Modèles de durée

Théorie et applications

Etienne Dagorn

Université de Lille - LEM

Objectifs de la séance

À l'issue de cette séance, vous serez capables de :

- ⇒ caractériser et interpréter les principales lois paramétriques en analyse de durée (exponentielle, Weibull, log-logistique) ;
- ⇒ distinguer clairement les modèles *proportionnalité des hasards* et *temps accéléré*, et comprendre dans quels contextes chacun est pertinent ;
- ⇒ sélectionner une loi paramétrique plausible à partir de la forme observée de la fonction de hasard ;
- ⇒ situer les modèles paramétriques dans l'ensemble des outils d'analyse de durée, notamment par rapport à l'estimateur de Kaplan-Meier et au modèle semi-paramétrique de Cox.

Exemple fil rouge : durée de chômage

Contexte empirique :

⇒ Population : jeunes sortant d'études en 2020.

⇒ Variable d'intérêt : durée (en mois) avant l'accès au premier emploi stable.

⇒ Censure :

- fin d'observation à 24 mois ;
- certains individus sont encore au chômage à 24 mois.

⇒ Covariables :

- diplôme, spécialité, sexe ;
- participation à une formation, logement parental, etc.

Tout au long de la séance, imaginez que l'on cherche à modéliser **cette durée de chômage** avec différentes lois (Expo/Weibull/log-logistique) et différents cadres (PH/AFT).

Rappels

Motivation

Lois de base

Extensions

Panorama & extensions

Rappels :
Analyse non-paramétrique des durées

Cas pratique : comprendre les durées

Individu	Temps t_i	Événement ? (δ_i)
1	3	1
2	5	0 (censuré)
3	6	1
4	6	1
5	9	0 (censuré)
6	10	1

1. Quels sont les **temps de défaillance uniques** $t_{(j)}$?
2. Quel est le **risk set** n_j juste avant chaque défaillance ?
3. Quelle est la contribution de chaque individu à :
   - la **Kaplan–Meier** : $\hat{S}(t)$?
   - la **Nelson–Aalen** : $\hat{H}(t)$?
4. Quelle serait l'**estimation du hasard instantané** entre 3 et 6 ?
5. deux groupes (hommes/femmes) : que testerait le **log-rank** ?

Cas pratique : corrigé (1/2)

1. Temps de défaillance uniques :

$$t_{(1)} = 3, \quad t_{(2)} = 6, \quad t_{(3)} = 10.$$

2. Risk sets juste avant chaque événement :

$$n_1 = 6, \qquad n_2 = 4 \ (\{3, 4, 5, 6\}), \qquad n_3 = 1 \ (\{6\}).$$

3. Contributions KM et Nelson–Aalen :

$t_{(j)}$	KM :	NA : $\Delta \hat{H}_j$
3	$1 - \frac{1}{6} = \frac{5}{6}$	$\frac{1}{6}$
6	$1 - \frac{2}{4} = \frac{1}{2}$	$\frac{2}{4} = \frac{1}{2}$
10	$1 - \frac{1}{1} = 0$	1

Cas pratique : corrigé (2/2)

4. Estimation du hasard instantané entre 3 et 6 :

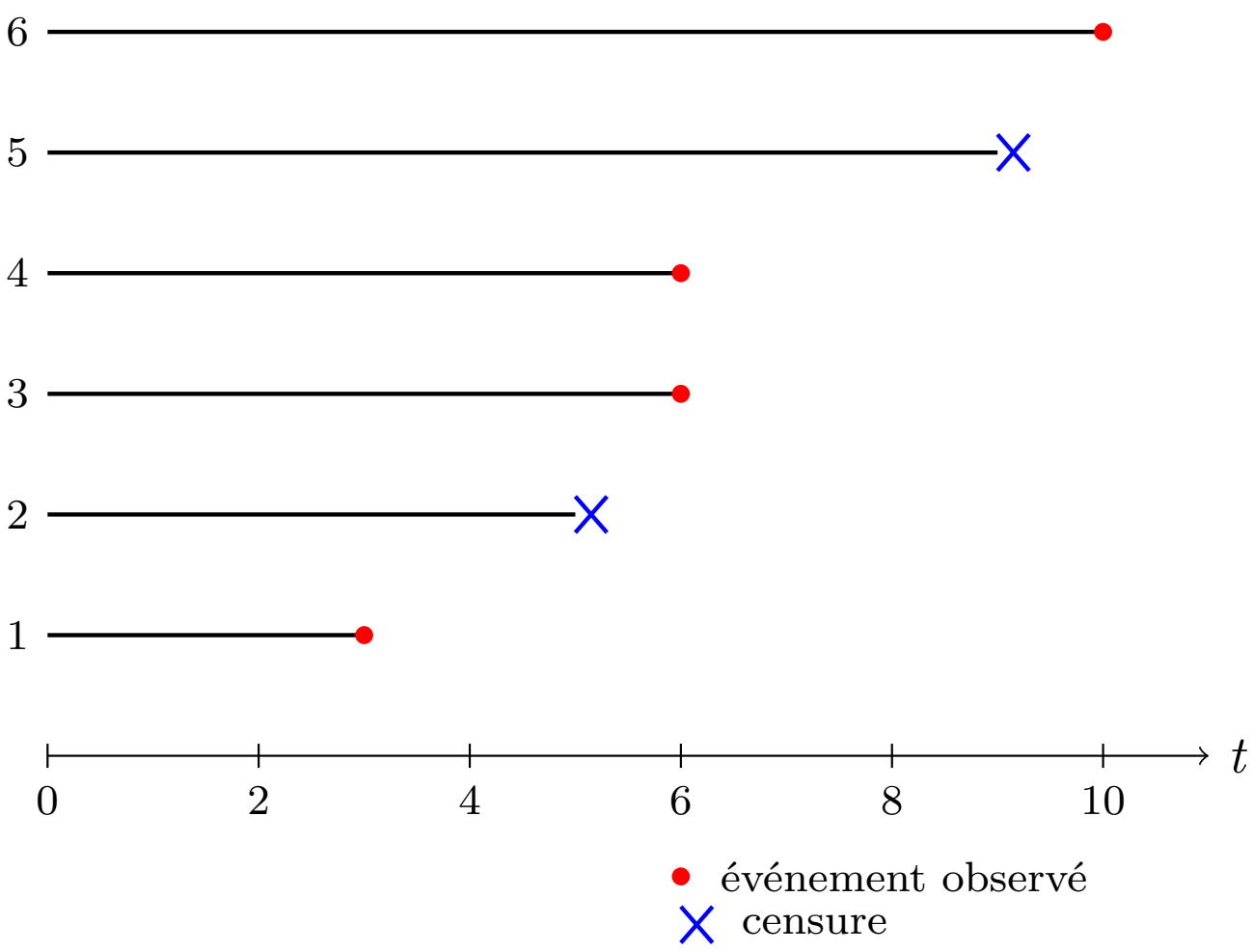
$$\hat{h}(6) \approx \frac{d_2}{n_2} = \frac{2}{4} = 0.5.$$

