MODÈLES DE PRÉDICTION
ANALYSE DE SURVIE

INTRODUCTION

Les méthodes d'analyse de survie sont des méthodes de référence 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
- D'observation d'un groupe de sujets : une **cohorte**

Point tut' : Les analyses de survie essaient de **modéliser la survenue d'un événement en fonction du temps**

On rencontre un très grand nombre de situations pratiques dans lesquelles **le centre d'intérêt est la survenue d'un évènement** (décès, survenue d'une complication après un geste opératoire, rechute d'une maladie après une période de rémission, disparition de symptômes sous traitement, ect.).

La méthodologie introduite dans ce cours s'appliquera sans modification à tout type d'événement à la survenue duquel on s'intéresse. Cependant, pour la commodité de l'expression, on parlera généralement dans la suite de survie, considérant ainsi que l'événement d'intérêt est le décès.

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).

En cas d'un décès pris comme évènement d'intérêt, dire d'un patient qu'il survit au moins un certain temps c'est dire que le délai de survenue du décès est supérieur à ce temps.

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

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

Point tut' : Les **facteurs pronostiques** sont à différencier des **facteurs de risques**

Au total, on s'intéresse à :

- la probabilité de **survivre au moins un certain temps *t*** à compter d'un instant de référence
- la probabilité pour que **l'événement d'intérêt survienne après un délai *t*** à compter de l'instant de référence.

*Exemple : On s'intéresse aux complications post opératoires en chirurgie digestive. On observe les patients pendant une semaine après leurs opérations. Au bout d'une semaine, on connaîtra le nombre de personnes avec complications et le nombre de patients sans complications. On pourra ensuite se demander quand les complications surviennent, leur dynamique, leur répartition. Ainsi, on pourrait trouver qu'au bout de 48h, 35% des patients ont eu des complications. C'est le fameux « **time to event** » : délai jusqu'à l'apparition de l'événement.*

DÉFINITIONS

Une cohorte est ensemble de sujets qui vivent les mêmes évènements au même moment. En matière de recherche médicale, c'est un ensemble de sujets inclus dans une étude au même moment et suivis dans des **conditions standardisées** pendant une **durée prédéfinie**.

Une cohorte « incipiente » (néologisme « inception cohort ») doit inclure des **patients observés au début de leur affection** à un point **uniforme** de l'évolution de leur maladie. Les sujets/patients sont les « cas incidents ».

Point tut' : Dans une **cohorte idéale**, tous les patients sont inclus au même moment, tous les patients sont « alignés ».

Contrairement à ce que le terme « survie » laisse penser, **l'événement d'intérêt n'est pas** forcément le décès, mais peut-être aussi la survenue d'une maladie, la récurrence de symptômes après traitement, ou encore, en dehors d'un contexte médical, la durée de vie des composants électroniques, ect.

- 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.** Dans ce cours, la terminologie « survie » sera utilisée quel que soit le type d'événement d'intérêt.

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

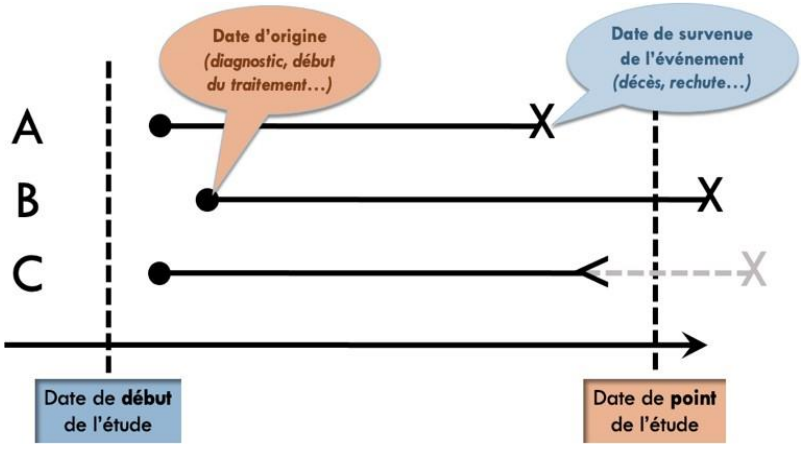
- Décès de **toutes causes**, où, chaque décès de patient compte comme un événement.
- Décès pour une **cause spécifique** *(par exemple, décès par accident coronaire)*, dans ce cas les décès d'autres causes *(par exemple, décès par cancer)* ne comptent pas comme un événement, mais comme une **censure**. Ceci n'est possible que lorsque les « autres causes de décès » sont indépendantes du phénomène étudié.

Point tut' : Une **cohorte** est un groupe d'individus qui suit les **mêmes contraintes** et qu'on suit dans le temps. Cela peut durer des années !
Par exemple, on peut s'intéresser à l'apparition de maladies cardiovasculaires dans la promo des LAS 2022-2023. C'est une cohorte, tout le monde arrive au même moment, est suivi et soumis aux mêmes contraintes.

