Données longitudinales et modèles de survie

3. Courbes de survie

André Berchtold

Département des sciences économiques, Université de Genève

Cours de Master

UNIVERSITÉ DE GENÈVE

FACULTÉ DES SCIENCES ÉCONOMIQUES ET SOCIALES
Département des sciences économiques

Plan du cours

Plan du cours

Objectif

- Représenter graphiquement l'évolution de la fonction de survie, $S(t)$, au fil du temps.
- Le temps est découpé en q intervalles :

$$[t_1, t_2[, [t_2, t_3[, \ldots, [t_k, t_{k+1}[, \ldots, [t_q, t_{q+1}[$$

- Les intervalles sont en général de longueur égale, $t_{k+1} - t_k = 12$ mois par exemple.
- Le moment exact où se produit un événement n'est pas considéré. Seul l'intervalle durant lequel il a eu lieu est utilisé dans les calculs.

Notations

- Toutes les données observées, censurées ou non, doivent être prises en compte dans les calculs.
- Notations :
  - n_k : nombre de personnes en séjour (survivantes) au début de l'intervalle k
  - d_k : nombre d'événements survenant dans l'intervalle k
  - w_k : nombre de données censurées à droite dans l'intervalle k

Estimation du hasard (risque instantané) (1)

- Le hasard h_k est calculé pour chaque intervalle de temps considéré : $\hat{h}_k = P(T < t_{k+1} | T \geq t_k)$
- Hypothèse 1 : La moitié des données censurées durant un intervalle restent exposées au risque dans cet intervalle.
- Cela revient à considérer que les censures sont distribuées de façon équiprobable tout au long de l'intervalle.
- Le nombre de personnes exposées au risque durant l'intervalle k passe donc déjà de n_k à $n_k - w_k/2$.

Estimation du hasard (risque instantané) (2)

- Hypothèse 2 : La moitié des données subissant un événement durant un intervalle ne sont plus exposées au risque dans cet intervalle.
- Cela revient à considérer que les événements sont distribués de façon équiprobable tout au long de l'intervalle.
- Le nombre de personnes devant être utilisées pour le calcul du hasard est donc $n_k - w_k/2 - d_k/2$ et on obtient

$$\hat{h}_k = \frac{d_k}{n_k - w_k/2 - d_k/2}$$

Estimation de la probabilité de survie

- S_k est la probabilité de survie au début de l'intervalle k.
- S_1=100%.
- Sachant que $P(T \geq t_{k+1} | T \geq t_k) = 1 - \hat{h}_k$, on en déduit

$$\begin{aligned}
\boxed{\hat{S}_k} &= P(T \geq t_k | T \geq t_{k-1}) \cdot P(T \geq t_{k-1} | T \geq t_{k-2}) \cdots \\
&= (1 - \hat{h}_{k-1}) \cdot (1 - \hat{h}_{k-2}) \cdots (1 - \hat{h}_1) \\
&= \prod_{i=1}^{k-1} (1 - \hat{h}_i) \\
&= \boxed{(1 - \hat{h}_{k-1}) \hat{S}_{k-1}}
\end{aligned}$$

Autres relations utiles

- De

$$\hat{S}_k = (1 - \hat{h}_{k-1})\hat{S}_{k-1}$$

on déduit

$$\hat{h}_{k-1} = \frac{\hat{S}_{k-1} - \hat{S}_k}{\hat{S}_{k-1}} = 1 - \frac{\hat{S}_k}{\hat{S}_{k-1}}$$

d'où

$$\hat{h}_k = \frac{\hat{S}_k - \hat{S}_{k+1}}{\hat{S}_k} = \frac{\hat{p}_k}{1 - \sum_{i=1}^{t-1} \hat{p}_i}$$

- Par ailleurs

$$\hat{p}_k = \hat{S}_k - \hat{S}_{k+1} = \hat{h}_k \, \hat{S}_k = \hat{h}_k \prod_{i=1}^{k-1} (1 - \hat{h}_i)$$

Exemple Breast Cancer (1)

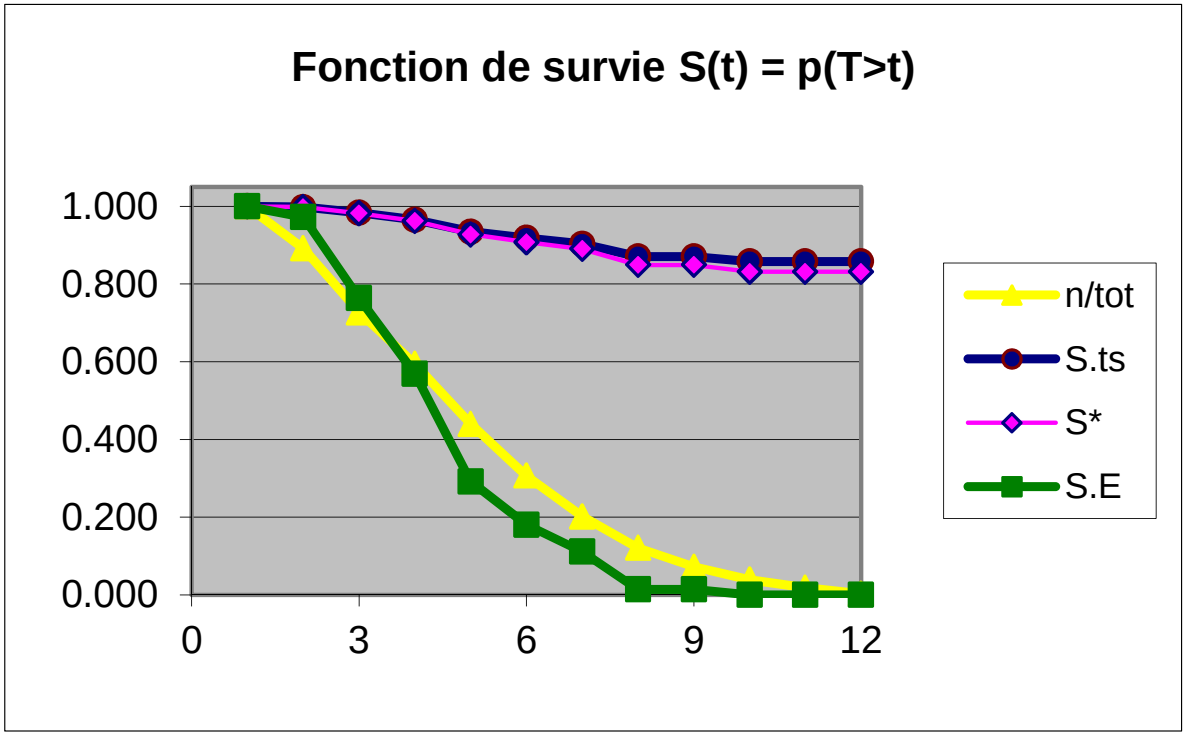
- Fichier livré avec SPSS concernant le risque de décéder d'un cancer du sein chez 1207 femmes.
- Variables :
  - AGE : Age (years)
  - PATHSIZE : Pathologic Tumor Size (cm)
  - LNPOS : Positive Axillary Lymph Nodes
  - HISTGRAD : Histologic Grade
  - ER : Estrogen Receptor Status
  - PR : Progesterone Receptor Status
  - STATUS : Status (0 : censored, 1 : died)
  - PATHSCAT : Pathological Tumor Size (Categories)
  - LN_YESNO : Lymph Nodes ?
  - TIME : Time (months)

Exemple Breast Cancer (2) : (n=1207)

t_k	n_k	d_k	w_k	$n_k - \frac{w_k}{2}$	$n_k - \frac{w_k}{2} - \frac{d_k}{2}$	$\hat{h}_k$	$1 - \hat{h}_k$	$\hat{S}_k$
1	1207	2	129	1142.5	1141.5	0.0018	0.9982	1.0000
2	1076	15	183	984.5	977	0.0154	0.9846	0.9982
3	878	14	147	804.5	797.5	0.0176	0.9824	0.9829
4	717	20	166	634	624	0.0321	0.9679	0.9657
5	531	8	153	454.5	450.5	0.0178	0.9822	0.9347
6	370	5	121	309.5	307	0.0163	0.9837	0.9181
7	244	7	91	198.5	195	0.0359	0.9641	0.9032
8	146	0	59	116.5	116.5	0.0000	1.0000	0.8707
9	87	1	39	67.5	67	0.0149	0.9851	0.8707
10	47	0	25	34.5	34.5	0.0000	1.0000	0.8577
11	22	0	19	12.5	12.5	0.0000	1.0000	0.8577
12	3	0	3	1.5	1.5	0.0000	1.0000	0.8577
		72	1135					0.8577

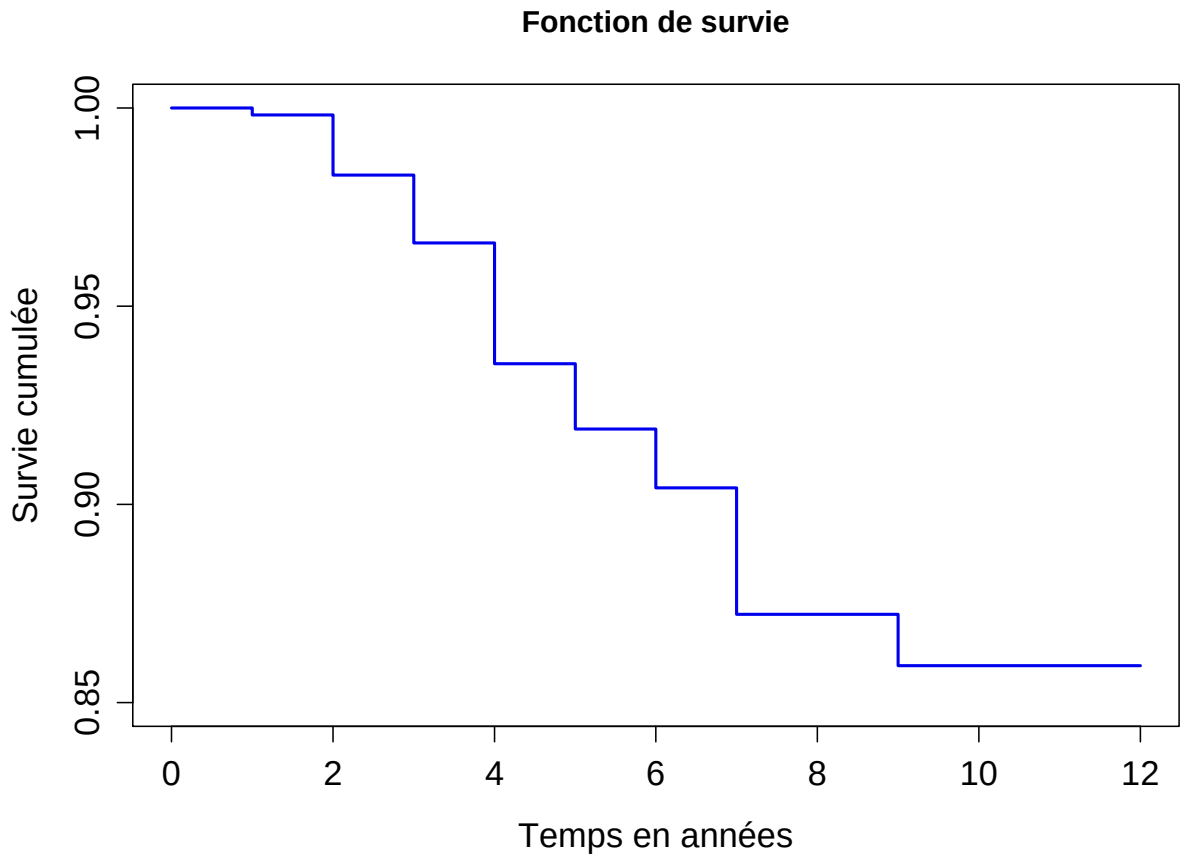
d_k et w_k nombre d'événements (décès) et de censures en k
$\hat{h}_k$ risque (hasard) par période (année)
$\hat{S}_k$ probabilité de survie en début de période

