Modèles de prédiction : Analyse de survie

I. INTRODUCTION :

Les méthodes d'analyse de survie sont des méthodes de références pour décrire les **données longitudinales** recueillies lors d'un **suivi** de sujets ou de groupes de sujets.

Une étude de survie est une étude :
- **Longitudinale**
- **Prospective**
- Observation d'un groupe de sujets = **cohorte**

On rencontre un très grand nombre de situations pratiques dans lesquelles le centre d'intérêt est la **survenue d'un évènement** (un décès, survenue d'une complication, la rechute d'une maladie, la disparition de symptômes). L'évènement considéré doit être **défini de la même manière** pour **tous les sujets.** Peu importe l'évènement, on utilisera le terme **« survie »**.

On s'intéresse à la survenue, dans le temps d'un évènement, c'est-à-dire au **délai** de survenue de cet évènement, délai compté à partir de **l'instant de référence** (ou **date d'origine**).

Les objectifs d'une analyse de survie sont **d'estimer** et **d'expliquer** la **durée de survie** en fonction de **facteurs pronostiques**, et, de **comparer la survie** entre 2 groupes de sujets ou plus.

+++ Les facteurs **pronostiques** sont à différencier des facteurs de **risques** +++

Définition : Un **facteur pronostique** est un facteur susceptible d'expliquer la survenue ou la non survenue du décès (ou d'un autre évènement) au cours du temps. Ils **influencent** de manière positive ou négative la survie.

Récap : on s'intéresse à :
- ⇨ La **probabilité de survivre au moins un certain temps t** à compter d'un instant de référence
- ⇨ La **probabilté que l'évènement d'intérêt survienne après un délai t** à compter de l'instant de référence

II. DÉFINITIONS :

Une cohorte : ensemble de sujets qui vivent les **mêmes évènements** au **même moment**, inclus dans une étude au même moment et suivis dans des **conditions standardisées** pendant une **durée prédéfinie**.

Une cohorte « incipiente » : la cohorte des patients qui rentrent dans l'étude doit inclure des **sujets observés au début de leur affection** à un point **uniforme** de l'évolution de leur maladie. Les sujets sont des « cas incidents ».

Une cohorte idéale : tous les patients sont inclus au même moment, tous les patients sont « allignés ».

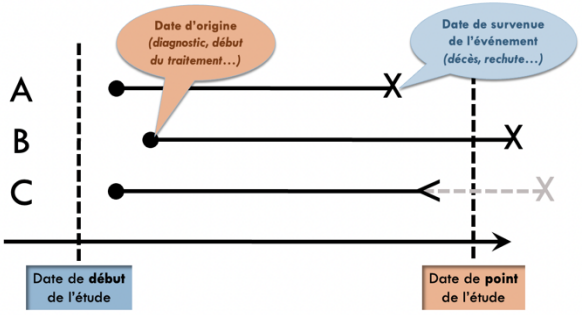
Évènement d'intérêt : pas forcément le décès, mais peut être la survenue d'une maladie, la récurrence de symptômes…

En pratique, les méthodes d'analyse de « survie » doivent donc être appliquées à chaque fois qu'il existe une notion de durée jusqu'à l'évènement d'intérêt.

Lorsque l'évènement d'intérêt est le décès :

- On peut s'intéresser au décès de toutes causes, et dans ce cas, chaque décès de patient compte comme un évènement
- On peut également ne s'intéresser qu'au décès pour une cause spécifique et, danc ce cas, les décès d'autres causes comptent comme une **censure**. Ceci n'est possible que lorsque les « autres causes de décès » sont indépendantes du **phénomène étudié**.

