael.genin@univ-lille2.fr

Université de Lille 2 DROIT ET SANTÉ

Plan

Définition

Analyse de données de survie : étude l'apparition d'un évènement au cours du temps.

Exemples :

- Temps de survie après le diagnostic d'un cancer du sein
- Durée de séropositivité sans symptôme de patients infectés par le VIH
- Durée de vie d'une ampoule, d'un pièce mécanique, ...

Distinguer l'évènement d'intérêt :

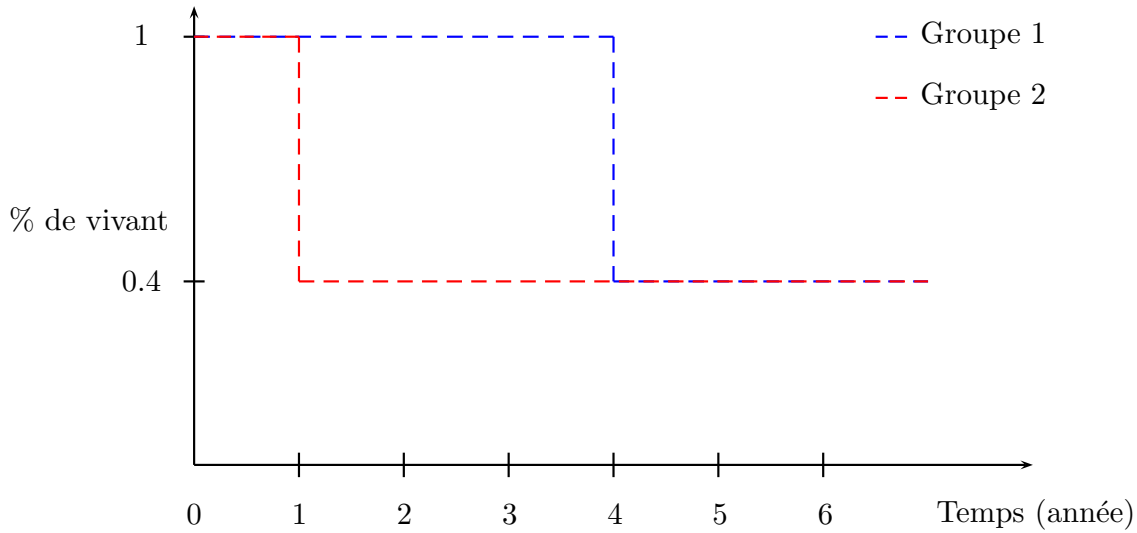
- Décès par cancer du sein
- Apparition de symptome
- Arrêt de fonctionnement de l'ampoule, pièce mécanique,...

de la variable à expliquer :

- Temps de survie
- Temps écoulé sans symptôme
- Temps de fonctionnement de l'ampoule, de la pièce mécanique,...

Spécificités des études de survie

1 Prise en compte du temps



Même % de décès à 5 ans (30%)

Une simple comparaison de % ne permet pas d'affirmer que le temps de survie dans le Groupe 1 $>$ à celui dans le groupe 2

Spécificités des études de survie

2 Prise en compte d'observations incomplètes

On peut pas attendre que l'évènement soit observé pour chaque sujet de l'échantillon (longueur de l'étude).

Dans certains cas, il se peut que l'évènement ne soit jamais observé chez certains sujets.

Mais on veut pouvoir prendre en compte toutes les observations y compris celles pour lesquelles l'évènement n'a pas été observé.

Exemple : 100 sujets suivis pendant 1 an - 6 évènements observés - 4 perdus de vue. Le % de survie à 1 an **n'est pas** :

- 6/100 $\rightarrow$ cela suppose que les perdus de vue sont indemnes au bout des 12 mois
- 6/96 $\rightarrow$ on ne prend pas en compte l'exposition sans déclenchement des perdus de vue. Perte d'information et source de biais (jamais sûr que les PV ont une évolution comparable aux autres)

Applications des méthodes d'analyse de survie

- **Descriptive** → Estimation de la durée de survie
- **Comparative** → comparaison de la survie entre plusieurs groupes
- **Prédictive** → Modèles multivariés
  - → Ajustement sur Facteurs de Confusion
  - → Détermination de facteurs pronostics de la survie

Définitions

Date d'Origine (DO)

Date d'entrée dans l'étude du patient (variable en fonction des individus).
Exemples :

