Modèles de durée

Théorie et applications

Etienne Dagorn

Université de Lille - LEM

Séance 2 :
Analyse non-paramétrique des durées

Objectifs de la séance

À l'issue de la séance, vous saurez :

1. **Construire** correctement (T, δ) et déclarer l'entrée retardée si besoin.
2. **Estimer** une courbe de survie **Kaplan–Meier** (KM) sous censure à droite.
3. **Lire/interpréter** une KM : marches, marques de censure, médiane/quantiles, N à risque.
4. **Comparer** des groupes avec le **log-rank** (idée, pas encore la dérivation).
5. **Tracer** et **interpréter** la **fonction de hasard cumulée** (Nelson–Aalen) à des fins diagnostiques.

Rappels

Hasard

De la durée brute à Kaplan–Meier

Greenwood

Breslow

Comparaison

Rappels

Rappels

Objectif des modèles de durée : analyser le *temps jusqu'à un événement* (en tenant compte de la **censure** et de la **troncature**).

Du calendrier aux données :

$$T = t_{\text{fin}} - t_0 \quad \text{(unité cohérente)}, \qquad \delta = \mathbf{1}\{\text{événement observé}\}.$$

Censure : événement non (entièrement) observé *pendant* la fenêtre.
Troncature : inclusion conditionnelle (ex. entrée retardée).

Hypothèse clé (pour les estimateurs standards) : **censure non-informative** (cond. aux covariables, le mécanisme de censure est indépendant de la durée vraie).

Les 5 objets fondamentaux

1. Répartition

$$F(t) = \Pr(T \leq t)$$

Part déjà sortie avant t.

2. Survie

$$S(t) = \Pr(T > t) = 1 - F(t)$$

Part encore "en vie" à t.

3. Densité

$$f(t) = F'(t) = -S'(t)$$

Vitesse d'arrivée des événements.

4. Hasard instantané

$$h(t) = \frac{f(t)}{S(t)}$$

Rythme instantané *parmi ceux encore à risque.*

5. Hasard cumulé

$$H(t) = \int_0^t h(u)\,du$$

Équations-pont

$$S(t) = e^{-H(t)} \iff H(t) = -\log S(t)$$

$$\Pr(t < T \leq t{+}\Delta t \mid T > t) \approx h(t)\Delta t$$

Rappel - enjeux de la censure

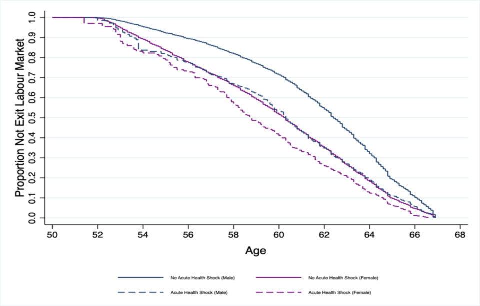
Problème de la moyenne naïve : exclure les censurés $\Rightarrow$ **biais vers le bas** (*on supprime les durées longues*).

Types fréquents

- $\Rightarrow$ **Droite** (fin d'étude/perte de vue avant l'événement) $\Rightarrow \delta = 0$.
- $\Rightarrow$ **Gauche** / **intervalle** (événement avant l'entrée / connu dans $[a, b]$).
- $\Rightarrow$ **Entrée retardée** (troncature gauche) : on n'observe que si $T > t_{\text{entrée}}$.

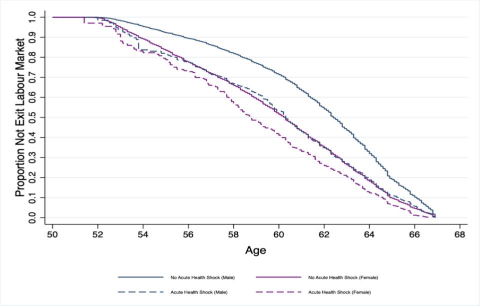
À vérifier avant d'estimer : t_0 & unité, (T, δ) corrects, % de censure, entrée retardée déclarée, plausibilité *non-informative*.

Figure 4 – Kaplan-Meier Survival Estimates of the Proportion of Labour Market Exits by Acute Health Shocks and Gender (all countries)



Source : Vieira, L. (2025) Too Sick for Working, or Sick of Working ? Impact of Acute Health Shocks on Early Labour Market Exits

Figure 4 – Kaplan-Meier Survival Estimates of the Proportion of Labour Market Exits by Acute Health Shocks and Gender (all countries)



Source : Vieira, L. (2025) Too Sick for Working, or Sick of Working ? Impact of Acute Health Shocks on Early Labour Market Exits

1. Vers quel temps se situe la **médiane** de survie ?
2. Le groupe traité semble-t-il **dominer** l'autre sur tout l'intervalle ?
3. Que signifie le fait que la courbe ne descende jamais sous 0,5 ?

Hasard $h(t)$: intuition, définition, unités

Idée clé

$h(t) =$ *rythme instantané* de réalisation de l'évènement **parmi ceux encore à risque** à t.

Définition

$$h(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \Pr\big(t < T \leq t + \Delta t \mid T > t\big) = \frac{f(t)}{S(t)} \quad \text{(si } f \text{ existe).}$$

Unités & lecture

- ⇒ Unités : (événements)/(individus à risque)/(unité de temps).
- ⇒ Lecture locale : $\Pr(t < T \leq t + \Delta t \mid T > t) \approx h(t)\,\Delta t$.

