Analyse de survie

Dr BOUNTOGO
Mai 2017

Plan

- **Objectif du cour**
- **Notion de données censurées**
- **Principe**
- **Les méthodes de construction des courbes de survie**
- **Comparaison de courbes de survie**

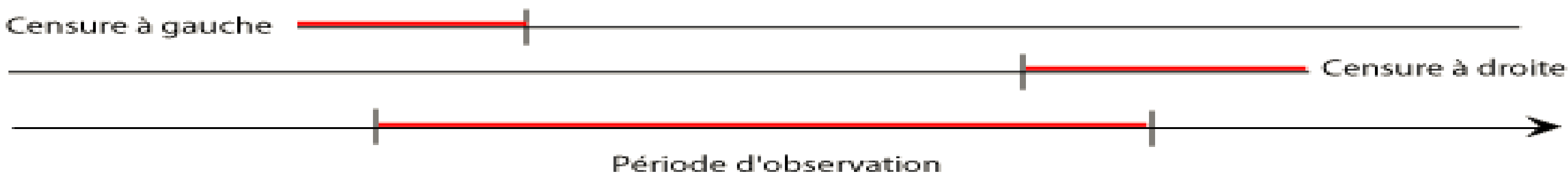
Objectif

- **Objectif général :**
  - Savoir quand et comment utiliser une analyse de données censurées en univariée
- **Objectifs spécifiques** :
  - Définir la notion de données censurées
  - Connaître les méthodes de construction paramétrique et non paramétrique des courbes de survie (Kaplan-Meier, Actuarielle), leurs intérêts et limites respectives
  - Mesurer le risque relatif dans la survie
  - Utiliser le test du Log-Rank et présenter la famille des tests du Log Rank pour comparer les courbes de survie et savoir interpréter ses résultats

Notion de données censurées

- Domaine: études épidémiologiques et **cliniques** qui étudient la **survenue d'événements** au cours du temps
- **Conceptualisation:**
  - Passé ➔ avenir: l'épidémiologiste enregistre les évènement en fonction du temps
  - En réalité, il y a des sujets chez qui l'évènement **survient** et échappera à l'observation, ce qui peut se produire de deux façons :
    - NON ENREGIISTRE = censure à gauche,
    - Suivi interrompu avant que l'événement ne survienne= censure à droite

Censure à gauche censure à droite



Censure à gauche

- Dans les études transversales d'estimation d'incidence, on procède à une recherche des cas incidents au cours d'une période passée (enquête rétrospective). Par définition, les cas survenus au cours de la période qui n'ont pas été diagnostiqués ne sont pas identifiés, et ne sont plus identifiables par exemple parce qu'ils sont décédés, guéris ou ont quitté la zone géographique de l'étude

Censure à droite

- Dans les études prospectives de cohorte, dont l'objectif est d'enregistrer l'évènement de santé d'intérêt chez tous les sujets qui en seront atteints au cours de la période d'observation dans une population donnée, il y a de fortes chances pour que tous les sujets ne soient pas tous suivis pour la durée exacte de la période. Certains vont décéder, d'autres vont quitter l'observation (être perdus de vue), On dit qu'ils sont censurés à droite

Principe

- Probabilité de survenue d 'un événement (B) chez des sujets ayant en commun un événement d 'origine (A), en tenant compte du délai écoulé entre ces 2 évènements.

➔survenue du décès (B) chez des sujets avec maladie grave (A)

- Méthode paramétrique qui utilise la fonction S(t) déjà vu dans le chapitre sur les indicateurs(ne sera pas étudier ici)
- Méthode non paramétrique

  Méthode de Kaplan-Meier et Méthode actuarielle

- 1 méthode de comparaison de survie

  1 Test du Log Rank

Champs d'application

- Les principales applications des courbes de survie en santé publique sont les suivantes :
  - *En épidémiologie* avec le suivi de cohortes permettant une évaluation à long Terme
  - *En recherche clinique* avec les essais thérapeutiques : évaluation de l'efficacité d'une thérapeutique et comparaison de l'efficacité de deux ou plusieurs traitements
  - *Détermination des facteurs pronostiques et des facteurs de risque* : il est important de connaître les facteurs influençant la survenue d'un événement.
  - NB:
  - **facteur de risque** lorsque l'événement est la survenue d'une maladie
  - **facteurs pronostiques** lorsque les événements (décès, rechute, toxicité) concernent des sujets malades.

Informations nécessaires pour la construction des courbes de survie

- ***La date d'origine (DO)***: point de départ du suivi du patient.
- ***La date de point 'DP)***: date à laquelle on arrête le suivi dans l'étude.
- ***La date des dernières nouvelles (DDN)*** : date à laquelle on a eu pour la dernière fois des nouvelles de l'état du patient:
  - a) le décès (ou l'événement auquel on s'intéresse) est survenu avant la date de point : la DDN = la date du décès
  - b) le décès est survenu après la date de point : la DDN = DP
  - c) le sujet n'est pas décédé et la dernière date connue est antérieure à la DP: la DDN = la dernière date connue
  - d) le sujet n'est pas décédé et la dernière date connue est postérieure à DP : la DDN = DP
- ***L'état aux dernières nouvelles*** **:** sujet décédé (situation a) ou sujet vivant (situations b, c et d).

