

Analyse de la survie

Qu'est ce qu'une analyse de survie et dans quel contexte l'utiliser ?

Par **survie** on définit un certain **laps de temps** avant qu'un **évènement** n'arrive à partir d'un **instant de référence**. Ça peut être la mort ou n'importe quel évènement (une rechute, une complication, la disparition des symptômes sous TTT...). Il faut que **l'évènement considéré soit défini de la même manière pour tous les sujets**. Peu importe l'évènement on utilisera le terme « survie ».

On s'intéresse donc à :

- La probabilité de survivre au moins un certain délai t à compter d'un instant de référence.
- La probabilité que l'évènement d'intérêt survienne après un délai t à compter de l'instant de référence.

Dans ce cours, on va s'intéresser à la manière de **mesurer ce laps de temps** avant que l'évènement n'arrive, parler en **probabilités**, **comparer** la survie de plusieurs groupes et trouver les **facteurs pronostiques**, c'est-à-dire les facteurs qui **influencent la durée de la survie**.

+++ Les facteurs pronostics sont bien différents des facteurs de risques ++
(il insiste sur les fc pronostics et la différence avec les fc de risque)

Il existe différents moyens de faire des études sur un groupe de patients :

- Etude **transversale** : à un instant T
- Etude **longitudinale** : implique la notion du temps que ce soit en revenant en arrière, dans le passé, avant l'instant T ou en suivant un groupe de personnes dans le futur à partir d'un instant T
- Etude **rétrospective** : revient dans le passé
- Etude **prospective** : va dans le futur

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

L'étude de survie est une étude longitudinale, prospective et étudiant une cohorte.

Risque : c'est une **probabilité**

Facteurs pronostics : (il insiste) élément qui modifie la probabilité, c'est un facteur **susceptible d'expliquer la survenue ou la non survenue de l'évènement au cours du temps**. Ce sont des facteurs qui influencent de manière positive ou négative la survie.

C'est différent des facteurs de risque (il insiste)

On va étudier la rapidité de survenue d'un évènement après une exposition à un facteur pronostic.

I) Théorie : définitions

Cohorte :

Une cohorte est un ensemble de sujets vivant les **mêmes évènements au même moment**, qui sont donc inclus dans une étude au même moment, suivis dans les **mêmes conditions standardisées** pendant une **même durée prédéfinie**.

Une cohorte qui inclut des patients observés au **début** de leur **affection**, à un **point uniforme** de l'évolution de leur maladie est une cohorte **incipiente**. Les sujets sont les **cas incidents**.

Cohorte idéale : tous les patients sont inclus au même moment, tous les patients sont « **alignés** ».

Choix de l'évènement d'intérêt :

L'évènement d'intérêt n'est **pas forcément un décès**. Il peut être la survenue d'une maladie, la récurrence de symptômes...

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

Lorsque l'événement d'intérêt est le décès, on peut s'intéresser à :

- **N'importe quel type de décès** peu importe la cause, qu'elle soit reliée ou non à notre sujet d'étude.
- **Décès pour cause spécifique**, censure des autres. Il faut cependant s'assurer que les autres causes du décès sont indépendantes du phénomène étudié.

Date d'origine, de point, de dernières nouvelles :

- La date **d'origine** : date à partir de laquelle on compte le délai de survenue de l'évènement. C'est le **point de départ** de la surveillance, l'instant de référence.

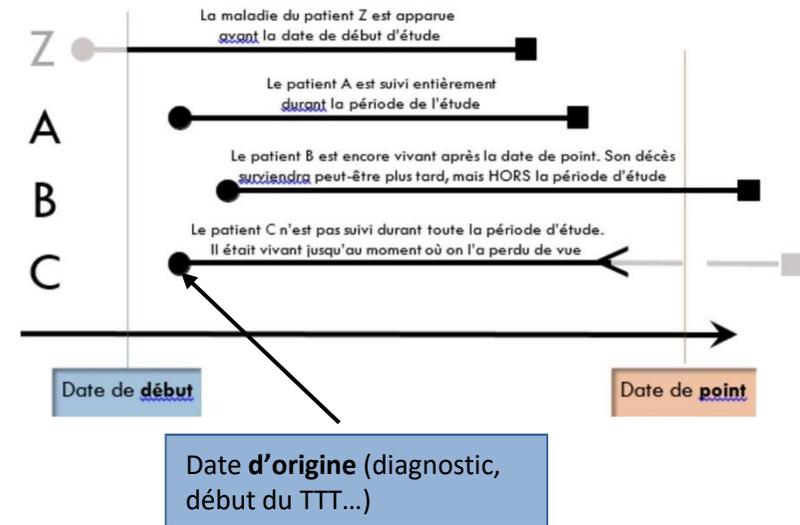
Ex : date de randomisation dans un essai thérapeutique.