Ordre de grandeur

Si $h(t) = 0{,}05$ par mois, alors, sur un petit mois $\Delta t = 1$, $\Pr(\text{sortie}|\text{ à risque}) \approx 5\%$.

Du hasard $h(t)$, hasard cumulé $H(t)$, survie $S(t)$

Relations structurantes

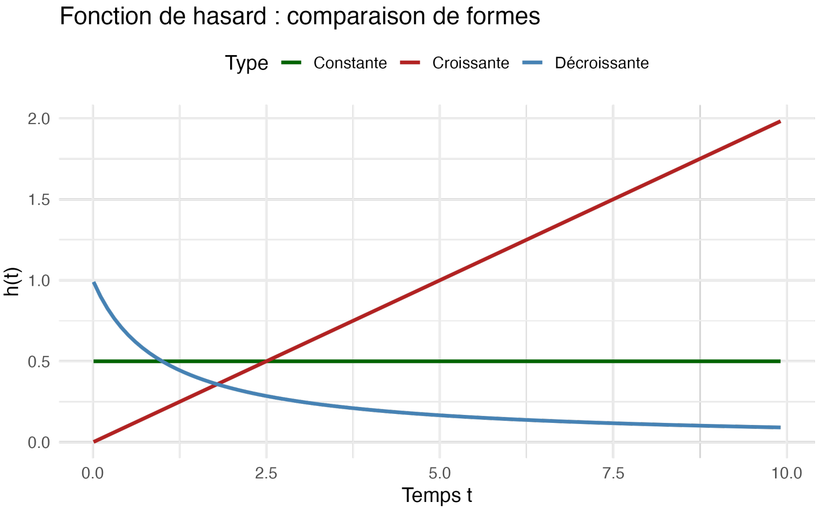
$$H(t) = \int_0^t h(u)\,du, \qquad S(t) = \exp\{-H(t)\}, \qquad f(t) = h(t)\,S(t).$$

- ⇒ $H(t)$ = **risque cumulé** : accumulation du risque instantané jusqu'à t.
- ⇒ $S(t)$ décroît quand $H(t)$ grandit : pente forte ⇒ risque élevé.
- ⇒ $f(t)$ combine rythme instantané (h) et part encore à risque (S).

Exemples de formes de hasard

- ⇒ $h(t) = \lambda$ constant ⇒ loi exponentielle (*pas de mémoire*).
- ⇒ $h(t)$ croissant : risque augmente avec le temps (usure/vieillissement).
- ⇒ $h(t)$ décroissant : sélection des plus résistants, queue lourde.

Comparaison de comportements de hasard



Hasard par morceaux (piecewise-constant)

Idée cle : on divise le temps en intervalles

$$0 = \tau_0 < \tau_1 < \cdots < \tau_J$$

et on suppose que le *hasard est constant* dans chaque intervalle :

$$h(t) = \lambda_j \quad \text{pour } t \in [\tau_{j-1}, \tau_j).$$

Ce que cela implique :

$$H(t) = \sum_{k<j} \lambda_k(\tau_k - \tau_{k-1}) \ + \ \lambda_j(t - \tau_{j-1}), \qquad S(t) = \exp\{-H(t)\}.$$

Lecture intuitive :

- ⇒ chaque intervalle a son **taux de sortie** propre : λ_j ;
- ⇒ $H(t)$ est la somme des “morceaux de risque” accumules ;
- ⇒ $S(t)$ est une courbe en **escaliers lisses** (segment exponentiel par intervalle).

Pourquoi la transformation complément log-log ?

Objectif : taux instantané continu $h(t)$ à une **probabilité discrète** de sortie $p(t)$.

Pourquoi la transformation complément log-log ?

Objectif : taux instantané continu $h(t)$ à une **probabilité discrète** de sortie $p(t)$.

Du continu au discret

$$p(t) = \Pr(T = t \mid T \geq t) = 1 - \exp\{-h(t)\Delta t\}$$

$$\begin{aligned} 1 - p &= \exp\{-h(t)\Delta t\} \\ \log(1-p) &= -h(t)\Delta t \\ \log\{-\log(1-p)\} &= \log h(t) + \log(\Delta t). \end{aligned}$$

La transformation *cloglog*

$$\underbrace{\log\{-\log(1-p)\}}_{\text{complément log-log}} = \alpha_t + X\beta$$

Hasard en temps discret

1. Definition du hazard discret

$$h(t) = P(T = t \mid T \geq t), \qquad S(t) = \prod_{k=1}^{t} \{1 - h(k)\}.$$

2. Modele person-period Chaque individu = une ligne par periode t. On modele

$$P(T = t \mid T \geq t, X)$$

comme une fonction de X et d'effets de duree.

3. Modele complement log-log (cloglog)

$$\log\{-\log(1 - h(t))\} = \alpha_t + X\beta.$$

Interpretation

- ⇒ e^{β} = hazard ratio (comme dans un modele PH continu).
- ⇒ α_t capture la baseline hazard (une valeur par periode).
- ⇒ Le cloglog est coherent avec un vrai processus continu sous-jacent.