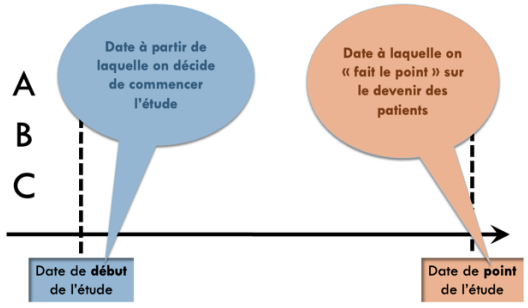
Durée de survie : délai entre 2 dates:

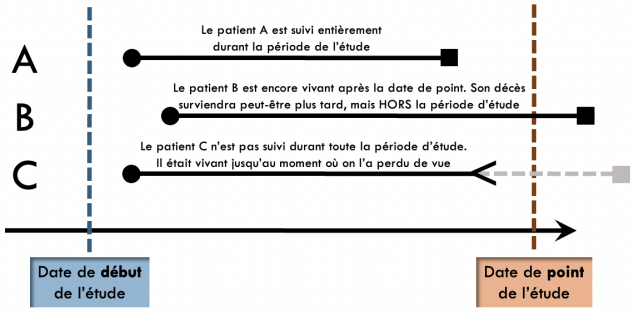


1. Date d'origine : date indiquant le point de départ de la surveillance. Elle peut être **identique** ou **différente** pour chaque sujet en fonction des modalités d'inclusion des sujets.
Ex : date de randomisation dans un essai.

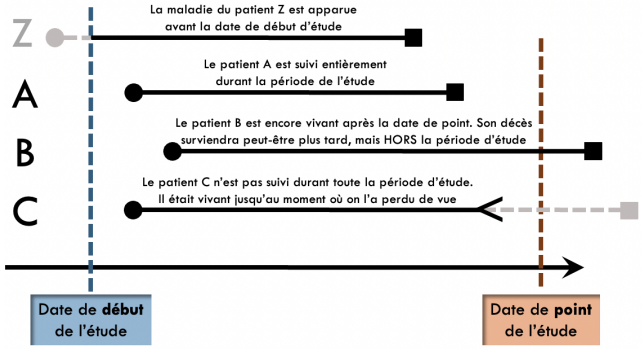
2. Date de point : date **fixe** calendaire et correspond à la **date choisie** pour faire le bilan et terminer l'étude.

2'. Date de dernières nouvelles : date la **plus récente** où on a **recueilli des infos** sur le patient (survenue ou non de l'évènement étudié).



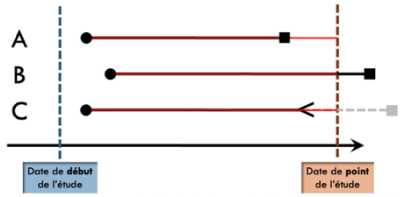


Cas particulier : dans certains cas, la date d'origine peut être **antérieure** à l'inclusion dans l'étude, on parle alors de **cohorte « historique »** *(patient Z).*



Cas particulier : un sujet est **perdu de vue** quand sa surveillance est interrompue avant la date de point et que l'évènement ne s'est pas produit. Ils sont considérés **exclus-vivants** (la raison de leur « disparition » peut être liée à l'évolution de la maladie).

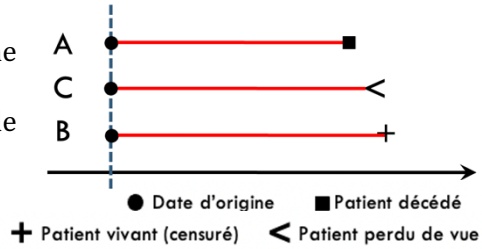
Censure : une durée de survie d'un individu est dite censurée lorsque **l'évènement d'intérêt n'a pas été observé** pour cet individu. Elle concerne donc les sujets perdus de vue et les sujets vivants à la date de point, 2 mécanismes de nature différente.



Temps de recul : délai entre la **date d'origine** et la **date de point**, c'est-à-dire le délai maximum potentiel de suivi pour un sujet. Les reculs minimum et maximum d'une série de sujets définissent donc l'ancienneté de cette série.