ATTENTION : la randomisation, en pratique, ne se fait pas d'un seul coup où on répartit tous les patients en 2 groupes. On inclut les patients 1 par 1 lorsqu'ils se présentent au cabinet, et ce faisant, on randomise à chaque fois qu'on inclut un patient.

Elle peut être différente d'un sujet à l'autre suivant leurs modalités d'inclusion.

Parfois la **date d'origine** est **antérieure** à la date **d'inclusion** dans l'étude : on parle alors de **cohorte historique**.

- La date de **point** : celle que l'on choisit pour faire le **bilan** et terminer l'étude
- La date de **dernières nouvelles** : date la plus récente à laquelle on a recueilli des informations sur un patient, notamment sur la survenue ou non de l'évènement étudié



Perdus de vue et censurés :

Il existe 2 types de censures qu'il faut distinguer :

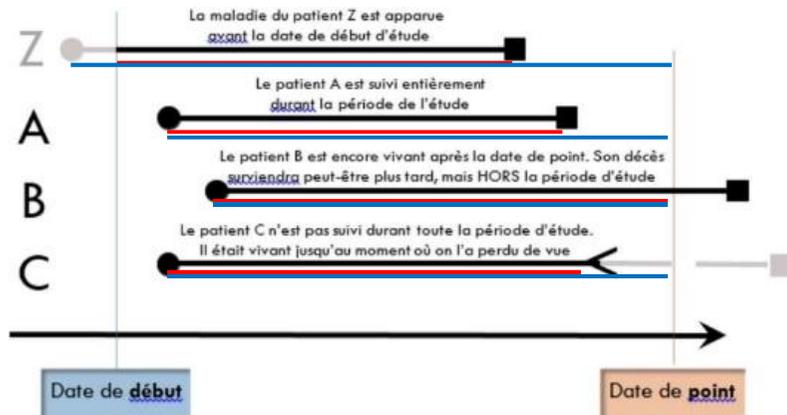
- L'évènement **ne se produit pas jusqu'à la date de dernières nouvelles**, donc aux dernières nouvelles il est encore vivant mais la surveillance du sujet est interrompue avant la date de point : c'est donc un **perdu de vue**.
- L'évènement **ne se produit pas** et on arrive à la **date de point** avec le patient **vivant**.

Attention : on parle de censure dès que l'évènement n'a pas eu lieu. Donc les perdus de vue et les vivants sont censurés, cependant on ne peut pas les assimiler. En effet, pour les vivants on est certains qu'ils sont encore en vie, mais pour les perdus de vue, on ne sait pas pourquoi ils ont été perdus de vue. Il est d'usage de vérifier que le processus de perte de vue n'est pas lié à l'évènement d'intérêt.

Temps de recul et de participation :

Le **temps de participation** se situe entre la date d'origine et :

- La date de **survenue** de l'évènement
- La date de **dernières nouvelles** pour un perdu de vue
- La date de **point** pour un sujet vivant à date de point



Le **temps de recul** est le temps entre la date d'origine et la date de **point**.

C'est le **délai maximum potentiel de suivi** pour un sujet.

Pour le patient Z on calcule bien son temps de recul à partir de sa date d'origine.

II) La fonction de survie

Rappels : la loi exponentielle :

La **loi de Poisson** régit la survenue d'un évènement en fonction d'une unité de mesure : ici, elle va régir la **survenue de la mort en fonction du temps**.