5. Pour deux groupes (H/F), le test du log-rank évalue :

$$H_0 : S_H(t) = S_F(t)\ \forall t \quad \Leftrightarrow \quad h_H(t) = h_F(t) \ \text{ (ratio de hasards} = 1).$$

Le test du log-rank compare, à chaque temps d'événement, les **défaillances observées** à celles **attendues sous** H_0.

Cas pratique : visualiser les durées et censures



Construction des **risk sets** et donc de $\hat{S}_{\mathrm{KM}}(t)$ et $\hat{H}_{\mathrm{NA}}(t)$.

Quiz : objets fondamentaux (1/6)

Question 1. Quelle(s) relation(s) est/sont toujours vraie(s) ?

a) $S(t) = 1 - F(t)$

b) $h(t) = \dfrac{f(t)}{S(t)}$

c) $S(t) = e^{-H(t)}$

d) $H(t) = \displaystyle\int_0^t f(u)\, du$

Quiz : objets fondamentaux (1/6)

Question 1. Quelle(s) relation(s) est/sont toujours vraie(s) ?

a) $S(t) = 1 - F(t)$

- Vrai pour toute variable de durée : définition générale de la fonction de survie.

b) $h(t) = \dfrac{f(t)}{S(t)}$

c) $S(t) = e^{-H(t)}$

d) $H(t) = \displaystyle\int_0^t f(u)\, du$

Quiz : objets fondamentaux (1/6)

Question 1. Quelle(s) relation(s) est/sont toujours vraie(s) ?

a) $S(t) = 1 - F(t)$

b) $h(t) = \dfrac{f(t)}{S(t)}$

- Vrai en temps continu : la densité $f(t)$ doit exister.
- Pour des durées discrètes, la définition du hazard est différente.

c) $S(t) = e^{-H(t)}$

d) $H(t) = \displaystyle\int_0^t f(u)\, du$

Quiz : objets fondamentaux (1/6)

Question 1. Quelle(s) relation(s) est/sont toujours vraie(s) ?

a) $S(t) = 1 - F(t)$

b) $h(t) = \dfrac{f(t)}{S(t)}$

c) $S(t) = e^{-H(t)}$

- Exact pour un hazard **continu** :
  $H(t) = \int_0^t h(u)\,du \;\Rightarrow\; S(t) = e^{-H(t)}$.
- En temps discret (ou en présence de ties), c'est une approximation.

d) $H(t) = \displaystyle\int_0^t f(u)\,du$

Quiz : objets fondamentaux (1/6)

Question 1. Quelle(s) relation(s) est/sont toujours vraie(s) ?

a) $S(t) = 1 - F(t)$

b) $h(t) = \dfrac{f(t)}{S(t)}$

c) $S(t) = e^{-H(t)}$

d) $H(t) = \displaystyle\int_0^t f(u)\, du$

- Faux : le hasard cumulé est

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

avec $h(u) = f(u)/S(u)$ en continu.

Quiz : censure à droite (2/6)

Question 2. Laquelle de ces affirmations est correcte ?

a) La censure réduit la taille de l'échantillon.

b) Un individu censuré apporte l'information : $T > t$.

c) La censure informative ne pose pas de problème pour KM.

d) La censure doit être indépendante du temps propre.