Hasards proportionnels : intuition de l'effet d'une covariable

Idée centrale (modèle PH) :

$$h(t \mid X = x) \;=\; h_0(t)\,\exp(x\beta).$$

⇒ $h_0(t)$: **hasard de base**, forme temporelle commune a tous les individus ;

⇒ $\exp(x\beta)$: **effet multiplicatif constant** dans le temps.

Interprétation du coefficient

⇒ e^{β} est un **hazard ratio** : ratio des risques instantanés.

⇒ Exemple : $e^{\beta} = 1.3 \Rightarrow$ "+30% de risque instantané *à tout moment*".

⇒ $e^{\beta} = 0.7 \Rightarrow$ "−30% de risque instantané" (effet protecteur).

Conséquence clé : les courbes de hasard restent **proportionnelles dans le temps**. La covariable *ne modifie pas la forme de* $h_0(t)$, seulement son échelle.

Hasards proportionnels : verifier l'hypothese

Idee generale. Si l'hypothese PH est vraie, alors les courbes

$$\log\{-\log \hat{S}(t \mid X)\}$$

sont a peu pres paralleles entre groupes.

Test graphique :

- ⇒ tracer log(-log S) pour chaque modalite de X ;
- ⇒ courbes paralleles : PH plausible ;
- ⇒ courbes qui se croisent : PH violee.

Test formel : residus de Schoenfeld

- ⇒ testent si l'effet de X varie avec le temps ;
- ⇒ tendance plate : PH OK ; tendance non plate : non-PH.

En cas de non-PH : interactions avec le temps, modeles stratifies, ou modele AFT.

Pourquoi h(t) decroit-il ?

Deux mecanismes possibles :

- ⇒ **Duree-dependance vraie** : pour un individu, rester plus longtemps dans l'etat reduit (ou augmente) son risque instantane.
- ⇒ **Selection (frailty)** : les individus a risque eleve sortent rapidement ; ceux qui restent sont les plus "robustes". Le hasard moyen decroit alors automatiquement, meme si le risque individuel est constant.

Implication importante : une courbe de hasard decroissante au niveau agregé *ne permet pas* de conclure que le risque individuel baisse avec la duree.

Comment distinguer les deux effets ?

- ⇒ modeles avec **frailty** (gamma, log-normal) ;
- ⇒ prise en compte de l'heterogeneite individuelle (covariables fines) ;
- ⇒ analyse par sous-groupes si la structure de risque est connue.

Comparaison de formes de densité

⇒ La densité $f(t)$ indique la **probabilité instantanée de sortie** à l'instant t.

⇒ Elle dépend de la forme de la distribution de durée sous-jacente :

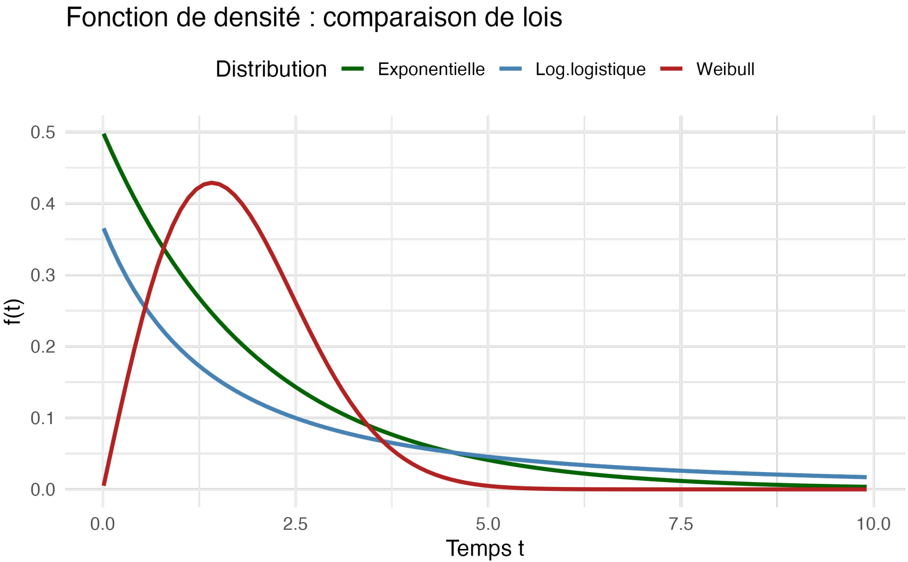
Exemples usuels

⇒ **Exponentielle** : pic immédiat, décroissance rapide (durée "sans mémoire").

⇒ **Weibull (shape > 1)** : montée puis décroissance (vieillissement).

⇒ **Log-logistique** : longue traîne (certaines durées très longues).

Comparaison de densités de durée



Erreurs fréquentes & corrections

1. **Exclure** les censurés ⇒ moyenne biaisée.
   ⇒ Utiliser médiane KM / quantiles, ou modèles de durée.
2. **Confondre** censure vs troncature.
   ⇒ Décrire l'entrée dans l'étude ; déclarer l'**enter time**.
3. **Unité/origine incohérentes** (jours vs mois).
   ⇒ Fixer t_0, harmoniser l'unité, documenter dans les figures.
4. **Ignorer** l'hypothèse de **censure non-informative**.
   ⇒ Discuter le mécanisme, tester des **sensibilités**.

Pourquoi une approche non-paramétrique ?

