

**MATH 60604**  
**Modélisation statistique**  
**§ 7c - Estimateur de Kaplan–Meier**

Léo Belzile

HEC Montréal  
Département de sciences de la décision

# Notation

On considère  $T$  une variable aléatoire continue et un échantillon de taille  $n$ .

- Supposons qu'il y a  $D$  temps distincts de défaillance.
- Soit  $0 \leq t_1 < t_2 < \dots < t_D$  ces  $D$  temps en ordre croissant.
- Soit  $r_j$  le nombre d'individus **à risque** d'expérimenter l'événement au temps  $t_j$ .
  - C'est-à-dire, ces individus n'ont toujours pas expérimenté l'événement (et n'ont pas été censuré) avant le temps  $t_j$ .
  - Donc,  $r_j$  est le nombre de survivants juste avant le temps  $t_j$  qui sont à risques d'expérimenter l'événement au temps  $t_j$ .
- Soit  $d_j \in \{0, \dots, r_j\}$  le nombre d'individus qui expérimentent l'événement au temps  $t_j$  (par exemple, il y a  $d_j$  décès au temps  $t_j$ ).

# Dérivation de l'estimateur de Kaplan–Meier

La probabilité de mourir dans l'intervalle  $(t_j, t_{j+1}]$  étant donné que l'individu a survécu jusqu'à  $t_j$  est

$$h_j = P(t_j < T \leq t_{j+1} \mid T > t_j) = \frac{S(t_j) - S(t_{j+1})}{S(t_j)}.$$

d'où une récursion qui donne

$$S(t) = \prod_{j:t_j < t} (1 - h_j).$$

L'estimateur de Kaplan–Meier est **non-paramétrique**,

- on ne fait pas d'hypothèse sur la loi de probabilité sous-jacente de  $T_i$ .
- on considère plutôt les  $\{h_j\}_{j=1}^D$  comme des paramètres du modèle.

# Vraisemblance pour données discrètes

- Chacun des décès au temps  $t_j$  contribue  $h_j$  à la vraisemblance
  - la probabilité de défaillance à  $t_j$  sachant qu'un individu a survécu jusque là.
- Les survivants au temps  $t_j$  contribuent  $1 - h_j$ .
- On peut donc écrire la log vraisemblance comme

$$\ell(\mathbf{h}) = \sum_{j=1}^D \{d_j \ln(h_j) + (r_j - d_j) \ln(1 - h_j)\},$$

soit la somme des contributions de variables binomiales du risque au temps  $t_j$ .

# Optimisation des différents probabilité de survie

- Si on différencie  $\ell(\mathbf{h})$  par rapport à  $h_j$ , on trouve  $\hat{h}_j = d_j/r_j$ .
- L'estimateur de Kaplan–Meier pour la fonction de survie est

$$\hat{S}(t) = \prod_{t_j < t} \left(1 - \frac{d_j}{r_j}\right)$$

- Intuition:  $d_j/r_j$  représente la probabilité conditionnelle de survivre jusqu'avant le temps  $t_j$  et d'expérimenter l'événement au temps  $t_j$ .

# Exemple

Les données cancersein tirées de Sedmak et *al.* (1989) traitent de survie de patients atteints du cancer du sein et contiennent les variables suivantes:

- temps: temps de survie ou temps écoulé à la fin de l'étude (en mois)
- mort: variable indicatrice pour la mort, 0 pour censure, 1 pour décès
- repimmuno: réaction à un examen immunohistochimique, soit négative (0) ou positive (1)

# Statistiques descriptives pour cancersein

Variable d'analyse : temps				
N	Moyenne	Ec-type	Minimum	Maximum
45	98.33	51.84	19.00	189.00

mort	Fréquence	Pourcentage
0	21	46.67
1	24	53.33

repimmuno	Fréquence	Pourcentage
1	36	80.00
2	9	20.00

En pratique, l'utilisation de l'estimateur de Kaplan–Meier avec si peu de données est déconseillée. La qualité de l'approximation dépend fortement du nombre d'observations (correct si  $n \gg 1000$ ).

Les observations censurées fournissent beaucoup moins d'information que les temps de défaillance observés.

# Estimation de la fonction de survie

## Code SAS pour ajuster l'estimateur de Kaplan–Meier

```
proc lifetest data=modstat.cancersein method=km plots=(s(cl));  
time temps*mort(0);  
run;
```

L'argument `time` indique la variable de temps  $T_i$  (temps) et l'indicateur de censure  $\delta_i$ , incluant la valeur de référence pour les observations censurées à droite (`mort=0`)