Exemple Breast Cancer (3)



- $S.ts = \hat{S}$: estimation selon table de survie
- S^* : sans tenir compte des deux hypothèses
- $S.E$: en supprimant les données censurées des calculs
- $n/tot = n_k/n$: proportion de survivants au début de l'intervalle k

Exemple Breast Cancer (4)



Version alternative de la fonction de survie

- Les logiciels statistiques considèrent généralement que même si le calcul du hasard doit tenir compte de la déduction $-d_k/2$, il n'en va pas de même avec la probabilité de survie qui elle ne tient pas compte de cette déduction.
- Par conséquent, la survie donnée par les logiciels statistiques est calculée comme si tous les événements d'un intervalle avaient lieu à la fin de celui-ci.
- Il s'agit donc d'une version conservatrice de la fonction de survie.

Exemple Breast Cancer (5)

	nsubs	nlost	nrisk	nevent	surv	pdf	hazard
0-1	1207	129	1142.5	2	1.000	0.00175	0.00175
1-2	1076	183	984.5	15	0.998	0.01521	0.01524
2-3	878	147	804.5	14	0.983	0.01711	0.01740
3-4	717	166	634.0	20	0.966	0.03047	0.03155
4-5	531	153	454.5	8	0.935	0.01647	0.01760
5-6	370	121	309.5	5	0.919	0.01485	0.01616
6-7	244	91	198.5	7	0.904	0.03188	0.03526
7-8	146	59	116.5	0	0.872	0.00000	0.00000
8-9	87	39	67.5	1	0.872	0.01292	0.01481
9-10	47	25	34.5	0	0.859	0.00000	0.00000
10-11	22	19	12.5	0	0.859	0.00000	0.00000
11-12	3	3	1.5	0	0.859	NA	NA

Exemple Breast Cancer (6)

	se.surv	se.pdf	se.hazard
0-1	0.00000	0.00124	0.00124
1-2	0.00124	0.00390	0.00393
2-3	0.00408	0.00453	0.00465
3-4	0.00605	0.00671	0.00705
4-5	0.00891	0.00577	0.00622
5-6	0.01048	0.00659	0.00722
6-7	0.01224	0.01184	0.01333
7-8	0.01672	NaN	NaN
8-9	0.01672	0.01283	0.01481
9-10	0.02087	NaN	NaN
10-11	0.02087	NaN	NaN
11-12	0.02087	NA	NA

Estimation des densités de probabilité

- Pour l'intervalle k, la probabilité de survenance de l'événement vaut

$$\boxed{\hat{p}_k = \hat{S}_k - \hat{S}_{k+1}}$$

- Si l'on considère que l'intervalle est lui-même composé de sous-intervalles (par exemple une année composée de 12 mois), alors pour un intervalle de longueur $t_{k+1} - t_k$, la densité de probabilité par sous-intervalle vaut

$$\boxed{\hat{f}_k = \frac{\hat{p}_k}{t_{k+1} - t_k}} = \frac{\hat{S}_k - \hat{S}_{k+1}}{t_{k+1} - t_k}$$

Calcul du risque par unité de temps

- Rique par unité de temps (exemple : mois) au sein d'intervalles de plus longue durée (exemple : année).

$$\begin{aligned}
\hat{h}_k^* &= \frac{\hat{f}_k}{(\hat{S}_k + \hat{S}_{k+1})/2} \\
&= \frac{2\hat{h}_k}{(t_{k+1} - t_k)(2 - \hat{h}_k)} \\
&= \frac{d_k/(t_{k+1} - t_k)}{n_k - (w_k + d_k)/2}
\end{aligned}$$

où $\hat{h}_k$ est le risque par intervalle.

Exemple Breast Cancer (7)

- La probabilité et le hasard sont calculés par mois et non par année.

	nsubs	nlost	nrisk	nevent	surv	pdf	hazard
0-12	1207	129	1142.5	2	1.000	0.000146	0.000146
12-24	1076	183	984.5	15	0.998	0.001267	0.001279
24-36	878	147	804.5	14	0.983	0.001426	0.001463
36-48	717	166	634.0	20	0.966	0.002539	0.002671
48-60	531	153	454.5	8	0.935	0.001372	0.001480
60-72	370	121	309.5	5	0.919	0.001237	0.001357
72-84	244	91	198.5	7	0.904	0.002657	0.002991
84-96	146	59	116.5	0	0.872	0.000000	0.000000
96-108	87	39	67.5	1	0.872	0.001077	0.001244
108-120	47	25	34.5	0	0.859	0.000000	0.000000
120-132	22	19	12.5	0	0.859	0.000000	0.000000
132-144	3	3	1.5	0	0.859	NA	NA

Durée moyenne de survie

- La durée moyenne de survie se calcule en admettant qu'une proportion d'individus correspondant à la moyenne des taux de survie en début et fin de période vit la période (intervalle) considérée :

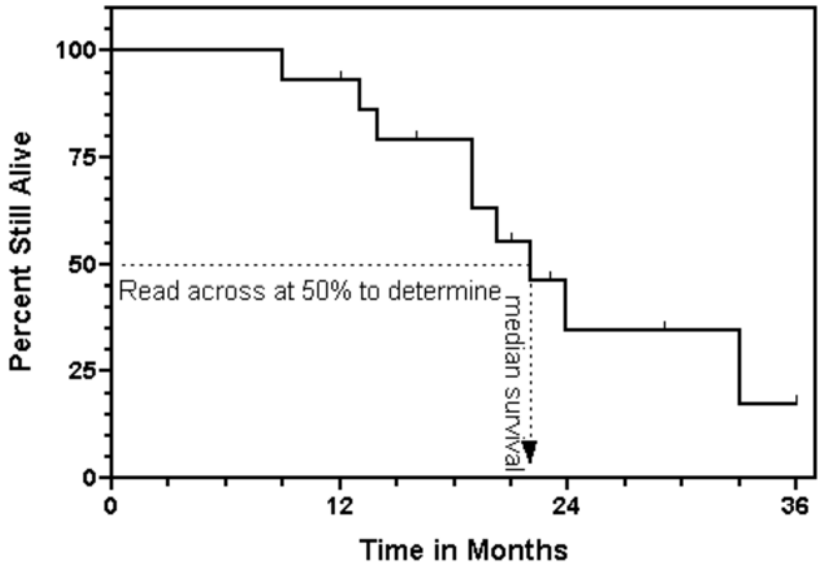
$$\hat{\mu} = \sum_{k=1}^{q} (t_{k+1} - t_k) \frac{\hat{S}_k + \hat{S}_{k+1}}{2}$$

- *Breast Cancer (selon arrondis des slides précédents)* :

$$\hat{\mu} = 12 \sum_{k=1}^{12} \frac{\hat{S}_k + \hat{S}_{k+1}}{2} = 12 \sum_{k=1}^{12} \frac{21.052}{2} = 126.3 \text{ mois}$$

Durée médiane de survie

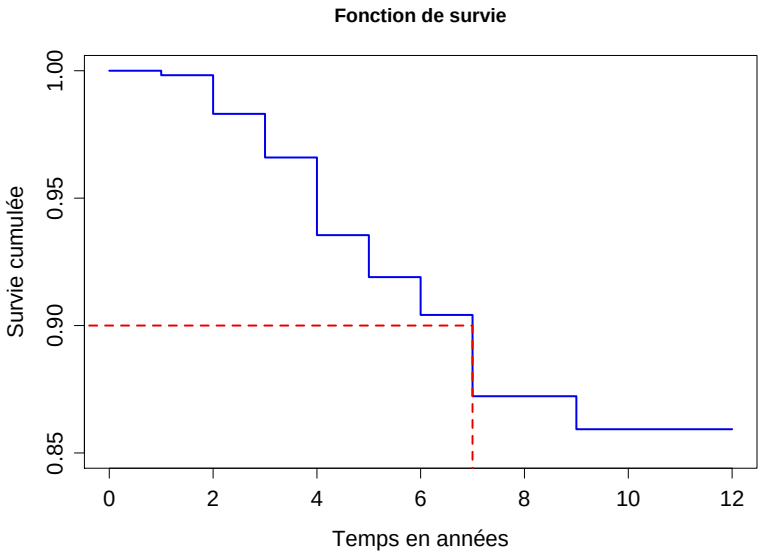
- Durée pour laquelle la valeur de la fonction de survie passe en dessous de 0.5.
- La valeur exacte peut être estimée par approximation linéaire.
- Selon le même principe, il est possible de déterminer la valeur de n'importe quantile.



Exemple Breast Cancer

- La probabilité de survie reste supérieure à 0.5 jusqu'à la durée maximale considérée.
  $\rightarrow$ On ne peut pas déterminer la durée médiane.
- Le quantile 90% est approximé comme

$$q_{0.9} = 7 + \frac{0.904 - 0.9}{0.904 - 0.872}(8 - 7) = 7.125$$



Plan du cours

Hypothèses

- L'échantillon étudié est représentatif de la population considérée.
- Le moment des censures est indépendant de la durée de survie.
  - Si cette hypothèse n'est pas vérifiée, alors la courbe de survie peut amener à des conclusions erronées.
  - L'exemple classique est celui de patients qui sortent d'une étude médicale parce qu'ils sont trop malades pour continuer à suivre le protocole de l'étude.
- Le temps moyen de survie des personnes étudiées doit être le même, indépendamment du moment d'entrée des personnes dans l'étude.

Interprétation d'une courbe

- Si les hypothèses précédentes sont vérifiées, la courbe de survie donne la probabilité qu'une personne survive au fil du temps.
- Les décès (événements) se visualisent par la présence de sauts (barres) verticales sur le graphique. S'il n'y a aucun décès durant un intervalle, alors il n'y a pas de saut à la fin de cet intervalle.