Idée générale : approche **souple** qui respecte les données avant d'imposer une forme de modèle.

- ⇒ **Pas d'hypothèse** sur la forme de la loi de durée.
- ⇒ Permet d'estimer $S(t)$ et $H(t)$ **avec censure à droite**.
- ⇒ Utile pour **comparer des groupes** (log-rank).
- ⇒ Sert de **référence visuelle** pour juger un modèle paramétrique.

Hypothèses de validité :

- ⇒ **Censure non-informative** : à caractéristiques observées identiques, la probabilité d'être censuré ne doit pas dépendre du temps de sortie "vrai".
- ⇒ **Indépendance** entre individus.
- ⇒ **Origine** t_0 **& unité cohérentes** : même départ et même unité pour tout le monde, **entrée retardée** correctement déclarée.

Étape 1 : de la durée vraie aux données observées

Pour chaque individu i

$$T_i = \text{durée vraie jusqu'à l'événement}, \qquad C_i = \text{temps de censure},$$

$$t_i = \min(T_i, C_i), \qquad \delta_i = \mathbf{1}\{T_i \le C_i\}.$$

On observe seulement (t_i, δ_i) dans les données.

Lecture rapide

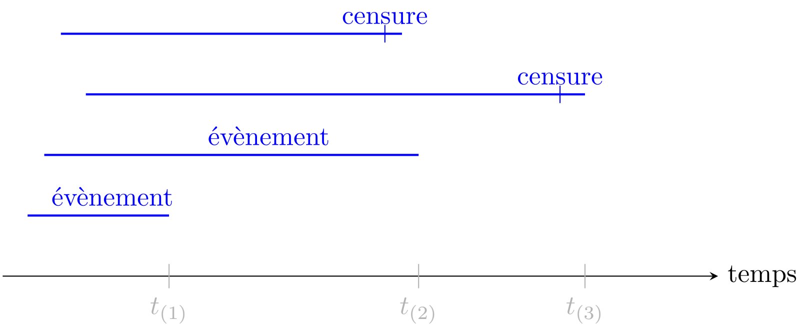
- ⇒ $\delta_i = 1$: événement observé à t_i.
- ⇒ $\delta_i = 0$: censuré à t_i (on sait seulement $T_i > t_i$).

Objectif : estimer $S(t) = \Pr(T > t)$ à partir de (t_i, δ_i).
À garder en tête : l'estimateur de Kaplan–Meier ne repose **que** sur ces deux objets.

Étape 2 : temps d'événements & risk set

On regroupe l'information par temps d'événements :

- ⇒ $t_{(1)} < \cdots < t_{(J)}$: instants où **au moins un événement** survient.
- ⇒ d_j : nombre d'événements à $t_{(j)}$.
- ⇒ n_j : nombre d'individus **à risque juste avant** $t_{(j)}$.



Idée : à chaque $t_{(j)}$, on regarde **combien sortent** (d_j) sur **combien étaient encore à risque** (n_j).

Étape 3 : sans censure, la survie empirique

Cas simple : aucune censure

$$\widehat{S}_{\text{naïve}}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(t_i > t) = 1 - \widehat{F}_{\text{emp}}(t).$$

Intuition : proportion d'individus encore “en vie” à t.

Où est le problème avec la censure ?

- ⇒ Si on traite une censure comme un événement (“sortie”) : on diminue artificiellement la survie.
- ⇒ ⇒ **biais vers le bas** : on sous-estime la durée typique.
- ⇒ On a besoin d'un estimateur qui **fait la différence** entre événement et censure.

Étape 4 : idée de Kaplan–Meier

Principe clé : la survie est un **produit** de probabilités conditionnelles :

$$S(t) = \Pr(T > t) = \prod_{t_{(j)} \le t} \Pr\left(T > t_{(j)} \mid T \ge t_{(j)}\right).$$

Kaplan–Meier remplace ces probabilités par les proportions observées :

$$\Pr(T > t_{(j)} \mid T \ge t_{(j)}) \approx 1 - \frac{d_j}{n_j}.$$

Idée en mots

À chaque temps d'événement $t_{(j)}$:

- ⇒ on regarde **la fraction qui sort** : d_j/n_j ;
- ⇒ on en déduit la fraction qui **survit au cran** : $1 - d_j/n_j$;
- ⇒ on **multiplie** tous ces facteurs de survie conditionnelle.

Étape 5 : formule de Kaplan–Meier

Estimateur de la fonction de survie

$$\hat{S}(t) = \prod_{t_{(j)} \leq t} \left(1 - \frac{d_j}{n_j}\right).$$

Lecture

- ⇒ Fonction en **marches** : elle ne bouge que lorsque survient un **événement**.
- ⇒ Les censures ⇒ diminuent n_j aux temps suivants, mais ne font **pas chuter** $\hat{S}(t)$.
- ⇒ Sans censure, on retrouve l'empirique : $\hat{S}(t) = 1 - F_n(t)$.

Etape 6 : calcul de Kaplan-Meier a la main

Donnees (5 individus) : 2*, 4, 4*, 5, 8 (* = censure)
Temps d'evenement : 4, 5, 8

t	n_t	d_t	Survie instantanee	S(t)
4	4	1	$1 - 1/4 = 0.75$	0.75
5	3	1	$1 - 1/3 = 0.67$	$0.75 \times 0.67 = 0.50$
8	1	1	$1 - 1/1 = 0$	$0.50 \times 0 = 0$

- ⇒ Censure a 2 : jamais dans le risk set.
- ⇒ Censure a 4 : dans le risk set a 4 juste avant l'instant, puis retire.
- ⇒ Kaplan-Meier : produit des probabilites de survivre a chaque evenement.

