

# Introduction à l'analyse de survie

## Modèle de Kaplan–Meier, modèle de Cox

Paul MINCHELLA  
[paul.minchella@lyon.unicancer.fr](mailto:paul.minchella@lyon.unicancer.fr)



- 1 Motivations et notations
- 2 Modélisation de la survie
- 3 Analyse de Kaplan–Meier
- 4 Modèle de Cox
- 5 Inférence et interprétation dans le modèle de Cox
- 6 Métriques de qualité
- 7 Synthèse de l'analyse de survie

- 1 Motivations et notations
- 2 Modélisation de la survie
- 3 Analyse de Kaplan–Meier
- 4 Modèle de Cox
- 5 Inférence et interprétation dans le modèle de Cox
- 6 Métriques de qualité
- 7 Synthèse de l'analyse de survie

### Motivation et contexte

De nombreuses situations réelles impliquent la mesure d'un **temps jusqu'à un évènement** :

- **Médecine** : date de diagnostic → date de décès ;
- **Criminologie** : sortie de prison → récidive ;
- **Ingénierie** : sortie d'usine → panne d'un appareil ;
- **Oncologie** : rémission → rechute.

### Problème fondamental

On s'intéresse à une variable aléatoire  $T$  représentant le **temps entre l'entrée dans l'étude et l'occurrence de l'évènement d'intérêt**. Pour certains individus, l'évènement n'est pas encore survenu à la fin du suivi : ils sont **censurés**.

⇒ Comment exploiter ces données incomplètes sans biais ?

## Objectifs du cours

- Estimer la probabilité de **survivre au-delà d'un horizon**  $t^*$  :

$$\mathbb{P}(T > t^*).$$

- Fournir :

une estimation **globale** (pour la cohorte) ;  
ou une estimation **individuelle** (tenant compte des covariables).

- Explorer deux familles d'approches :

**Non paramétriques** : empiriques, sans hypothèse sur la loi de  $T$  ;

**(Semi-)paramétriques** : interprétables, mais fondées sur un modèle.

*L'enjeu : apprendre à modéliser, estimer et interpréter le risque dans le temps, malgré la censure.*

### Éléments constitutifs d'une étude de survie

- (i) une **population source**,
- (ii) des **critères d'inclusion/exclusion**,
- (iii) une **origine du temps** (mise en service, diagnostic, randomisation),
- (iv) un **évènement d'intérêt**,
- (v) une **période d'observation**  $[T_0, T_{\text{end}}]$ ,
- (vi) des règles de **censure** (perte de vue, fin de suivi).

## Cadre conceptuel

Dans une étude de survie, on s'intéresse à un **temps aléatoire d'évènement**  $T$  (correspondant à la survenue d'une défaillance, d'une rechute ou d'un décès). Comme l'évènement peut ne pas être observé, on introduit :

$$Y = \min(T, C), \quad D = \mathbb{1}_{\{T \leq C\}},$$

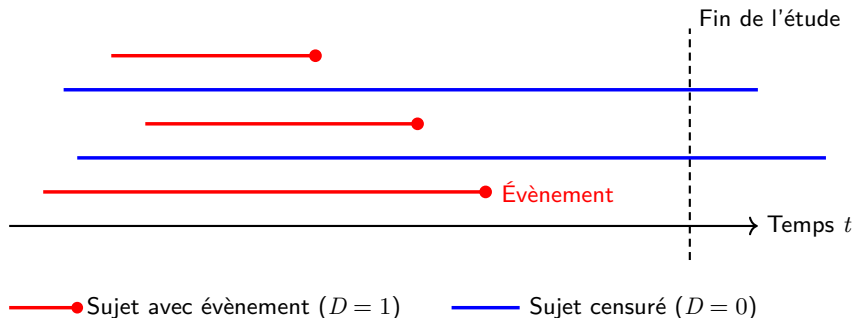
où :

- $C$  : temps de *censure*,
- $Y$  : durée observée,
- $D$  : indicateur d'observation de l'évènement.

ainsi que des covariables  $X \in \mathcal{X}$  (fixes ou dépendantes du temps). Pour  $i = 1, \dots, n$ , on note

$$(X_i, Y_i, D_i)$$

les observations individuelles, supposées indépendantes ou conditionnellement indépendantes.



**Figure:** Représentation schématique des temps de suivi dans une étude de survie. Les segments rouges indiquent les sujets ayant connu l'évènement ( $D = 1$ ), tandis que les segments bleus représentent les sujets censurés ( $D = 0$ ). La ligne verticale pointillée marque la fin de l'étude.

**Attention :** Pour que l'étude ait un sens, il faut **recentrer les temps d'observation** : à chaque instant, pour chaque individu, on doit retrancher son *temps d'entrée dans l'étude* (ou temps de départ). Cela permet d'exprimer toutes les durées relativement à une même origine temporelle ( $t = 0$ ) et de rendre les observations comparables entre individus. Sans cette harmonisation, les temps  $Y_i$  ne correspondraient pas à une même échelle, et la fonction de survie serait incohérente.



### Interprétation et intuition

- Dans une étude de survie, on ne connaît pas toujours  $T$  : si l'évènement ne s'est pas produit avant la fin du suivi, on sait seulement que  $T > C$ .
- On parle alors de **censure à droite**.
- La variable  $Y$  est la durée *observée*, et  $D$  indique si l'évènement a été observé ( $D = 1$ ) ou censuré ( $D = 0$ ).