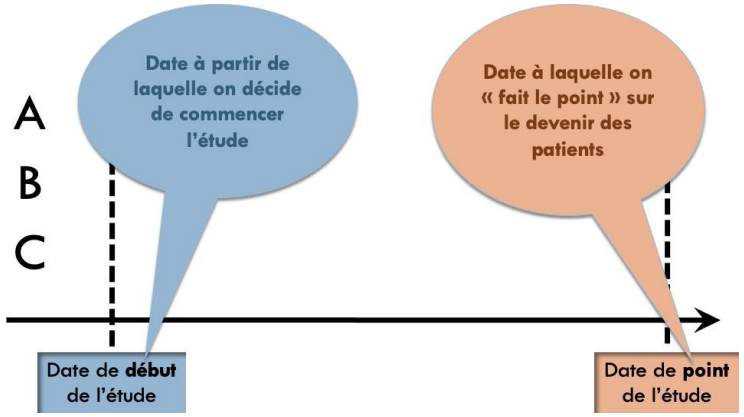
La durée de survie est le délai entre deux dates :

Prenons un exemple schématique : trois patients A, B et C sont inclus dans une étude de suivi longitudinal (étude de cohorte…)

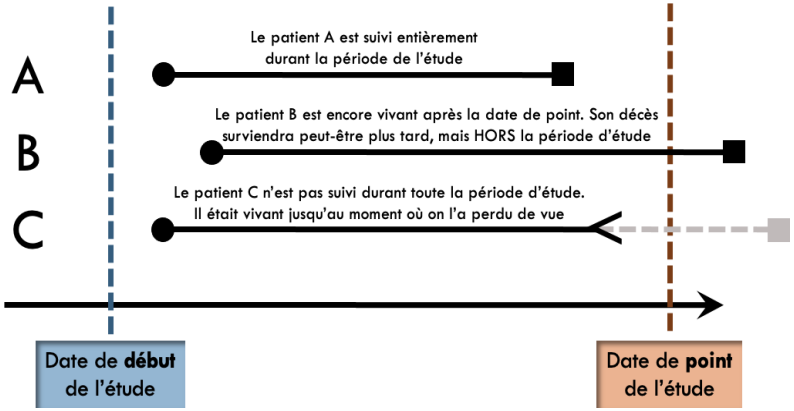
- La **date d'origine** est une date calendaire indiquant le **point de départ de la surveillance** : *par exemple, la date de randomisation dans un essai thérapeutique*.
  Cette date d'origine peut être identique ou différente pour chaque sujet en fonction des modalités d'inclusion des sujets.



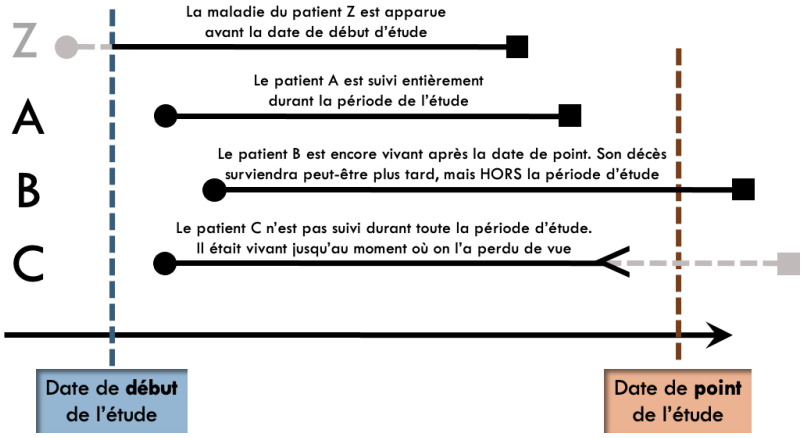
- La **date de point** est une date fixe calendaire et correspond à la **date choisie pour faire le bilan**, au-delà de laquelle les informations recueillies ne sont plus considérées dans l'analyse.



- La **date de dernières nouvelles** est la date la plus récente à laquelle on a recueilli des informations sur le patient, notamment sur la survenue ou non de l'événement 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)*. *Par exemple, il peut s'agir de la date de découverte d'une hypertension artérielle dans une étude de cohorte portant sur les facteurs de risque de mortalité cardio-vasculaire.*



Point tut' : Dans l'exemple sur les complications d'interventions chirurgicales. La date de début d'étude commence le jour de l'intervention et on suit le patient les jours qui suivent.

Cependant, autre exemple, dans le cas de la **survenue d'AVC chez des patients hypertendus**, on va être dans le cas du **patient Z**. Ces patients arrivent dans l'étude à un certain moment mais peuvent être hypertendus depuis quelques années déjà. Ainsi, si on prend dans notre étude des gens hypertendus depuis 40 ans, ils ont de fortes chances de faire un AVC peu de temps après **l'inclusion dans l'étude**. Mettons que notre patient fasse un AVC au bout de 2 mois, peut-on dire que le délai de survenue des AVC est de 2 mois ? NON : il va falloir remonterà **l'origine de la maladie** pour les patients Z.

Perdu de vue (lost of follow-up) :

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 *(patient C)*.

Un cas particulier concerne les sujets inclus dans l'étude mais n'ayant fait l'objet d'aucun suivi. Ces sujets ne seront pas comptabilisés dans l'analyse. On parle alors de ***« perte de vue » d'emblée****.*

Dans tous les cas, il est d'usage de vérifier que le processus de perte de vue pour l'ensemble des sujets (perte de vue d'emblée, ou après une durée de suivi) n'est pas lié à l'événement d'intérêt.

Par exemple en comparant les caractéristiques de ces patients à celles des sujets ayant fait l'objet d'un suivi complet.

Censure (censored data) :

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** *(patient C)* et les sujets vivants à la date de point (souvent appelés **exclus-vivants**) *(patient B)*.

Ces deux mécanismes de censure sont de **nature différente**. En effet, on ne peut assimiler les perdus de vue aux exclus-vivants, car la raison de leur « disparition » peut être liée à l'évolution de la maladie (décès méconnu de l'investigateur par exemple).

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.