- Date de tirage au sort (essai thérapeutique)
- Date de diagnostic (étude prospective)

Date de Dernières Nouvelles (DDN)

Date la plus récente où le sujet à été revu. (Rq : si patient décédé $\rightarrow$ DDN = date de décès)

Délai de surveillance

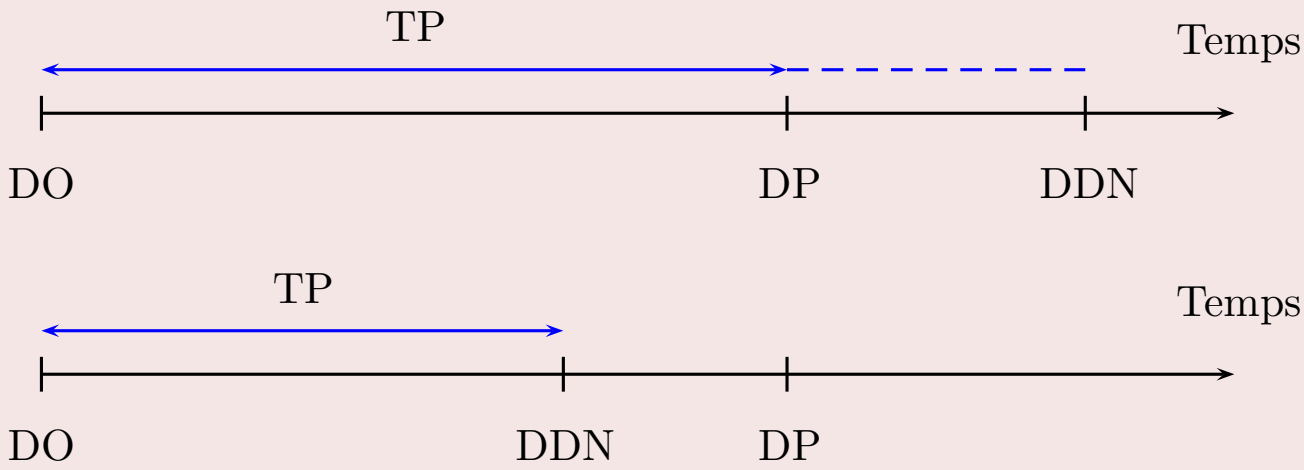
Délai entre DO et DDN.

Définitions

Temps de Participation (TP) - I

Analyse des résultats → on ne peut attendre la survenue de l'évènement pour tous les sujets.
2 cas possibles :

1. On fixe *a priori* (dans le protocole) la date du bilan de l'étude → au delà de cette date, on ne tient plus compte des informations éventuellement recueillies (patient décédé). Cette date butoir est appelée **date de point (DP)**.

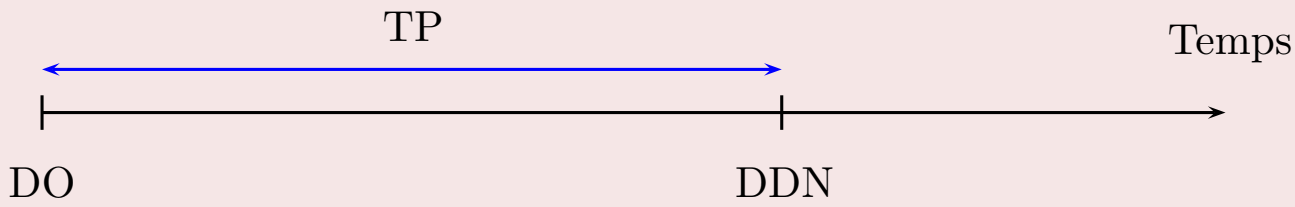


Définitions

Temps de Participation (TP) - II

Analyse des résultats → on ne peut attendre la survenue de l'évènement pour tous les sujets.
2 cas possibles :

2. On fixe *a priori* la durée d'observation potentielle unique pour chaque sujet observé.
   Exemple : survenue d'une hépatite B après avoir reçu un vaccin. Chaque sujet est suivi pendant 1 an.