### Pourquoi la censure est cruciale ?

- Tous les sujets ne connaissent pas nécessairement l'évènement avant la fin de l'étude : certains sont *perdus de vue* ou leur suivi s'interrompt prématurément.
- Exclure ces sujets introduirait un **biais d'analyse**. Même un individu censuré fournit de l'information, au moins jusqu'à son temps de censure !
- L'analyse de survie intègre donc à la fois :
  - les observations complètes (évènements observés),
  - et les observations incomplètes (censurées).

- 1 Motivations et notations
- 2 Modélisation de la survie**
- 3 Analyse de Kaplan–Meier
- 4 Modèle de Cox
- 5 Inférence et interprétation dans le modèle de Cox
- 6 Métriques de qualité
- 7 Synthèse de l'analyse de survie

### Définition

La **fonction de survie** de  $T$  est définie par :

$$\mathcal{S}(t) = \mathbb{P}(T > t), \quad t \geq 0.$$

La fonction de répartition correspondante est :

$$F(t) = \mathbb{P}(T \leq t) = 1 - \mathcal{S}(t).$$

Si  $T$  admet une densité  $f$ , alors :

$$f(t) = F'(t) \quad \text{presque partout.}$$

### Interprétation

- $\mathcal{S}(t)$  donne la **probabilité de survivre au-delà du temps  $t$** .
- Elle **décroît de 1** (au temps 0) vers 0 **à long terme**.
- Graphiquement, elle représente la proportion de sujets encore « en vie » au fil du temps.
- Sa dérivée  $f(t)$  exprime la **vitesse de survenue** des évènements autour de  $t$ .

## Définition

Le **hasard instantané** (ou *taux de risque*) est défini par :

$$h(t) = \lim_{dt \rightarrow 0} \frac{\mathbb{P}(t \leq T < t + dt \mid T \geq t)}{dt}, \quad (t > 0).$$

Le **risque cumulé** associé est :

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

### Interprétation

- $h(t)$  mesure la **probabilité instantanée de survenue** d'un évènement à l'instant  $t$ , *sachant que le sujet a survécu jusqu'à  $t$* .
- Il s'agit d'un **risque conditionnel instantané**, non d'une probabilité sur une durée finie.
- $H(t)$  est la **somme continue des micro-risques** subis au fil du temps :

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

- $H(t)$  représente donc le **risque total accumulé** entre le début du suivi et le temps  $t$ .

### Théorème : Lien fondamental entre $\mathcal{S}$ , $h$ et $H$

Si  $h$  est localement intégrable, alors pour tout  $t \geq 0$  :

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

### Remarque

Le lien  $f(t) = h(t) \mathcal{S}(t)$  traduit l'équilibre entre la *vitesse de survenue* des évènements et la *proportion encore à risque*.

*Preuve du théorème ?*

### Intérêt d'une approche infinitésimale

Cette formulation dépasse le cadre habituel de la régression ou de la classification :

- elle permet de modéliser des phénomènes évolutifs dans le temps, où le risque varie de façon continue ;
- elle offre une lecture fine du comportement des individus dans une cohorte, en capturant la **dynamique du risque instantané** plutôt qu'un simple état binaire (survie / échec) ;
- elle fournit un cadre unifié reliant la probabilité cumulative, la densité et le taux de risque.

Ainsi, la perspective infinitésimale est le cœur mathématique de l'analyse de survie, et fonde les modèles modernes tels que le modèle de Cox ou ses extensions neuronales.

- 1 Motivations et notations
- 2 Modélisation de la survie
- 3 Analyse de Kaplan–Meier**
- 4 Modèle de Cox
- 5 Inférence et interprétation dans le modèle de Cox
- 6 Métriques de qualité
- 7 Synthèse de l'analyse de survie



### Intérêts

- Pas de forme paramétrique imposée pour la loi de  $T$  ;
- Objectif : estimer **empiriquement** la fonction de survie à partir des données  $(Y_i, D_i)_{i=1}^n$  ;
- Prendre en compte la **censure** (observations incomplètes) ;
- Solution : l'**estimateur Kaplan and Meier 1958**.

### Cadre de l'observation

On dispose d'un échantillon  $(Y_i, D_i)$  pour  $i = 1, \dots, n$ , où :

$$Y_i = \min(T_i, C_i), \quad D_i = \mathbb{1}_{\{T_i \leq C_i\}}.$$

Les individus sont supposés **indépendants** et la **censure non-informative** :

$$T \perp\!\!\!\perp C.$$

Cette hypothèse garantit que la censure ne dépend pas du risque de l'évènement lui-même.

### Définition

Soient  $t_{(1)} < t_{(2)} < \dots < t_{(m)}$  les instants où au moins un évènement est observé ( $D_i = 1$ ). On définit :

- $d_j$  : nombre d'évènements observés à l'instant  $t_{(j)}$  ;
- $n_j$  : nombre d'individus encore à **risque** juste avant  $t_{(j)}$  (ceux tels que  $Y_i \geq t_{(j)}$ ).

L'**estimateur de Kaplan–Meier** de la fonction de survie est :

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

### Interprétation intuitive

- Chaque facteur  $\left( 1 - \frac{d_j}{n_j} \right)$  : probabilité empirique de **survivre à l'instant**  $t_{(j)}$ , sachant qu'on était encore à risque juste avant ;
- Le produit successif de ces facteurs : **probabilité de n'avoir connu aucun évènement jusqu'à**  $t$  ;
- $\hat{S}_{\text{KM}}(t)$  : **fonction en escalier décroissante**, constante entre deux évènements et chutant à chaque nouvelle occurrence.