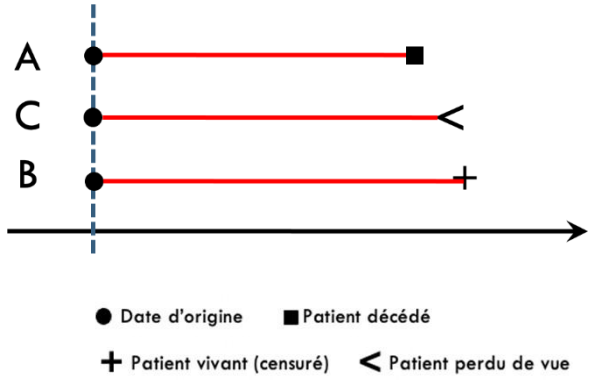
Temps de participation :

Le temps de participation est la **durée de surveillance pour chaque sujet** utilisé dans l'estimation de la survie.

On distingue 3 situations :

- L'événement s'est produit au cours de la surveillance : le temps de participation est le délai entre la date d'origine et la survenue de l'événement (patient A)
- le sujet est vivant à la date de point : son temps de participation est le délai entre la date d'origine et la date de point (patient B)
- le sujet est perdu de vue : dans ce cas, son temps de participation est défini par le délai entre la date d'origine et la date de dernières nouvelles (patient C)

Point tut' : Sur le graphique, on a une échelle de temps avec une date de début d'étude et une date de fin d'étude. On va observer **la survenue d'un événement** particulier (qui n'est pas forcément le décès).



Point tut' :

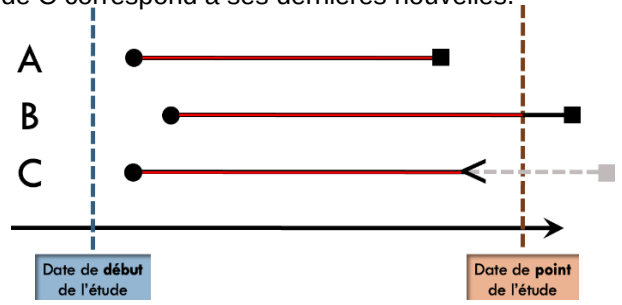
On a 3 individus : A, B et C dont **l'évolution est différente**.

Patient A : arrivée à une certaine date (date d'origine), puis suivi jusqu'à la survenue d'un événement.
L'information est complète chez le patient A car l'événement survient pendant le temps de surveillance (avant la date de point).

Patient B : date d'origine différente du patient A. L'événement n'est pas survenu pendant toute la durée de suivi, de la date d'origine à la date de point. On n'a pas d'information sur l'événement pour le patient B. Le patient B ne nous renseigne pas sur l'événement, bien qu'il puisse survenir ultérieurement.

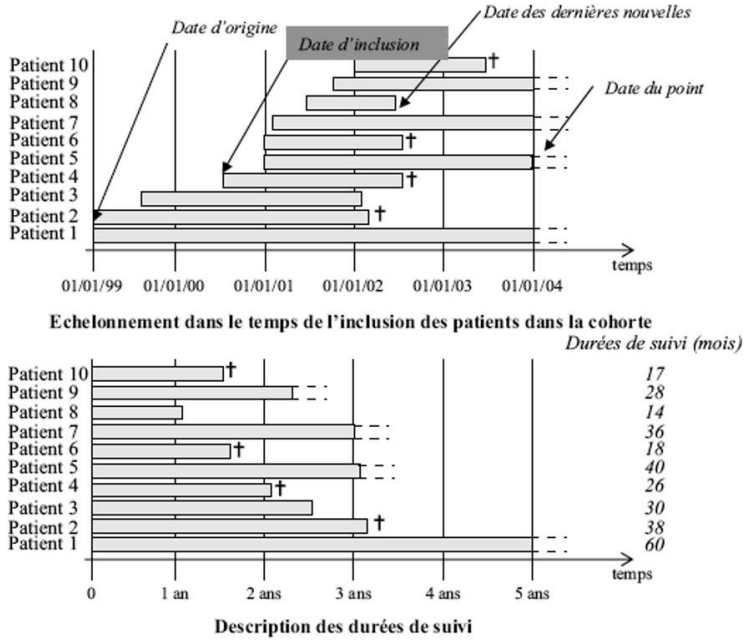
Patient C : suivi jusqu'à un point antérieur à la date de point, sans aucun événement observé. Il est perdu de vue (le patient ne vient plus faire son suivi régulier). On n'a pas d'information sur l'événement pour ce patient. On ne sait pas ce qu'il se passe après (grisé), peut-être que l'événement va survenir mais on ne voit plus le patient.

L'état aux dernières nouvelles est l'état du patient au moment du recueil des dernières informations ; si c'est le décès (patient A), alors on n'a plus de nouvelles le lendemain de sa mort. Les dernières nouvelles pour B sont la date de point. La dernière visite de C correspond à ses dernières nouvelles.



Point tut' : +++

Ce type de graphique est important pour comprendre que tout le monde n'a **pas le même temps de participation**. Dans une cohorte idéale, on voudrait que tous les individus aient le même temps de participation. *Prenons l'exemple de la cohorte de LAS 2022-2023 de Nice, en 5ème année de médecine, tout le monde n'aura pas le même temps de participation, des étudiants entrent sur passerelle ou ne sont pas présents au début de l'étude.*



Le graphique du haut présente un **calendrier de type grégorien**, celui du bas un **calendrier relatif**. Pour le calendrier relatif, tous les patients sont alignés sur 0. On voit directement la durée de participation de chaque individu avec la survenue de l'événement (ici le décès) ou bien les perdus de vue.