Quiz : censure à droite (2/6)

Question 2. Laquelle de ces affirmations est correcte ?

a) La censure réduit la taille de l'échantillon.

- Faux : un individu censuré est **conservé** dans l'analyse ; il quitte seulement le risk set après son temps de censure.

b) Un individu censuré apporte l'information : $T > t$.

c) La censure informative ne pose pas de problème pour KM.

d) La censure doit être indépendante du temps propre.

Quiz : censure à droite (2/6)

Question 2. Laquelle de ces affirmations est correcte ?

a) La censure réduit la taille de l'échantillon.

b) Un individu censuré apporte l'information : $T > t$.

- Vrai : la censure fournit une **borne inférieure** sur la durée réelle.

c) La censure informative ne pose pas de problème pour KM.

d) La censure doit être indépendante du temps propre.

Quiz : censure à droite (2/6)

Question 2. Laquelle de ces affirmations est correcte ?

a) La censure réduit la taille de l'échantillon.

b) Un individu censuré apporte l'information : $T > t$.

c) La censure informative ne pose pas de problème pour KM.

- Faux : la censure **informative** viole l'hypothèse clé de KM/NA et entraîne un **biais systématique**.

d) La censure doit être indépendante du temps propre.

Quiz : censure à droite (2/6)

Question 2. Laquelle de ces affirmations est correcte ?

a) La censure réduit la taille de l'échantillon.

b) Un individu censuré apporte l'information : $T > t$.

c) La censure informative ne pose pas de problème pour KM.

d) La censure doit être indépendante du temps propre.

- Correct, mais incomplet : la condition générale est

$$T \perp C \mid X,$$

c'est-à-dire indépendance conditionnelle vis-à-vis des covariables.

Condition indispensable : censure **non-informative**, sinon les estimateurs KM/NA sont biaisés.

Quiz : risk sets (3/6)

Question 3. Le risk set n_j à l'instant $t_{(j)}$ contient :

a) seulement les individus défaillants à $t_{(j)}$;

b) les individus "en vie" juste avant $t_{(j)}$;

c) les individus censurés **après** $t_{(j)}$;

d) les individus censurés **avant** $t_{(j)}$;

Quiz : risk sets (3/6)

Question 3. Le risk set n_j à l'instant $t_{(j)}$ contient :

a) seulement les individus défaillants à $t_{(j)}$;

- Faux : le risk set contient **tous les individus encore à risque juste avant** $t_{(j)}$, pas seulement ceux qui vont défaillir.

b) les individus "en vie" juste avant $t_{(j)}$;

c) les individus censurés **après** $t_{(j)}$;

d) les individus censurés **avant** $t_{(j)}$;

Quiz : risk sets (3/6)

Question 3. Le risk set n_j à l'instant $t_{(j)}$ contient :

a) seulement les individus défaillants à $t_{(j)}$;

b) les individus "en vie" juste avant $t_{(j)}$;

- Vrai : un individu est dans n_j s'il n'a ni été censuré ni eu d'événement avant $t_{(j)}$.

c) les individus censurés **après** $t_{(j)}$;

d) les individus censurés **avant** $t_{(j)}$;

Quiz : risk sets (3/6)

Question 3. Le risk set n_j à l'instant $t_{(j)}$ contient :

a) seulement les individus défaillants à $t_{(j)}$;

b) les individus "en vie" juste avant $t_{(j)}$;

c) les individus censurés **après** $t_{(j)}$;

- Vrai : un individu censuré plus tard est encore à risque à $t_{(j)}$, donc bien dans le risk set.

d) les individus censurés **avant** $t_{(j)}$;

Quiz : risk sets (3/6)

Question 3. Le risk set n_j à l'instant $t_{(j)}$ contient :

a) seulement les individus défaillants à $t_{(j)}$;

b) les individus "en vie" juste avant $t_{(j)}$;

c) les individus censurés **après** $t_{(j)}$;

d) les individus censurés **avant** $t_{(j)}$;

- Faux : s'ils ont été censurés avant, ils ont quitté le risk set et ne contribuent plus à n_j.

Quiz : Kaplan–Meier (4/6)

Que représente $\hat{S}(t)$ (estimateur de Kaplan–Meier) ?

a) $\Pr(T > t \mid T > 0)$

b) $\Pr(T > t)$

c) La proportion d'individus encore suivis en fin de période.

d) Une probabilité ajustée des covariables.

Quiz : Kaplan–Meier (4/6)

Que représente $\hat{S}(t)$ (estimateur de Kaplan–Meier) ?

a) $\Pr(T > t \mid T > 0)$

- Vrai : c'est l'interprétation correcte.
- $T > 0$ est une condition triviale (aucun individu n'a encore échoué à $t = 0$).

b) $\Pr(T > t)$

c) La proportion d'individus encore suivis en fin de période.

d) Une probabilité ajustée des covariables.

Quiz : Kaplan–Meier (4/6)

Que représente $\hat{S}(t)$ (estimateur de Kaplan–Meier) ?

a) $\Pr(T > t \mid T > 0)$

b) $\Pr(T > t)$

- Presque vrai : équivalent car $\Pr(T > 0) = 1$.
- Mais l'écriture conditionnelle est la forme rigoureuse.

c) La proportion d'individus encore suivis en fin de période.

d) Une probabilité ajustée des covariables.