Nelson–Aalen : une autre facon de lire la survie

1. Hasard cumule estime

$$\widehat{H}(t) = \sum_{t_{(j)} \leq t} \frac{d_j}{n_j}, \qquad \widehat{S}_{\mathrm{NA}}(t) = \exp\{-\widehat{H}(t)\}.$$

2. Interpretation

- ⇒ Chaque terme $\frac{d_j}{n_j}$ est un petit morceau de risque instantane.
- ⇒ $\widehat{H}(t)$ = accumulation de ces risques au cours du temps.
- ⇒ Une pente forte indique un risque eleve ; une pente faible indique un risque faible.

3. Pourquoi utile ?

- ⇒ La courbe NA est souvent plus stable (moins de sauts) que la courbe KM.
- ⇒ Donne une meilleure intuition de la dynamique du risque dans le temps.
- ⇒ Les deux estimateurs utilisent exactement les memes valeurs d_j et n_j.

Lire ensemble KM et Nelson–Aalen (1/2)

Kaplan–Meier (KM)

- ⇒ Estime le **niveau de survie** $S(t)$.
- ⇒ Courbe en **marches** : chute uniquement aux **événements**.
- ⇒ Les censures ne font **pas** chuter la courbe.

Nelson–Aalen (NA)

- ⇒ Estime le **hasard cumulé** $\widehat{H}(t)$.
- ⇒ La **pente locale** ≈ **niveau instantané du risque**.
- ⇒ Courbe plus lisse → meilleure lecture de la dynamique du risque.

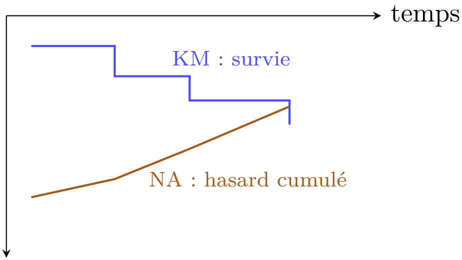
Pourquoi regarder les deux ?

KM = "**combien restent**" ; NA = "**à quelle vitesse ils sortent**".

Lire ensemble KM et Nelson–Aalen (2/2)

Heuristique de lecture du NA

Segments quasi-linéaires → **risque constant**.
Segments convexes → **risque qui augmente**.
Segments concaves → **risque qui diminue**.



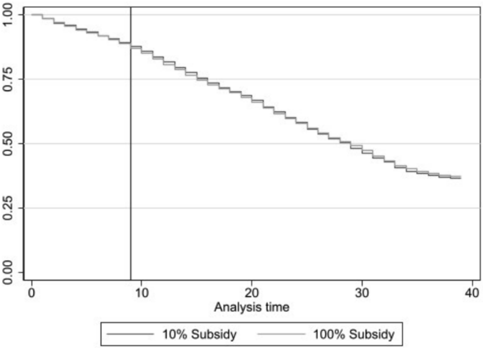
KM = niveau de survie ; NA = pente du risque cumulé.

Feuille de route d'estimation (KM/NA/log-rank)

1. **Préparer les données** : construire (T, δ) ; déclarer l'**entrée retardée** si besoin.
2. **Compter par temps** : lister $t_{(j)}$, calculer d_j et n_j (**risk set** cohérent).
3. **Estimer KM** : fonction en marches, **pas de chute** aux censures.
4. **Estimer NA** : lire la dynamique du **risque cumulé** $\widehat{H}(t)$.
5. **Comparer des groupes** : log-rank (test global) + inspection visuelle des courbes.

Courbe de Kaplan-Meier (exemple)

Figure 2: Duration until next birth, by subsidy status



The figure shows a Kaplan-Meier Survival curve for the 10% subsidy group and the 100% subsidy group. The y-axis shows the fraction of people who have not given birth, starting from the date of the baseline survey (t=0) and up to 36 months later (endline survey). The intervention could not have affected fertility in the first 9 months (area left of the vertical line). A test that the survival curves are identical cannot reject the null (p-value = 0.55), thus the fertility behavior post intervention is the same regardless of the contraception subsidy received.

Source : Dupas, P., Jayachandran, S., Lleras-Muney, A., & Rossi, P. (2025). The negligible effect of free

Courbe de Kaplan-Meier (exemple)

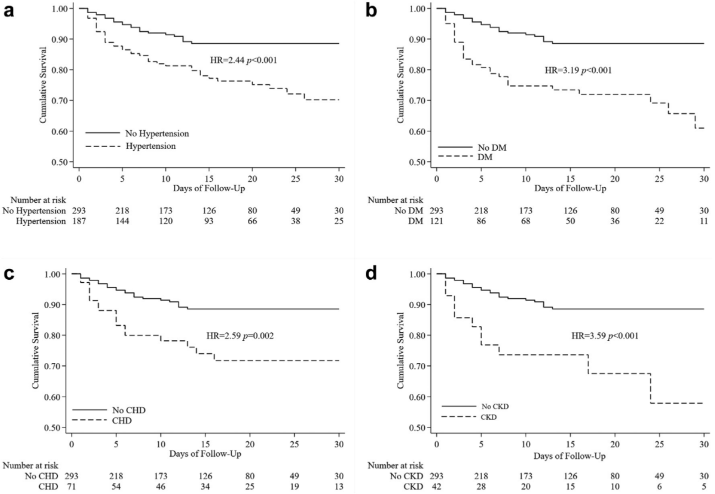


Fig. 4: Comorbidities as risk factors for mortality. Kaplan-Meier curve of mortality related to a) hypertension b) diabetes mellitus (DM), c) coronary heart disease (CHD) and d) chronic kidney disease (CKD). HR = hazard ratio.