I. FONCTION DE SURVIE

1. La loi exponentielle

La l**oi de Poisson** régit la survenue d'un évènement par unité de mesure (temps, volume, surface …). Ici elle régit la survenue de la mort en fonction du temps.

On démontre que si un évènement se réalise selon une loi de Poisson (de paramètre **λ=μ=σ²**), le temps entre deux réalisations consécutives de l'événement considéré est distribué selon une l**oi exponentielle d'espérance 1/λ** (λ est appelé le taux de défaillance instantané)

Point tut' : Ainsi les événements sont régis selon une **loi de Poisson** et les délais entre 2 événements par une **loi exponentielle**.

Avec ces lois, on est dans le domaine/modèle du **paramétrique** car on fait une hypothèse sur le comportement des événements (et variable).

La **loi exponentielle** est utilisée couramment pour représenter la durée de vie de composants ou d'équipements pour lesquels l'hypothèse d'un taux de défaillance constant au cours du temps peut être justifiée. Cela implique que les défaillances sont dues uniquement au hasard et qu'elles se produisent selon un processus de Poisson.

Fonction de densité de la loi exponentielle :

pour tout **x≥0,** $f(\mathrm{x}) = \lambda \mathrm{e} - \lambda \mathrm{x}$

Fonction de répartition de la loi exponentielle :

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

$$\boldsymbol{donc : F(t) = 1 - e^{-\lambda t}}$$

$\boldsymbol{F(t)}$ représente la proportion d'équipements (de composants, ect.) qui tombent en panne avant le temps *t* (c'est la **fonction de « défaillance »**).

Ainsi, la quantité $\boldsymbol{1 - F(t)}$ représente la quantité d'équipements qui fonctionnent pendant une durée de temps au moins égale à t. Cette quantité est notée $\boldsymbol{S(t)}$ et s'appelle la ***fonction de survie*** : $\boldsymbol{S(t) = e^{-\lambda t}}$**.**

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

Point tut' : On peut voir la fonction de survie comme une fonction de répartition complémentaire à la fonction de répartition qui s'occupe/ identifie les décès. On regarde d'abord les décès puis on va s'intéresser à la survie.

2. La 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* :

- $\boldsymbol{S(t) = \Pr(T > t) = 1 - F(t)}$
- 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 une fonction de répartition**. On la note également **S(t).** Elle représente :

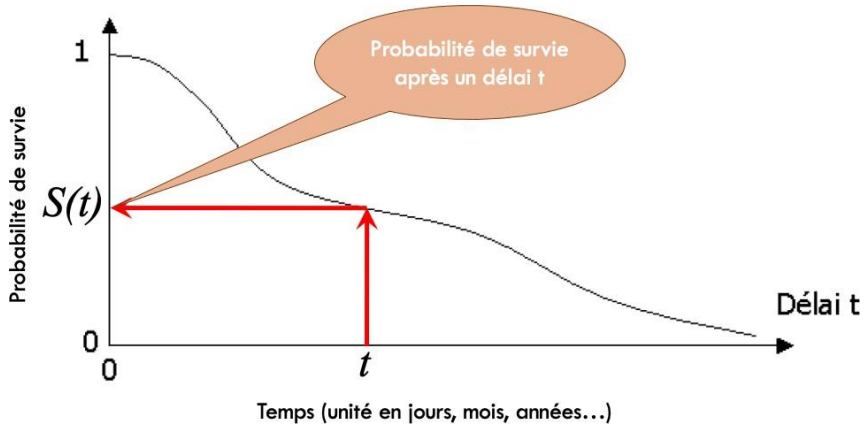
- la probabilité pour qu'un patient soit encore vivant après un délai *t*
- la proportion « vraie » des survivants après un délai *t*

►La **fonction de survie**, **S(t)**, est la **probabilité que l'événement d'intérêt ne survienne pas avant la date *t***.

⇨ *S(t) = Pr (délai de survenue de l'événement d'intérêt à compter de l'instant de référence > t)*

Ainsi, si l'événement d'intérêt est le décès, c'est la probabilité de survivre au moins jusqu'à la date *t*. Si l'événement d'intérêt est la récurrence de symptômes après traitement, c'est la probabilité de survivre sans symptômes jusqu'à la date *t* (on parle alors de *disease free survival).*

►La **fonction de survie** est représentée graphiquement par une **courbe desurvie**.



Point tut' : Face à ce graphique, on peut se demander quelle est probabilité de survie au bout d'un certain temps mais aussi :au bout de combien de temps un certain nombre de patients sont décédés ?

►La **fonction de survie** permet de calculer la probabilité pour que ledécès survienne après un délai t_1 et avant le délai t_2 (avec $t_2 > t_1$).

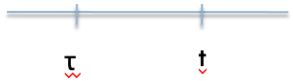
Il s'agit de calculer $\Pr(T \in]t_1; t_2])$.