A une loi de Poisson (qui décrit la survenue d'un évènement) de paramètre λ correspond à une loi **exponentielle** (qui donne le temps entre 2 survenues consécutives de l'évènement) de paramètre $1/\lambda$.

La fonction de répartition de la loi exponentielle $F(t)$ est appelée fonction de défaillance.

Fonction de densité de la loi exponentielle :

$$\lambda e^{-\lambda x}$$

Fonction de répartition de la loi exponentielle :

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

Exemple : on peut aussi bien prendre comme évènement le fait que dans un groupe d'appareils électroménagers, certains tombent en panne. Cette définition définira, en fonction du temps, le nombre d'appareils étant tombés en panne.

Utiliser la loi exponentielle pour décrire la durée de vie implique que les défaillances soient uniquement dues au hasard, et qu'elles se produisent selon un processus de Poisson.

La fonction mathématique de survie :

En épidémiologie clinique, la durée résiduelle de la vie d'un patient, à compter de l'instant de référence (date d'origine), est une caractéristique variable d'un patient à un autre ; c'est donc une variable aléatoire que nous noterons T.

La probabilité que le décès intervienne après un délai supérieur à t est donc la probabilité que T soit supérieur à t :

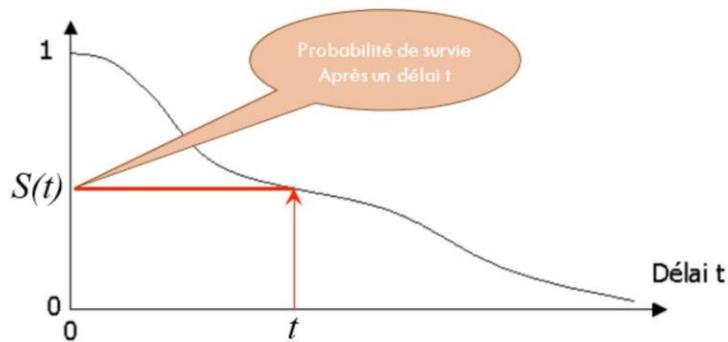
$$S(t) = P(T > t) = 1 - F(t)$$

Puisque $F(t)$ est la proportion de patients déjà décédés au temps t, alors la quantité **1-F(t)** est la **proportion** de patients qui **vivent** pendant une durée au moins égale à t. Cette nouvelle fonction $1-F(t)$ est appelée **S(t)** et est la fonction de **survie**. Il s'agit d'une fonction de répartition.

$$S(t) = 1 - F(t) = P(X > t) = e^{-\lambda t}$$

Attention : F(t) et S(t) sont compris **entre 0 et 1** : donc S(t) ne représente pas un nombre de survivants au temps t mais un **proportion** de survivants ou une **probabilité** d'être encore vivant à ce temps t.

Elle est représentée graphiquement par une **courbe de survie** :



La fonction de survie permet de calculer la **probabilité** pour que le décès survienne **entre un temps t₁ et un temps t₂** :

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

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

- On note S(t) la probabilité que le décès survienne après le temps t
- S(τ) la probabilité que le décès survienne après le temps τ
- s l'écart entre t et τ

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

Au final :

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

Calcul de la survie : analyse actuarielle et méthode de Kaplan-Meier :

Si **aucun patient n'est censuré**, c'est-à-dire que personne n'est perdu de vue et que personne n'est vivant à la date de point, alors la **fonction de survie** est tout simplement le **pourcentage de survivants en fonction du temps**, on peut directement tracer la courbe.

En pratique, ça ne se produit jamais. Dans ce cas on a 2 méthodes pour obtenir la fonction de survie : **l'analyse actuarielle** et la **méthode de Kaplan-Meier**. Elles ont 2 caractéristiques en commun :

- Elles sont **non paramétriques** : elles ne nécessitent donc aucune hypothèse de distribution des temps de survie.
- Elles supposent une hypothèse forte d'équiprobabilité : 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.

L'inconvénient de l'analyse actuarielle et la raison pour laquelle on ne l'utilise **que dans de grands échantillons**, c'est qu'on estime la survie à chaque **borne supérieure** des intervalles constitués a priori, et considère **chaque censure**, survenant dans un intervalle, de **manière équivalente** c'est-à-dire que si on veut calculer la survie au bout de 30 jours, un sujet censuré au bout de 21 jours aura autant de poids qu'un autre censuré au bout de 29 jours.