Source : Hartantri, Y., Debora, J., Widyatmoko, L., Giwangkancana, G., Suryadinata, H., Susandi, E., ... &

Greenwood : intuition générale

Idée de départ : Kaplan–Meier est un *produit* de probabilités de survie :

$$\hat{S}(t) = \prod_{t_{(j)} \leq t} \left(1 - \frac{d_j}{n_j}\right).$$

Chaque facteur

$$X_j = 1 - \frac{d_j}{n_j}$$

est estimé à partir d'un échantillon fini de taille n_j avec d_j défaillances.

Idée clé :

⇒ plutôt que la variance d'un *produit*, on regarde la variance du *log-produit* :

$$\log \hat{S}(t) = \sum_j \log(X_j);$$

⇒ chaque terme $\log(X_j)$ a une variance propre,

⇒ la variance de la somme s'obtient en sommant ces contributions.

Variance de Kaplan–Meier : formule de Greenwood

Formule de Greenwood

$$\widehat{Var}(\hat{S}(t)) = \hat{S}^2(t) \sum_{t_{(j)} \leq t} \frac{d_j}{n_j(n_j - d_j)}.$$

Ingrédients :

- ⇒ $t_{(j)}$: instants de défaillance observés,
- ⇒ d_j : nombre de défaillances à $t_{(j)}$,
- ⇒ n_j : nombre à risque juste avant $t_{(j)}$.

But pratique :

$$\hat{S}(t) \pm z_{1-\alpha/2} \sqrt{\widehat{Var}(\hat{S}(t))}$$

pour construire des intervalles ou bandes de confiance.

Greenwood : lecture de la variance

$$\widehat{Var}(\hat{S}(t)) = \hat{S}^2(t) \sum_{t_{(j)} \leq t} \frac{d_j}{n_j(n_j - d_j)}.$$

Lecture des termes :

- ⇒ chaque défaillance ($d_j > 0$) ajoute de l'incertitude ;
- ⇒ un grand risk set n_j *atténue* cette incertitude ;
- ⇒ quand n_j devient petit, le terme $\frac{d_j}{n_j(n_j-d_j)}$ devient grand.

Conséquence :

- ⇒ au début du suivi : variance modérée (beaucoup d'individus à risque),
- ⇒ en fin de suivi : variance élevée (peu d'individus à risque),
- ⇒ les derniers points de $\hat{S}(t)$ sont les plus incertains.

Bandes de confiance : échelle log(-log)

Problème :

$$\hat{S}(t) \pm z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{S}(t))}$$

peut donner des bornes hors de $[0, 1]$ (IC symétrique sur une quantité bornée).

Solution : travailler sur une échelle non bornée.

$$Y(t) = -\log\{-\log \hat{S}(t)\}.$$

Procédure :

1. calculer $Y(t)$ et sa variance (via Greenwood et Delta method),
2. construire l'IC normal sur $Y(t)$,
3. revenir à $S(t)$ par $\hat{S}(t) = \exp\{-\exp(-Y(t))\}$.

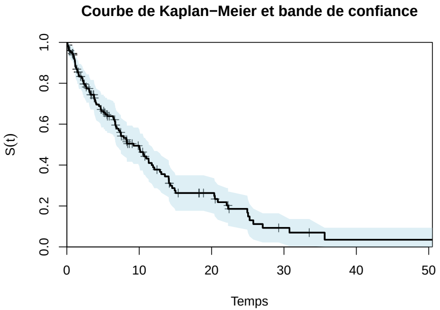
Avantage : bandes toujours dans $[0, 1]$, souvent mieux comportées.

Approche graphique : courbe de Kaplan-Meier

⇒ Variance de Greenwood, pour chaque temps t : **incertitude** de $\hat{S}(t)$.

⇒ Graphiquement, cela se traduit par une **bande de confiance** :

- chaque saut est associé à une **barre verticale d'incertitude**,
- la bande est plus **étroite** quand n_j est grand,
- elle devient plus **large** quand il reste peu d'individus à risque.



Approche graphique : pourquoi les bandes s'élargissent ?

⇒ Au début du suivi : n_j est grand ⇒ les termes $\dfrac{d_j}{n_j(n_j - d_j)}$ sont petits.

⇒ Au fur et à mesure :

- certains individus sont censurés, & d'autres ont déjà connu l'événement,
- ⇒ le nombre à risque n_j diminue.