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

►La **fonction de survie** donne aussi une information essentielle pourla suite : la probabilité de survivre encore après un délai *t* sachant que l'on est survivant après un délai r *avec* $r < t$, que l'on notera $S(t/r)$. On a la démonstration suivante *(pas à retenir) :*

$S(t) = \Pr(X > t)$ et $S(r) = \Pr(X > r)$

$t = r + s$ avec $s > 0$



Or nous avons l'égalité d'événements suivante :

$$\{X > r + s\} \cap \{X > r\} = \{X > r + s\}$$

En appliquant la formule des probabilités composées il vient aisément que :

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

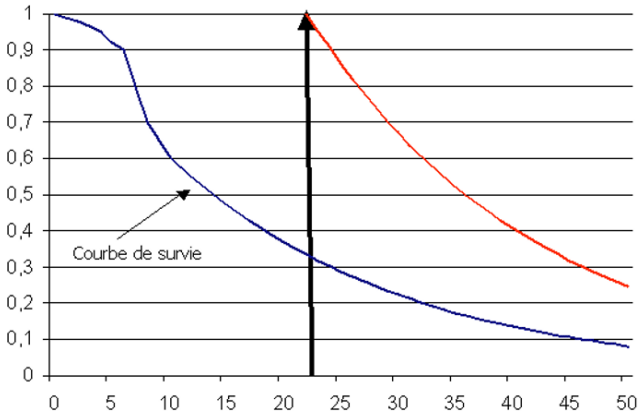
Au final :

$$\boldsymbol{S(t/r) = \frac{S(t)}{S(r)}}$$

Point tut' : On peut voir la fonction de survie comme une **probabilité conditionnelle**. *Par exemple, quelle est la probabilité d'être en vie à 3 ans sachant que j'étais en vie à 2 ans ?*

On dit que la fonction de survie est une fonction « sans mémoire », ici la **survie conditionnelle** prend en compte le fait qu'on était encore en vie à un temps *t*.

Exemple : Supposons que l'on veuille calculer la probabilité de survivre après (un délai de) $t = 33$ ans sachant que l'on est vivant à $t = 23$ ans.



À la lecture de la courbe de survie on remarque qu'il y a 33% de survivants à 23 ans. On lit également qu'à 33 ans, la proportion de survivants dans la population initiale est de 20%.

Mais, ne nous intéressant qu'aux **survivants à 23 ans**, ces 20%

représentent $\frac{3,5}{3,66}$ de la population d'intérêt, c'est-à-dire $\frac{S(33ans)}{S(23ans)}$.

II. ESTIMATION DE LA SURVIE

1. Recueil de données

Date d'origine : date à laquelle a débuté l'observation. Ex : date de diagnostic d'un cancer. Cette date doit avoir un sens clinique, afin que la « survie » analysée puisse être interprétée facilement par les lecteurs

Date des dernières nouvelles : date de décès pour les patients décédés ou date à laquelle on dispose des 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. Tout patient chez qui l'événement d'intérêt n'a pas été observé à la date de point est censuré à cette date. Un sujet perdu de vue à la date de point sera censuré à la date des dernières nouvelles.

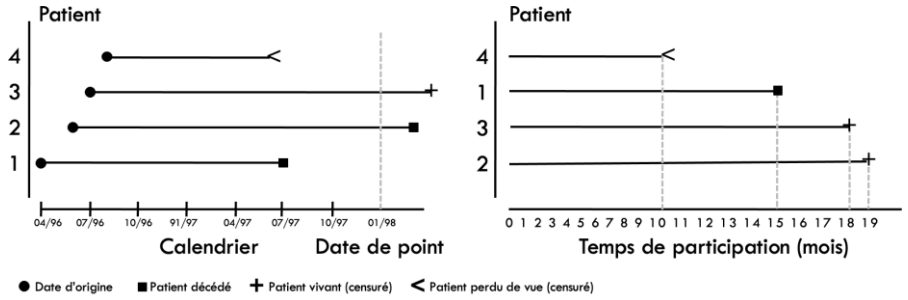
Un événement « en tout ou rien » (binaire) correspond à l'état du patient en deux éventualités (vivant ou décédé) à la date des dernières nouvelles. Tout événement binaire autre que le décès à un délai de survenue peut être analysé en délai de survie. Par exemple, on peut étudier la survenue de la rechute ou de la récidive tumorale après un traitement ou la survenue de métastases.

2. Calcul des durées de suivi

À partir de ces données, les **durées de suivi** (ou temps de participation) de chaque patient sont calculés par différence. Ainsi, les durées de suivi correspondent au délai entre la **date d'origine** et la **date des dernières nouvelles** qui sera soit :

- La date de décès en cas de décès
- La date de point pour les patients vivants pour lesquels le suivi est assuré
- Ou la date de perte de vue pour les patients vivants n'étant plus suivi dans la cohorte à la date de point.

Exemple :



Patient	Date d'origine	Date événement ou DDN	Etat aux dernières nouvelles	Etat à la date de point	Temps de participation	Temps de recul
1	04/96	07/97	Décédé	Décédé	15	21
2	06/96	03/98	Décédé	Vivant	19	19
3	07/96	04/98	Vivant	Vivant	18	18
4	08/96	06/97	Vivant	?	10	17

Point tut' : À gauche un calendrier grégorien (durées absolues) et à droite un calendrier statisticien (durées relatives) où on regarde le temps de participation. À partir de leurs données on peut construire le tableau ci-dessus. *(DDN : date des dernières nouvelles)*

3. 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. Cependant, en pratique, cela ne se produit jamais, car un certain nombre de sujets seront perdus de vue, et un certain nombre seront encore vivants à la date de point.

Deux méthodes d'analyse de survie sont de préférence utilisées : **l'analyse actuarielle** et la **méthode de Kaplan-Meïer**, qui sont deux **méthodes non paramétriques** (*non-parametric* ou *distribution-free*), puisqu'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) et de nombreux événements +++

La méthode de Kaplan-Meïer est donc la méthode de choix pour les *échantillons de taille plus réduite*.