### Idées clés

- Approche purement **non paramétrique** : aucune hypothèse sur la forme de la loi de  $T$ , seulement une agrégation empirique des risques observés.
- Survie = produit de probabilités de « continuer à survivre » à chaque évènement. Il s'agit en fait de la proportion de ceux restant dans l'étude à  $t_{(j)}$  ;
- Approche **multiplicative** : chaque évènement fait décroître  $\mathcal{S}(t)$  ;
- On ne modélise pas  $f(t)$  ni  $h(t) \rightarrow$  on estime directement  $\mathcal{S}(t) = \mathbb{P}(T > t)$  ;
- Chaque saut de la courbe correspond à un **évènement observé** (décès, panne, rechute, etc.) ;
- Méthode intuitive, non paramétrique, adaptée aux données censurées.

### Contexte

On suit  $n = 9$  patients traités pour un même type de cancer. Pour chacun, on observe la durée  $Y_i$  (en mois) jusqu'à la *rechute* ou jusqu'à la *censure* (fin de suivi, perte de vue).

Objectif : construire à la main la **courbe de survie empirique** à partir des données observées  $(Y_i, D_i)$ .

Patient $i$	$Y_i$ (mois)	$D_i$
1	1.0	1
2	1.7	0
3	2.2	1
4	3.5	1
5	4.3	0
6	5.0	1
7	6.2	0
8	7.0	1
9	7.5	0

### Travail demandé

- Identifier les instants  $t_{(1)} < t_{(2)} < \dots < t_{(m)}$  où  $D_i = 1$  ;
- Pour chaque  $t_{(j)}$ , déterminer :

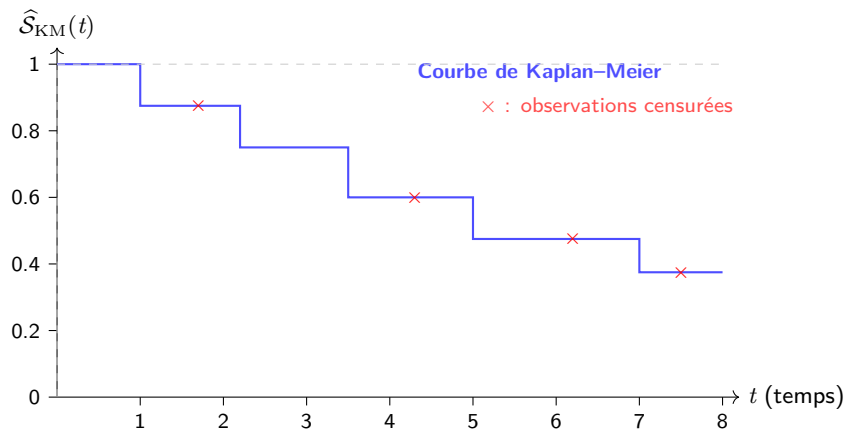
$$n_j = \text{nb. de sujets à risque juste avant } t_{(j)}, \quad d_j = \text{nb. d'évènements à } t_{(j)};$$

- Calculer  $\hat{S}(t_{(j)}) = \prod_{k: t_{(k)} \leq t_{(j)}} \left(1 - \frac{d_k}{n_k}\right)$  ;
- Tracer à la main la **courbe de survie en escalier** correspondante.

$i$	$Y_i$	$D_i$	$n_i$	$d_i$	$1 - \frac{d_i}{n_i}$	$\hat{S}_{\text{KM}}(t)$
1	1.0	1	9	1	$1 - \frac{1}{9} = 0.8889$	$1 \times 0.8889 = 0.8889$
2	1.7	0	8	0	1	0.8889
3	2.2	1	7	1	$1 - \frac{1}{7} = 0.8571$	$0.8889 \times 0.8571 = 0.7619$
4	3.5	1	6	1	$1 - \frac{1}{6} = 0.8333$	$0.7619 \times 0.8333 = 0.6349$
5	4.3	0	5	0	1	0.6349
6	5.0	1	4	1	$1 - \frac{1}{4} = 0.7500$	$0.6349 \times 0.7500 = 0.4762$
7	6.2	0	3	0	1	0.4762
8	7.0	1	2	1	$1 - \frac{1}{2} = 0.5000$	$0.4762 \times 0.5000 = 0.2381$
9	7.5	0	1	0	1	0.2381

## Lecture du tableau

- $n_i$  : nb. de sujets encore à risque juste avant  $Y_i$  ;
- $d_i$  : nb. d'évènements observés à cet instant ;
- La survie reste constante aux censures ( $D_i = 0$ ) ;
- Les valeurs finales suivent :  $1 \rightarrow 0.8889 \rightarrow 0.7619 \rightarrow 0.6349 \rightarrow 0.4762 \rightarrow 0.2381$ .



**Figure:** Exemple de fonction de survie estimée par l'approche de Kaplan-Meier avec indicateurs de censure (croix rouges).