Informations nécessaires pour la construction des courbes de survie

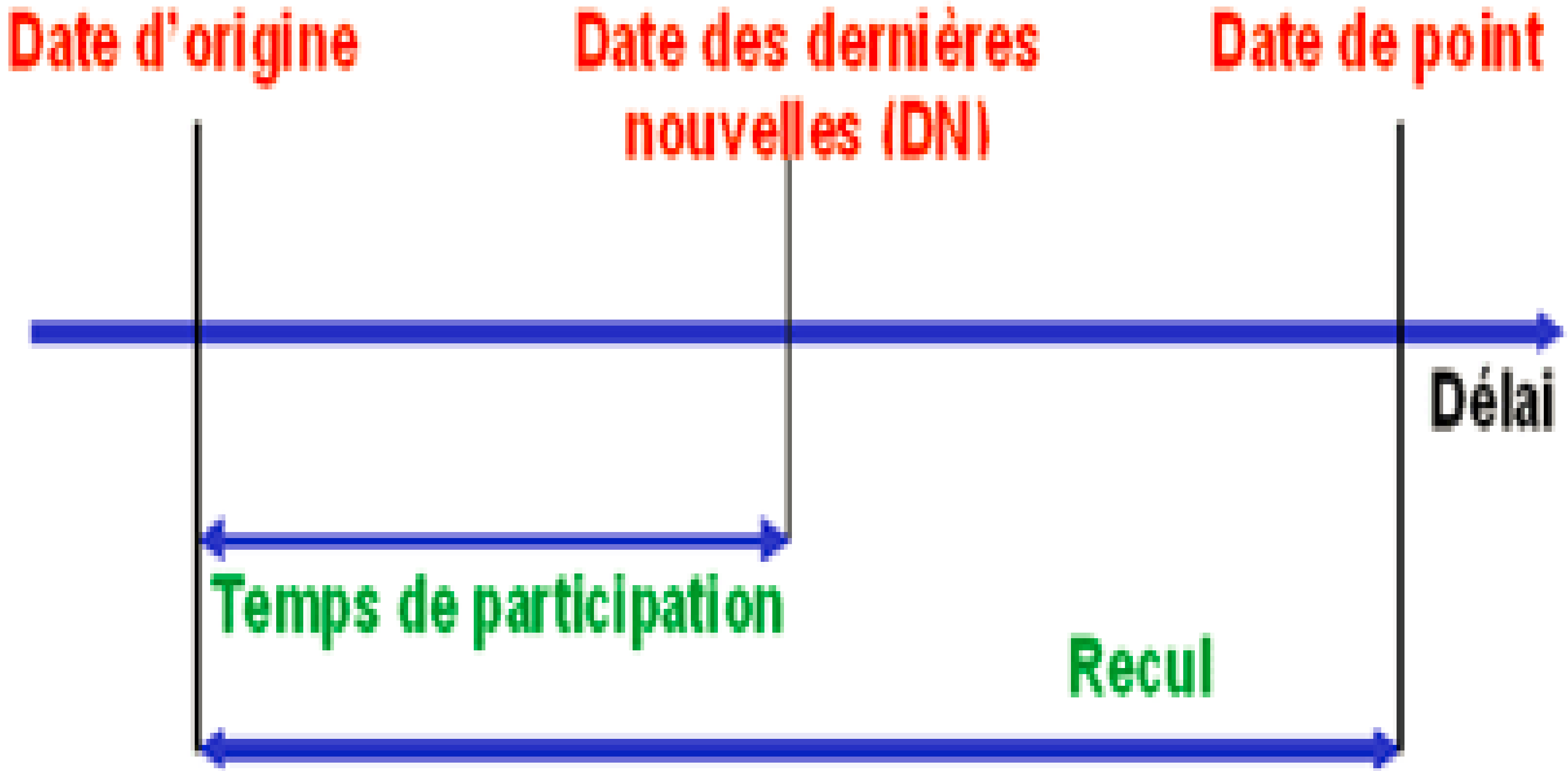