Temps de participation : durée de surveillance pour chaque sujet utilisée dans l'estimation de la survie. 3 situations peuvent se produire :

- L'évènement s'est produit au cours de la surveillance : date d'origine → survenue de l'évènement *(patient A)*
- Le sujet est vivant à la date de point : date d'origine → date de point *(patient B)*
- Le sujet est perdu de vue : date d'origine → date de dernières nouvelles *(patient C)*



III. FONCTION DE SURVIE :

A. Loi exponentielle :

La Loi de Poisson *(cf. cours variables)* de paramètre $\lambda = \mu = \sigma^2$, régit la survenue d'un évènement par unité de mesure, ici elle régit la survenue de la mort en fonction du temps. L'évènement considéré est distribué selon une **loi exponentielle d'espérance 1/λ**. Utiliser la loi exponentielle pour décrire des durées de vie implique que les défaillances sont dues uniquement au hasard.

Fonction de **densité** de la loi exponentielle : pour tout x ≥ 0, $\mathbf{f(x) = \lambda e^{-\lambda x}}$

La fonction de **répartition** de la loi exponentielle, appellée fonction de défaillance est donnée par l'équation :

$$F(t) = P(X \leq t) = \int_0^t \lambda e^{-\lambda x} dx = 1 - e^{-\lambda t}$$

Ex : on peut prendre comme évènement le fait que dans un groupe d'appareils électroménagers, certains tombent en panne ; cette définition définit en fonction du temps le nombre d'appareils étant tombés en panne.

B. Fonction de survie :

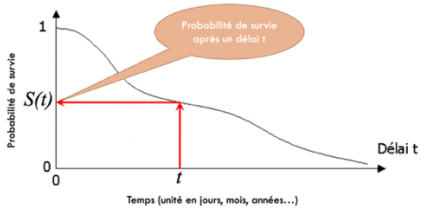
En épidémiologie clinique, la durée résiduelle de vie d'un patient, à compter de l'instant de référence (date d'origine), est une caractéristique variable d'un patient à l'autre ; c'est donc une variable aléatoire, que nous noterons T.
La probabilité pour que le décès (« la défaillance ») intervienne après un délai supérieur à t est donc la probabilité pour que T soit supérieur à t :

- **S(t) = P(T > t) = 1 – F(t)**
- Où F est la fonction de répartition de la durée de vie résiduelle (proportion de patients décédés au temps t)

En épidémiologie clinique, la fonction de survie est notée **S(t)**. Elle représente :

- La probabilité pour qu'un patient soit encore vivant après un délai t
- Ou encore la proportion « vraie » des survivants après un délai t

F(t) et S(t) sont compris entre **0 et 1**, donc S(t) ne représente pas un nombre de surivants mais une **proportion** ou **probabilité**. La fonction de survie est représentée graphiquement par une **courbe de survie**.



La fonction de survie permet de calculer la probabilité pour que le décès survienne après un délai t_1 et avant un délai t_2 ($t_2 > t_1$) :

$$Pr(T \in]t_1; t_2]) = F(t_2) - F(t_1) = S(t_1) - S(t_2)$$

La fonction de survie permet aussi de calculer la probabilité de survivre encore après un délai t sachant que l'on est survivant après un délai τ ($\tau < t$), que l'on notera $S(t|\tau)$:

$$S(t/_\tau) = \frac{Pr((X > t) \cap (X > \tau))}{Pr(X > \tau)} = \frac{Pr((X > \tau + s) \cap (X > \tau))}{Pr(X > \tau)} = \frac{Pr(X > \tau + s)}{Pr(X > \tau)} = \frac{S(\tau + s)}{S(\tau)}$$

Au final :

$$S(t/_\tau) = \frac{S(t)}{S(\tau)}$$

IV. ESTIMATION DE LA SURVIE :

A. Recueil des données :

Date d'origine : date à laquelle a **débuté l'observation.** *Ex : date de diagnostic d'un cancer.*

Date des dernières nouvelles : date de décès ou **date de dernières données** relatives à l'état du patient sachant qu'il n'est pas décédé.

Date de point : date à laquelle on fait le point ou **date de fin d'observation**.

Un évènement « en tout ou rien » (binaire) : correspond à l'état du patient en 2 éventualités (vivant ou décédé à la date de dernières nouvelles). Tout évènement binaire autre que le décès associé à un délai de survenue peut être analysé en **délai de survie**.