TP = au + le délai d'observation fixé à l'avance (ex : 1 an)

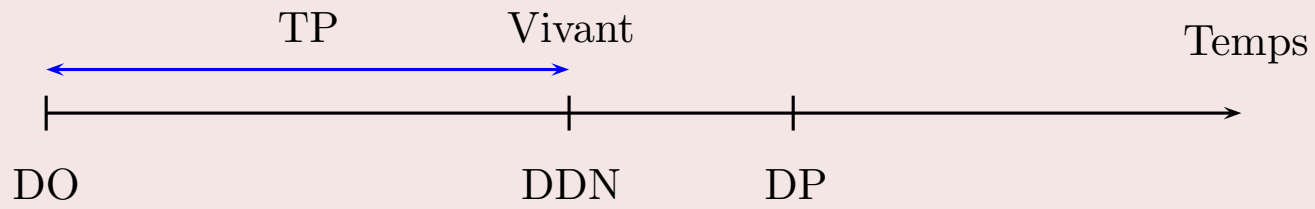
Recul

Délai entre DO et DP

Définitions

Perdu de Vue (PV)

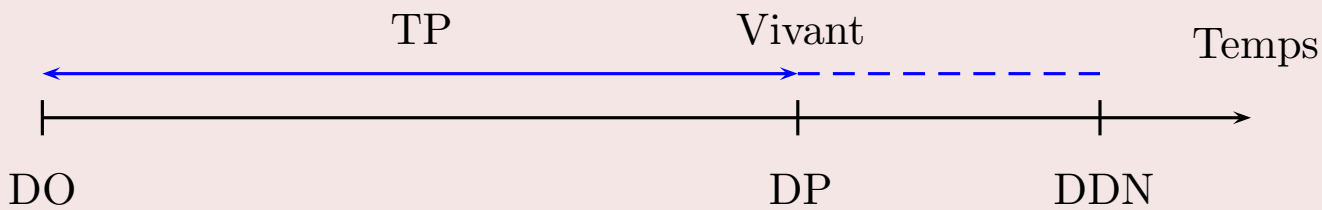
Sujet dont on ne connait pas l'état à la DP ($\equiv$ vivant à la DDN).



En pratique : $< 10\%$ dans les études prospectives.

Exclu-Vivant (EV)

Sujet n'ayant pas présenté l'évènement à la DP.

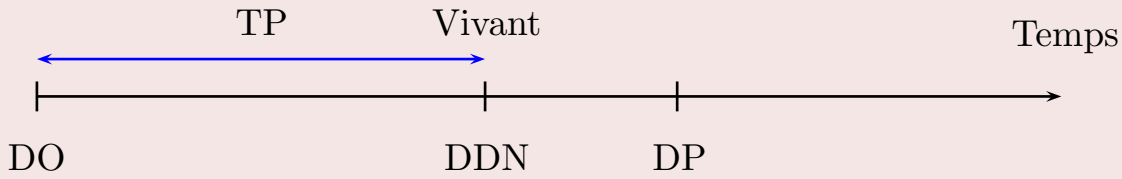


Donnée censurée à droite

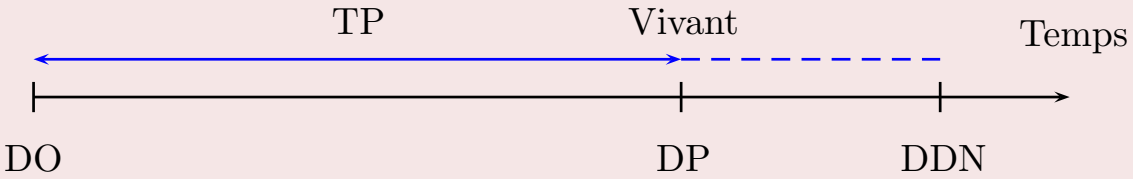
Une durée de vie (observation) est dite **censurée à droite** si l'individu n'a pas présenté l'évènement à sa dernière observation.

2 cas de figure :

1. Perdu de vue



2. Exclu-vivant



Observations incomplètes $\rightarrow$ la durée de vie n'est pas totalement observée.