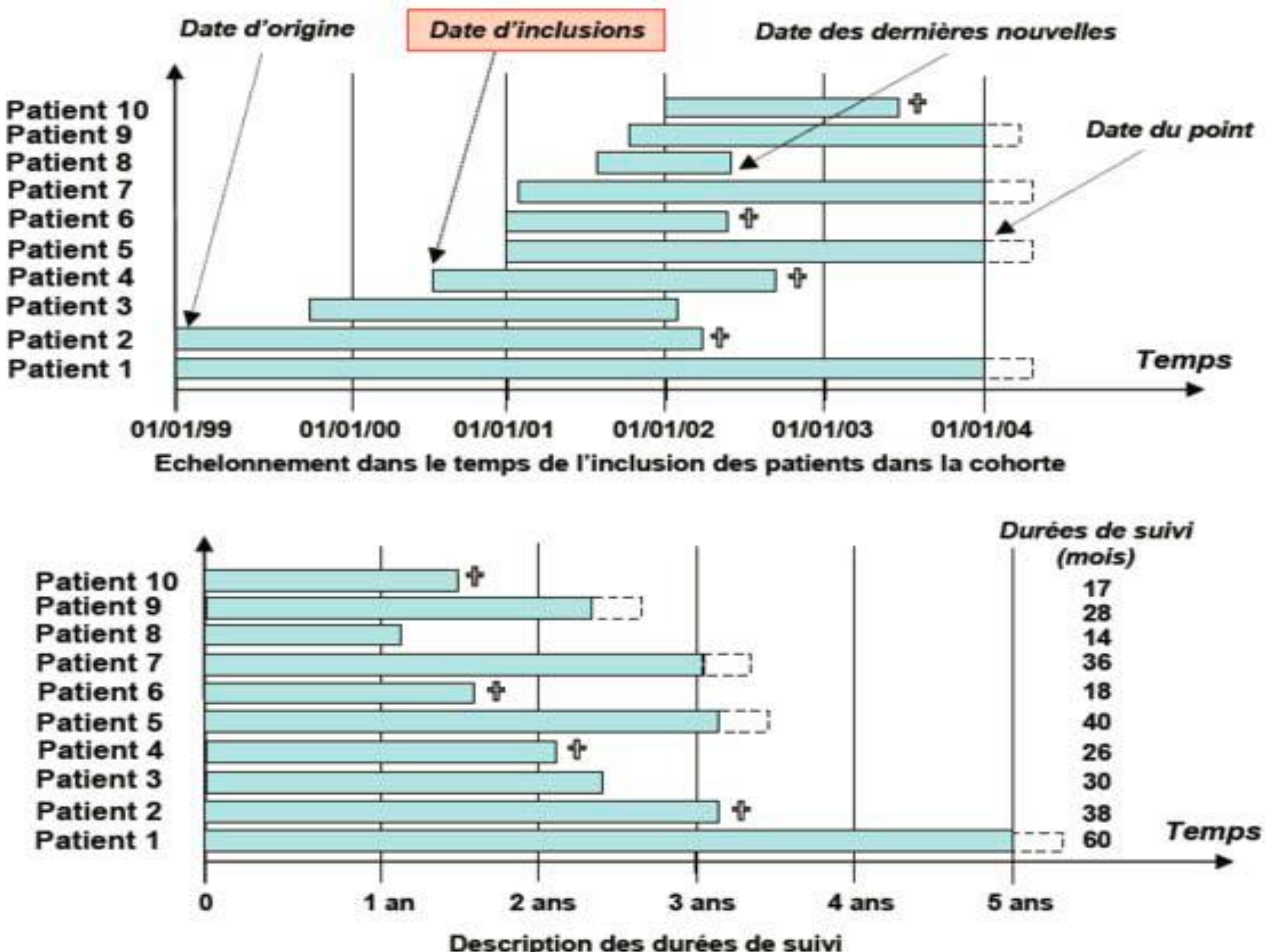
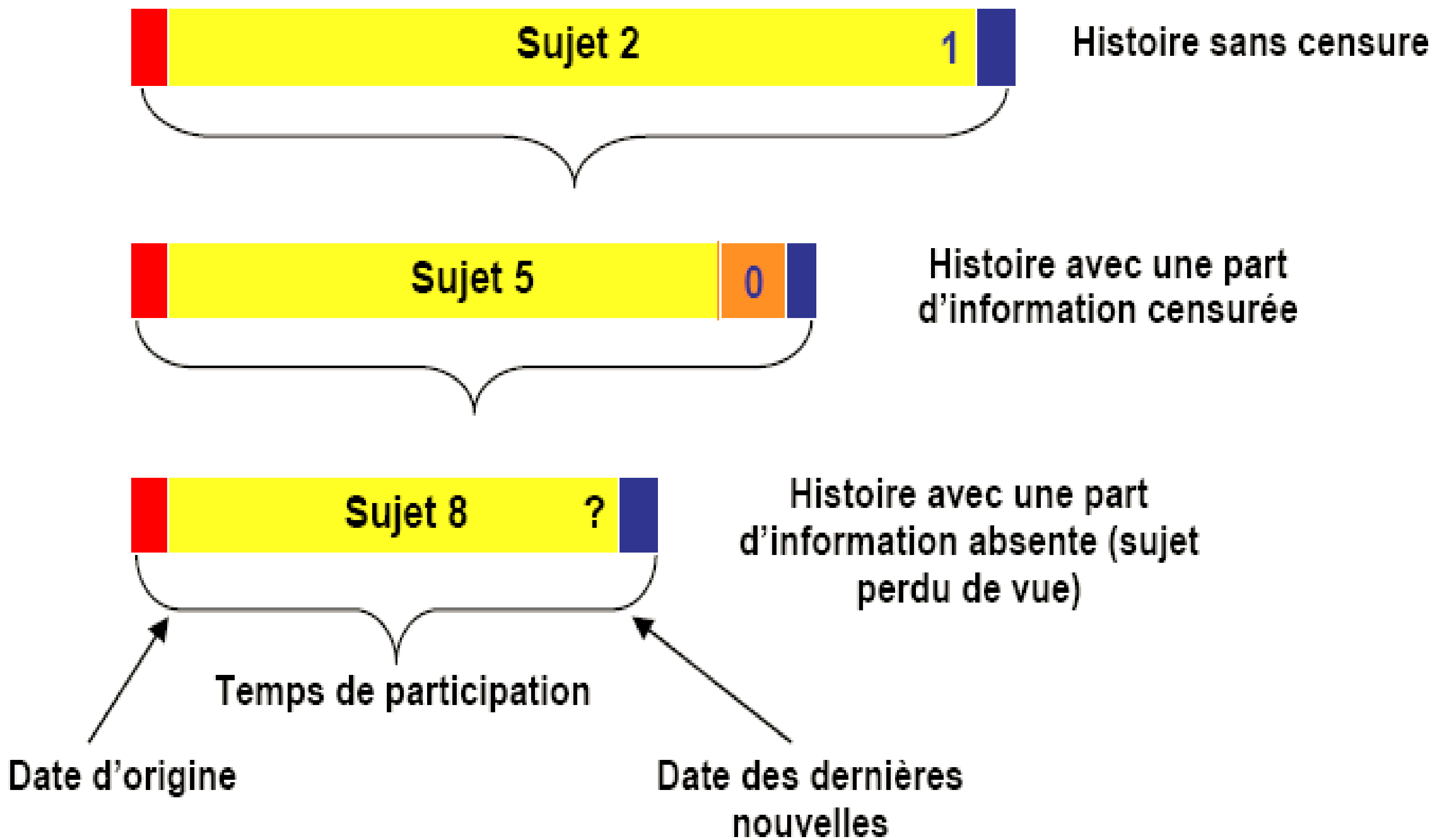