B. Calcul des durées de suivi :

Les durées de suivi (ou temps de participation) correspondent au délai entre la **date d'origine** et la **date de dernières nouvelles** (date de décès, date de point pour les patients vivants pour lequel le suivi est assuré ou **la date de perte de vue** pour les patients vivants n'étant plus suivis dans la cohorte à la date de point).

C. Calcul de la survie :

Si aucune variable n'est censurée, la fonction de survie se calcule par le **pourcentage de survivants** en fonction du temps, et on peut directement tracer la courbe.

En pratique, cela ne se produit jamais. Deux méthodes d'analyse de survie sont de préférence utilisées : **l'analyse actuarielle** et la **méthode Kaplan-Meïer**, qui sont 2 méthodes **non paramétriques**, car elles ne nécessitent aucune hypothèse sur la distribution des temps de survie.

L'analyse actuarielle est moins utilisée que la méthode de Kaplan-Meïer, et s'applique principalement lorsqu'il y a un **grand nombre de sujets** (plus de 200 par groupe) / évènements. La méthode de **Kaplan-Meïer** est donc la méthode de choix pour les échantillons de taille réduite.

Ces 2 méthodes supposent une **hypothèse forte** : les probabilités de survie sont supposées indépendantes du calendrier. Elles partent du principe qu'il n'y a **pas de progrès thérapeutique** le long de l'étude.

La fonction de survie estimée peut être résumée soit par le **taux de survie à un délai fixé** (1 an, 5 ans, etc..), soit par une **valeur de durée** : médiane de survie et quantiles.

D. Analyse actuarielle :

Pour chaque intervalle de temps, on définit :

- Le nombre de sujets vivants au début de l'intervalle : **V**
- Le nombre de sujets décés dans l'intervalle : **D**
- Le nombre de sujets vivants aux dernières nouvelles : **C**
- Le nombre de sujets exposés au risque de décès : **N = V – (C/2)**

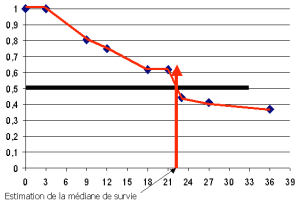
Instants	V	C	D	N = V – C/2	(N - D) / N	S(t)
0	-	-	-	-	-	1
3	210	0	0	210	1	1x1 = 1
6	210	10	40	210 - 5 = 205	(205-40)/205 = **0,805**	0,805 x 1 = **0,805**
9	160	30	10	160 - 15 = 145	(145-10)/145 = **0,931**	0,931 x 0,805 = **0,749**
12	120	10	20	120 – 5 = 115	(115-20)/115 = **0,826**	0,826 x 0,749 = **0,619**
15	90	20	0	90 – 10 = 80	1	1 x 0,619 = **0,619**
18	70	0	20	70	(70-20)/70 = **0,714**	0,714 x 0,619 = **0,442**
21	50	18	3	50 – 9 = 41	(41-3)/41 = **0,927**	0,927 x 0,442 = **0,410**
24	29	8	2	29 – 4 = 25	(25-2)/25 = **0,920**	0,920 x 0,410 = **0,377**

Les tableaux à remplir comme ça ça peut tomber !

La **probabilité d'évènements** durant l'intervalle est simplement estimée par le rapport du nombre d'évènements sur le nombre de sujets à risque : **D/N**

La survie sur cet intervalle est : **(N – D) /N**, cette probabilité est appellée **survie instantanée**.

La fonction de survie est obtenue en faisant le produit des survies instantanées sur l'ensemble des intervalles. On suppose que les sujets soient exposés au risque d'évènement sur la **moitié de l'intervalle**.



Pour chaque intervalle de temps, on représente l'estimation de la survie S(t) par un point. Tous les points consécutifs sont reliés par un segment de droite.