Exemple Yamaguchi (1)

- Données tirées de Yamaguchi (1991). Enquête sur l'après études secondaires dans les pays du bassin du Pacifique.
- 265 étudiants qui ont commencé un cursus universitaire de 4 ans en 1980.
- Durée jusqu'à l'abandon des études universitaires (risque d'abandonner).
- Les variables sont :

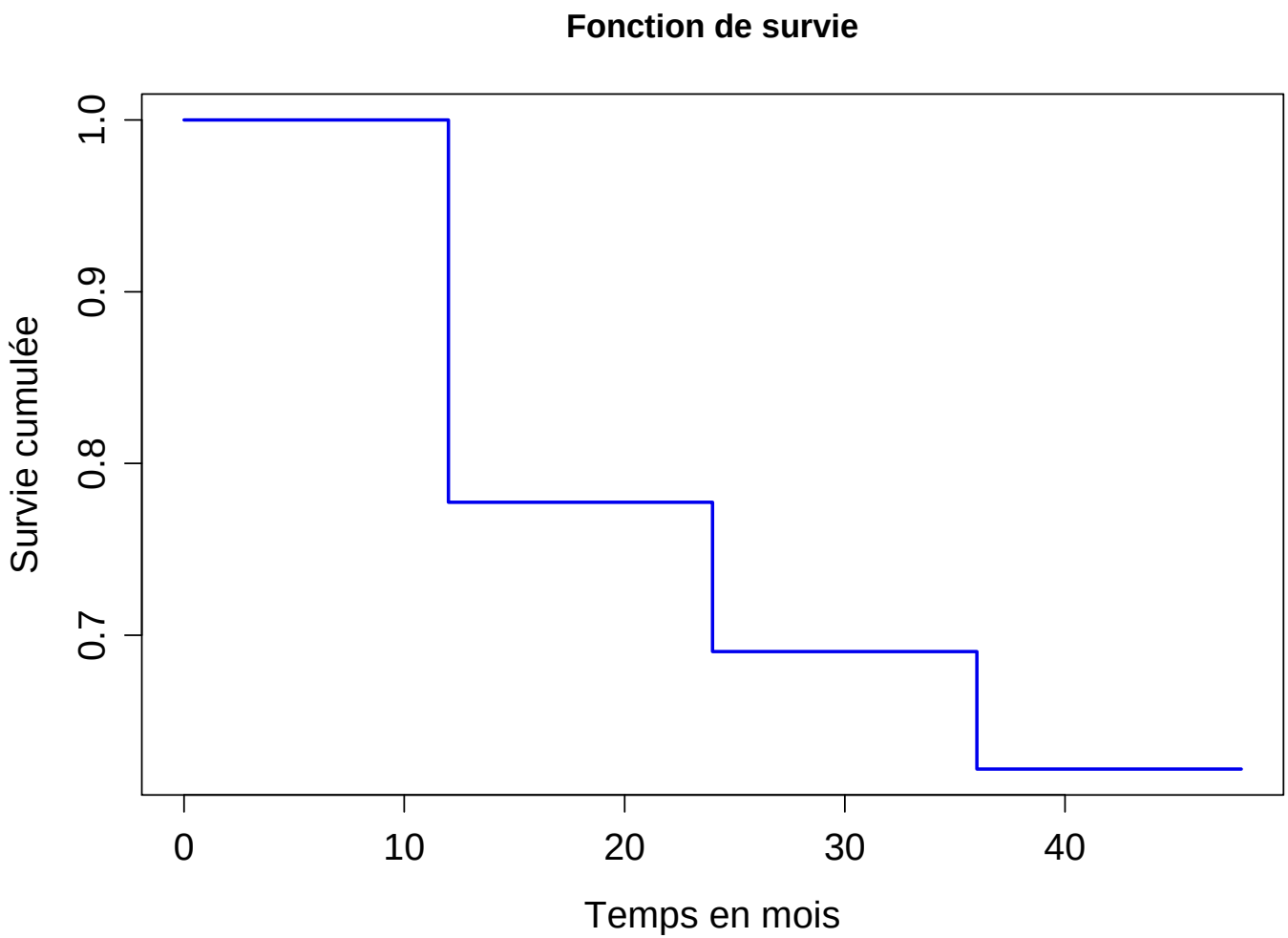
dur	durée jusqu'à l'obtention du diplôme ou l'abandon
evt	abandon (1, ou 0 sinon)
sex	genre : homme (0) ou femme (1)
grd	note moyenne à l'école secondaire : A (meilleure), A-B, B, B-C ou C
prt	études à temps partiel (1, ou 0 sinon)
lag	temps entre fin école secondaire et début études (en mois)
mrg	temps jusqu'au mariage depuis début 80 (en mois)
tms	temps jusqu'au début des études depuis début 80 (en mois)

Exemple Yamaguchi (2)

	nsubs	nlost	nrisk	nevent	surv	pdf
0-12	265	0	265.0	59	1.000	0.01855
12-24	206	0	206.0	23	0.777	0.00723
24-36	183	3	181.5	18	0.691	0.00571
36-48	162	155	84.5	7	0.622	NA

	hazard	se.surv	se.pdf	se.hazard
0-12	0.02088	0.0000	0.00213	0.00270
12-24	0.00985	0.0256	0.00144	0.00205
24-36	0.00870	0.0284	0.00130	0.00205
36-48	NA	0.0298	NA	NA

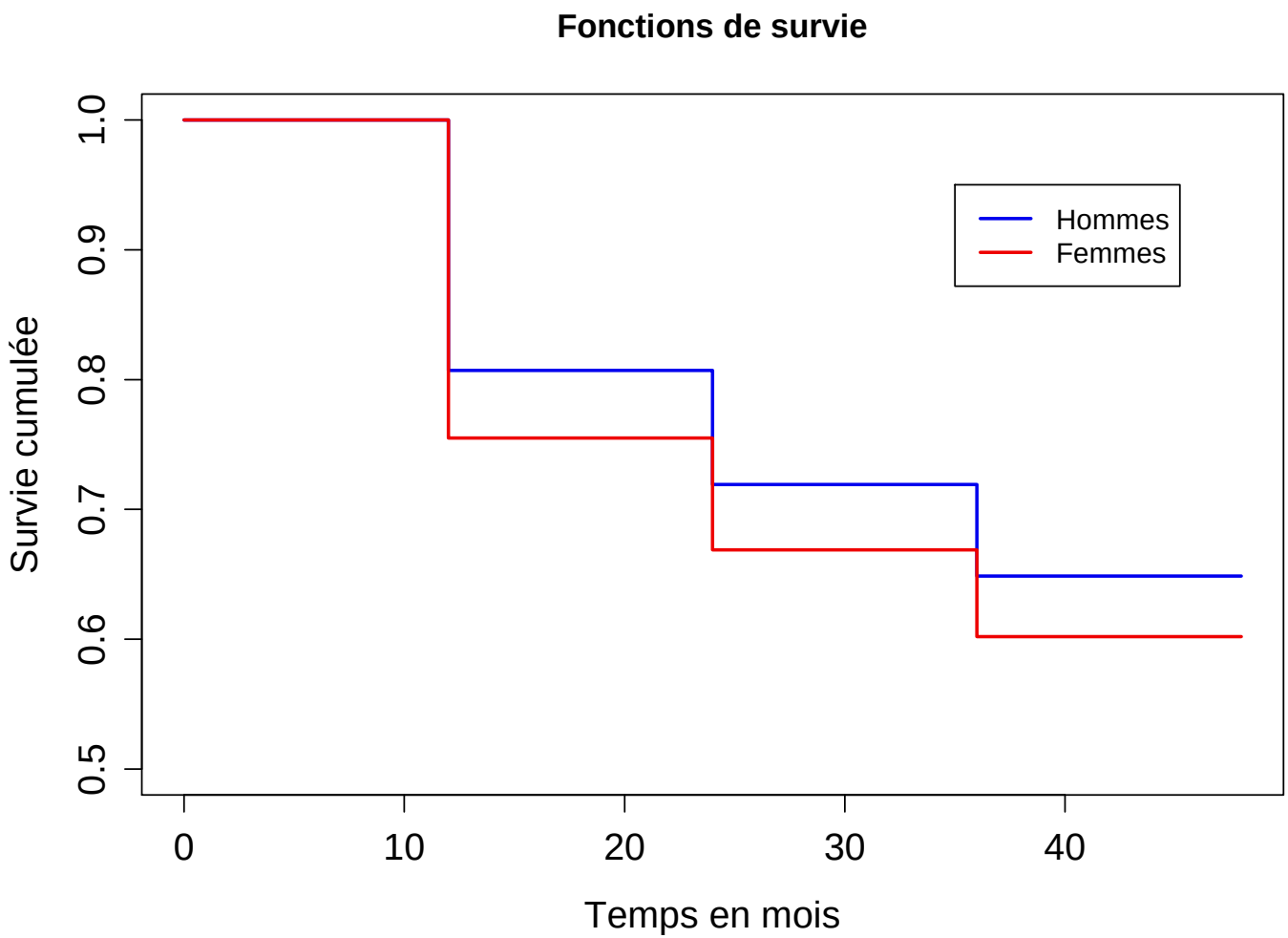
Exemple Yamaguchi (3)



Comparaison de courbes

- Lorsque deux ou plusieurs courbes de survie sont représentées sur un même graphique, la différence verticale entre les courbes donne, à un instant donné, la différence de probabilité de survie entre les groupes.
- La différence horizontale au niveau de la médiane donne la différence des temps médians de survie des deux courbes. Il est aussi possible de considérer d'autres quantiles.
- Attention : Dans le cas d'un événement comme le décès, il est faux d'interpréter une courbe de survie plus haute comme une diminution de la mortalité. Il ne s'agit en réalité que d'un décalage dans le temps de quelque chose qui finira inévitablement par survenir !