Tableau 1 : Présentation des données de survie (*schéma issu de l'article de C. Aberti - Rev Mal Respir 2005 ; 22 : 333-7*)



1

Tableau 2 : Représentation des informations pour trois situations



ETAT AUX DERNIERES NOUVELLES
0 : l'événement ne s'est pas produit
1 : l'événement s'est produit

CENSURES
- Tout délai situé après la date de point est tronqué
- Tout état du sujet enregistré après la date de point est censuré

Méthode de Kaplan-Meier

- Calcul de la probabilité de survie chaque fois qu'au moins un « décès » est enregistré.
- Soit:

V_i : nombre de vivants au début de l 'intervalle t_i - t_{i-1}

D_i : nombre de décès pendant l 'intervalle t_i - t_{i-1}

E_i : nombre de d 'exclus au début de l 'intervalle t_i - t_{i-1}

q_i : probabilité de décès pendant l'intervalle t_i-t_{i-1} : $q_i = D_i / (V_i - E_i)$

p_i : probabilité de survie pendant l 'intervalle t_i - t_{i-1} : $p_i = 1 - q_i$

S_i : fonction de survie à l 'instant t_i : $S_i = p_0 p_1 \ldots p_i = p_i S_{i-1}$

Méthode de Kaplan-Meier

- Les valeurs des taux de survie S(ti) sont estimées à partir des données d'un échantillon de n sujets suivis jusqu'à la date de point. Comme toute estimation il est nécessaire d'en évaluer sa précision à l'aide de la variance et si possible d'un intervalle de confiance. Le calcul de la variance a été proposé par Greenwood et permet de calculer l'intervalle de confiance.
- Variance de Greenwood de S(t) pour tout t compris dans l'intervalle [ti et ti+1[Ni = nombre d'individu à ti
- Intervalle de confiance au risque choisi, sous l'hypothèse que **S(t) est distribué normalement** (loi de Gauss), est égal à :

$$\mathrm{Var}\ S_{(t)} = S^2_{(t)} \times \left[\frac{D_1}{(N_1 - D_1).N_1} + \ldots + \frac{D_i}{(N_i - D_i).N_i} \right]$$

$$I.C.(1-\alpha) = S_{(t)} \pm Z(\alpha/2) . \sqrt{\mathrm{Var}\ S_{(t)}}$$

Méthode de Kaplan-Meier

- Cette **hypothèse n'est pas toujours vérifiée** en particulier si le nombre d'événements est faible ou lorsque la valeur de S(t) est voisin de 1 ou de 0. Si l'on utilisait la formule précédente les bornes de l'intervalle de confiance (IC) dépassent 0 ou 1.
- Rothman propose une meilleure estimation permettant d'éviter ce problème. L'IC de Rothman conduit à des valeurs des bornes asymétriques par rapport à la valeur de S(t). **La formule de l'IC de Rothman**, est la suivante :

$$\frac{M}{M+Z^2_{(\alpha/2)}}\left[S_{(t)}+\frac{Z^2_{(\alpha/2)}}{2\,M}\pm Z^2_{(\alpha/2)}\sqrt{\operatorname{Var} S_{(t)}+\frac{Z^2_{(\alpha/2)}}{4\,M^2}}\right]$$

avec $\quad M=\frac{S_{(t)}\left(1-S_{(t)}\right)}{\operatorname{Var} S_{(t)}}$

1

Méthode de Kaplan-Meier exemple

Il s'agit d'un essai clinique randomisé comparant deux traitements A et B dont le critère principal d'efficacité clinique est la survie globale (

- Plusieurs centres participent à l'essai
- Le nombre de patients recevant le traitement A est de 340
- Le nombre de patients recevant le traitement B est de 370
- La date d'origine = date de randomisation
- Durée de suivi = 1 an
- Les données sont présentées dans les tableaux de calcul des taux de survie

On classe les temps de participation (en jours) par ordre croissant pour estimer les
taux de survie.