Calcul de la survie (« je n'ai jamais fait de QCM là-dessus » mais apprenez quand même, ça ne veut pas dire qu'il n'en fera pas cette année...)

Analyse actuarielle	Méthode Kaplan-Meier
Moins utilisée	Plus utilisée
S'applique principalement pour un grand nombre de sujets $n > 200$	Pour des échantillons de taille plus réduite
Suppose que les sujets soient exposés au risque d'évènement sur la moitié de l'intervalle	Les intervalles sont définis par les instants où les évènements sont observés (on change d'intervalle à chaque décès)
La fonction de survie est calculée sur des intervalles de temps fixés à priori	
Pour chaque intervalle de temps entre 2 dates on définit : <ul style="list-style-type: none"> Le nb V de sujets vivants au début de l'intervalle Le nb D de sujets décédés dans l'intervalle Le nb C de censurés car perdus de vue Le nb N de sujets exposés au risque d'évènement sur l'intervalle 	
$N = V - C/2$	$N = V - C$
Probabilité d'évènement durant l'intervalle :	
D / N	
Survie sur l'intervalle □ survie instantanée :	
$(N - D) / N$	

La **fonction de survie** s'obtient dans les deux cas en faisant le **produit des survies instantanées** sur l'ensemble des intervalles.

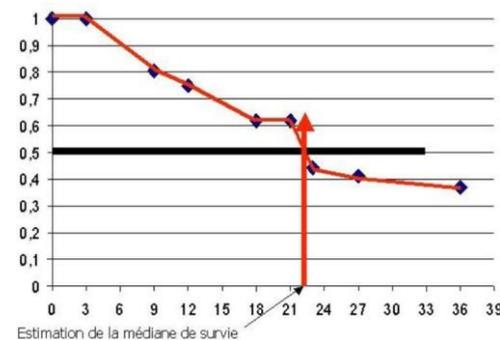
Autrement dit, à chaque intervalle, on **multiplie** la **survie instantanée** de cet intervalle par :

- Celle de **chaque** intervalle **précédent**
- Ou, par conséquent, par la **survie de l'intervalle précédent** seulement, puisque cette survie est le produit de toutes les survies instantanées précédentes

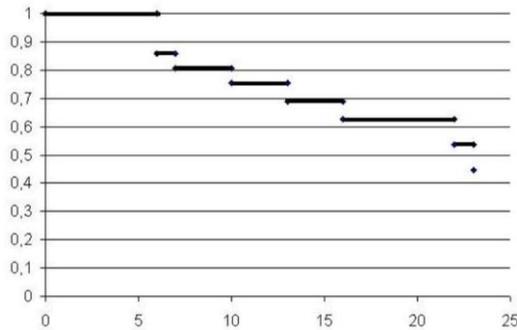
Instants	V	C	D	$N = V - C/2$	$(N - D) / N$	$S(t)$
0	-	-	-	-	-	1
3	210	0	0	210	1	$1 \times 1 = 1$
6	210	10	40	$210 - 5 = 205$	$(205-40)/205 = 0,805$	$0,805 \times 1 = 0,805$
9	160	30	10	$160 - 15 = 145$	$(145-10)/145 = 0,931$	$0,931 \times 0,805 = 0,749$
12	120	10	20	$120 - 5 = 115$	$(115-20)/115 = 0,826$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

V : nombre de sujets vivants au début de l'intervalle
 C : nombre de sujets vivants censurés dans l'intervalle
 D : nombre de sujets décédés dans l'intervalle
 N : nombre de sujets exposés au risque de décès

Courbes de survie obtenues :



Pour l'analyse **actuarielle**, à chaque fin d'intervalle, on met un point à l'ordonnée de la **survie calculée** puis on relie les points par des segments de droite.



Pour la méthode **Kaplan-Meier**, on a une **courbe en paliers** avec des **intervalles complètement inégaux** puisqu'ils sont délimités par chaque décès.