Exemple Yamaguchi (4)



Exemple Yamaguchi (5)

- Hommes

	nsubs	nlost	nrisk	nevent	surv
0-12	114	0	114.0	22	1.000
12-24	92	0	92.0	10	0.807
24-36	82	1	81.5	8	0.719
36-48	73	71	37.5	2	0.649

- Femmes

	nsubs	nlost	nrisk	nevent	surv
0-12	151	0	151	37	1.000
12-24	114	0	114	13	0.755
24-36	101	2	100	10	0.669
36-48	89	84	47	5	0.602

Intervalles de confiance

- Une autre méthode permettant 1) de s'assurer de la fiabilité d'une courbe de survie et 2) de comparer des courbes entre-elles consiste à calculer des intervalles de confiance autour de chaque courbe.
- Altman (1991) propose le calcul approximé suivant :

$$\hat{S}_k \pm z_{1-\alpha/2}\sqrt{\frac{1-\hat{S}_k}{n_k}}$$

avec

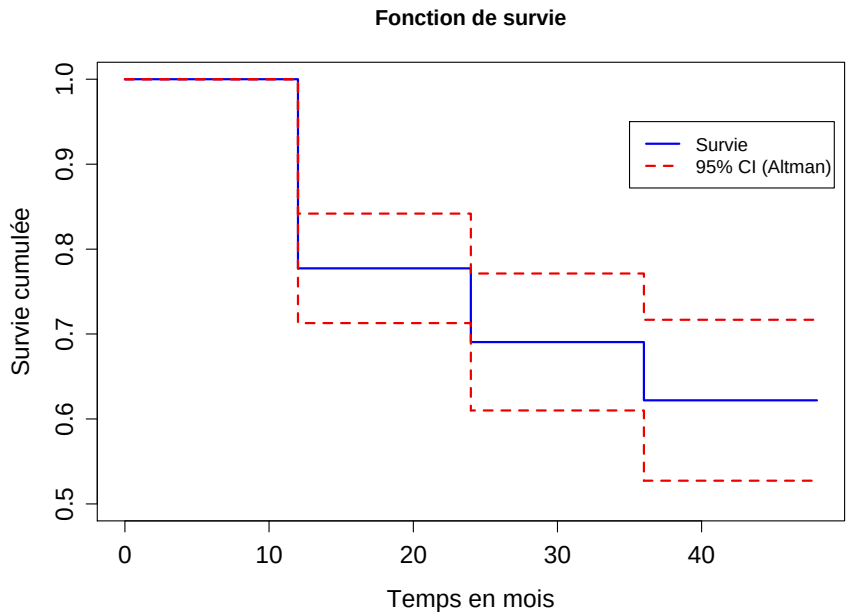
- $\hat{S}_k$: valeur de la courbe de survie au temps k
- α : risque de première espèce
- $z_{1-\alpha/2}$: seuil de la loi normale (1.96 pour α=5%)
- n_k : nombre de survivants au temps k

Disponibilité

- R ne propose pas le calcul automatique d'intervalles de confiance pour les courbes de survie calculées selon la méthode actuarielle.
- Il faut les calculer à part, soit par la formule d'Altman, soit en se basant sur les erreurs standards fournies par R.
- Selon le logiciel utilisé, les résultats peuvent être légèrement différents. Stata par exemple utilise une autre formule de calcul que celle donnée par Altman.

Exemple Yamaguchi (6)

Interval	n_k	d_k	w_k	Survie	R [95% CI]		Altman [95% CI]	
0-12	265	59	0	1.000	1.000	1.000	1.000	1.000
12-24	206	23	0	0.777	0.727	0.827	0.713	0.842
24-36	183	18	3	0.691	0.635	0.746	0.610	0.771
36-48	162	7	155	0.622	0.564	0.681	0.527	0.717



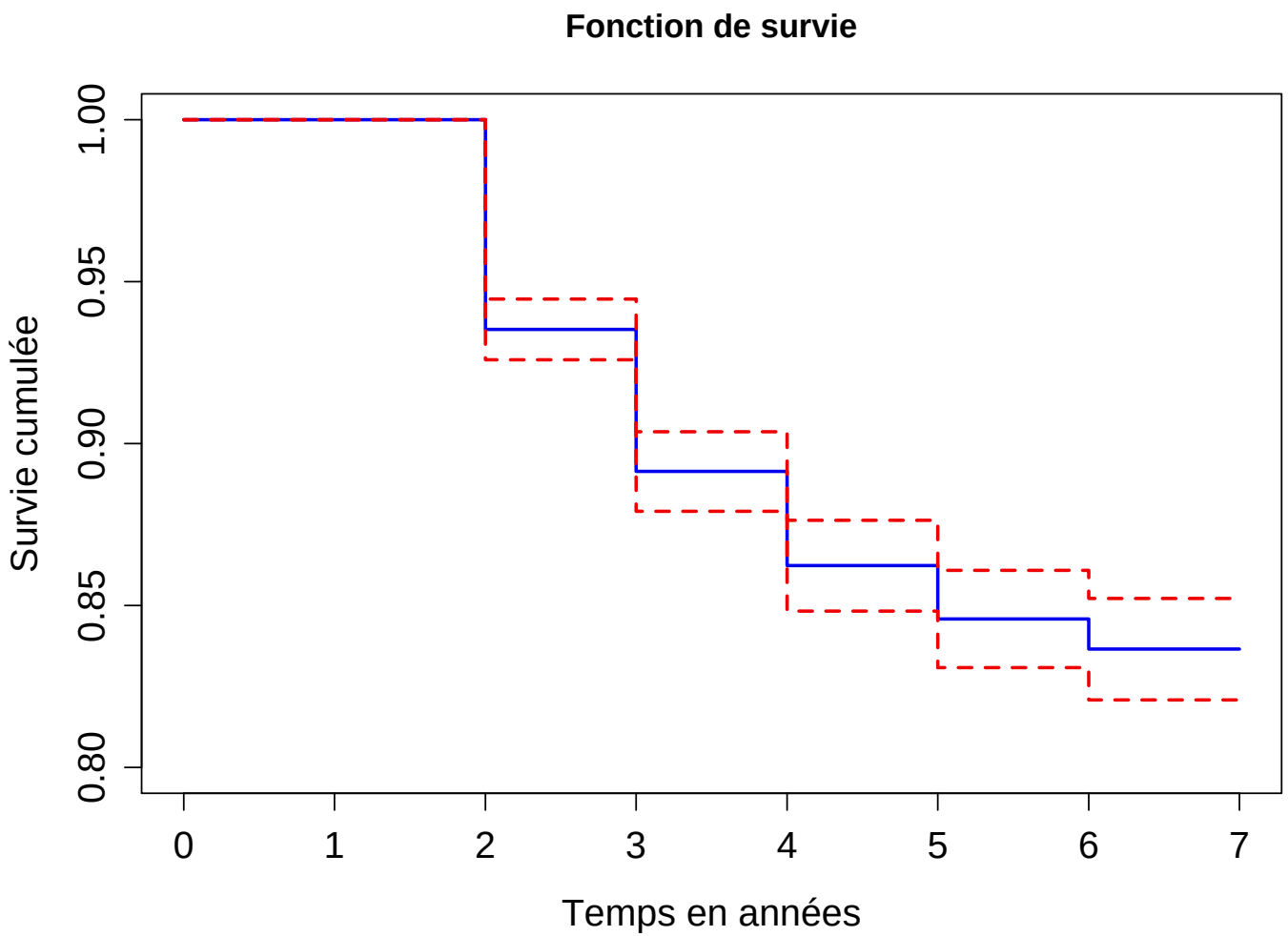
Exemple TREE (1)

- L'enquête TREE (TRansitions de l'Ecole à l'Emploi) est une enquête longitudinale menée en Suisse à partir de 2001 auprès de jeunes qui venaient de terminer l'école obligatoire.
- Nous considérons les participants à l'enquête TREE qui ne consommaient ni tabac, ni cannabis en 2001 (n=3283).
- Nous aimerions déterminer la probabilité de commencer à consommer du cannabis (événement) au fil des années chez ces participants.
- On étudie tout d'abord l'échantillon complet, puis on calcule des courbes par sexe.
- Un intervalle de confiance à 95% est calculé pour chaque courbe à partir des erreurs standard données par R.

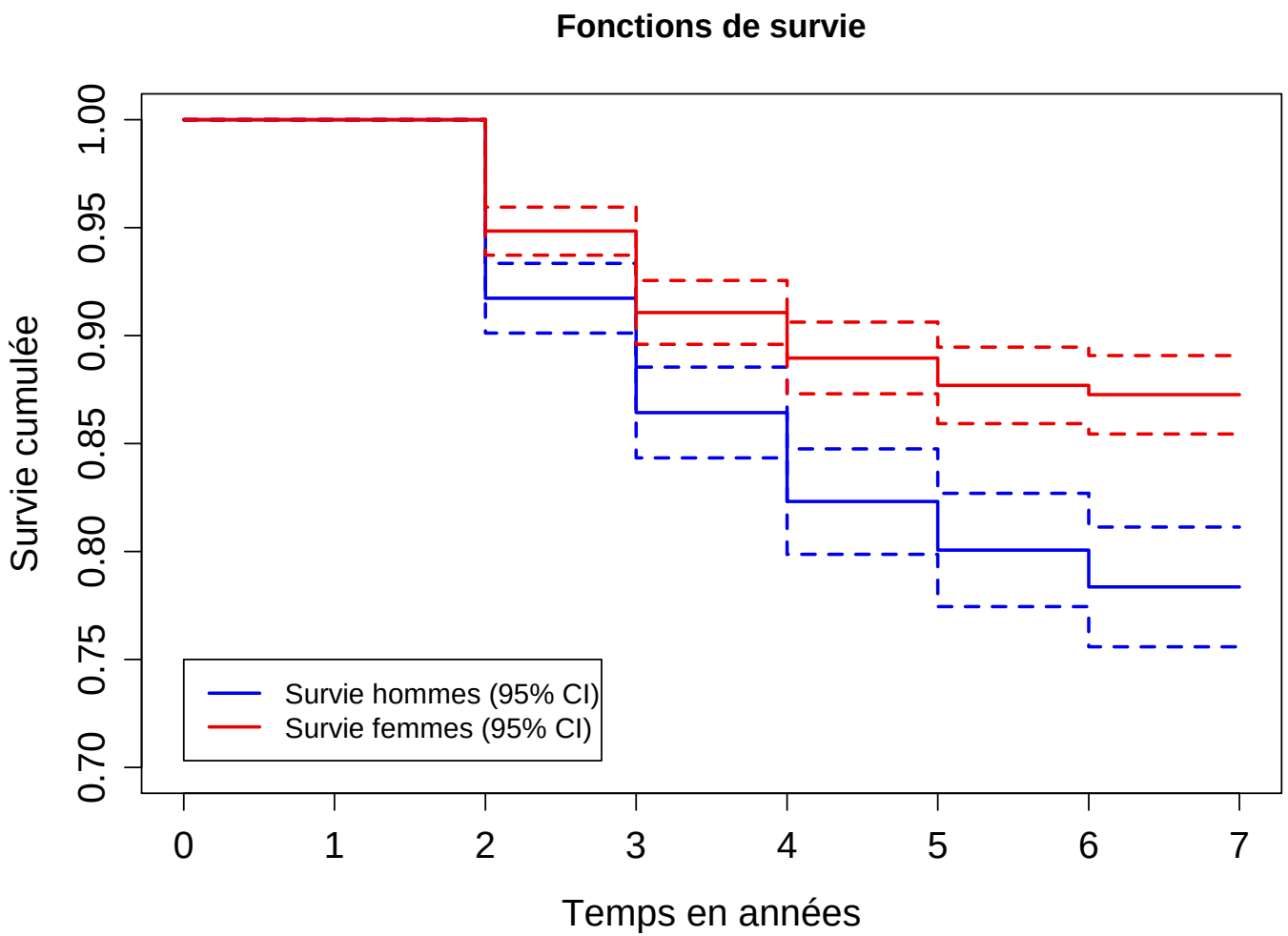
Exemple TREE (2)

	nsubs	nlost	nrisk	nevent	surv	pdf	hazard	se.surv
0-1	3283	358	3104	0	1.000	0.00000	0.0000	0.00000
1-2	2925	541	2654	172	1.000	0.06480	0.0670	0.00000
2-3	2212	242	2091	98	0.935	0.04383	0.0480	0.00478
3-4	1872	246	1749	57	0.891	0.02905	0.0331	0.00628
4-5	1569	206	1466	28	0.862	0.01647	0.0193	0.00716
5-6	1335	129	1270	14	0.846	0.00932	0.0111	0.00767
6-7	1192	1178	603	14	0.837	NA	NA	0.00798

Exemple TREE (3)



Exemple TREE (4)



Principe des tests

- On stratifie la population en sous-groupes en fonction d'un facteur comme le sexe, la classe d'âge, ...
- On estime la fonction de survie séparément pour chaque sous-groupe.
- On compare les différentes fonctions afin de déterminer si le facteur choisi affecte la survie.