Méthode de Kaplan-Meier TTT A

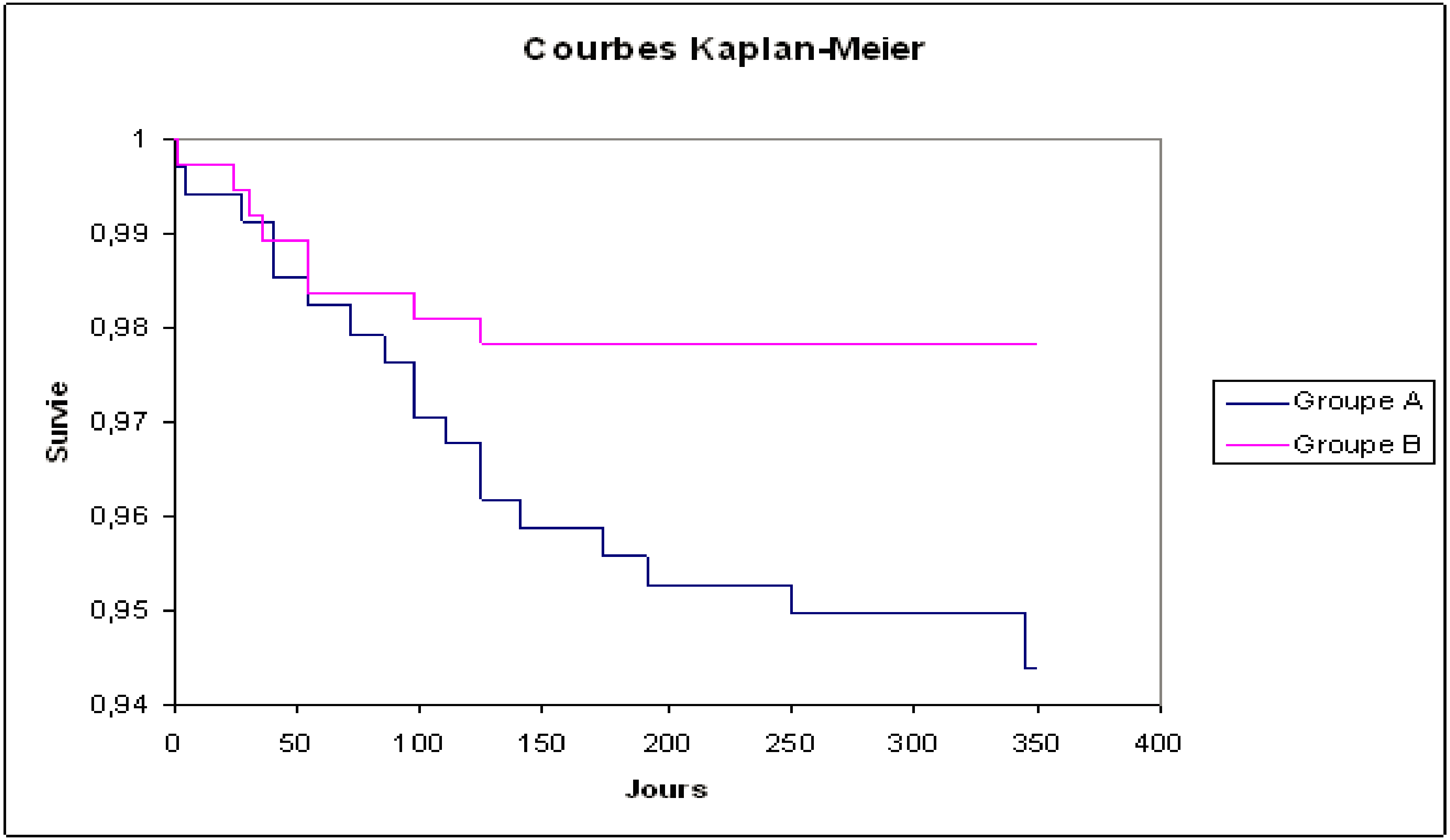
Durée en jours	Etat	Nb de sujets exposés $n(t_i)$	Nb de décès $d(t_i)$	Nb de PV	$S_{t_i/t_{i-1}}$	S_{t_i}	Var $S(t_i)$ x 104
2	DCD	340	1	0	0.9971	0.9971	0.0894
5	DCD	339	1	0	0.9971	0.9941	0.4780
25	VV	338	0	1	1		
28	DCD	337	1	0	0.9970	0.9912	0.266
34	VV	336	0	1	1		
40	DCD	335					
40	DCD		2	0	0.9911	0.9853	0.4440
55	DCD	333	1	0	0.9970	0.9823	0.5315
72	DCD	332	1	0	0.9970	0.9791	0.6190
85	DCD	331	1	0	0.9970	0.9764	0.7678
97	DCD	330					
97	DCD		2	0	0.9940	0.9705	0.8765
110	DCD	328	1	0	0.9970	0.9676	0.9616
116	VV	327	0	1	1		
125	DCD	326		0			
125	DCD		2		0.9909	0.9617	1.316
140	DCD	324	1	0	0.9969	0.9587	1.2156
162	VV	323	0	1	1		
174	DCD	322	1	0	0.9969	0.9558	1.2901
192	DCD	321	1	0	0.9969	0.9528	1.3840
220	VV	320	0	1	1		
250	DCD	319	1	0	0.9969	0.9498	1.4658
310	DCD	318	1	0	0.9969	0.9498	1.5485
321	VV	317	0	1	1		
345	DCD	316	1	0	0.9968	0.9438	1.6321