Quiz : Kaplan–Meier (4/6)

Que représente $\hat{S}(t)$ (estimateur de Kaplan–Meier) ?

a) $\Pr(T > t \mid T > 0)$

b) $\Pr(T > t)$

c) La proportion d'individus encore suivis en fin de période.

- Faux : ce serait un simple ratio $\frac{\text{non-événements}}{\text{effectif}}$.
- Kaplan–Meier **pondère** chaque intervalle par les tailles de risk sets et **corrige des censures** - ce n'est pas un comptage brut.

d) Une probabilité ajustée des covariables.

Quiz : Kaplan–Meier (4/6)

Que représente $\hat{S}(t)$ (estimateur de Kaplan–Meier) ?

a) $\Pr(T > t \mid T > 0)$

b) $\Pr(T > t)$

c) La proportion d'individus encore suivis en fin de période.

d) Une probabilité ajustée des covariables.

- Faux : KM est un estimateur **non-paramétrique sans covariable**.
- Les covariables n'apparaissent que dans les modèles de Cox, AFT, etc.

Quiz : KM vs Nelson–Aalen (5/6)

Quel est le lien entre Nelson–Aalen et Kaplan–Meier ?

a) NA estime la fonction de survie.

b) NA estime le hazard cumulé.

c) $\hat{S}(t) \approx e^{-\hat{H}(t)}$.

d) NA = KM exactement.

Quiz : KM vs Nelson–Aalen (5/6)

Quel est le lien entre Nelson–Aalen et Kaplan–Meier ?

a) NA estime la fonction de survie.

- Faux : l'estimateur de Nelson–Aalen calcule le **hasard cumulé** : $\hat{H}(t) = \sum_{t_j \leq t} d_j / n_j$.
- La survie n'est obtenue qu'indirectement, via une transformation.

b) NA estime le hazard cumulé.

c) $\hat{S}(t) \approx e^{-\hat{H}(t)}$.

d) NA = KM exactement.

Quiz : KM vs Nelson–Aalen (5/6)

Quel est le lien entre Nelson–Aalen et Kaplan–Meier ?

a) NA estime la fonction de survie.

b) NA estime le hazard cumulé.

- Vrai : c'est précisément sa définition.
- La survie correspond ensuite à l'idée : "ne pas avoir subi d'événement jusqu'à t" $\Rightarrow S(t) \approx e^{-\hat{H}(t)}$.

c) $\hat{S}(t) \approx e^{-\hat{H}(t)}$.

d) NA = KM exactement.

Quiz : KM vs Nelson–Aalen (5/6)

Quel est le lien entre Nelson–Aalen et Kaplan–Meier ?

a) NA estime la fonction de survie.

b) NA estime le hazard cumulé.

c) $\hat{S}(t) \approx e^{-\hat{H}(t)}$.

- Vrai : en **temps continu**, la relation $S(t) = e^{-H(t)}$ est exacte.
- En temps discret, $\hat{S}_{\mathrm{KM}}$ et $e^{-\hat{H}}$ diffèrent légèrement car KM applique une **multiplication de risques conditionnels**.

d) NA = KM exactement.

Quiz : KM vs Nelson–Aalen (5/6)

Quel est le lien entre Nelson–Aalen et Kaplan–Meier ?

a) NA estime la fonction de survie.

b) NA estime le hazard cumulé.

c) $\hat{S}(t) \approx e^{-\hat{H}(t)}$.

d) NA = KM exactement.

- Faux : KM utilise les probabilités conditionnelles $(1 - d_j/n_j)$, alors que NA additionne les taux d_j/n_j.
- Les deux estimateurs sont proches aux temps précoces mais peuvent diverger en **queue de distribution**.

Quiz : log-rank (6/6)

Le test du log-rank est le plus puissant lorsque :

a) les risques sont proportionnels ;

b) les courbes s'entrecroisent ;

c) il y a beaucoup de censures ;

d) le hasard dépend fortement du temps.

Quiz : log-rank (6/6)

Le test du log-rank est le plus puissant lorsque :

a) les risques sont proportionnels ;

- Vrai : le log-rank est optimal lorsque

  $h_1(t) = c\, h_2(t), \quad c > 0$ constant (hazards proportionnels, $c = 1$ sous H_0

  Il détecte surtout un **décalage global** entre les deux hazards.

b) les courbes s'entrecroisent ;

c) il y a beaucoup de censures ;

d) le hasard dépend fortement du temps.

Quiz : log-rank (6/6)

Le test du log-rank est le plus puissant lorsque :

a) les risques sont proportionnels ;

- Vrai : le log-rank est optimal lorsque

  $h_1(t) = c\, h_2(t), \quad c > 0$ constant (hazards proportionnels, $c = 1$ sous H_0

  Il détecte surtout un **décalage global** entre les deux hazards.

b) les courbes s'entrecroisent ;

- Faux : les croisements violent la proportionnalité $\rightarrow$ le test perd **fortement** en puissance.

c) il y a beaucoup de censures ;

d) le hasard dépend fortement du temps.