⇒ En fin de suivi :

- Peu d'individus à risque & nouvelle défaillance ajoute de l'incertitude.

$$\widehat{\mathrm{Var}}(\hat{S}(t)) = \hat{S}^2(t) \cdot \underbrace{\sum_{t_{(j)} \leq t} \frac{d_j}{n_j(n_j - d_j)}}_{\text{augmente fortement quand } n_j \text{ est petit}}$$

Conséquence pratique : les derniers points de la courbe de survie sont très incertains et doivent être interprétés avec prudence.

Greenwood : exemple numérique

Considérons deux instants de défaillance successifs :

$$n_1 = 10, \quad d_1 = 2 \qquad \Rightarrow \qquad \hat{S}(t_1) = 1 - \frac{2}{10} = 0{,}8$$

$$n_2 = 8, \quad d_2 = 1 \qquad \Rightarrow \qquad \hat{S}(t_2) = 0{,}8\left(1 - \frac{1}{8}\right) = 0{,}7$$

Variance à t_1 :

$$\widehat{\mathrm{Var}}(\hat{S}(t_1)) = 0{,}8^2 \cdot \frac{2}{10 \cdot 8}$$

Variance à t_2 :

$$\widehat{\mathrm{Var}}(\hat{S}(t_2)) = 0{,}7^2 \left(\frac{2}{10 \cdot 8} + \frac{1}{8 \cdot 7}\right)$$

Observation : le second événement ajoute proportionnellement plus d'incertitude car n_j est plus faible.

Estimateur de Breslow : intuition

Objectif : estimer la fonction de risque cumulee

$$\Lambda(t) = \int_0^t \lambda(s)\,ds.$$

Idee cle : aux temps de defaillance $t_{(j)}$, on observe : n_j individus encore a risque juste avant $t_{(j)}$, d_j evenements au temps $t_{(j)}$.
hazard instantane moyen au temps $t_{(j)}$ comme

$$\lambda(t_{(j)}) \approx \frac{d_j}{n_j}.$$

Chaque defaillance ajoute alors un *petit morceau* de risque cumule :

$$\Delta\hat{\Lambda}(t_{(j)}) = \frac{d_j}{n_j}.$$

$$\hat{\Lambda}(t) = \sum_{t_{(j)} \leq t} \frac{d_j}{n_j}.$$

Estimateur de Breslow : proprietes et lien avec KM

Proprietes de $\hat{\Lambda}(t)$:

- ⇒ fonction en escalier, croissante dans le temps ;
- ⇒ elle ne saute qu'aux temps de defaillance observes $t_{(j)}$;
- ⇒ la censure ne cree pas de saut, mais diminue le n_j disponible.

Lien avec Kaplan–Meier :

$$\hat{S}(t) = \prod_{t_{(j)} \leq t} \left(1 - \frac{d_j}{n_j}\right).$$

En prenant le logarithme :

$$\log \hat{S}(t) = \sum_{t_{(j)} \leq t} \log \left(1 - \frac{d_j}{n_j}\right) \approx - \sum_{t_{(j)} \leq t} \frac{d_j}{n_j} = -\hat{\Lambda}(t).$$

$$\hat{S}(t) \approx \exp\{-\hat{\Lambda}(t)\}.$$

A retenir : Breslow est l'estimateur "logarithmique" sous-jacent a la courbe de Kaplan–Meier.

Estimateur de Breslow : variance

Estimateur du risque cumule :

$$\hat{\Lambda}(t) = \sum_{t_{(j)} \leq t} \frac{d_j}{n_j}.$$

Comme $\hat{\Lambda}(t)$ est une **somme** de termes, sa variance s'ecrit (sous hypotheses usuelles) :

$$\widehat{Var}(\hat{\Lambda}(t)) = \sum_{t_{(j)} \leq t} \frac{d_j}{n_j^2}.$$

Comparaison :

- ⇒ Kaplan–Meier : produit ⇒ variance plus complexe (Greenwood),
- ⇒ Breslow : somme ⇒ variance additive, forme simple.

A retenir : il est souvent plus simple d'analyser $\hat{\Lambda}(t)$, meme si $\hat{S}(t)$ est plus naturel a presenter graphiquement.

Kaplan–Meier vs methodes parametriques

⇒ **Kaplan–Meier** (non parametrique) :
- ne suppose aucune forme fonctionnelle pour la distribution de T ;
- gere naturellement la censure a droite ;
- ideal pour **decrire** et **explorer** les donnees ;
- moins efficace si l'on connait (ou suppose) une loi bien adaptee.

⇒ **Methodes parametriques** (exponentielle, Weibull, log-logistique) :
- reposent sur l'hypothese d'une loi specifique pour T ;
- plus precises si cette hypothese est correcte ;
- permettent d'extrapoler, de calculer un temps moyen/median, et d'integrer des covariables dans le modele ;
- plus sensibles aux erreurs de specification.

Regle pratique : utiliser Kaplan–Meier pour explorer et verifier, puis eventuellement ajuster un modele parametrique si c'est justifie.