Remarques

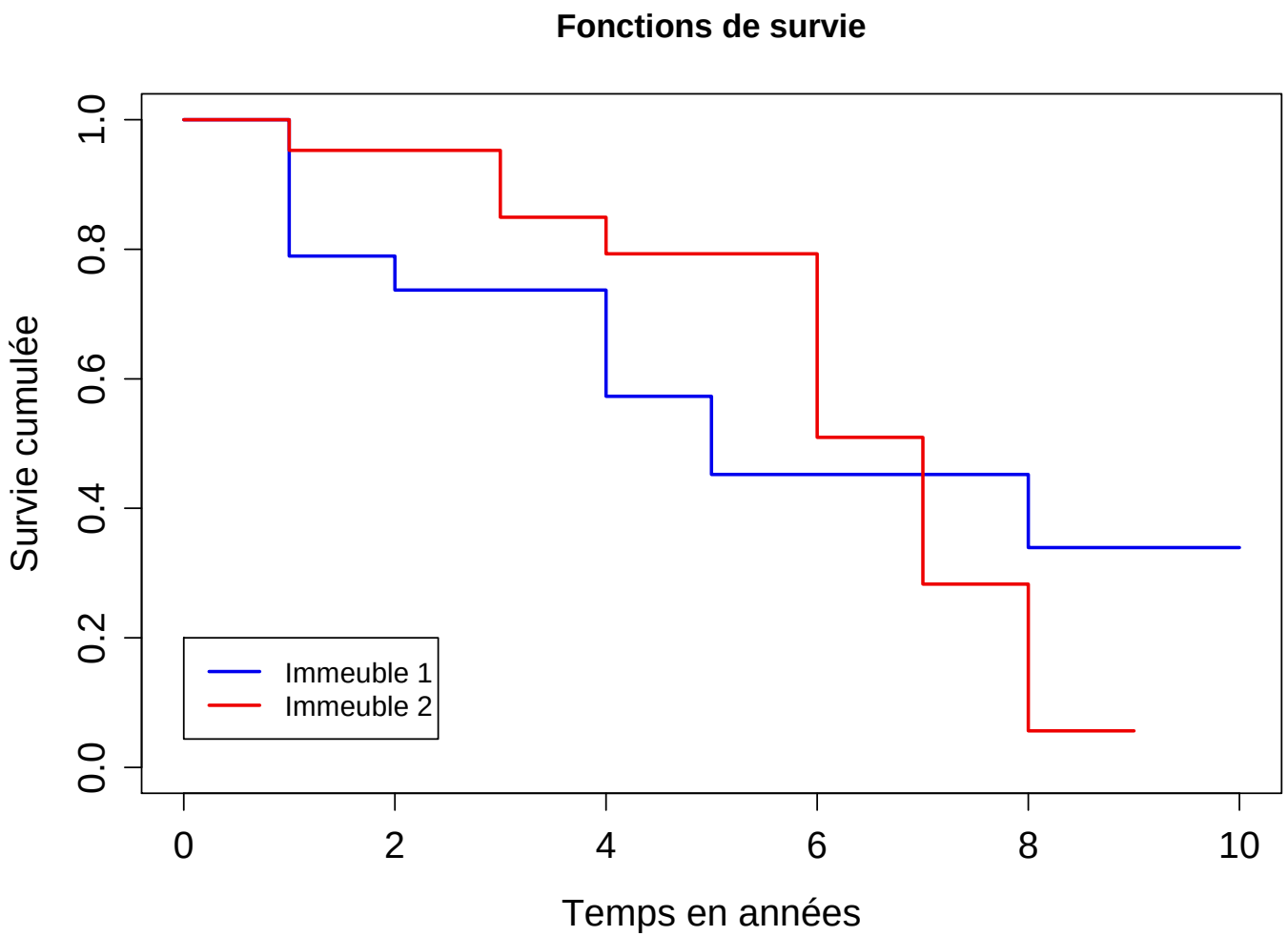
1. Le critère de stratification doit bien entendu être fixe dans le temps (genre, pays de naissance, ...). Si le facteur variait au fil du temps, alors les personnes inclues dans le calcul d'une courbe de survie changeraient d'un endroit à l'autre de la courbe, ce qui ne ferait aucun sens !
2. S'il n'y a pas de données censurées, on peut appliquer des méthodes classiques de comparaison d'échantillons (Wilcoxon, Kruskal-Wallis, etc.), mais en cas de censures il faut appliquer une approche particulière.
3. On ne considère dans un premier temps que le cas de deux sous-groupes.

Exemple des 2 immeubles (1)

- Emigration des habitants de deux immeubles (source : Courgeau & Lelièvre, 1989, p. 66).
- Données observées :

Durée	Immeuble 1			Immeuble 2		
t_i	n_{1i}	d_{1i}	w_{1i}	n_{2i}	d_{2i}	w_{2i}
t1	19	4	0	21	1	0
t2	15	1	0	20	0	0
t3	14	0	0	20	2	3
t4	14	3	1	15	1	0
t5	10	2	1	14	0	0
t6	7	0	1	14	5	0
t7	6	0	0	9	4	0
t8	6	1	4	5	4	0
t9	1	0	0	1	1	0
t10	1	1	0	-	-	-

Exemple des 2 immeubles (2)



Données ordonnées

- La première étape consiste à ordonner dans le temps tous les événements qui se sont produits dans l'un ou l'autre des deux groupes :
  - Soit $t_{11}, t_{12}, t_{13}, \ldots, t_{1m_1}$ les durées de survie observées avant que l'événement ne se produise au sein du groupe 1, et soit $t_{21}, t_{22}, t_{23}, \ldots, t_{2m_2}$ les durées de survie observées avant que l'événement ne se produise au sein du groupe 2.
  - $m = m_1 + m_2$ est le nombre total d'événements observés au sein des deux groupes.
  - L'ensemble de toutes les durées s'écrit alors $(t_1, t_2, \ldots, t_m)$ et il est ordonné de manière à ce que

$$t_1 \leq t_2 \leq \ldots \leq t_m$$

Nombre de cas exposés

- Pour chaque durée t_i, on calcule séparément pour chaque groupe
  - Le nombre de cas exposés au risque à ce moment :

$$n_{1i} + n_{2i} = n_i$$

  - Le nombre de cas qui subissent l'événement à ce moment :

$$d_{1i} + d_{2i} = d_i$$

- On calcule ensuite le nombre théorique de cas qui devraient subir l'événement à chaque moment t_i si les deux courbes de survie étaient identiques, c'est-à-dire s'il y avait indépendance entre la distribution de la survie et l'appartenance à l'un ou l'autre groupe :

$$e_{1i} + e_{2i} = e_i = d_i$$

Exemple des 2 immeubles (3)

Durée	Groupe 1			Groupe 2			Ensemble	
t_i	d_{1i}	n_{1i}	e_{1i}	d_{2i}	n_{2i}	e_{2i}	d_i	n_i
t1	4	19	2.375	1	21	2.625	5	40
t2	1	15	0.429		20	0.571	1	35
t3		14	0.824	2	20	1.176	2	34
t4	3	14	1.931	1	15	2.069	4	29
t5	2	10	0.833		14	1.167	2	24
t6		7	1.667	5	14	3.333	5	21
t7		6	1.600	4	9	2.400	4	15
t8	1	6	2.727	4	5	2.273	5	11
t9		1	0.500	1	1	0.500	1	2
t10	1	1	1.000			0.000	1	1

Calcul des effectifs théoriques (1)

- Sous l'hypothèse H_0 de distributions identiques,

$$d_{gi} \sim \text{hypergéométrique}(n_{gi}, n_i, d_i)$$

où $g = 1, 2$ numérote les groupes.

- Par analogie avec l'utilisation habituelle d'une loi hypergéométrique :
  - n_{gi} : taille de l'échantillon
  - n_i : taille de la population
  - d_i : nombre d'événements dans la population

Calcul des effectifs théoriques (2)

- On a alors les caractéristiques suivantes :

$$e_{gi} = \mathrm{E}(d_{gi}|H_0, d_i, n_i) = n_{gi}\, d_i/n_i$$

$$\sigma^2_{d_{gi}} = \mathrm{Var}(d_{gi}|H_0, d_i, n_i) = n_{gi}\, \frac{d_i(n_i - n_{gi})(n_i - d_i)}{n_i^2(n_i - 1)}$$

- Remarque : le rapport d_i/n_i est le risque instantané h_i.

Exemple des 2 immeubles (4)

Durée	Groupe 1			Groupe 2		
t_i	e_{1i}	$d_{1i} - e_{1i}$	σ^2_{1i}	e_{2i}	$d_{2i} - e_{2i}$	σ^2_{2i}
t1	2.375	1.625	1.119	2.625	-1.625	1.119
t2	0.429	0.571	0.245	0.571	-0.571	0.245
t3	0.824	-0.824	0.470	1.176	0.824	0.470
t4	1.931	1.069	0.892	2.069	-1.069	0.892
t5	0.833	1.167	0.465	1.167	-1.167	0.465
t6	1.667	-1.667	0.889	3.333	1.667	0.889
t7	1.600	-1.600	0.754	2.400	1.600	0.754
t8	2.727	-1.727	0.744	2.273	1.727	0.744
t9	0.500	-0.500	0.250	0.500	0.500	0.250
t10	1.000	0.000	-	0.000	0.000	-
Somme	13.885	-1.885	5.827	16.115	1.885	5.827

Statistique de test (2 groupes)

- La statistique de test compare les nombres de cas observés, d_{gi}, avec les effectifs théoriques correspondant, e_{gi} :

$$Q = \frac{\left(\sum_{i=1}^{m}(d_{1i} - e_{1i})\right)^2}{\sum_{i=1}^{m}\sigma^2_{d_{1i}}} \sim \chi^2_{(1)}$$

- Seul le groupe 1 est utilisé dans le calcul, car la somme des écarts entre effectifs observés et théoriques est identique au signe près dans les 2 groupes. De même, les variances sont les mêmes pour les deux groupes. On aurait pu utiliser de façon équivalente le groupe 2.

Exemple des 2 immeubles (5)

- La statistique vaut

$$\begin{aligned} Q &= \frac{-1.885^2}{5.827} \\ &= 0.610 \end{aligned}$$

- Pour un risque $\alpha = 5\%$, le seuil de rejet de la loi du chi-2 à 1 degré de liberté vaut 3.84.
- L'hypothèse nulle d'égalité des 2 courbes de survie est donc acceptée.