Ces deux méthodes supposent une **hypothèse forte** : les probabilités de survie sont supposées indépendantes du calendrier. *Ceci revient à supposer, par exemple, que la survie à 1 an d'un groupe de patients inclus en 1970 est identique à celle d'un groupe de patients inclus en 1990*. Cette hypothèse n'est pas forcément vérifiée pour les études disposant d'un recul maximum très important, notamment en raison des progrès thérapeutiques vis-à-vis de la maladie étudiée. Elles partent du principe qu'il n'y a pas de progrès thérapeutique tout au 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 (*median survival time*) et quantiles (*percentiles*).

4. Analyse actuarielle

La fonction de survie est calculée sur des **intervalles de temps fixés a priori** (mois, trimestre, semestre, année...). Schématiquement, le mode de calcul est le suivant. Pour chaque intervalle de temps (*par exemple [0,1an], [1,2ans], etc.*), on définit :

- le nombre de sujets vivants au début de l'intervalle : **V**
- le nombre de sujets décédés dans l'intervalle : **D**
- le nombre de sujets vivants aux dernières nouvelles, dont le temps de participation s'arrête dans l'intervalle (censure), c'est-à-dire le nombre de sujets vivants censurés dans l'intervalle : **C**
    - L'hypothèse actuarielle (*actuarial assumption*) suppose que ces sujets sont exposés au risque d'événement sur la moitiéde l'intervalle (*6 mois dans notre exemple*).

Le **nombre de sujets exposés au risque d'événement** (ex : décès) sur l'intervalle est : **N = V - (C/2)**

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

D / N.

La survie sur cet intervalle est : **(N - D) / N**. Cette probabilité est appelée **survie instantanée**.

La **fonction de survie** est obtenue en faisant le **produit des survies instantanées sur l'ensemble des intervalles** : *par exemple, la survie à 3 ans = (survie instantanée entre 2 et 3 ans) x (survie instantanée entre 1 et 2 ans) x (survie instantanée entre 0 et 1 an).*

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**

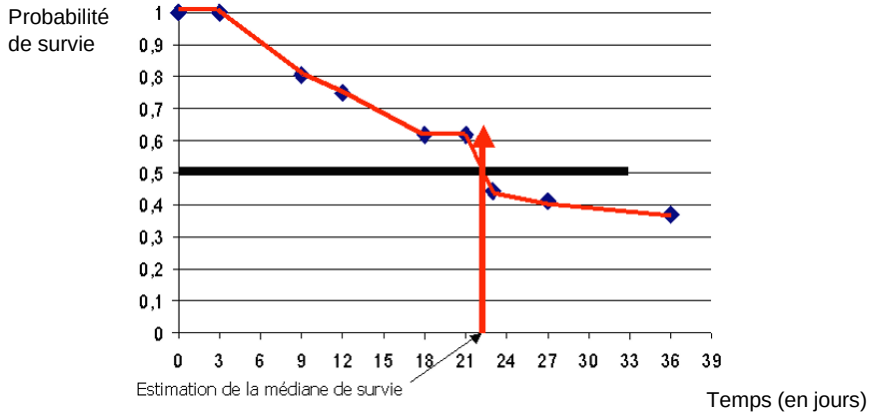
Point tut' : On appelle ce type tableau une **table de mortalité** si l'événement d'intérêt est la mort.

La première colonne du tableau est une échelle de temps (0, 3, 6…) qui va permettre de faire la différence (en partie) entre la méthode actuarielle où les intervalles sont toujours les mêmes (réguliers) et la méthode de Kaplan-Meïer qui va prendre comme intervalle les moments où les événements surviennent.

Point tut' : À 6 mois, on a 210 vivants, il y a eu 40 décès dans l'intervalle et 10 sujets sont censurés.

Pour calculer la probabilité de survie, il faut connaître le dénominateur des **survies instantanées** (**N**). À 6 mois N=205, cela va permettre de calculer le taux de survie instantanée qui va être de 0,805. Enfin, on fait le **produit** des survies instantanées sur les intervalles précédents. C'est toujours la **même mécanique**, être de vivant à 6 mois c'est aussi être vivant à 3 mois donc on fait le produit de la survie instantanée à 6 mois et de la probabilité de survie à 3 mois : 0,805 x 1 (*cf flèches*)

Pour chaque intervalle de temps, on représente **l'estimation de la survie S(t)** par un point. Les coordonnées du premier point sont 0 (à j_0) en abscisse, et 1 (100 %) en ordonnée. Tous les points consécutifs sont reliés par un segment de droite.



L'inconvénient majeur de cette méthode est qu'elle 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 qu'un sujet suivi pendant 21 jours apporte la même information qu'un sujet suivi pendant 29 jours pour la survie à 30 jours dans l'exemple présenté*. C'est la raison pour laquelle cette méthode est **à réserver à de grands échantillons**.

5. Méthode de Kaplan-Meïer

Contrairement à l'analyse actuarielle, les intervalles ne sont **pas fixés a priori**, mais sont définis par les **instants** auxquels les événements sont observés. *Ex : on change d'intervalle à chaque décès.* Ces intervalles sont donc inégaux, débutent à l'instant d'un événement et s'arrêtent juste avant l'événement suivant.