## Propriété (Formule de Greenwood 1926)

Une estimation classique de la variance de  $\hat{S}_{\text{KM}}(t)$  est donnée par :

$$\widehat{\text{Var}}[\hat{S}_{\text{KM}}(t)] = \hat{S}_{\text{KM}}(t)^2 \sum_{j: t_{(j)} \leq t} \frac{d_j}{n_j (n_j - d_j)}.$$

## Construction d'un intervalle de confiance

Cette variance permet de construire un **intervalle de confiance**  $(1 - \alpha)$  pour la fonction de survie. Une approche courante repose sur une **transformation logarithmique** pour stabiliser la variance :

$$\left[ \log(-\log \hat{S}_{\text{KM}}(t)) \pm z_{1-\alpha/2} \frac{\sqrt{\widehat{\text{Var}}[\hat{S}_{\text{KM}}(t)]}}{\hat{S}_{\text{KM}}(t) \log \hat{S}_{\text{KM}}(t)} \right],$$

où  $z_{1-\alpha/2}$  est le quantile d'ordre  $1 - \alpha/2$  de la loi  $\mathcal{N}(0, 1)$ .

- L'intervalle est ensuite **retransformé par exponentielle inverse** pour revenir à l'échelle de  $\mathcal{S}(t)$ .
- Cette approche est préférée à un intervalle linéaire car elle garantit des bornes comprises entre 0 et 1.



### Définition

Supposons que la population soit divisée en  $K$  groupes distincts (par exemple selon un traitement). Pour chaque groupe  $k$ , on estime une fonction de survie  $\mathcal{S}_k(t)$  à l'aide de l'estimateur de Kaplan–Meier, calculé sur les observations de ce groupe uniquement :

$$\widehat{\mathcal{S}}_k(t) = \prod_{j: t_{k,(j)} \leq t} \left( 1 - \frac{d_{k,j}}{n_{k,j}} \right),$$

où :

- $d_{k,j}$  = nb. d'évènements observés dans le groupe  $k$  à  $t_{k,(j)}$ ,
- $n_{k,j}$  = nb. d'individus encore à risque juste avant  $t_{k,(j)}$ .

### Motivation

Comparer plusieurs courbes de survie permet d'évaluer l'effet d'un traitement, d'un facteur de risque ou d'une exposition. La question centrale est alors : *les fonctions de survie des groupes sont-elles significativement différentes ?*

### Référence

Test introduit par [Mantel 1966](#), puis formalisé par les frères [Peto and Peto 1972](#).

### Définition (Hypothèses du test)

On souhaite tester l'hypothèse selon laquelle deux groupes présentent la même fonction de survie :

$$(H_0) : \quad \forall t \geq 0, S_1(t) = S_2(t),$$

$$(H_1) : \quad \exists t \geq 0, S_1(t) \neq S_2(t).$$

### Idée clé

Le test compare, à chaque instant d'évènement  $t_{(j)}$ , le **nombre observé d'évènements** dans le groupe 1 ( $d_{1j}$ ) avec le **nombre attendu**  $e_{1j}$  sous  $H_0$  (mêmes taux de risque instantané).

## Statistique du test

$$Z = \frac{\sum_j (d_{1j} - e_{1j})}{\sqrt{\sum_j v_{1j}}}, \quad Z^2 \sim \chi_1^2 \quad (\text{asymptotiquement sous } H_0).$$

où  $v_{1j}$  est la variance de  $(d_{1j} - e_{1j})$  sous  $H_0$ .

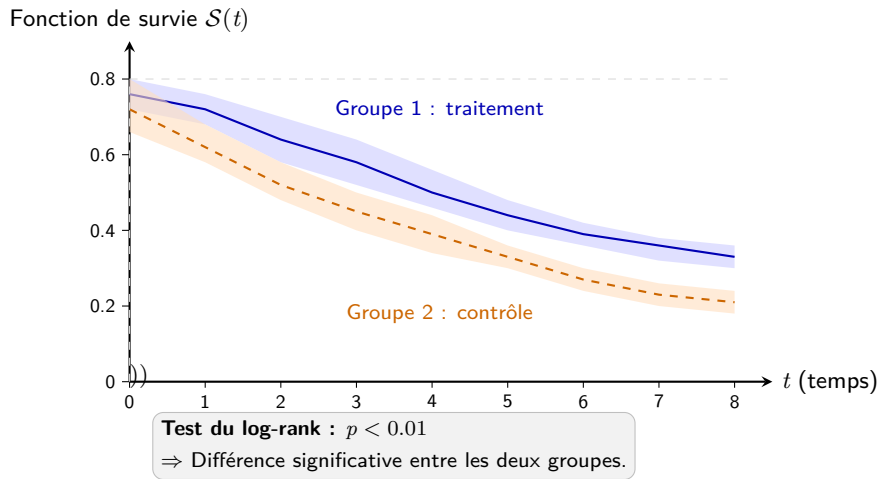
## Décision

Étant donnée  $p\text{-value} = \mathbb{P}(\chi_1^2 \geq z_{\text{obs}}^2)$ , si  $p\text{-value} < \alpha$ , on rejette  $(H_0)$  (différence significative entre les courbes).

## Remarque pratique

Le test du log-rank est :

- **non paramétrique** : aucune hypothèse sur la forme du risque,
- **robuste** : s'applique sous censure non-informative,
- **limité** : sensible aux différences globales, moins puissant si les courbes se croisent.



**Figure:** Comparaison de deux courbes de survie estimées (Kaplan–Meier) avec intervalles de confiance à 95%. Le résultat renvoyé par le test du log-rank permet de conclure quant à l'efficacité du traitement.