Statistique de test approximée (2 groupes)

- La statistique précédente donne le résultat exact du test. Cependant, il est aussi possible de l'approximer à l'aide de la statistique suivante :

$$Q = \sum_{g=1}^{2} \left(\frac{\left(\sum_{i=1}^{m_g} (d_{gi} - e_{gi}) \right)^2}{\sum_{i=1}^{m_g} e_{gi}} \right) \sim \chi^2_{(1)}$$

- Pour l'exemple des 2 immeubles, on obtient

$$\begin{aligned} Q &= \frac{-1.885^2}{13.885} + \frac{1.885^2}{16.115} \\ &= 0.477 \end{aligned}$$

Trois tests particuliers (1)

- Le test présenté précédemment est en fait seulement l'un des 3 tests couramment utilisés pour comparer les courbes de survie.
- La forme générale de la statistique utilisée par ces 3 tests s'écrit

$$Q = \frac{\left(\sum_{i=1}^{m} \omega_i (d_{1i} - e_{1i})\right)^2}{\sum_{i=1}^{m} \omega_i^2 \sigma_{d_{1i}}^2} \sim \chi^2_{(1)}$$

où les ω_i sont des pondérations permettant de donner plus ou moins d'importance aux différents événements entrant dans le calcul.

Trois tests particuliers (2)

- Différents choix possibles pour les poids ω_i :
  - $\omega_i = 1, \forall i \rightarrow$ **Log Rank**
    (autres noms : Mantel-Haenszel ou Mantel-Cox)
    Tous les événements ont le même poids.
  - $\omega_i = n_i, \forall i \rightarrow$ **Breslow**
    (Wilcoxon, ou Gehan dans le cas de 2 groupes)
    Accorde plus d'importance aux événements précoces, notamment lorsque l'échantillon est grand.
  - $\omega_i = \sqrt{n_i}, \forall i \rightarrow$ **Tarone-Ware**

Quel test choisir ?

- **Log rank** est plus puissant que **Breslow** pour tester les différences lorsque la mortalité est proportionnelle ($h_{1i} = \lambda h_{2i}, \forall i$).
- Sinon, **Breslow** est en général plus puissant, sauf s'il y a beaucoup de données tronquées, en raison de la domination de ces dernières.
- **Tarone-Ware** est un compromis qui s'avère plus efficace dans un grand nombre de situations.

Exemple des 2 immeubles (6)

- Log-rank : Q=0.610
- Breslow : Q=0.318
- Tarone-Ware : Q<0.001
- Le seuil de rejet de ces tests vaut 3.84.
- Dans les 3 cas, l'hypothèse nulle d'égalité des deux courbes de survie est fortement acceptée.
- Il n'y a pas de différence significative entre les deux immeubles en ce qui concerne l'émigration de leurs habitants.
- Attention cependant, car la taille d'échantillon (40) est très petite !

Généralisation à plus que deux courbes

- Dans le cas où l'on veut comparer simultanément 3 courbes de survie ou plus, l'hypothèse nulle signifie l'égalité de toutes les courbes.
- Il y a alors *# de courbes -1* degrés de liberté.
- Les trois tests présentés précédemment peuvent être généralisés à cette situation, mais la statistique de test elle-même devient très complexe, car il est nécessaire de tenir compte non seulement des variances, mais aussi des covariances entre les différentes courbes.
- Il est alors préférable de laisser l'ordinateur faire le calcul ! Dans le cas où l'on désire quand même faire le calcul à la main, il vaut mieux se contenter de la formule simplifiée donnée précédemment, en effectuant la somme sur le nombre de courbes comparées.

Plan du cours

Introduction

- Pas de découpage du temps en intervalles fixes.
- On considère successivement chaque instant t_k où survient soit un événement, soit une censure ($k = 1, 2, \ldots, m$).
- On construit $m + 1$ intervalles de la forme

$$[t_0, t_1[, [t_1, t_2[, \ldots, [t_m, \infty[$$

- Si à l'instant t_k des censures surviennent en même temps que des événements ($d_k > 0$ et $w_k > 0$), on fait l'hypothèse que les censures surviennent juste avant les événements.
$\rightarrow$ nombre d'individus exposés $= n_k - w_k$

Hasard et probabilité de survie

- Le hasard h_k se calcule comme

$$\hat{h}_k = \frac{d_k}{n_k - w_k}$$

- Situations particulières (pas d'événements simultanés) :
  - $d_k = 1$ et $w_k = 0 \rightarrow \hat{h}_k = 1/n_k$
  - $d_k = 0$ et $w_k = 1 \rightarrow \hat{h}_k = 0$

- Estimation de S_k : **Estimateur produit limite**

$$\hat{S}_k = \prod_{i=1}^{k-1} (1 - \hat{h}_i), \quad k > 1$$

- $\hat{S}_1 = 1$, car aucun événement n'a lieu avant t_1.

Variance

- Estimation de la variance (asymptotique) de $\hat{S}_k$:

$$\hat{\sigma}^2_{\hat{S}_k} = \hat{S}_k^2 \sum_{i<k} \frac{d_i}{n_i(n_i - d_i)}$$

- Ecart-type :

$$\hat{\sigma}_{\hat{S}_k} = \hat{S}_k \sqrt{\sum_{i<k} \frac{d_i}{n_i(n_i - d_i)}}$$

Durées moyenne et médiane de survie

- Estimation de la durée moyenne de survie :

$$\hat{\mu}_T = \sum_{k=1}^{m} (t_k - t_{k-1})\, \hat{S}_k$$

  Somme des longueurs d'intervalles pondérées par la probabilité de survie en début d'intervalle.

- Estimation de la durée médiane de survie : C'est la durée correspondant à $S_k = 0.5$. Dans la plupart des cas, elle peut être calculée par approximation linéaire.

Exemple miniature (1)

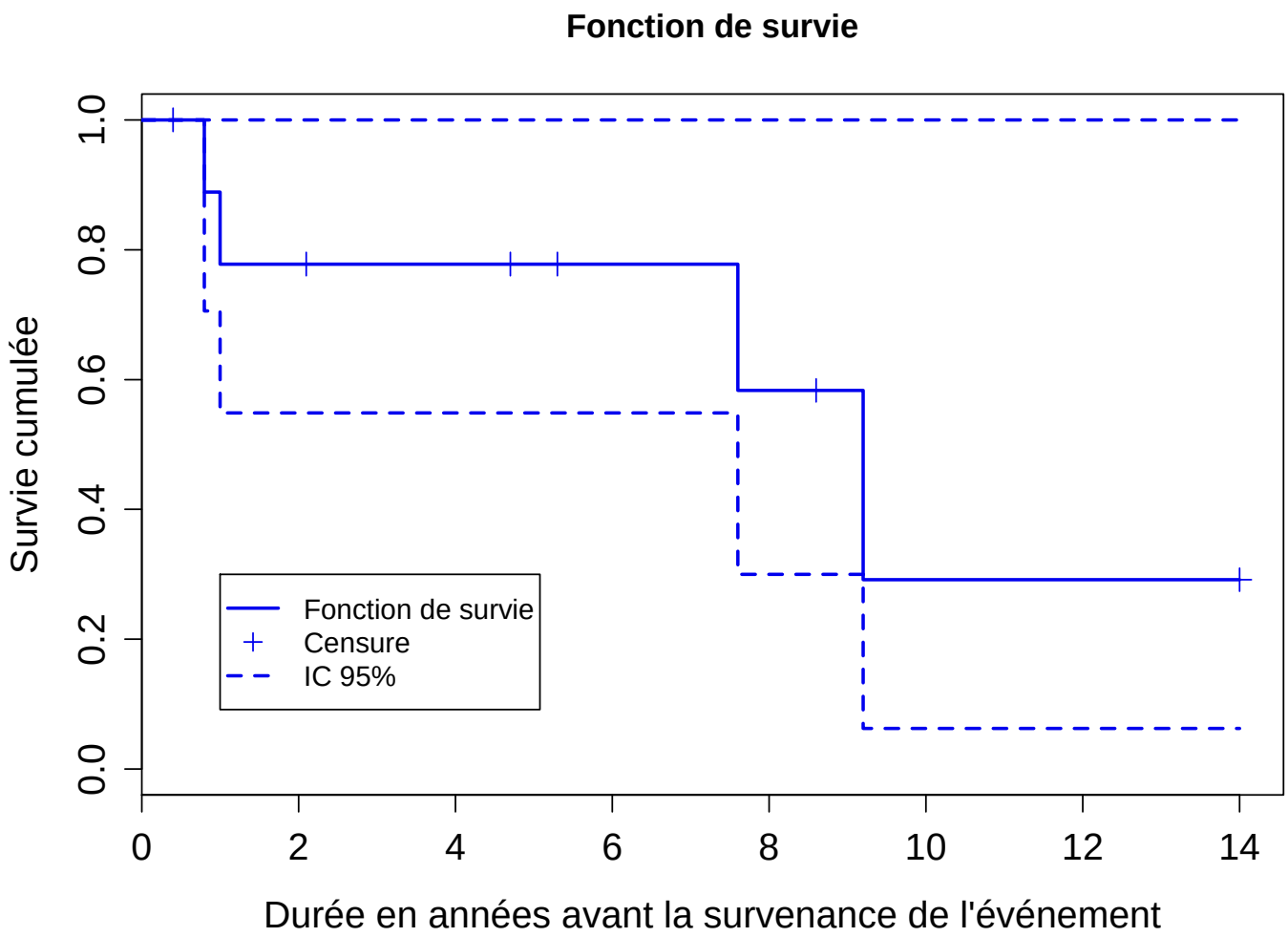
- On considère les décès au sein d'une population de 10 personnes.
- Il n'y a aucun événement simultané, à savoir qu'à chaque temps t_k on observe soit un décès (événement), soit une censure, mais pas les deux.
- Source : Courgeau & Lelièvre, 1989, p 49.