Quiz : log-rank (6/6)

Le test du log-rank est le plus puissant lorsque :

a) les risques sont proportionnels ;

- Vrai : le log-rank est optimal lorsque

  $h_1(t) = c\, h_2(t), \quad c > 0$ constant (hazards proportionnels, $c = 1$ sous H_0

  Il détecte surtout un **décalage global** entre les deux hazards.

b) les courbes s'entrecroisent ;

- Faux : les croisements violent la proportionnalité $\rightarrow$ le test perd **fortement** en puissance.

c) il y a beaucoup de censures ;

- Faux : un volume important de censures réduit l'information et dégrade la puissance de tout test de survie, log-rank inclus.

d) le hasard dépend fortement du temps.

Quiz : log-rank (6/6)

Le test du log-rank est le plus puissant lorsque :

a) les risques sont proportionnels ;

- Vrai : le log-rank est optimal lorsque

  $h_1(t) = c\, h_2(t), \quad c > 0$ constant (hazards proportionnels, $c = 1$ sous H_0

  Il détecte surtout un **décalage global** entre les deux hazards.

b) les courbes s'entrecroisent ;

- Faux : les croisements violent la proportionnalité $\rightarrow$ le test perd **fortement** en puissance.

c) il y a beaucoup de censures ;

- Faux : un volume important de censures réduit l'information et dégrade la puissance de tout test de survie, log-rank inclus.

d) le hasard dépend fortement du temps.

- Faux : le log-rank ne cible pas la forme de $h(t)$, mais les **différences persistantes** entre deux hazards.

Rappel : le log-rank est le test adapté pour comparer des groupes sous

Rappels : objets de base & censure

⇒ **Durée étudiée** : temps jusqu'à un événement

$$T = t_{\text{fin}} - t_0, \qquad \delta = \mathbb{1}\{\text{événement observé}\}.$$

⇒ **Censure à droite** : on sait seulement que $T > t$ (événement non observé pendant la fenêtre).

⇒ **Troncature / entrée retardée** : inclusion conditionnelle (« on n'entre dans l'échantillon que si $T > t_{\text{entrée}}$ »).

⇒ **Hypothèse clé** : censure **non-informative** (conditionnellement aux covariables).

5 objets fondamentaux : $F(t)$, $S(t)$, $f(t)$, $h(t)$, $H(t)$ avec les ponts

$$S(t) = 1 - F(t), \quad H(t) = \int_0^t h(u)\,du, \quad S(t) = e^{-H(t)}.$$

Rappels : Kaplan–Meier & Nelson–Aalen

⇒ $t_{(1)} < \cdots < t_{(J)}$: instants où au moins une défaillance survient.

⇒ d_j : nb de défaillances à $t_{(j)}$; n_j : nb à risque juste avant $t_{(j)}$.

Kaplan–Meier (survie)

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

⇒ Fonction en **marches** : chute seulement aux **événements**.

⇒ Censures : pas de chute mais réduisent les risk sets.

Nelson–Aalen (hasard cumulé)

$$\hat{H}(t) = \sum_{t_{(j)} \le t} \frac{d_j}{n_j}, \qquad \hat{S}(t) \approx e^{-\hat{H}(t)}.$$

KM : « combien restent ? » / NA : « à quelle vitesse sortent-ils ? »

Rappels : variance (Greenwood) & Breslow

Kaplan–Meier : produit de facteurs $1 - \dfrac{d_j}{n_j}$ $\Rightarrow$ variance plus complexe.

Variance de Greenwood

$$\widehat{\mathrm{Var}}\big(\hat{S}(t)\big) = \hat{S}^2(t) \sum_{t_{(j)} \le t} \frac{d_j}{n_j\,(n_j - d_j)}$$

$\Rightarrow$ Chaque défaillance ajoute de l'incertitude (terme en d_j).

$\Rightarrow$ Quand n_j diminue $\Rightarrow$ la variance augmente fortement en fin de suivi.

Estimateur de Breslow pour le risque cumulé

$$\hat{\Lambda}(t) = \sum_{t_{(j)} \le t} \frac{d_j}{n_j}, \qquad \widehat{\mathrm{Var}}\big(\hat{\Lambda}(t)\big) = \sum_{t_{(j)} \le t} \frac{d_j}{n_j^2}.$$

Travailler sur $\hat{\Lambda}(t)$ (Breslow) est **plus simple analytiquement**, mais $\hat{S}(t)$ est plus intuitif à commenter.

Rappels : comparaison de courbes (log-rank)

Objectif : tester l'égalité de deux fonctions de survie

$$H_0 : S_1(t) = S_2(t) \quad \forall t.$$

Principe du log-rank

$\Rightarrow$ À chaque temps de défaillance t_j :

- on observe d_{1j}, d_{2j} dans chaque groupe,
- on compare à ce qu'on attendrait si les risques étaient identiques.

$\Rightarrow$ On cumule ces écarts dans une statistique χ^2 globale (approx. loi $\chi^2(1)$).

À retenir : test **global**, le plus puissant sous risques proportionnels, à **compléter par la visualisation** des courbes de Kaplan–Meier.