1

Méthode de Kaplan-Meier TTTB

Durée en jours	Etat	Nb de sujets exposés $n(t_i)$	Nb de décès $d(t_i)$	Nb de PV	S_i	St_i	Var $S(t_i)$ x 104
2	DCD	370	1	0	0.9973	0.9973	0.0762
10	VV	369	0	1	1		
24	DCD	368	1	0	0.9973	0.9946	0.1523
31	DCD	367	1	0	0.9973	0.9919	0.2285
36	DCD	366	1	0	0.9973	0.9891	0.4438
48	VV	365	0	1	1		
55	DCD						
55	DCD	364	2	0	0.9945	0.9837	0.4542
97	DCD	362	1	0	0.9972	0.9809	0.5286
125	DCD	361	1	0	0.9972	0.9782	0.6030
130	VV	360	0	1	1		
175	VV	359	0	1	1		
182	VV	358	0	1	1		
330	VV	357	0	1	1		
350	VV	356	0	1	1		

Courbe Kaplan-Meier TTTB



1

Notion de médiane de survie

- Dans les publications scientifiques lorsque les courbes de survie ne sont pas fournies on indique dans les tableaux présentant les résultats des valeurs de S(ti) correspondant à des valeurs de ti ayant une signification clinique
- D'autre part on précise parfois la valeur de la « médiane de survie » qui correspond à la **valeur de ti pour lequel S(ti) = 0.5** c'est-à-dire pour lequel la probabilité d'être vivant est égale à 50%

Méthode actuarielle

- Le principe de base = Kaplan-Meier mais les intervalles de temps ne sont plus définis par la survenu des événements (décès) et **le calcul des probabilités conditionnelles reposent sur des hypothèses**.
- **Les intervalles de temps sont ici déterminés** *a priori* et non fixés par la survenue d'un décès. Le pas de l'intervalle est déterminé a priori et doit être pertinent par rapport à **l'objectif** et à la **durée (recul)** de l'étude.
- Le calcul des probabilités conditionnelles repose sur les **hypothèses** suivantes :
  - Les sujets « **censurés** » et les **décès** se **distribuent uniformément** dans l'intervalle de temps.
  - Les sujets « **censurés** » sont donc considérés **exposés au risque** en moyenne **pendant la moitié de l'intervalle**

Méthode actuarielle

- On situe, la fin de l'« histoire » de chaque sujet dans l'un des intervalles. Pour chaque intervalle [ti, ti+1[on a donc
- Vi = nombre de sujets vivants au début de l'intervalle
- Di = nombre de décès dans l'intervalle
- Li = nombre de sujets dont le temps de participation se termine dans l'intervalle (Perdus de Vue (PV) ou Exclus vivant (EV) c'est-à-dire "censurés à droite")
- Ni = nombre de sujets exposés au risque pendant l'intervalle

$$\boxed{N_i = V_i - \frac{1}{2} L_i}$$

Méthode actuarielle

- **Variance et intervalle de confiance**
- L'expression de la variance de Greenwood est la suivante :

$$\text{Var } S_{ti+1} = S^2_{ti+1}\left[\frac{1-S_{t1}}{N_0 x S_{t1}} + \frac{1-S_{t2}}{N_1 x S_{t2}} + \ldots + \frac{1-S_{ti+1}}{N_i x S_{ti+1}}\right]$$

- Ou N0 représente le nombre de sujets pour le premier intervalle [t0 - t1[, N1 le nombre de sujets pour l'intervalle suivant [t1 - t2[et ainsi de suite jusqu'à l'intervalle [ti - ti+1[.
- Les intervalles de confiance sont calculés comme pour la méthode de Kaplan-Meier en utilisant la formule de Rothman ou en supposant que S(t) est gaussienne.