Exemple miniature (2)

k	d_k ou w_k	date t_k	statut	n_k	n_k/n	$\hat{h}_k = d_k/n_k$	$1-\hat{h}_k$	$\hat{S}_k$
1	1	0.4	0	10	1	0	1	1
2	1	0.8	1	9	0.9	0.111	0.889	1
3	1	1	1	8	0.8	0.125	0.875	0.889
4	1	2.1	0	7	0.7	0	1	0.778
5	1	4.7	0	6	0.6	0	1	0.778
6	1	5.3	0	5	0.5	0	1	0.778
7	1	7.6	1	4	0.4	0.250	0.750	0.778
8	1	8.6	0	3	0.3	0	1	0.583
9	1	9.2	1	2	0.2	0.500	0.500	0.583
10	1	14	0	1	0.1	0	1	0.292
total	10		4					

Durée moyenne jusqu'à l'événement : 8.44
Durée médiane : entre 9.2 et 14 $\rightarrow$ env. 10.57

Exemple miniature (3)



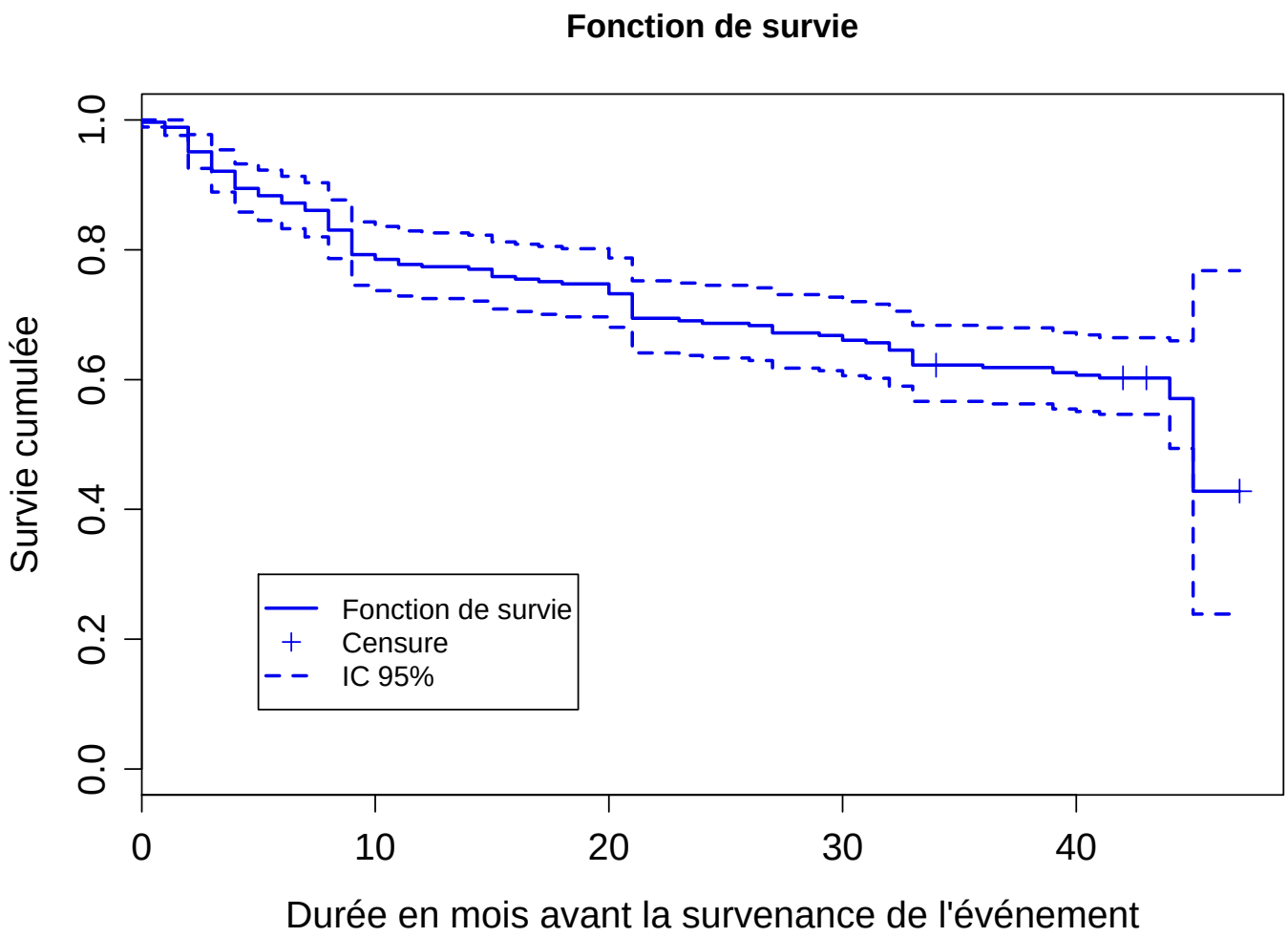
Exemple miniature (4)

- Remarque : R ne fait pas d'interpolation linéaire pour la médiane et donne la valeur inférieure de l'intervalle.

```
records   n.max n.start  events  median 0.95LCL 0.95UCL
   10.0    10.0    10.0     4.0     9.2     7.6      NA


time n.risk n.event survival std.err lower 95% CI upper 95% CI
 0.8      9       1    0.889   0.105       0.7056            1
 1.0      8       1    0.778   0.139       0.5485            1
 7.6      4       1    0.583   0.198       0.3000            1
 9.2      2       1    0.292   0.229       0.0627            1
```

Exemple Yamaguchi (1)



Exemple Yamaguchi (2)

records	n.max	n.start	events	median	0.95LCL	0.95UCL
265	265	265	107	45	44	NA

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
0	265	1	0.996	0.00377	0.989	1.000
1	264	2	0.989	0.00650	0.976	1.000
2	262	10	0.951	0.01327	0.925	0.977
3	252	8	0.921	0.01659	0.889	0.954
4	244	7	0.894	0.01888	0.858	0.932
5	237	3	0.883	0.01974	0.845	0.923
6	234	3	0.872	0.02054	0.832	0.913
7	231	3	0.860	0.02129	0.820	0.903
8	228	8	0.830	0.02306	0.786	0.877
9	220	10	0.792	0.02491	0.745	0.843
10	210	2	0.785	0.02524	0.737	0.836
11	208	2	0.777	0.02556	0.729	0.829
12	206	1	0.774	0.02571	0.725	0.826
14	205	1	0.770	0.02586	0.721	0.822
...						

Exemple Yamaguchi (3)

- Nous voulons maintenant comparer les courbes de survie des hommes (sex=0) et des femmes (sex=1).

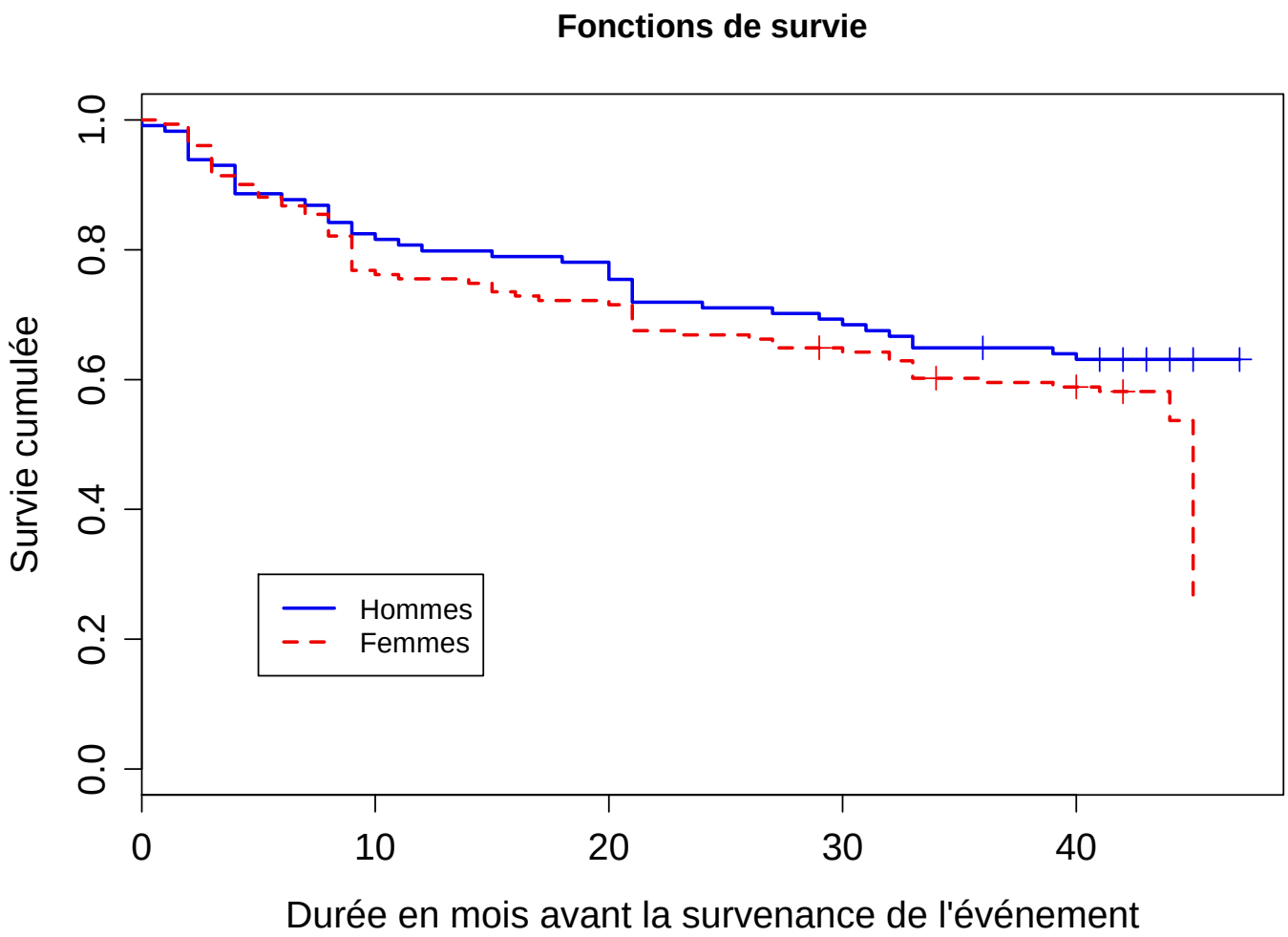
```
        records n.max n.start events median 0.95LCL 0.95UCL
Hommes      114   114     114     42     NA      NA      NA
Femmes      151   151     151     65     45      44      NA
```

- Log-rank test :

```
                  N Observed Expected (O-E)^2/E (O-E)^2/V
Dataset$sex=0 114        42     46.8     0.498     0.906
Dataset$sex=1 151        65     60.2     0.387     0.906

 Chisq= 0.9  on 1 degrees of freedom, p= 0.341
```

Exemple Yamaguchi (4)