### Ce que permet l'approche

L'estimateur de Kaplan–Meier offre une description empirique complète de la cohorte :

- estimation directe de la **fonction de survie**  $\hat{S}(t)$ , sans hypothèse paramétrique ;
- prise en compte naturelle de la **censure à droite** ;
- possibilité d'estimer la probabilité de survie à un horizon fixé  $t^*$  ;
- visualisation simple et interprétable via la **courbe en escalier**.

### Intérêt pratique

- Fournit un résumé clair du suivi de la cohorte ;
- Sert de référence pour comparer plusieurs groupes (test du log-rank) ;
- Constitue la base des représentations graphiques de survie utilisées en pratique clinique et en recherche.

### Limite

Kaplan–Meier ne tient pas compte directement des **covariables explicatives** (âge, traitement, etc.). → Pour modéliser ces effets, on utilisera les **modèles de régression**, notamment le **modèle de Cox**.

- 1 Motivations et notations
- 2 Modélisation de la survie
- 3 Analyse de Kaplan–Meier
- 4 Modèle de Cox**
- 5 Inférence et interprétation dans le modèle de Cox
- 6 Métriques de qualité
- 7 Synthèse de l'analyse de survie

### Définition (Cox, 1972)

Dans un contexte  $(\mathbb{X}_i, Y_i, D_i)_{i=1}^n$ , le modèle [Cox 1972](#) postule une **séparabilité multiplicative** du risque instantané :

$$h(t \mid \mathbb{X}) = h_0(t) \exp(\beta^\top \mathbb{X}),$$

où :

- $h_0(t)$  : **fonction de risque de base** (non paramétrique), décrivant la dynamique temporelle globale ;
- $\exp(\beta^\top \mathbb{X})$  : **facteur multiplicatif** propre à chaque individu ;
- $\beta^\top \mathbb{X}$  : **score de risque** ou *risk score*, noté  $\eta$ .

### Interprétation

Cette hypothèse exprime une idée de **séparation des effets** : le risque dépend d'un facteur *temporel global*  $h_0(t)$  et d'un facteur *structurel* propre à l'individu. Autrement dit, la dynamique temporelle est commune, mais modulée par les covariables.

### Propriété : risque relatif

Pour deux individus  $i$  et  $j$  :

$$\frac{h(t \mid \mathbb{X}_i)}{h(t \mid \mathbb{X}_j)} = \exp(\beta^\top (\mathbb{X}_i - \mathbb{X}_j)),$$

indépendamment du temps  $t$ .

### Conséquence

Les rapports de risques instantanés sont **constants dans le temps**. Ainsi :

- les courbes de survie peuvent se croiser verticalement (différence de niveau) ;
- mais jamais horizontalement (même forme temporelle).

### Remarque

Le modèle de Cox est dit à **risques proportionnels** : les effets des covariables s'expriment uniquement via un facteur multiplicatif sur le risque de base.



### Idée de Cox (1975)

Plutôt que de modéliser  $h_0(t)$ , Cox introduit la **vraisemblance partielle** :

$$L_p(\beta) = \prod_{j=1}^m \frac{\exp(\beta^\top \mathbb{X}_{(j)})}{\sum_{i \in \mathcal{R}(t_{(j)})} \exp(\beta^\top \mathbb{X}_i)},$$

où  $\mathcal{R}(t_{(j)}) = \{i \in \{1, \dots, n\}, t_{(j)} \leq T_i\}$  est l'ensemble des individus encore à risque juste avant  $t_{(j)}$ .

### Interprétation intuitive

À chaque instant d'évènement :

- on compare la probabilité que l'individu défaillant soit celui observé,
- parmi tous ceux encore en compétition ;
- la fonction de risque de base  $h_0(t)$  s'annule dans le rapport.

## Forme logarithmique

En pratique, on maximise :

$$\ell_p(\beta) = \sum_{j=1}^m \left[ \beta^\top \mathbb{X}_{(j)} - \log \left( \sum_{i \in \mathcal{R}(t_{(j)})} \exp(\beta^\top \mathbb{X}_i) \right) \right].$$

## Lien avec le *softmax*

L'expression précédente est **équivalente à une log-vraisemblance de type softmax** : chaque individu à risque reçoit un score  $\exp(\beta^\top \mathbb{X}_i)$ , et la probabilité d'être l'évènement observé est une normalisation exponentielle sur  $\mathcal{R}(t_{(j)})$ .

## Estimation du vecteur de coefficients

$$\hat{\beta} = \arg \max_{\beta \in \mathbb{R}^p} \ell_p(\beta),$$

souvent obtenue par des méthodes numériques (gradient, Newton–Raphson, etc.).

## Estimateur de Breslow 1974

Une fois  $\hat{\beta}$  estimé, la **fonction de risque cumulée de base**  $H_0(t)$  est estimée par :

$$\hat{H}_0(t) = \sum_{j: t_{(j)} \leq t} \frac{d_j}{\sum_{i \in \mathcal{R}(t_{(j)})} \exp(\hat{\beta}^\top \mathbb{X}_i)},$$

où :

- $d_j$  : nombre d'évènements observés à  $t_{(j)}$ ,  $d_j = \sum_{i=1}^n \mathbb{1}_{\{T_i=t_{(j)}, D_i=1\}}$  ;
- $\mathcal{R}(t_{(j)})$  : ensemble des individus encore à risque juste avant  $t_{(j)}$ .