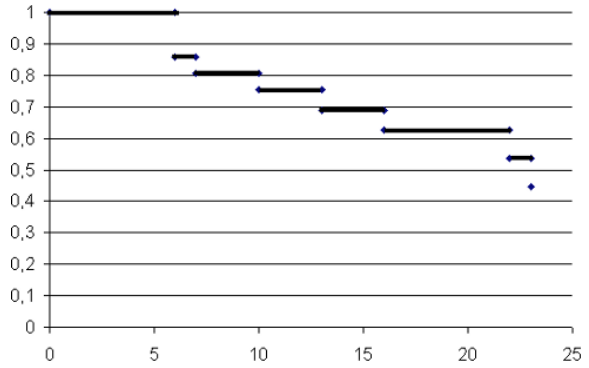
Pour chaque intervalle entre deux événements, on définit **V**, **D** et **C** comme précédemment (avec la particularité que D vaut souvent 1, sauf dans le cas où plusieurs événements surviennent au même temps de participation).

Dans l'analyse de Kaplan-Meïer, **N = V - C** et la probabilité de survie instantanée calculée sur cet 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**.

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**

La courbe de survie se compose de **paliers successifs**, où les probabilités de survie sont constantes entre deux temps d'événements consécutifs. Le **premier palier vaut 1** depuis l'origine jusqu'au délai de survenue du premier événement. Il s'abaisse ensuite à la première valeur calculée pour constituer un second palier jusqu'au délai de survenue de l'événement suivant, etc. Il est possible de relier les paliers successifs par des segments verticaux, mais il n'est pas correct de les relier par des segments obliques. La courbe ainsi obtenue présente une **allure en « marches d'escalier »**.



Tut'Récap :

Principales différences entre la méthode actuarielle et Kaplan-Meïer :

1) Pour la méthode actuarielle les intervalles sont réguliers alorsque pour la méthode de Kaplan-Meïer les intervalles dépendent du moment de survenue des événements. (regarder la 1ère colonne de la table)
2) Dans la méthode actuarielle, les censures sont prises en compte à moitié alors que dans la méthode de Kaplan-Meïer les censures sont prises en entier.
3) Allure des courbes différente

6. Choix d'une valeur résumée

Médiane de survie

La courbe de survie apporte des renseignements importants, mais il est utile de disposer d'indicateurs synthétiques ou résumés de cette courbe. La **moyenne de survie** n'est pas un bon indicateur, pour des raisons d'ordre statistique, notamment liées à l'existence de censures.

La **médiane de survie** lui est préférée. Elle représente la **durée t** pour laquelle la probabilité de survie **S(t) est de 50 %.** À cause de la distribution par paliers de la fonction de survie, il est souvent impossible de connaître la durée correspondant à une survie exacte de 50 %. En pratique, la médiane est estimée par la plus petite durée pour laquelle la survie est inférieure à 50 %.

Il arrive que la fonction de survie soit toujours supérieure à 50 %. Dans ce cas, la médiane ne peut être estimée. On estime alors les **quantiles** (NB : un quartile = 25 %) : pour le p^{ième} quantile on estime la durée pour laquelle la probabilité de survie est de 100-p. *Par exemple, le 25^{e} quantile (ou 1er quartile) correspond à la plus petite durée pour laquelle la survie est inférieure à 75 %.*

Survie à date fixée

Un autre indicateur fréquemment utilisé pour résumer l'information d'une courbe de survie est l'estimation de la **survie à un temps donné** *(survie à 5 ans par exemple…)*

III. COMPARAISON DE DEUX FONCTIONS DE SURVIE

1. Contexte

Il arrive fréquemment que l'on souhaite montrer qu'une action (intervention, traitement) ou une classification ont un lien avec la survie. Il s'agira de conduire une étude comparative et de mettre en œuvre un test d'hypothèses.

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

Attention !

Une erreur importante, et souvent retrouvée, consiste à assimiler l'efficacité du traitement à la réponse des patients à ce traitement, et à **comparer la survie**, non plus entre les **patients traités** et les **patients non traités**, mais entre les sujets qui répondent au traitement et les sujets qui ne répondent pas (*comparison of survival by response*).

Cette méthode est à proscrire et peut provoquer des biais et des conclusions fausses :

- les sujets répondeurs sont en général en meilleure santé que les sujets non répondeurs et sont donc susceptibles - indépendamment de tout traitement - de vivre plus longtemps ;
- la comparaison de la survie par la réponse au traitement peut être biaisée puisque les patients doivent vivre suffisamment longtemps pour avoir la possibilité de répondre au traitement (*guarantee-time bias*).

2. Principe du test du log-rank

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

Le test du log-rank, évaluant l'écart entre le **nombre observé** et le **nombre attendu d'événements** sur les deux groupes, est un Khi² 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 :

- H_0 : les fonctions de survie sont les mêmes dans les deux populations d'où sont issus les groupes A et B.
  On a donc $S_A(t) = S_B(t)$
- H_1 : les deux fonctions de survie diffèrent

Exemple : Imaginons que l'on souhaite faire la **preuve qu'un traitement adjuvant** à la chirurgie dans le carcinome hépatocellulaire **améliore la survie des patients**. Les grands traits de l'étude sont les suivants :

- *la survie sera comptée à partir de la date de la chirurgie.*
- *des patients ont été inclus pendant une année dans une étude qui a duré 3 ans et répartis par tirage au sort dans un des deux groupes de traitement : chirurgie seule (groupe A) ou chirurgie + traitement adjuvant (groupe B).*
- *la durée de suivi des patients (durée de participation à l'étude ou recul) varie d'un patient à l'autre*

A la fin de l'étude on dispose pour chaque patient :

- *du groupe auquel il a appartenu, A ou B*
- *des temps de suivi pour chaque patient selon son groupe et selon le fait que le patient soit décédé ou bien que le patient soit censuré, qu'il soit encore vivant ou perdu de vue.*

Supposons que l'on dispose des observations suivantes.