Rappels

Hasard

De la durée brute à Kaplan–Meier

Greenwood

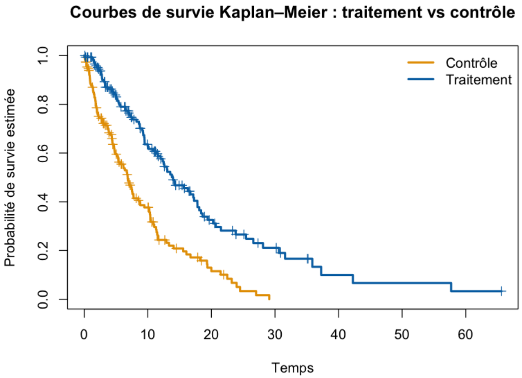
Breslow

Comparaison

Comparer deux courbes : intuition visuelle

Lire deux courbes de survie, c'est regarder :

- ⇒ qui sort (ou survit) plus vite ;
- ⇒ a quels moments les differences apparaissent ;



Pourquoi comparer deux groupes ?

Question empirique classique : est-ce qu'un groupe "survit" plus longtemps qu'un autre ?

- ⇒ traitement vs controle ;
- ⇒ hommes vs femmes ;
- ⇒ entreprises subventionnees vs non subventionnees.

Objectif statistique : tester si les deux fonctions de survie sont identiques sur *l'ensemble du suivi.*

Test du log-rank : intuition

A chaque instant de defaillance t_j :

- ⇒ on observe combien de defaillances dans chaque groupe ;
- ⇒ on calcule combien on *attendrait* si les groupes etaient identiques ;
- ⇒ on additionne les ecarts dans le temps.

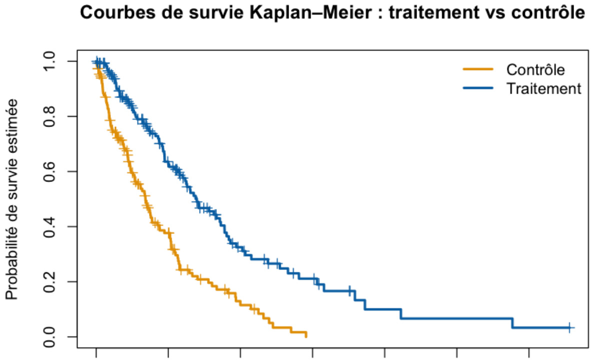
Idee clee : le log-rank detecte des differences *persistantes* entre les courbes sur toute la periode.

Log-rank : lecture graphique

⇒ Si une courbe est toujours en dessous de l'autre : ⇒ difference forte et coherente.

⇒ Si les courbes se croisent : ⇒ signes opposes des contributions, test moins puissant.

⇒ Les differences tardives comptent autant que les differences precoces.



Test du log-rank : principe formel

Hypothese nulle :

$$H_0 : S_1(t) = S_2(t) \quad \text{pour tout } t.$$

Idee : a chaque instant de defaillance t_j, on compare *ce que l'on observe* dans le groupe 1 a *ce que l'on attendrait* si les deux groupes etaient identiques.

Pour chaque t_j :

- ⇒ d_{1j} : defaillances observees dans le groupe 1 ;
- ⇒ d_j : defaillances totales (groupes 1+2) ;
- ⇒ $E_{1j} = \frac{n_{1j}}{n_j} d_j$: defaillances attendues sous H_0 ;
- ⇒ on regarde l'ecart $d_{1j} - E_{1j}$.

Statistique du test :

$$\chi^2_{\mathrm{LR}} = \frac{\left[\sum_j (d_{1j} - E_{1j})\right]^2}{\sum_j Var(d_{1j})} \quad \overset{H_0}{\sim} \quad \chi^2(1).$$

Comparer a un instant donne ou sur tout le suivi ?

⇒ **Question ponctuelle** : a t_0, $S_1(t_0)$ et $S_2(t_0)$ sont-elles differentes ?
$\rightarrow$ comparaison des IC.

⇒ **Question globale** : differencent-elles sur l'ensemble du suivi ?
$\rightarrow$ test du log-rank.

Un test peut etre non significatif, alors que les differences ponctuelles sont importantes (et inversement).

Limites du test du log-rank

Le log-rank est optimal si :

- ⇒ les courbes ne se croisent pas ;
- ⇒ les risques sont proportionnels (PH) ;
- ⇒ l'effet est a peu pres constant dans le temps.

Perte de puissance quand :

- ⇒ les courbes se croisent ;
- ⇒ l'effet est fort au debut puis disparait (ou l'inverse) ;
- ⇒ forte censure en fin de suivi.

Pourquoi utiliser des tests alternatifs ?

Le log-rank donne le meme poids a tous les instants :

$$w_j^{LR} = 1.$$

Mais dans beaucoup d'applications :

- ⇒ les differences sont precoces ;
- ⇒ le traitement agit tardivement ;
- ⇒ les courbes se croisent.

Solution : introduire des **poids** w_j pour cibler certains moments.

Tests ponderes : Wilcoxon et Tarone–Ware

Wilcoxon / Breslow :

$$w_j = n_j$$

⇒ poids fort au debut ;

⇒ differences precoces detectees efficacement.

Tarone–Ware :

$$w_j = \sqrt{n_j}$$

⇒ compromis entre debut et fin ;

⇒ utile si non-PH modere.

Regle pratique :

⇒ differences precoces → Wilcoxon ;

⇒ differences tardives → log-rank ;

⇒ courbes qui se croisent → Tarone–Ware.

Programme de la prochaine séance

1. **Rappel : modèles de Cox (PH continu)**
2. **Introduction aux modèles *paramétriques* de durée**
3. **Familles de modèles paramétriques**