## Interprétation

- Le **numérateur**  $d_j$  correspond aux évènements réellement observés ;
- Le **dénominateur** représente le risque total « exposé » à  $t_{(j)}$ , pondéré par les effets estimés des covariables ;
- Chaque fraction  $\frac{d_j}{\sum_{i \in \mathcal{R}(t_{(j)})} \exp(\hat{\beta}^\top \mathbb{X}_i)}$  mesure la contribution empirique au risque cumulatif de base.

### Fonction de survie prédite

À partir de l'estimateur de Breslow, la fonction de survie d'un individu  $i$  de covariables  $\mathbb{X}_i$  s'écrit :

$$\hat{\mathcal{S}}(t | \mathbb{X}_i) = \exp \left[ - \hat{H}_0(t) \exp(\hat{\beta}^\top \mathbb{X}_i) \right].$$

### Interprétation

Cette formule illustre la **séparabilité du modèle de Cox** :

- $\hat{H}_0(t)$  : partie **non paramétrique**, capturant la dynamique temporelle ;
- $\exp(\hat{\beta}^\top \mathbb{X}_i)$  : partie **paramétrique**, représentant l'effet des covariables ;
- leur produit : **interaction multiplicative** entre temps et caractéristiques individuelles.

### Conclusion

Cette dualité — flexibilité temporelle ( $\hat{H}_0$ ) et linéarité covariable ( $\beta$ ) — explique la **longévité et la popularité du modèle de Cox**, encore aujourd'hui la référence des modèles semi-paramétriques de survie.

- 1 Motivations et notations
- 2 Modélisation de la survie
- 3 Analyse de Kaplan–Meier
- 4 Modèle de Cox
- 5 Inférence et interprétation dans le modèle de Cox**
- 6 Métriques de qualité
- 7 Synthèse de l'analyse de survie

### Motivation

L'intérêt ici est de **quantifier l'incertitude** sur les coefficients estimés  $\hat{\beta}$  et de proposer une **étude inférentielle rigoureuse**.

### Idée générale

Après estimation du vecteur de coefficients  $\hat{\beta}$  par vraisemblance partielle :

- on cherche à **évaluer la précision** de ces estimateurs ;
- on veut pouvoir construire des **intervalles de confiance** et des **tests d'hypothèses** ;
- cela repose sur la **normalité asymptotique** de  $\hat{\beta}$  lorsque  $n$  est grand.

### Principe

Sous les hypothèses de régularité et de proportionnalité des risques :

$$\hat{\beta} \xrightarrow{\text{asympt.}} \mathcal{N}(\beta, \mathcal{I}(\hat{\beta})^{-1}).$$

- $\mathcal{I}(\hat{\beta})$  : matrice d'information observée (Hessienne négative) ;
- $\mathcal{I}^{-1}$  : matrice de variance-covariance asymptotique de  $\hat{\beta}$ .

### Matrice d'information observée

La **matrice d'information de Fisher observée** est définie comme :

$$\mathcal{I}(\hat{\beta}) = - \frac{\partial^2 \ell_p(\beta)}{\partial \beta \partial \beta^\top} \Big|_{\beta = \hat{\beta}}.$$

Elle représente la **courbure locale** de la log-vraisemblance partielle autour du maximum :

- une courbure forte  $\Rightarrow$  estimateur précis ;
- une courbure faible  $\Rightarrow$  grande incertitude.

### Variance et interprétation

$$\widehat{\text{Var}}(\hat{\beta}) = \mathcal{I}(\hat{\beta})^{-1}, \quad \widehat{\text{Var}}(\hat{\beta}_j) = [\mathcal{I}(\hat{\beta})^{-1}]_{jj}.$$

Chaque variance diagonale traduit l'**incertitude sur un effet covariable particulier**.

### Résumé pédagogique

- L'inférence du modèle de Cox repose sur la **normalité asymptotique** de  $\hat{\beta}$ .
- La **matrice d'information observée** mesure la stabilité numérique de l'estimation.

### Objectif

Quantifier l'incertitude associée à chaque coefficient estimé  $\hat{\beta}_j$  dans le modèle de Cox, afin de juger la significativité des covariables.

### Propriété : intervalle de confiance asymptotique

Sous l'hypothèse de normalité asymptotique :

$$\hat{\beta}_j \pm z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{\beta}_j)},$$

où :

- $z_{1-\alpha/2}$  est le quantile de la loi normale standard  $\mathcal{N}(0, 1)$  ;
- par exemple :  $z_{0.975} \approx 1.96$  pour un niveau de confiance de 95%.

### Interprétation pratique

- Si l'intervalle **ne contient pas 0**, la covariable a un effet significatif sur le risque ;
- La largeur de l'intervalle reflète la **précision de l'estimation** : plus il est étroit, plus la covariable est estimée avec confiance.



## Test de Wald

On souhaite tester, pour chaque coefficient  $\beta_j$  :

$$\begin{cases} H_0 : \beta_j = 0, \\ H_1 : \beta_j \neq 0. \end{cases}$$

La statistique de test est :

$$Z_j = \frac{\hat{\beta}_j}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_j)}}, \quad \text{avec} \quad Z_j^2 \sim \chi_1^2 \text{ sous } H_0.$$

## Interprétation

- Le test évalue si le rapport **estimation / incertitude** est suffisamment grand pour conclure à un effet réel ;
- Un  $|Z_j|$  élevé ou une  $p$ -valeur faible  $\Rightarrow$  effet significatif de la covariable  $X_j$  ;
- Si  $p < \alpha$  (souvent 0.05)  $\Rightarrow$  rejet de  $H_0$ , la variable est jugée pertinente.