- *Dans le groupe A, les tA_i et tA_i* sont : 1; 1; 2; 2; 3; 4; 4; 5; 5; 8; 8; 8; 8; 11; 11; 12; 12; 15; 17; 22; 23*
- *Dans le groupe B, les tB_i et tB_i* sont : 6; 6; 6; 6,1*; 7; 9*; 10; 10,1*; 11,2*; 13; 16; 17,3*; 19*; 20*; 22; 23; 25*; 32*; 32*; 34*; 35**

Les ensembles des tA_i et tB_i (patients décédés) constituent l'ensemble des temps de décès observés, quel que soit le groupe ; on les notera t_i et on les considérera ordonnés par valeurs croissantes. Ici les t_i sont : 1; 2; 3; 4; 5; 6; 7; 8; 10; 11; 12; 13; 15; 16; 17; 22; 23

3. 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 $(1 - S(t_i / t_{i-1}))$ 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 ainsi 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
4	37		37	2	0,946	0,054
5	35		35	2	0,943	0,057
6	33		33	3	0,909	0,091
7	30	1	29	1	0,966	0,034
8	28		28	4	0,857	0,143
10	24	1	23	1	0,957	0,043
11	22	1	21	2	0,905	0,095
12	19	1	18	2	0,889	0,111
13	16		16	1	0,938	0,062
15	15		15	1	0,933	0,067
16	14		14	1	0,929	0,071
17	13		13	1	0,923	0,077
22	12	3	9	2	0,778	0,222
23	7		7	2	0,714	0,286

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

4. Calcul des décès attendus

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

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 $\mathbf{E_{Ai}}$ et $\mathbf{E_{Bi}}$. On remarque que l'on utilise ici, comme toujours, la justesse supposée de l'hypothèse nulle (H_0) puisque les probabilités de décès, et donc la survie, sont supposées ne pas dépendre du groupe.

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
5	35		35	2	0,943	0,057	14	21	0,799	1,201
6	33		33	3	0,909	0,091	12	21	1,092	1,988
7	30	1	29	1	0,966	0,034	12	17	0,408	0,579
8	28		28	4	0,857	0,143	12	16	1,714	2,286
10	24	1	23	1	0,957	0,043	8	15	0,344	0,656
11	22	1	21	2	0,905	0,095	8	13	0,760	1,240
12	19	1	18	2	0,889	0,111	6	12	0,666	1,334
13	16		16	1	0,938	0,062	4	12	0,249	0,751
15	15		15	1	0,933	0,067	4	11	0,268	0,732
16	14		14	1	0,929	0,071	3	11	0,214	0,786
17	13		13	1	0,923	0,077	3	10	0,230	0,770
22	12	3	9	2	0,778	0,222	2	7	0,445	1,555
23	7		7	2	0,714	0,286	1	6	0,286	1,714

Sous l'hypothèse nulle (H_0) 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 selon le groupe) doit être voisin du nombre total de décès observés (noté D_A et D_B selon le groupe), et ceci dans chacun des groupes.

Dans l'exemple, on obtient : E_A=10,74 ; E_B=19,26 ; D_A=21 ; D_B=9.

Point tut' : Au temps t avant le décès, on a :

- N_A : Effectif du groupe A
- N_B : Effectif du groupe B
- N : Effectif global, avant le décès (-lescensurés) : **N = NA + NB**
- DA : Nombre de décès observés dans le groupe A
- DB : Nombre de décès observés dans le groupe B
- D : Nombre de décès observés : **D = DA + DB**
- E_A : Nombre de décès attendus dans le groupe A : $E_A = (D. N_A)/N_A$
- E_B : Nombre de décès attendus dans le groupe A : $E_B = (D. N_B)/N_B$

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

5. Test Khi²

Le paramètre du test est construit à partir de ces quatre valeurs/paramètres (aléatoires normalement à ce stade de la construction) :

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

Sous H0, Q suit une distribution de χ^2 à 1 degré de liberté

Condition de validité : E_A et $E_B \geq 5$

On construit l'intervalle de pari de niveau 0,95 : $IP_{),+5} = [0\,;\,3{,}84]$

On met en place la règle de décision. Si la valeur calculée $Q_c \in [0\,;\,3{,}84]$, on ne pourra conclure à une différence entre les fonctions de survie dans les deux populations considérées. Si la valeur Q_c excède 3,84 on conclura au risque de 5% que les fonctions de survie diffèrent.

Dans l'exemple traité, on obtient $Q_c = 15{,}26$. On rejette donc l'hypothèse d'égalité des fonctions de survie. La survie est meilleure dans le groupe dans lequel $D < E$, c'est le groupe B. La preuve est faite (au risque d'erreur de 5%) que le traitement adjuvant améliore la survie des patients à compter de la date de chirurgie.

Tout d'abord énorme dédi à Sap1ens, ex tut de biostat, cette fiche s'inspire énormément de la sienne

Conseils généraux pour aborder cette fiche :
- retenez bien la partie définition au début +++
- faites bien la distinction en analyse par Kaplan Meier et analyse actuarielle, ça tombe souvent
-ne négligez par les graphiques et les schémas, ils sont hyper importants pour comprendre et certains tombent +++
- ce cours a énormément de lien avec tous les autres cours de biostat, soyez sûr de les maîtriser pour essayer de tout comprendre
- ne vous focalisez pas sur les démo et les formules sauf les plus importantes