Rappels

Motivation

Lois de base

Extensions

Panorama & extensions

Rappel : ce que nous savons déjà

Objectifs de la séance précédente : KM, NA, censure, risk set, log-rank, hasard instantané.

5 objets fondamentaux :

$$F(t), \quad S(t) = \Pr(T > t), \quad f(t), \quad h(t) = \frac{f(t)}{S(t)}, \quad H(t) = \int_0^t h(u)\, du$$

$$S(t) = e^{-H(t)} \qquad \text{(relation clé des modèles paramétriques)}$$

⇒ Méthodes **non-paramétriques** : ne supposent rien sur la forme de la loi (KM/NA).

⇒ Limites : pas d'extrapolation, pas d'interprétation paramétrique, efficacité moindre.

Motivation : besoin d'un modèle structurel

Motivation : besoin d'un modèle structurel

En pratique, les décideurs veulent :

- ⇒ des **prédictions** : durée attendue au chômage, temps médian avant défaut ;
- ⇒ des **effets de politiques** : impact d'un traitement/formation/contrat ;
- ⇒ des **comparaisons robustes** entre groupes ;
- ⇒ une **interprétation économique** (vieillissement, sélection, décrochage, épuisement des opportunités...).

Modèles paramétriques = formaliser une hypothèse sur le cycle du risque :

$h(t)$ croît ? décroît ? reste constant ? change de signe ?

Approche : on encode des mécanismes plausibles dans la forme de $h(t)$, puis on l'estime statistiquement.

Conséquence : un modèle paramétrique est une *théorie* sur la dynamique des durées.

Pourquoi des modèles paramétriques ?

Idée clé

Imposer une loi sur les durées permet d'obtenir des estimateurs plus efficaces et interprétables.

Avantages :

⇒ Estimation directe de la **moyenne**, de la **médiane**, des **quantiles**.

⇒ Possibilité de **prédire** au-delà de la fenêtre d'observation (*extrapolation*).

⇒ Effets des covariables via modèles **PH** ou **AFT**.

⇒ Lien clair entre forme du hasard et structure du modèle.

Inconvénients :

⇒ Risque de **mauvaise spécification** (shape du hasard incorrecte).

⇒ Diagnostics cruciaux : comparaison à KM, log(-log), résidus.

Limites pratiques des méthodes non-paramétriques

1. Impossibilité de synthétiser la durée par un paramètre simple

⇒ KM donne une **courbe**, pas une moyenne ni un modèle.

⇒ Impossible d'écrire des quantités comme sans lois :

$$\mathbb{E}[T], \quad \mathrm{Var}(T), \quad m_{0.5}, \quad \Pr(T > t^\star)$$

2. Comparaison de groupes limitée

⇒ Log-rank = test global mais **pas d'effet marginal interprétable**.

⇒ Pas de mesure du traitement : pas de **hazard ratio**, pas de **time ratio**.