### Autres tests (brièvement)

- **Test du rapport de vraisemblance** : compare la log-vraisemblance du modèle complet à celle du modèle restreint ( $\beta_j = 0$ ) ;
- **Test du score (Rao)** : basé sur la dérivée première de la log-vraisemblance, utile avant ajustement complet ;
- Ces trois tests sont **asymptotiquement équivalents** sous les hypothèses du modèle de Cox.

### Définition

Dans le modèle de Cox

$$h(t | \mathbb{X}) = h_0(t) \exp(\beta^\top \mathbb{X}),$$

le **hazard ratio (HR)** associé à une variation d'une unité de la covariable  $X_j$  est :

$$\text{HR}_j = \exp(\beta_j).$$

### Interprétation du HR

- $\beta_j > 0 \Rightarrow \text{HR}_j > 1$  : la covariable **augmente le risque instantané**.
- $\beta_j < 0 \Rightarrow \text{HR}_j < 1$  : la covariable **réduit le risque instantané**.
- $\beta_j = 0 \Rightarrow \text{HR}_j = 1$  : la covariable n'a **pas d'effet** sur le risque.

Le HR mesure l'**effet multiplicatif d'une covariable sur le risque**, constant dans le temps (hypothèse de proportionnalité).

Intervalle de confiance du HR à  $100(1 - \alpha)\%$

$$\left[ \exp(\hat{\beta}_j - z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{\beta}_j)}), \exp(\hat{\beta}_j + z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{\beta}_j)}) \right].$$

### Lecture pratique

Un  $\text{HR} = 1.5$  indique une **augmentation de 50% du risque instantané**. Cependant, si l'intervalle de confiance inclut 1, l'effet n'est **pas statistiquement significatif**.

- 1 Motivations et notations
- 2 Modélisation de la survie
- 3 Analyse de Kaplan–Meier
- 4 Modèle de Cox
- 5 Inférence et interprétation dans le modèle de Cox
- 6 Métriques de qualité**
- 7 Synthèse de l'analyse de survie

## Contexte

Évaluer un modèle de survie = étape clé. Objectif : mesurer **ajustement** + **prédiction**. Difficulté : présence de **censure**  $\Rightarrow$  contributions inégales aux métriques.

### Définition

L'indice de concordance mesure la capacité du modèle à ordonner correctement les individus selon leur risque prédit :

$$\text{C-index} = \frac{\sum_{i,j} \mathbb{1}_{\{T_j < T_i\}} \cdot \mathbb{1}_{\{\hat{\eta}_j > \hat{\eta}_i\}} \cdot \delta_j}{\sum_{i,j} \mathbb{1}_{\{T_j < T_i\}} \cdot \delta_j}.$$

### Interprétation

- Représente la proportion de paires correctement classées par le modèle ;
- $C = 0.5$  : modèle aléatoire ;  $C = 1$  : discrimination parfaite.

### Lien avec l'AUC

En absence de censure, le C-index se confond avec l'**AUC ROC** : c'est une généralisation naturelle de l'AUC au cadre de la survie, adaptée aux comparaisons partielles.

### Définition

Pour mesurer la discrimination à un instant  $t$  donné :

$$\widehat{\text{AUC}}(t) = \frac{\sum_{i,j} \mathbb{1}_{\{T_i > t\}} \mathbb{1}_{\{T_j \leq t\}} \mathbb{1}_{\{\hat{\eta}_j > \hat{\eta}_i\}} \delta_j(t)}{\sum_{i,j} \mathbb{1}_{\{T_i > t\}} \mathbb{1}_{\{T_j \leq t\}} \delta_j(t)}.$$

### Intuition

- À chaque instant  $t$ , on compare les sujets encore en vie ( $T_i > t$ ) à ceux ayant connu l'évènement ( $T_j \leq t$ ) ;
- $\widehat{\text{AUC}}(t)$  évalue la capacité du modèle à distinguer ces deux groupes ;
- On peut tracer la courbe  $t \mapsto \widehat{\text{AUC}}(t)$  pour observer la discrimination au fil du temps.



## Définition

Le Brier Score mesure la précision des probabilités de survie prédites à un horizon  $t$  :

$$\text{BS}(t) = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \left[ \mathbb{1}_{\{T_i \leq t\}} \frac{(0 - \hat{S}(t | \mathbb{X}_i))^2}{\hat{G}(T_i)} \delta_i + \mathbb{1}_{\{T_i > t\}} \frac{(1 - \hat{S}(t | \mathbb{X}_i))^2}{\hat{G}(t)} \right],$$

où  $\hat{G}$  est la fonction de survie estimée de la censure.