Censure aléatoire - I

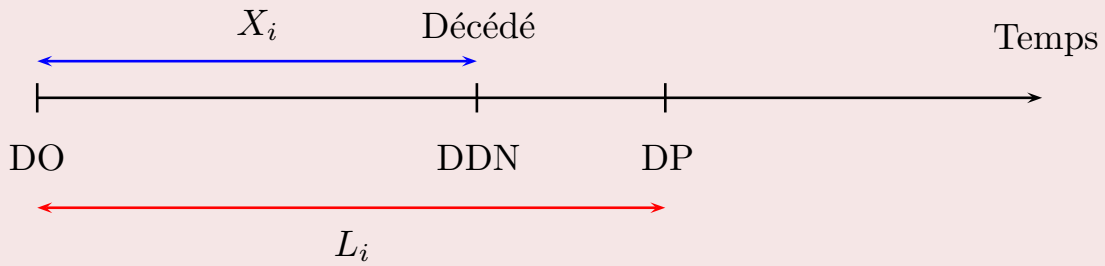
Délai entre la DO et la DP est considéré comme aléatoire (les sujets entrent dans l'étude de manière aléatoire).

Posons L_i, $1 \leq i \leq n$ la v.a.r. qui associe à un individu i sa durée maximale d'observation (appelée également délai de censure).

Posons X_i la v.a.r. qui associe à un individu i son temps de survie. (Rq : $X_i \geq 0$)

Le délai exact de survie est connu uniquement si

$$X_i \leq L_i$$



Nous sommes en présence d'observations complètes (non-censurées).

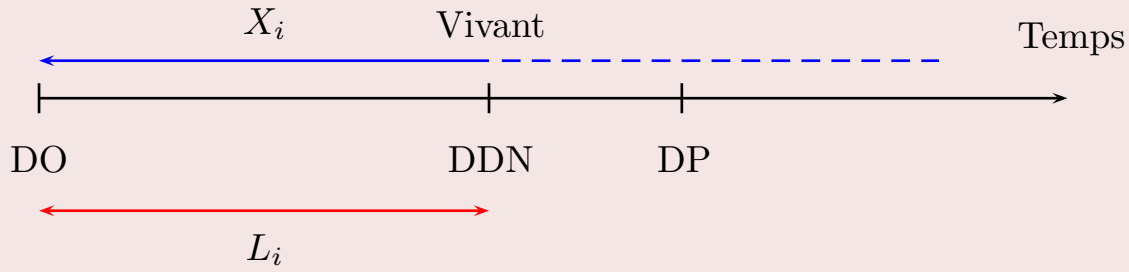
Censure aléatoire - II

Le délai exact de survie n'est pas connu si

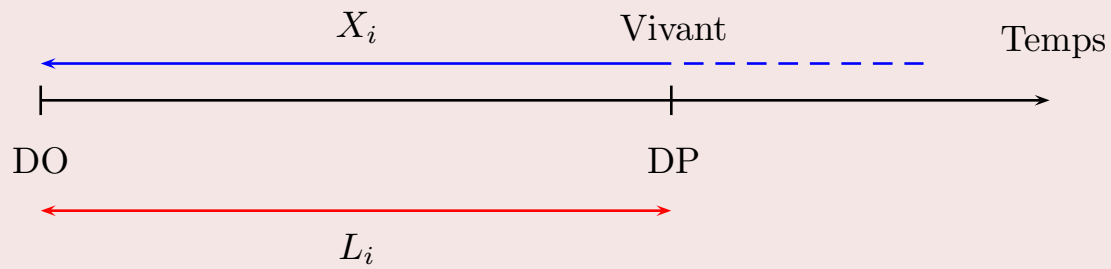
$$X_i > L_i$$

2 cas de figure :

1. Perdu de vue



2. Exclu-vivant



Nous sommes en présence d'observations incomplètes : données censurées.

Censure aléatoire - III

Aussi le temps de survie est définie par

$$T_i = \min\{X_i, L_i\} \text{ et } \delta_i = \mathbb{1}_{X_i \leq L_i}$$

- Si $X_i > L_i \Rightarrow T_i = L_i$ et $\delta_i = 0$. Donnée censurée (évènement non observé).
- Si $X_i \leq L_i \Rightarrow T_i = X_i$ et $\delta_i = 1$. Donnée complète (évènement observé).

Remarque : le couple (T_i, δ_i) est suffisant pour réaliser les analyses de survie.

La censure aléatoire est le mécanisme de censure le plus courant.

Lors d'un essai thérapeutique elle peut être due à :

- La perte de vue (déménagement, soins dans un autre hôpital, ...)
- Arrêt ou changement du traitement (effets secondaires ou inefficacité)
- Exclus-vivants

Censure aléatoire - IV

Hypothèse fondamentale : en cas de censure aléatoire, on considère que le délai de censure L_i est une v.a.r. indépendante du temps de survie X_i.