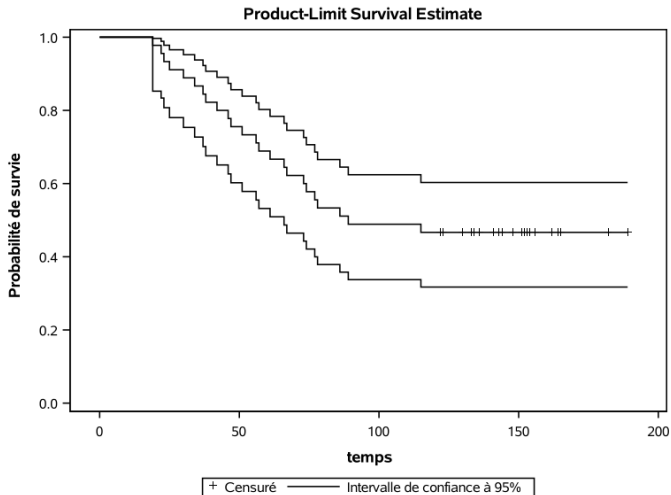


# Estimé de la fonction de survie

Valeurs estimées de survie de Kaplan-Meier					
temps	Survie	Echec	Erreur type de survie	Nombre d'échecs	Nombre restant
0.000	1.0000	0	0	0	45
19.000	0.9778	0.0222	0.0220	1	44
22.000	0.9556	0.0444	0.0307	2	43
23.000	0.9333	0.0667	0.0372	3	42
25.000	0.9111	0.0889	0.0424	4	41
			⋮		
165.000	*	.	.	24	2
182.000	*	.	.	24	1
189.000	*	.	.	24	0

**Note:** The marked survival times are censored observations.

# Graphique de la fonction de survie



La courbe de survie estimée est déficiente:  $\hat{S}(t)$  ne descend jamais à 0 parce que le temps de survie le plus long dans les données est censuré à droite.

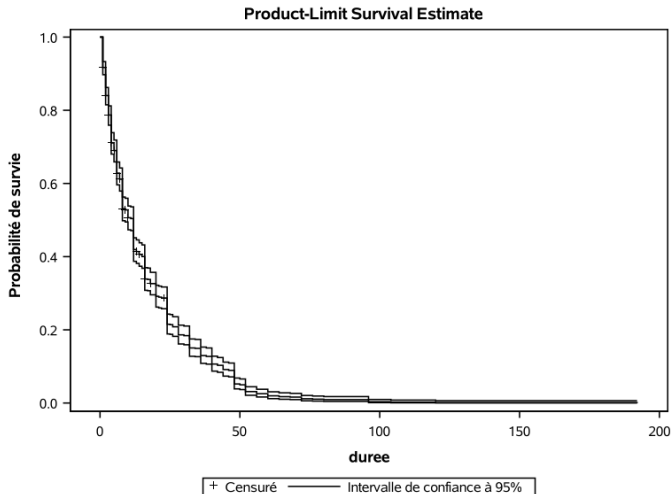
# Durée de l'allaitement

La base de données allaitement contient des données provenant de l'Enquête longitudinale nationale sur les jeunes sur la durée de la période d'allaitement de mères depuis la naissance de leur bébé. On se concentre sur les variables suivantes:

- `duration`: durée de l'allaitement (en semaines)
- `delta`: variable indicatrice de la fin de l'allaitement,
  - soit observée (1)
  - soit censurée (0)

Récapitulatif du nombre de valeurs censurées et non censurées			
Total	A échoué	Censuré	Pourcentage censuré
927	892	35	3.78

# Courbe de survie pour les données allaitement



$\hat{S}(t)$  atteindra zéro puisque le plus grand temps de survie est observé.

# La médiane de survie

La médiane du temps de survie est le temps  $t_M$  auquel  $S(t_M) = 0.5$ .

- C'est-à-dire, le temps médian  $t_M$  est tel que 50% des individus survivent jusqu'au temps  $t_M$ .

On peut facilement trouver la médiane du temps de survie en cherchant le temps  $t$  où la ligne horizontale  $\hat{S}(t) = 0.5$  croise la courbe de survie.

Estimations du quartile				
Intervalle de confiance à 95%				
Pourcentage	Valeur estimée du point	Transformation	[Inférieur	Supérieur)
75	.	LOGLOG	.	.
50	89.000	LOGLOG	66.000	.
25	51.000	LOGLOG	34.000	67.000

# Moyenne de la survie

Pour une variable aléatoire positive,  $T > 0$ , on peut démontrer que

$$E(T) = \int_0^{\infty} S(t)dt$$

On peut estimer l'espérance du temps de survie  $E(T)$  en calculant l'aire sous la courbe de survie estimée  $\hat{S}(t)$ .

- Par exemple, le temps de survie moyen pour les données d'allaitement est 16.89 semaines avec erreur-type 0.614 semaines.
- Si le temps observé le plus long est **censuré**, la courbe de survie estimée  $\hat{S}(t)$  va atteindre un plateau et ne descendra jamais à 0. L'aire sous la courbe est infinie.
- Dans ce cas, on peut plutôt estimer le temps de survie moyen limité:  $E(\min\{T, \tau\})$  pour une valeur choisie  $\tau$ . C'est-à-dire, nous calculerons le temps de survie moyen comme si la courbe descendait à 0 au temps  $\tau$  (option `rmst` dans SAS).