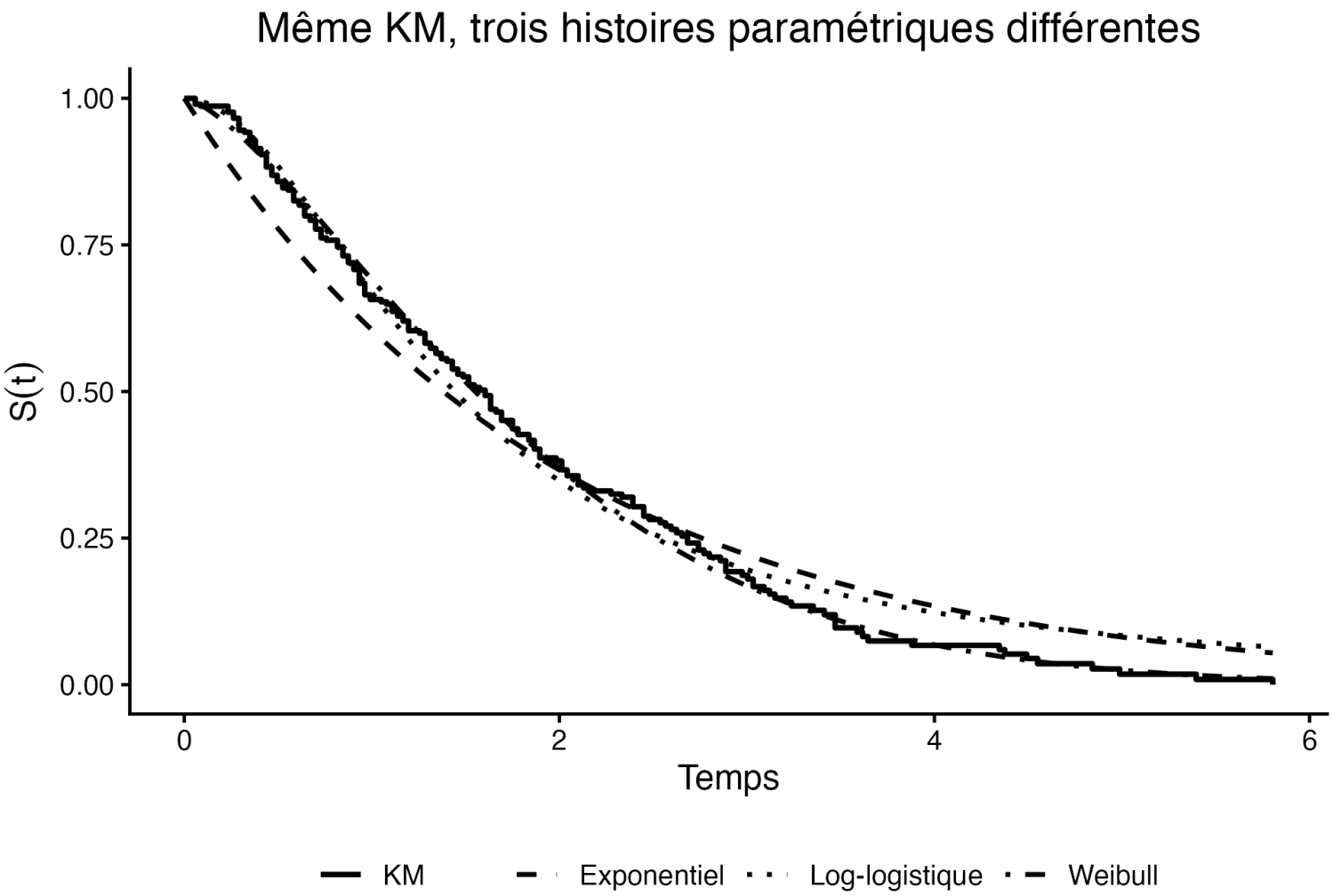
3. Pas d'extrapolation

⇒ KM s'arrête au dernier événement observé.

⇒ impossible d'évaluer : $S(60 \text{ mois})$ si la fenêtre s'arrête à 36 mois.

Conclusion : KM/NA décrivent bien les données, mais n'expliquent ni ne prédisent.

Trois histoires différentes pour une même KM



Une même KM, trois histoires différentes

⇒ La **même** courbe de Kaplan–Meier peut être très bien ajustée par :

- une **exponentielle** ⇒ hasard **constant** (arrivées purement aléatoires, pas de durée-dépendance) ;
- une **Weibull croissante** ⇒ hasard **croissant** (vieillissement, usure, découragement progressif) ;
- une **log-logistique** ⇒ hasard **en cloche** (phase d'entrée, puis accélération des sorties, puis saturation).

⇒ **Même qualité d'ajustement sur** $S(t)$, mais :

- formes de $h(t)$ **radicalement différentes** ;
- donc **histoires économiques différentes** derrière les données.

Moralité : une bonne KM ne suffit pas. Il faut regarder la forme de $h(t)$ et choisir un modèle cohérent avec le mécanisme.

Le rôle central du *hasard* dans la modélisation

Dans un modèle paramétrique, tout part de $h(t)$:

$$h(t) \xrightarrow{\text{intégration}} H(t) \xrightarrow{S(t)=e^{-H(t)}} S(t) \xrightarrow{f(t)=h(t)S(t)} f(t)$$

Pourquoi choisir $h(t)$?

⇒ C'est le **rythme instantané du phénomène**.

⇒ Il révèle la mécanique du système :

- **Croissant** : usure, apprentissage, accumulation de risques.
- **Décroissant** : sélection des robustes.
- **En cloche** : phases d'entrée puis de sortie accélérée.

Modéliser un phénomène de durée = modéliser son risque.

Pourquoi plusieurs lois ?

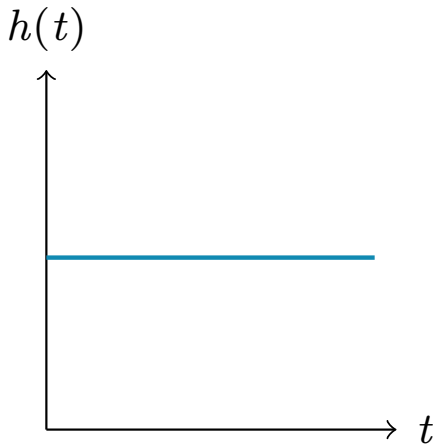
Chaque loi impose une géométrie spécifique au risque :

Loi	**Forme de $h(t)$**	**Interprétation**
Exponentielle	constante	Poisson/arrivées indépendantes
Weibull ($k > 1$)	croissant	vieillissement / usure
Weibull ($k < 1$)	décroissant	sélection des survivants
Log-logistique	en cloche	diffusion / contagion / cycles
Log-normale	en cloche	forte hétérogénéité
Gompertz	expo. croissant	mortalité / risques accumulés

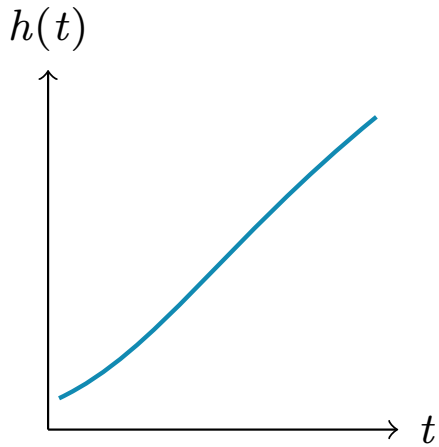
Choisir un modèle paramétrique = choisir un mécanisme économique ou biologique.