L'inconvénient de cette méthode est qu'elle estime la survie à chaque **borne supérieure** des intervalles constitués a priori (mois, semaines…), et considère **chaque censure**, survenant dans un intervalle, de **manière équivalente**. *Ex : un sujet suivi pendant 21 jours apporte la même information qu'un sujet suivi pendant 29 jours pour la survie à 30 jours.*
C'est la raison pour laquelle cette méthode est **à réserver à de grands échantillons.**

E. Méthode kaplan-Meïer :

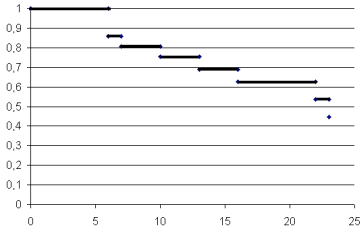
Contrairement à l'analyse actuarielle, les intervalles ne sont **pas fixés a priori**, mais sont définis par les **instants** auxquels les évènements sont observés. *Ex : on change d'intervalle à chaque décès.*

Pour chaque intervalle, on définit **V**, **D** et **C** (avec la particularité que D vaut souvent 1, sauf dans le cas où plusieurs évènements surviennent au même temps de participation).

Instants	V	C	D	N = V - C	(N - D) / N	S(t)
0	21	-	-	-	-	1
6	21	0	3	21	**0,857**	**0,857**
7	18	1	1	17	**0,941**	**0,807**
10	16	1	1	15	**0,933**	**0,753**
13	14	2	1	12	**0,917**	**0,690**
16	11	0	1	11	**0,909**	**0,627**
22	10	3	1	7	**0,857**	**0,537**
23	6	0	1	6	**0,833**	**0,448**

Dans l'analyse Kaplan-Meïer, **N = V – C** et la probabilité de survie instantannée calculée sur cette intervalle vaut : **(N – D) / N**
L'estimation de Kaplan-Meïer de la fonction de survie s'obtient, comme dans l'analyse actuarielle, en faisant le **produit des survies instantanées**.

La courbe de survie se compose de **paliers successifs**, où les probabilités de survie sont constantes entre deux temps consécutifs. Le premier palier **vaut 1** depuis l'origine jusqu'au délai de survenue du premier évènement. Il s'abaisse ensuite à la première valeur calculée pour constituer un second palier jusqu'au délai de survenue de l'évènement suivant, etc. La courbe ainsi obtenue présente une allure en « **marches d'escalier** ».



V. COMPARAISON DE 2 FONCTIONS DE SURVIE :

A. Contexte :

Il arrive que l'on souhaite montrer qu'une action (intervention, traitement) ou une classification ont un lien avec la survie.

Le principe du **test log-rank** (ou test Mantel-Cox ou de Peto-Mantel-Haenszel) est de comparer, dans chaque groupe, le nombre observé et le nombre attendu d'évènements si la survie était identique dans les 2 groupes, sur l'ensemble de la période étudiée.

B. Principe du test du log-rank :

Pour chaque **intervalle de temps** (qu'il s'agisse de l'analyse actuarielle ou de Kaplan-Meïer), le nombre attendu d'évènements, sous l'hypothèse nulle d'égalité de la survie entre deux groupes, s'obtient en appliquant, au nombre de sujets exposés au risque d'évènements, la proportion d'évènements observés sur l'ensemble des deux groupes.

On peut alors faire une **étude comparative entre 2 groupes**, où pour chacun on a calculé la fonction de survie et tracé la courbe. Dans ces 2 groupes, on varie l'exposition au facteur pronostic à étudier. On observe une différence entre les 2 groupes et on cherche si elle est due au **hasard**. On ne peut pas faire un Khi 2 à chaque fois, on va utiliser une méthode globale = le **log rank**.

Le test du log-rank, évaluant l'écart entre le **nombre observé** et le **nombre attendu d'évènements** sur les deux groupes, est un **Khi 2 à 1 degré de liberté** (ddl).