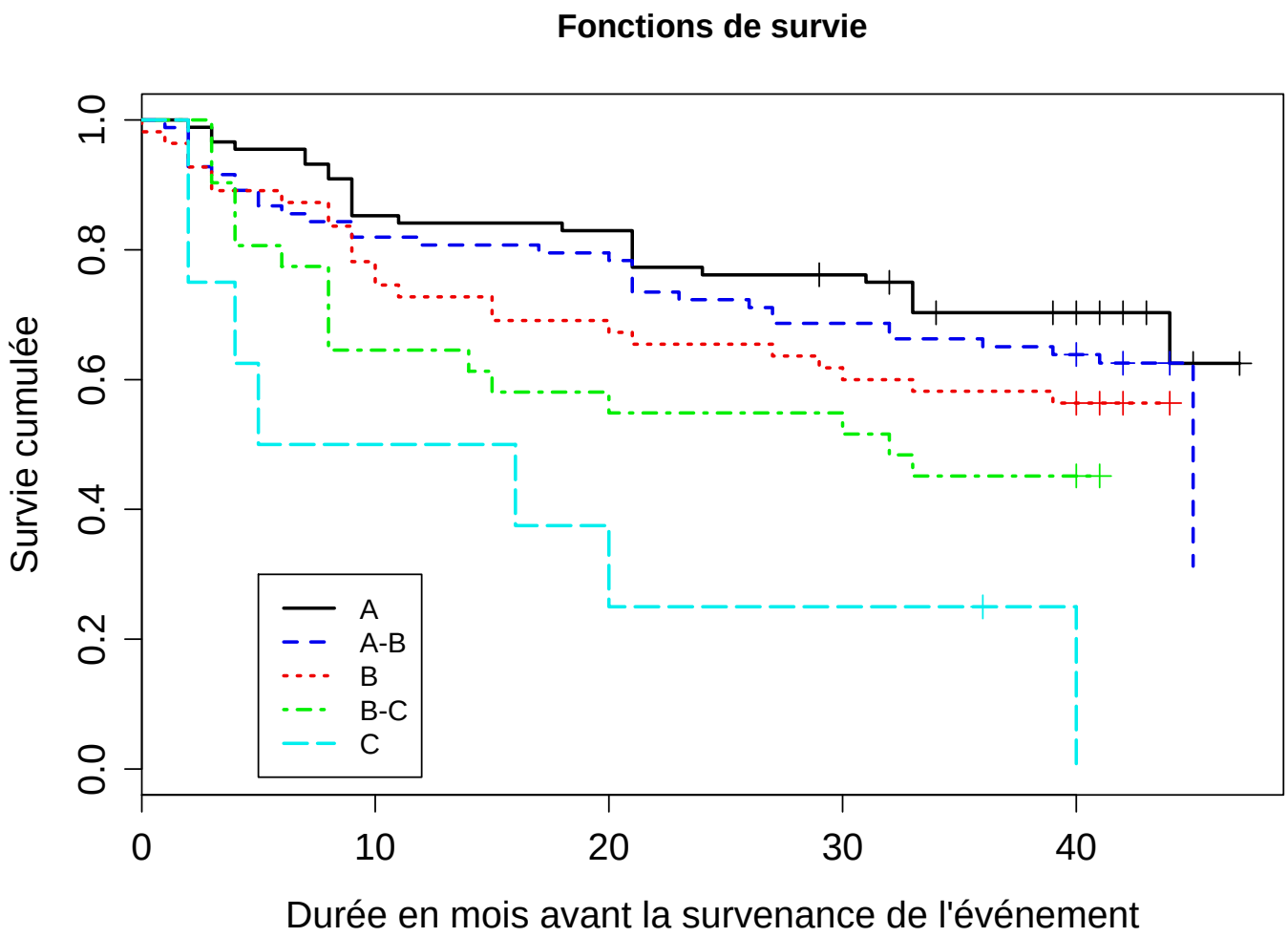


Exemple Yamaguchi (5)

- Nous voulons maintenant comparer les courbes de survie en fonction de la note moyenne (grd) obtenue à la fin des études secondaires (de 1 (=A, la meilleure) à 5(=C)).

	records	n.max	n.start	events	median	0.95LCL	0.95UCL
grd=1	88	88	88	27	NA	44	NA
grd=2	83	83	83	32	45.0	45	NA
grd=3	55	55	55	24	NA	30	NA
grd=4	31	31	31	17	32.0	8	NA
grd=5	8	8	8	7	10.5	4	NA

Exemple Yamaguchi (6)



Exemple Yamaguchi (7)

- Log-rank test :

```
                 N Observed Expected (O-E)^2/E (O-E)^2/V
Dataset$grd=1 88        27    38.84     3.610     5.805
Dataset$grd=2 83        32    34.56     0.189     0.286
Dataset$grd=3 55        24    21.15     0.384     0.491
Dataset$grd=4 31        17    10.57     3.915     4.460
Dataset$grd=5  8         7     1.88    13.896    14.521

 Chisq= 22.7  on 4 degrees of freedom, p= 0.000147
```

Plan du cours

Approche de Nelson-Aalen

- Méthode alternative pour estimer la fonction de survie $S(t)$ en temps continu (approche des processus de comptage).
- Soit $H(t)$ la fonction de hasard cumulée. Dans le cas continu

$$H(t) = \int_{t_0}^{t} h(s)ds = \int_{t_0}^{t} \frac{f(s)}{S(s)}ds = -\ln S(t)$$

d'où

$$S(t) = \exp(-H(t))$$

Estimateur

- L'idée est alors d'estimer $S(t)$ à partir d'un estimateur de $H(t)$.
- En considérant tous les instants t_k où des événements surviennent jusqu'à l'instant t, nous avons

$$\hat{H}(t) = \sum_{t_k \leq t} \frac{d_k}{n_k} \Rightarrow \hat{S}(t) = \exp(-\hat{H}(t))$$

- La variance de cet estimateur vaut

$$\hat{\sigma}^2_{\hat{H}(t)} = \sum_{t_k \leq t} \frac{d_k}{n_k^2}$$

Comparaison KM - NA

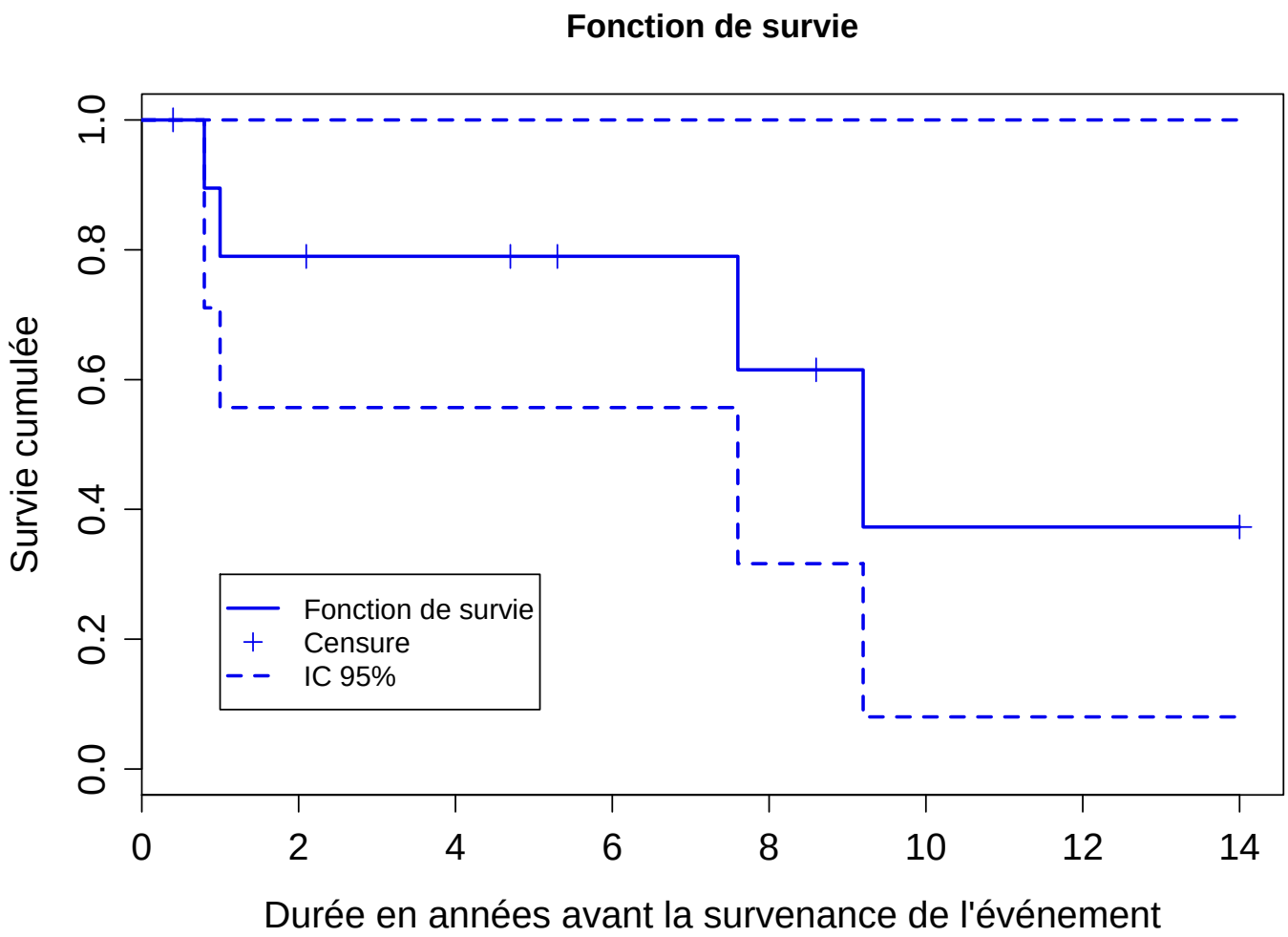
- Asymptotiquement, Kaplan-Meier et Nelson-Aalen sont équivalents.
- Sur de petits échantillons, Kaplan-Meier serait meilleur lorsque le hasard diminue au fil du temps, alors que Nelson-Aalen serait meilleur lorsque le hasard augmente au fil du temps.
- Source : Colosimo & al. (2002).

Exemple miniature (1)

k	d_k ou w_k	date t_k	statut	n_k	n_k/tot	$\hat{h}_k = d_k/n_k$	$\hat{H}_k$	$\hat{S}_k$
1	1	0.4	0	10	1	0	0	1
2	1	0.8	1	9	0.9	0.111	0	1
3	1	1	1	8	0.8	0.125	0.111	0.895
4	1	2.1	0	7	0.7	0	0.236	0.790
5	1	4.7	0	6	0.6	0	0.236	0.790
6	1	5.3	0	5	0.5	0	0.236	0.790
7	1	7.6	1	4	0.4	0.250	0.236	0.790
8	1	8.6	0	3	0.3	0	0.486	0.615
9	1	9.2	1	2	0.2	0.500	0.486	0.615
10	1	14	0	1	0.1	0	0.986	0.373
total	10		4					

Durée moyenne jusqu'à l'événement : 8.97
Durée médiane : entre 9.2 et 14 $\rightarrow$ env. 11.48

Exemple miniature (2)



Exemple miniature (3)

records	n.max	n.start	events	median	0.95LCL	0.95UCL
10.0	10.0	10.0	4.0	9.2	7.6	NA

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
0.8	9	1	0.895	0.105	0.7103	1
1.0	8	1	0.790	0.141	0.5569	1
7.6	4	1	0.615	0.209	0.3163	1
9.2	2	1	0.373	0.293	0.0802	1

Bibliographie

- Altman DG (1991) *Practical Statistics for Medical Research.* Chapman & Hall : London.
- Colosimo E, Ferreira F, Oliveira M, Sousa C (2002) Empirical Comparisons Between Kaplan-Meier and Nelson-Aalen Survival Function Estimators. *Journal of Statistical Computation and Simulation*, 72 : 299-308.
- Courgeau D, Lelièvre E (1989) *Analyse démographique des biographies.* Paris : Editions de l'INED.
- Yamaguchi K (1991) *Event history analysis.* ASRM 28. Newbury Park and London : Sage.