III) Le log-rank : un test comparatif

Ce qui est intéressant pour les chercheurs, c'est de voir si des caractéristiques (environnement, physiques, fonctionnelles, thérapeutiques) vont modifier la dynamique de survie de la population (retarder ou précipiter la survenue de l'événement) : ce sont les **facteurs pronostics ++**

On peut être amené à **calculer des taux de survie en fonction de ces modalités**.

Ex : on pourrait comparer les buveurs de lait et les non-buveurs de lait, avec comme événement un développement d'ostéoporose, ou encore de fracture ou accident osseux quel qu'il soit.

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 va se demander si cette différence est **due ou non au hasard**. On compare les **pourcentages** en fonction du temps (à 3, 6, 12 mois...). Mais on ne peut pas faire un Khi 2 à chaque fois : on va plutôt utiliser une **méthode globale**. Cette méthode globale s'appelle le **log-rank**.

C'est un test de comparaison de courbes de survie.

Il s'agit d'un test du **X²** avec une **ddl = 1**

H0 = les fonctions de survie sont les mêmes dans les 2 populations

d'où sont issus les groupes A et B, soit

SA(t)SB(t) H1 = les deux fonctions de survie différent

Principe du Log-Rank :

On a un certain nombre attendu d'événements (= de décès), et un nombre observé d'événements, qui se produisent en réalité. Le log-rank permet **d'évaluer l'écart entre le nombre d'événements attendus et observés**.

Attention aux biais :

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

Ex : on donne à certains patients de la vitamine C, en prévention de la fragilité osseuse. Les deux groupes seront « patients supplémentés en vitamine C » et « patients non supplémentés en vitamine C », mais même les patients sur lesquels le TTT ne marche pas devront rentrer dans le groupe des patients supplémentés.

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

Estimation du décès :

On estime la **probabilité de décéder à t sachant qu'on est vivant à τ** ; on le fait pour tous les patients sans distinction de groupe, supposant donc la justesse de l'hypothèse H0.

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

t_j	V	C	N = V - C	D	Survie instantanée $S(t_j / t_{j-1}) = (N - D) / N$	Probabilité de décès $1 - S(t_j / t_{j-1})$
1	42		42	2	0,952	0,048
2	40		40	2	0,950	0,050

Calcul des décès attendus :

On estime, de même, le **nombre de décès attendus** dans chaque groupe en **supposant que la probabilité de décès s’applique identiquement aux deux groupes.**

On obtient ainsi les dernières colonnes du tableau :

t_i	V	C	$N=V-C$	D	$S(t_i/t_i) = (N-D)/N$	$1-S(t_i/t_i)$	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

Selon l’hypothèse H0, ces nombres doivent être **voisins** des nombres de décès réellement **observés**, en particulier le **nombre de décès au cours du temps** (estimés EA et EB, observés DA et DB).

Estimation du nombre de décès théorique E, que l’on attend dans chacun des groupes A et B, à chaque temps t (proportionnel aux effectifs à risque de chaque groupe), puis observation du nombre réel de décès D dans chacun des groupes à chaque temps t.

NA : Effectifs du groupe A au temps t, avant que se produisent les décès

NB : Effectifs du groupe B au temps t, avant que se produisent les décès

N : Effectifs global au temps t, avant que se produisent les 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 (DA et DB n’apparaissent pas dans le tableau)

D : Nombre de décès global au temps t (groupes A et B confondus).

$$D = DA + DB$$

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).

t_i	V	C	$N=V-C$	D	$S(t_i/t_i) = (N-D)/N$	$1-S(t_i/t_i)$	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

$$EA + EB = D$$

Ex : Pour t_2 : NA = 19 et NB = 21
 EA = 2 x 1940 = 0,950EB = 2x 2140 = 1,050

Test du X^2 : (seuls les vrais ninjas maîtrisent le chi2)

On calcule le paramètre calculé à partir de DA, DB, EA et EB

$$Qc = \frac{(DA - EA)^2}{EA} = \frac{(DB - EB)^2}{EB}$$

Condition de validité du calcul et du test : EA et EB > 5.

On a 1 ddl.

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