Ce test est généralisable au cas de k groupes et permet de tester si globalement la survie est différente entre les groupes :

- H0 : les fonctions de survie sont les mêmes dans les 2 populations
- H1 : les 2 fonctions de survie diffèrent

⚠ Attention aux biais ⚠

⇨ Il ne faut pas comparer la survie entre les sujets qui répondent au traitement (TTT) et ceux qui n'y répondent pas, mais entre les sujets traités et les non traités

⇨ Il ne faut pas assimiler l'efficacité du TTT à la réponse des patients à ce TTT

C. Estimation des décès :

Le principe est d'abord d'estimer, tous groupes confondus, **la probabilité de décéder à t_i sachant que l'on est vivant à t_{i-1}**, c'est-à-dire estimer la **probabilité de décès $(1- S(t_i / t_{i-1}))$** et ceci pour chacun des temps de décès observés t_i.

On utilise ici l'estimateur de **Kaplan-Meïer** de **$S(t_i / t_{i-1})$** : on obtient la dernière colonne du tableau :

t_i	V	C	N = V - C	D	$S(t_i / t_{i-1}) = (N - D) / N$	$1 - S(t_i / t_{i-1})$
1	42		42	2	0,952	0,048
2	40		40	2	0,950	0,050
3	38		38	1	0,974	0,026

Ex : on compare deux groupes de 21 patients chacun ; on calcule cette probabilité pour les 42 patients.

D. Calcul de décès attendus :

On estime ensuite le **nombre de décès** que l'on attend dans chacun des groupes, à chaque t_i, en supposant que la **probabilité conditionnelle** de décès estimée s'applique identiquement à chacun des deux groupes.

t_i	V	C	N = V - C	D	$S(t_i / t_{i-1}) = (N - D) / N$	$1 - S(t_i / t_{i-1})$	N_A	N_B	E_A	E_B
1	42		42	2	0,952	0,048	21	21	1,000	1,000
2	40		40	2	0,950	0,050	19	21	0,950	1,050
3	38		38	1	0,974	0,026	17	21	0,447	0,553
4	37		37	2	0,946	0,054	16	21	0,864	1,136

Pour cela on évalue à chaque t_i **l'effectif à risque** à cette date. On obtient les deux dernières colonnes du tableau suivant : Ces nombres sont notés E_A et E_B.

Sous l'hypothèse H0 ces nombres doivent être voisins des nombres de décès réellement **observés**. En particulier le total de ces **nombres de décès au cours du temps** (noté E_A et E_B) doit être voisin du **nombre total de décès observés** (noté D_A et D_B), et ceci dans chacun des groupes.

- NA : Effectif du groupe A au temps t, avant le décès
- NB : Effectif du groupe B au temps t, avant le décès
- N : Effectif global au temps t, avat le décès (-les censurés) : **N = NA + NB**
- DA : Nombre de décès observés dans le groupe A au temps t
- DB : Nombre de décès observés dans le groupe B au temps t
- D : Nombre de décès observés global au temps t : **D = DA + DB**
- EA : Nombre de décès attendus dans le groupe A au temps t : **$EA = D.\frac{NA}{N}$**
- EB : Nombre de décès attendus dans le groupe B au temps t : **$EB = D.\frac{NB}{N}$**

EA et EB impliquent que les fonctions de survie soient les mêmes dans les 2 groupes (H0 accepté = pas de différence entre les groupes A et B) : **EA + EB = D**

E. Test du Chi² :

Le paramètre du test est construit à partir de ces 4 paramètres :

$$Q_c = \frac{(D_A - E_A)^2}{E_A} + \frac{(D_B - E_B)^2}{E_B}$$

Condition de validité du calcul du test : EA et EB > 5, et 1ddl

On compare donc Qc, le paramètre calculé au paramètre théorique observé dans la table du X^2 à 1ddl, par défaut au **risque 5%.**

Voilaaaaaa ce cours est terminé !!! ❤️ *Je sais qu'il est difficile, mais pas de souci vous allez gérer les boss !* 😎 *Pour touuuutes vos questions direction le fofo ! Je vais aussi vous mettre une fiche récap pour les méthodes actuarielle et Kaplan-Meïer histoire de jamais les confondre !* 😜 *Faites plein plein plein de QRU, vous verrez la biostat c'est easyyy !!* 🔥 *Bon courage, lachez rien !!* 💪

Et biensur ptite dédi à mes vieux, mes cotuts et mes fillots, que du LOVEEE ! 🥰