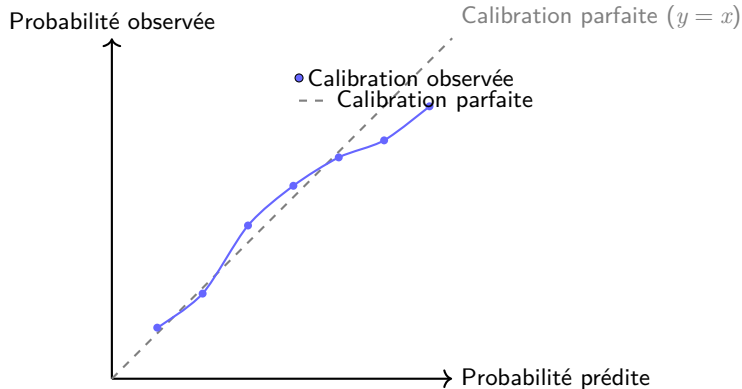
## Interprétation

- Mesure l'écart quadratique entre la survie prédite et observée ;
- $\text{BS}(t) \in [0, 1]$  — plus il est faible, meilleure est la calibration ;
- Un modèle parfaitement calibré aurait  $\text{BS}(t) = 0$  pour tout  $t$ .

### Principe

Pour un horizon  $t^*$ , on compare les survies prédites  $\hat{S}(t^* | \mathbb{X}_i)$  aux proportions observées d'individus effectivement en vie :

Probabilité prédite vs. Probabilité observée.



## Comparaison empirique

Pour évaluer la qualité du modèle de Cox, on compare sa survie marginale moyenne :

$$\hat{S}_{\text{Cox}}(t) = \frac{1}{n} \sum_{i=1}^n \hat{S}(t \mid \mathbb{X}_i),$$

à la courbe non paramétrique de Kaplan–Meier  $\hat{S}_{\text{KM}}(t)$  [Hosmer, Lemeshow, and May 2008](#).

**Intérêt de la comparaison :**

- Bonne concordance  $\Rightarrow$  modèle bien ajusté ;
- Divergence systématique  $\Rightarrow$  hypothèse de proportionnalité non respectée ou mauvais ajustement global.

## Synthèse des métriques

Deux dimensions essentielles de la performance :

- **Discrimination** : le modèle distingue-t-il bien les individus selon leur risque ? (C-index,  $\text{AUC}(t)$ )
- **Calibration** : les probabilités prédites reflètent-elles la réalité observée ? (Brier Score, calibration plot)

*Une évaluation complète combine discrimination, calibration et ajustement global (AIC/BIC, validation croisée).*

- 1 Motivations et notations
- 2 Modélisation de la survie
- 3 Analyse de Kaplan–Meier
- 4 Modèle de Cox
- 5 Inférence et interprétation dans le modèle de Cox
- 6 Métriques de qualité
- 7 Synthèse de l'analyse de survie**

## Bilan conceptuel

On a ainsi présenté deux modèles de survie *gérant la censure*. L'objectif est d'**interpréter**, de **valider** et d'**exploiter** les résultats :







- **Kaplan–Meier** : estimation non paramétrique de la survie globale, intégrant la censure ;
- **Cox** : modèle semi-paramétrique reliant le risque instantané aux covariables via  $h(t|\mathbb{X}) = h_0(t) \exp(\mathbb{X}^\top \hat{\beta})$ .

## Applications principales.

- **Survie globale** : comparer  $\hat{S}_{\text{Cox}}(t)$  et  $\hat{S}_{\text{KM}}(t)$  pour juger de la calibration du modèle ;
- **Stratification du risque** : classer les individus selon le score  $\eta_i = \mathbb{X}_i^\top \hat{\beta}$  ;
- **Interprétation des coefficients** : à travers les *Hazard Ratios*  $\text{HR}_j = \exp(\hat{\beta}_j)$  ;
- **Prédiction individuelle** : estimer  $\hat{S}(t|\mathbb{X}^*)$  pour un profil donné.

## Enjeux pratiques

- Combiner **interprétabilité** (Cox) et **robustesse empirique** (Kaplan–Meier) ;
- Quantifier l'**incertitude** (IC, tests) et valider la **proportionnalité des risques** ;
- Utiliser les résultats pour la **décision clinique**, la **stratification de cohortes** et la **prévision de survie**.

-  Breslow, Norman E. (1974). "Covariance Analysis of Censored Survival Data". In: *Biometrika* 61.3, pp. 579–594. DOI: [10.2307/2334738](https://doi.org/10.2307/2334738).
-  Cox, David Roxbee (1972). "Regression Models and Life-Tables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2, pp. 187–202. DOI: <https://doi.org/10.1111/j.2517-6161.1972.tb00899.x>.
-  Greenwood, Major (1926). *The Natural Duration of Cancer*. Reports on Public Health and Medical Subjects 33. London: His Majesty's Stationery Office (H.M.S.O.)
-  Hosmer, David W., Stanley Lemeshow, and Susanne May (2008). *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. 2nd. Hoboken, NJ: Wiley-Interscience. DOI: [10.1002/9780470258019](https://doi.org/10.1002/9780470258019).
-  Kaplan, Edward L. and Paul Meier (1958). "Nonparametric Estimation from Incomplete Observations". In: *Journal of the American Statistical Association* 53.282, pp. 457–481. DOI: [10.1080/01621459.1958.10501452](https://doi.org/10.1080/01621459.1958.10501452).
-  Mantel, Nathan (1966). "Evaluation of Survival Data and Two New Rank Order Statistics Arising in its Consideration". In: *Cancer Chemotherapy Reports* 50, pp. 163–170.



Peto, Richard and Julian Peto (1972). "Asymptotically Efficient Rank Invariant Test Procedures". In: *Journal of the Royal Statistical Society: Series A (General)* 135.2, pp. 185–207. DOI: [10.2307/2344317](https://doi.org/10.2307/2344317).