En pratique, il faut observer la distribution des PV :

- Répartition uniforme dans le temps
- Equilibrée selon la gravité de la maladie, selon les groupes
  - $\rightarrow$ Exclusion des patients les plus à risque $\rightarrow$ surestimation de la survie
  - $\rightarrow$ La censure apporte une information sur le temps de survie des sujets
  - $\rightarrow$ L_i n'est plus indépendante à X_i

Remarque : il existe d'autres mécanismes de censure

- Censure de type I ($L_1 = L_2 = \ldots = L_n = L$) avec L une constante
- Censure de type II : attente d'observation de K évènements.
- ...

Posons T_i la v.a.r. qui associe à un individu i, $1 \leq i \leq n$, son temps de survie ($T_i \geq 0$). On cherche à déterminer sa distribution $f(t)$ ou encore sa fonction de répartition $F(t) = \mathbb{P}(T < t)$

Fonction de survie $S(t)$

Probabilité de survivre au temps t (appelée également courbe de survie).

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

$S(t)$ est une fonction monotone décroissante

$$S(0) = 1, \qquad \lim_{t \to \infty} S(t) = 0$$

Remarquons que

$$S(t) = \mathbb{P}(T \geq t) = 1 - \mathbb{P}(T < t) = 1 - F(t)$$

Fonction quantile du temps de survie

Cette fonction est définie par

$$Q(p) = \inf\{t : S(t) \leq p\}, \ p \in]0, 1[$$

On cherche à estimer $S(t) \rightarrow$ 3 types d'analyse de survie :

1. Méthodes non-paramétriques (Kaplan-Meier, ...)
2. Méthodes semi-paramétriques (Modèle de Cox, ...)
3. Méthodes paramétriques (Modèle exponentiel, Weibul, ...)

Les méthodes 2 et 3 permettent la prise en compte de variables explicatives (X_j).

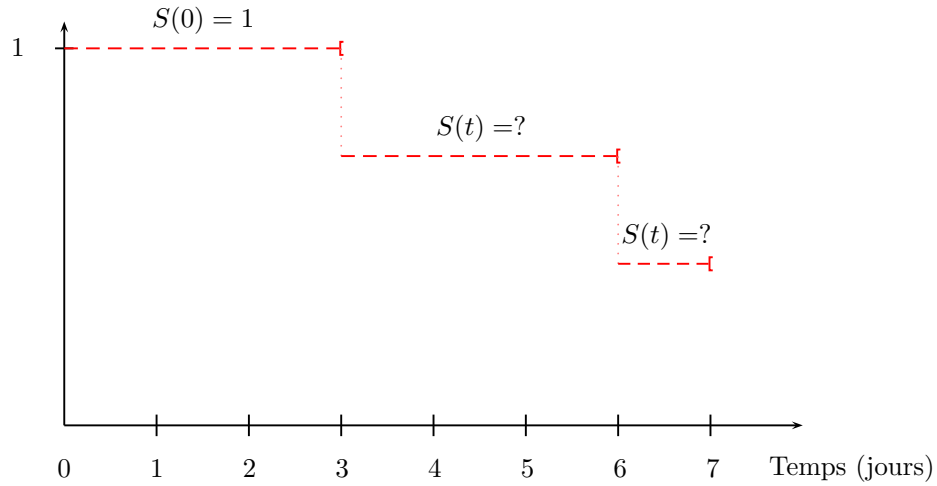
Cours uniquement basé sur les méthodes non-paramétriques (1).

Données utilisées pour l'analyse : (T_i, δ_i).

Exemple :

Sujet	Temps de survie (T_i) en jours	Etat (δ_i)
1	3	DC (1)
2	4	V (0)
3	6	DC (1)
4	6	DC (1)
5	7	DC (1)

- On cherche à estimer $S(t)$
- Possible uniquement à chaque temps de décès observé (variation de la probabilité)
- On suppose que $S(t)$ est constante entre chaque temps de décès.



Méthode de Kaplan-Meier

Idée : Etre encore à vie à l'instant t c'est être encore en vie juste avant t et ne pas mourir à l'instant t.