1

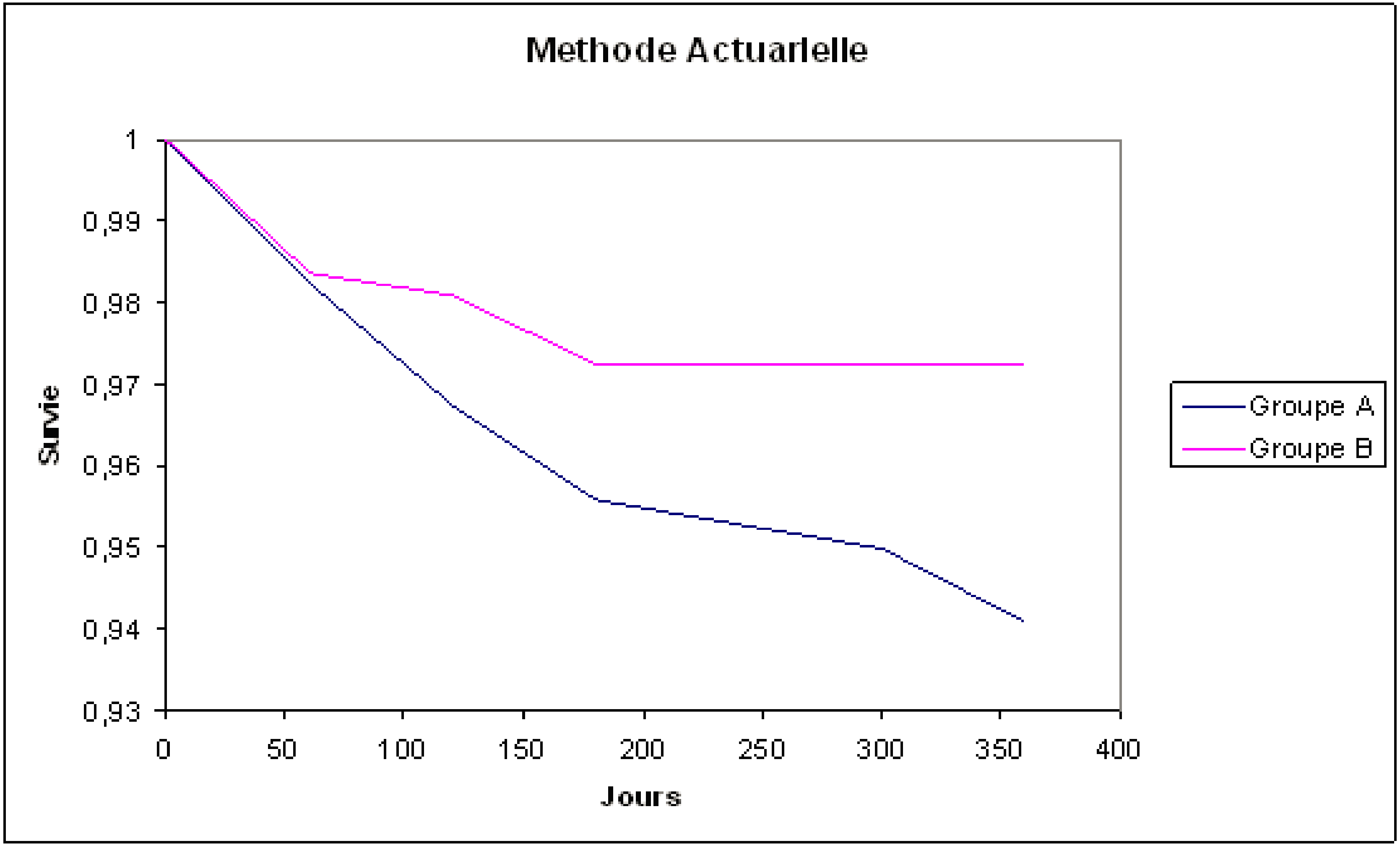
Méthode actuarielle TTTA

Durée (en jours)	Exposés au début de l'intervalle	Décès dans l'intervalle	PV dans l'intervalle	Exposés pendant l'intervalle	$S(t_i/t_{i-1})$	$S(t_i)$	Var $S(t_i)$ x 104
0-60	340	6	2	339	0.9823	0.9823	0.3436
61-120	332	5	1	331.5	0.9849	0.9676	0.9616
121-180	326	4	1	325.5	0.9877	0.9558	1.2991
181-240	321	1	1	320.5	0.9969	0.9528	1.3830
241-300	319	1	0	319	0.9969	0.9498	1.726
301-360	318	3	1	317.5	0.9906	0.9408	1.7140
361	314	0	0	314	1	0.9408	1.7140

Méthode actuarielle TTTB

Durée (en jours)	Exposés au début de l'intervalle	Décès dans l'intervalle	PV dans l'intervalle	Exposés pendant l'intervalle	St/t_{i-1}	St_i	Var $S(t) \times 10^4$
0-60	370	6	2	369	0.9837	0.9837	0.4542
61-120	362	1	0	362	0.9972	0.9810	0.5286
121-180	361	1	2	360	0.9972	0.9725	0.6030
181-240	358	0	0	358	1	0.9725	0.6030
241-300	358	0	0	358	1	0.9725	0.6030
301-360	358	0	2	357	1	0.9725	0.6030
361-365	357	0	0	357	1	0.9725	0.6030