Considérons $t_1, t_2, \ldots, t_k$ les temps de décès observés ordonnés de manière croissante.

$$S(t_j) = \mathbb{P}(\{T \geq t_j\} \cap \{T \geq t_{j-1}\}) = \mathbb{P}(T \geq t_j / T \geq t_{j-1})\mathbb{P}(T \geq t_{j-1})$$

Par récurrence,

$$S(t_j) = \mathbb{P}(T \geq t_j / T \geq t_{j-1})\mathbb{P}(T \geq t_{j-1} / T \geq t_{j-2}) \ldots \mathbb{P}(T \geq t_1 / T \geq t_0) \underbrace{\mathbb{P}(T \geq t_0)}_{=S(0)=1}$$

Posons

$$Q_j = \mathbb{P}(T \geq t_j / T \geq t_{j-1})$$

Aussi

$$S(t_j) = \prod_{i:t_i \leq t_j} \mathbb{P}(T \geq t_i / T \geq t_{i-1}) = \prod_{i:t_i \leq t_j} Q_i$$

Méthode de Kaplan-Meier

Objectif : pour estimer $S(t_j)$ on va estimer, à partir des données, les Q_i, $i \leq j$.

Considérons t_j et t_{j-1}. Posons

- n_j : nombre de sujets exposés en t_j (encore vivants avant t_j)
- m_j : nombre de sujets décédés en t_j

Donc en t_j il reste encore $n_j - m_j$ personnes exposées encore vivantes.

Une estimation de Q_j est donnée par

$$q_j = \frac{n_j - m_j}{n_j}$$

Aussi, une estimation de $S(t_j)$ est donnée par

$$\widehat{S(t_j)} = \prod_{i:t_i \leq t_j} \frac{n_i - m_i}{n_i}$$

Méthode de Kaplan-Meier

Les données censurées sont prises en compte dans les n_i

Posons c_{i-1} le nombre de données censurées entre t_{i-1} et t_i.

$$n_i = n_{i-1} - m_{i-1} - c_{i-1}$$

Remarque 1 : En l'absence de données censurées, $\widehat{S(t)}$ correspond à la proportion de sujet encore en vie à t.

Remarque 2 : $\widehat{S(t)}$ n'est qu'une estimation ponctuelle. Un intervalle de confiance asymétrique a été fourni par Rothman.

Méthode de Kaplan-Meier

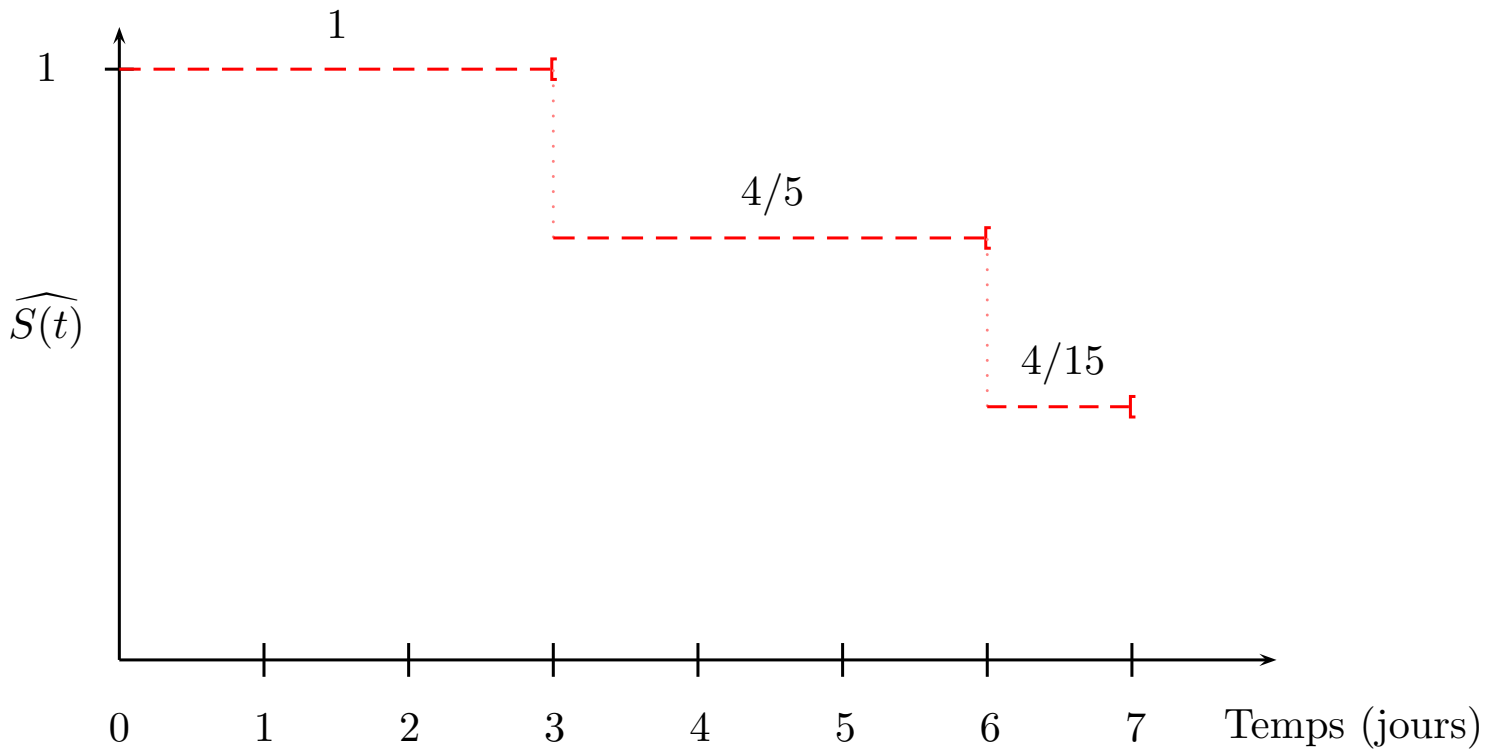
Retour à l'exemple

Sujet	Temps de survie (T_i) en jours	Etat (δ_i)
1	3	DC (1)
2	4	V (0)
3	6	DC (1)
4	6	DC (1)
5	7	DC (1)

$$
\begin{aligned}
q_1 &= q_2 = 5/5 = 1 \\
q_3 &= \mathbb{P}(T \geq t_3 / T \geq t_2) = (5-1)/5 = 4/5 \\
q_4 &= \mathbb{P}(T \geq t_4 / T \geq t_3) = 4/4 = 1 \\
q_5 &= \mathbb{P}(T \geq t_5 / T \geq t_4) = (4-1)/(4-1) = 3/3 = 1 \\
q_6 &= \mathbb{P}(T \geq t_6 / T \geq t_5) = (3-2)/3 = 1/3 \\
q_7 &= \mathbb{P}(T \geq t_7 / T \geq t_6) = (1-1)/1 = 0
\end{aligned}
$$

$$
\begin{aligned}
\widehat{S(1)} &= \widehat{S(2)} = 1 \\
\widehat{S(3)} &= 4/5 \\
\widehat{S(4)} &= 4/5 \\
\widehat{S(5)} &= 4/5 \\
\widehat{S(6)} &= 4/5 \times 1/3 = 4/15 \\
\widehat{S(7)} &= 4/5 \times 1/3 \times 0 = 0
\end{aligned}
$$

Représentation graphique de $\widehat{S(t)}$



On peut calculer la médiane de survie en utilisant la fonction quantile de la durée de survie ($p = 0.5$)

$$Q(0.5) = \inf\{t : S(t) \leq 0.5\} = 6$$

Méthode actuarielle

Utile si beaucoup d'évènements observés (bcp de t_j) car le graphique de Kaplan-Meier devient illisible.

Le principe de cette méthode est proche de celui de KM.

Différence :

- On ne dispose pas des dates précises de décès
- L'échelle de temps est découpée en intervalles de temps égaux fixés *a priori*
- Exemple : Evaluation tous les mois, semestres, années,...

Les probabilités conditionnelles Q_j sont estimées pour chaque intervalle de temps.

$\rightarrow$ Méthode moins courante en médecine $\leftarrow$

Objectifs

Comparaison de la durée de survie entre deux groupes.

Exemples :

- Comparaison de l'efficacité de 2 traitements (Essai thérapeutique)
- Comparaison de la durée de survie en fonction du sexe, âge, ...
- ...

En l'absence de données censurées, l'analyste pourrait utiliser :

- Un test de Kolmogorov - Smirnov
- Un test de Mann-Withney-Wilcoxon

La présence de données censurées nécessite l'utilisation d'autres tests.

$\rightarrow$ Test du Log-Rank (version approchée)

Test du Log-Rank - Hypothèses

Considérons deux groupes G_A et G_B.