Courbe de survie actuarielle TTTB



1

comparaison des courbes de survie

- **Taux survie à** un temps donné, cliniquement pertinent. Pour comparer par exemple le taux de survie à 3 ans de deux traitements, SA(t=3ans) et SB(t=3ans), on pourra utiliser le test de l'écart réduit sous l'hypothèse de distribution gaussienne des taux de survie

$$z = \frac{S_A - S_B}{\sqrt{Var(S_A) + Var(S_B)}}$$

- **test du Log-Rank**
- *Hypothèse nulle (H0)* : SA(t) = SB(t) pour tout t>0
- *Hypothèse alternative (H1)* : SA(t) ≠ SB(t)

Paramètre et loi statistique du test

- Le test statistique est Khi-deux. Sous l'hypothèse H0 on calculera le nombre de décès (ou événements) « attendus » dans chacun des groupes (nombre de décès estimés noté E).
- On comparera ensuite le nombre de décès (ou événements) observés (O) aux nombre de décès attendus (E)

Exemple Test du Log rank

- Ni=nBi+nAi nAi et nBi sont des effectifs au temps ti dans le groupe de traitement A et B;
- OAi et Obi sont le nombre d'événement d'intérêt observer dans le groupe de traitement A et B au temps ti.
- Eai et Ebi sont le nombre d'évenement attendu ou théorique dans le groupe de traitement A et B au temps ti

$$E_{A_i} = \frac{n_{A_i}}{n_{A_i} + n_{B_i}} \times \left(O_{A_i} + O_{B_i}\right) = D_i \times \frac{n_{A_i}}{N_i}$$

$$E_{B_i} = \frac{n_{B_i}}{n_{A_i} + n_{B_i}} \times \left(O_{A_i} + O_{B_i}\right) = D_i \times \frac{n_{B_i}}{N_i}$$

$$\frac{D_i}{N_i} = \frac{O_{A_i} + O_{B_i}}{n_{A_i} + n_{B_i}}$$

avec $$\underbrace{E_{A_i} + E_{B_i}}_{\text{Décès attendus à } t_i} = \underbrace{O_{A_i} + O_{B_i}}_{\text{Décès observés à } t_i}$$

Exemple Test du Log rank

- On effectue les calculs des EAi et EBi pour tous les ti. On peut alors obtenir le nombre total de décès attendus dans chaque groupe en faisant la somme des EAi et des Ebi

$$E_A = \sum_i E_{A_i} \qquad O_A = \sum_i O_{A_i}$$

$$E_B = \sum_i E_{B_i} \qquad O_B = \sum_i O_{B_i}$$

- On calcul ensuite un Khi2 a k dégré de liberté k= nombre de courbes à comparer
- *Conditions de validité du test:* EA ≥ 5 ; EB ≥ 5

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

1

Calcul du risque relatif

- On peut quantifier cette différence en utilisant le risque relatif de décéder d'un groupe par rapport à un autre

$$RR = \frac{\sum_{i=1}^{k} o_{B_i}\left(n_{A_i} - o_{A_i}\right)/N_i}{\sum_{i=1}^{k} o_{A_i}\left(n_{B_i} - o_{B_i}\right)/N_i}$$

1

Exemple Test du Log rank

Durée en jours	Décès observés		Sujets exposés		Décès attendus	
	O_{Bi}	O_{Ai}	n_{Bi}	n_{Ai}	E_{Bi}	E_{Ai}
2	1	1	370	340	1.042	0.958
5	0	1	369	339	0.521	0.479
24	1	0	368	338	0.521	0.479
28	0	1	367	337	0.521	0.479
31	1	0	367	336	0.521	0.479
36	1	0	366	335	0.522	0.479
40	0	2	365	335	1.043	0.957
55	2	1	364	333	1.567	1.133
72	0	1	362	332	0.522	0.478
85	0	1	362	331	0.522	0.478
97	1	2	362	330	1.569	1.431
110	0	1	361	328	0.524	0.476
125	1	2	361	326	1.576	1.424
140	0	1	359	324	0.526	0.474
174	0	1	359	322	0.527	0.473
192	0	1	357	321	0.526	0.474
250	0	1	357	319	0.528	0.472
310	0	1	357	318	0.529	0.471
345	0	1	356	316	0.530	0.470
360	0	1	355	315	0.530	0.470
Total	**8**	**20**			**14.667**	**13.333**

1

Exemple Test du Log rank

$$E_{A_i} = \frac{n_{A_i}}{n_{A_i} + n_{B_i}} \times \left(O_{A_i} + O_{B_i}\right) = D_i \times \frac{n_{A_i}}{N_i} = \frac{333}{333+364} \times (1+2) = 1{,}133$$

$$E_{B_i} = \frac{n_{B_i}}{n_{A_i} + n_{B_i}} \times \left(O_{A_i} + O_{B_i}\right) = D_i \times \frac{n_{B_i}}{N_i} = \frac{364}{333+364} \times (1+2) = 1{,}567$$

$$O_A + O_B = E_A + E_B = 20 + 8 = 13{,}\ 333 + 14{,}667 = 28$$

$$\boxed{\chi^2_{cal} = \frac{(8-14.667)^2}{14.667} + \frac{(20-13.333)^2}{13.333} = 6.364}$$

Merci