Posons :

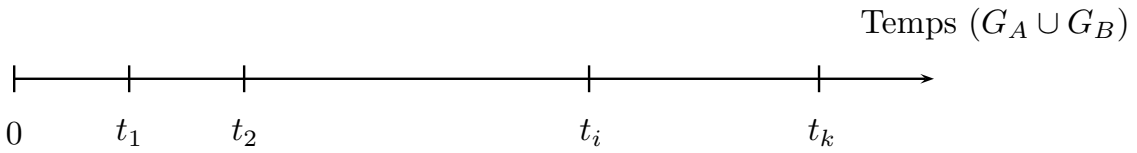
- $S_A(t)$ la fonction de survie du groupe G_A
- $S_B(t)$ la fonction de survie du groupe G_B

Les hypothèses de test sont les suivantes :

$$\begin{cases} \mathcal{H}_0 : S_A(t) = S_B(t) & \text{La survie est identique entre les groupes} \\ \mathcal{H}_1 : S_A(t) \neq S_B(t) & \text{La survie est différente entre les groupes} \end{cases}$$

Test du Log-Rank - Principe et statistique de test

Considérons $t_1, t_2, \ldots, t_k$ les temps de décès observés dans $G_A \cup G_B$.



Principe : à chaque temps de décès observés t_i, construction d'un tableau de contingence GROUPE $\times$ ETAT.

	Décédé	Vivant	
G_A	m_{Ai}	$n_{Ai} - m_{Ai}$	n_{Ai}
G_B	m_{Bi}	$n_{Bi} - m_{Bi}$	n_{Bi}
	m_i	$n_i - m_i$	n_i

- m_{Ai} : Nombre de décès observés dans G_A en t_i
- m_{Bi} : Nombre de décès observés dans G_B en t_i
- m_i : Nombre de décès observés en t_i

- n_{Ai} : Nombre de sujets exposés dans G_A en t_i
- n_{Bi} : Nombre de sujets exposés dans G_B en t_i
- n_i : Nombre de sujets exposés en t_i

Test du Log-Rank - Principe et statistique de test

A t_i, sous $\mathcal{H}_0$, le pourcentage de décès est identique dans les deux groupes
$\rightarrow$ indépendance entre Groupe et Etat (indépendance les lignes et des colonnes)
Mais impossibilité de faire un test statistique à chaque t_i car dans l'idéal

$$m_i = 1$$

si on échantillonne finement. Posons

- e_{Ai} : le nombre décès attendus en t_i dans le groupe G_A sous $\mathcal{H}_0$

$$e_{Ai} = \frac{m_i n_{Ai}}{n_i}$$

- e_{Bi} : le nombre décès attendus en t_i dans le groupe G_B sous $\mathcal{H}_0$

$$e_{Bi} = \frac{m_i n_{Bi}}{n_i}$$

Test du Log-Rank - Principe et statistique de test

Posons

- $E_A = \sum_{i=1}^{k} e_{Ai}$: le nombre total de décès attendus dans G_A sous $\mathcal{H}_0$
- $E_B = \sum_{i=1}^{k} e_{Bi}$: le nombre total de décès attendus dans G_B sous $\mathcal{H}_0$

- $O_A = \sum_{i=1}^{k} m_{Ai}$: le nombre total de décès observés dans G_A
- $O_B = \sum_{i=1}^{k} m_{Bi}$: le nombre total de décès observés dans G_B

Sous $\mathcal{H}_0$, on montre que

$$\chi^2 = \frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B} \sim \chi^2_{1\ ddl}$$

Remarque : il existe une extension du log-rank à K groupes.

Bibliographie

Livres

1. *Analyse statistique des données de survie*. Hill, C. ; Com-Nougue, C. ; Kramar, A. ; Moreau, T. ; O'Quigley, J. ; Senoussi, R. ; Chastang, C. . Collection : Statistique en biologie et en médecine. Flammarion Sciences 1996 ; 2ème édition 2009.
2. *Survival analysis*. Kleinbaum, D. G. et Klein, M. (2005). Statistics for biology and Health. Springer.

Supports en ligne

3. *Introduction à l'analyse des durées de survie*. P. Saint-Pierre.

   `http://www.lsta.upmc.fr/psp/Cours_Survie_1.pdf`

4. *Durées de survie*. M-L.Taupin.

   `http://stat.genopole.cnrs.fr/_media/members/mtaupin/CoursSurvieENSIIE.pdf`