Forme du hasard

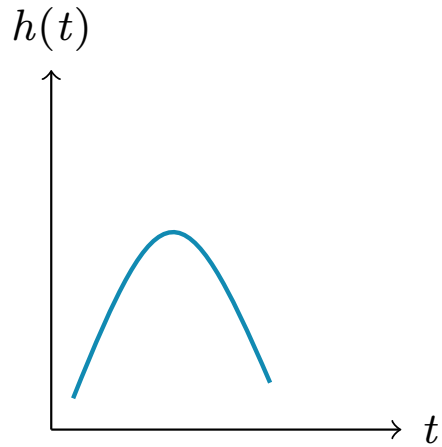
Expo



Weibull $k > 1$



Log-logistique



Expo : risque constant. **Weibull** : risque monotone (croissant ou décroissant). **Log-logistique** : risque en cloche (entrée, pic, saturation).

Rappels

Motivation

Lois de base

Extensions

Panorama & extensions

Exponentielle : hasard constant (1/2)

Hypothèse structurelle (définition) :

$$h(t) = \lambda \qquad (\lambda > 0)$$

⇒ Le risque instantané est **constant dans le temps**.

⇒ Aucune durée-dépendance : ni usure, ni apprentissage, ni sélection.

Objets associés :

$$H(t) = \lambda t, \qquad S(t) = e^{-\lambda t}, \qquad f(t) = \lambda e^{-\lambda t}.$$

Propriété fondamentale : absence de mémoire

$$\Pr(T > t + s \mid T > s) = \Pr(T > t)$$

Le fait d'avoir "déjà attendu" ne change rien : le processus "redémarre" à chaque instant.

Conséquence intuitive : le passé n'influence jamais le risque futur (défaillance, sortie, embauche…)

Exponentielle : usages et limites (2/2)

Quantités fermées :

$$\mathbb{E}[T] = \frac{1}{\lambda}, \qquad m_{0.5} = \frac{\ln 2}{\lambda}$$

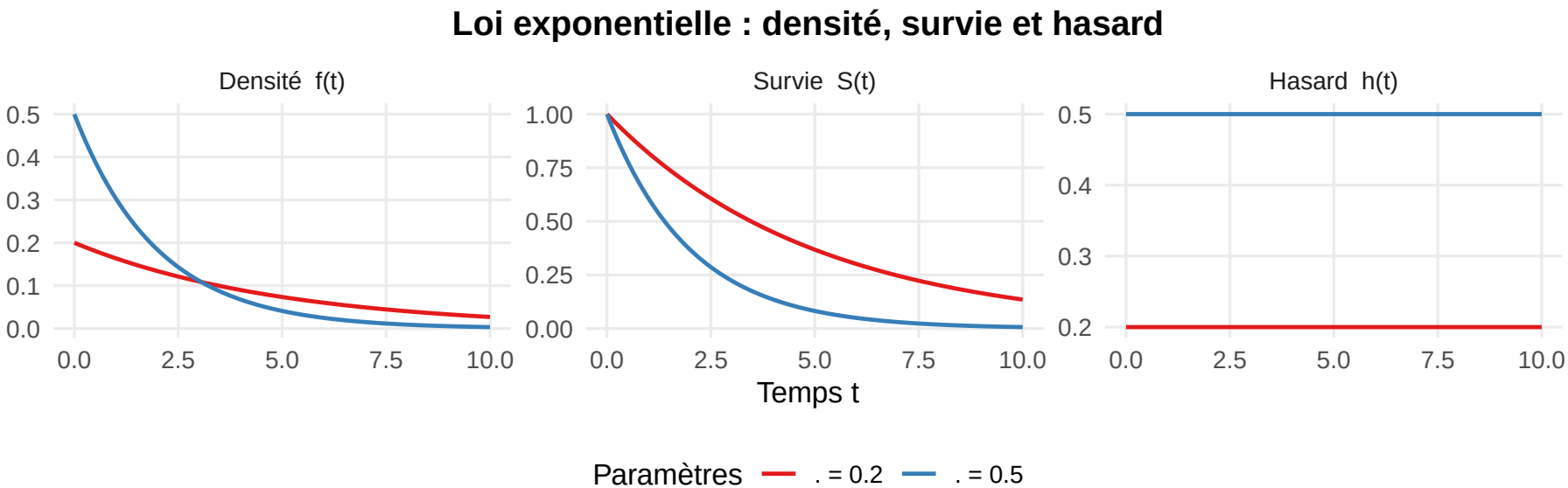
Applications typiques :

- ⇒ files d'attente (M/M/1),
- ⇒ pannes simples / fiabilité,
- ⇒ arrivées indépendantes (Poisson).

Limites :

- ⇒ très peu réaliste pour les comportements humains ;
- ⇒ ne capture ni apprentissage, ni usure, ni sélection ;
- ⇒ risque constant = hypothèse très forte.

Exponentielle : densité, survie, hasard



Weibull : hasard monotone (1/4)

Hasard :

$$h(t) = \lambda k\, t^{k-1}$$

Objets associés :

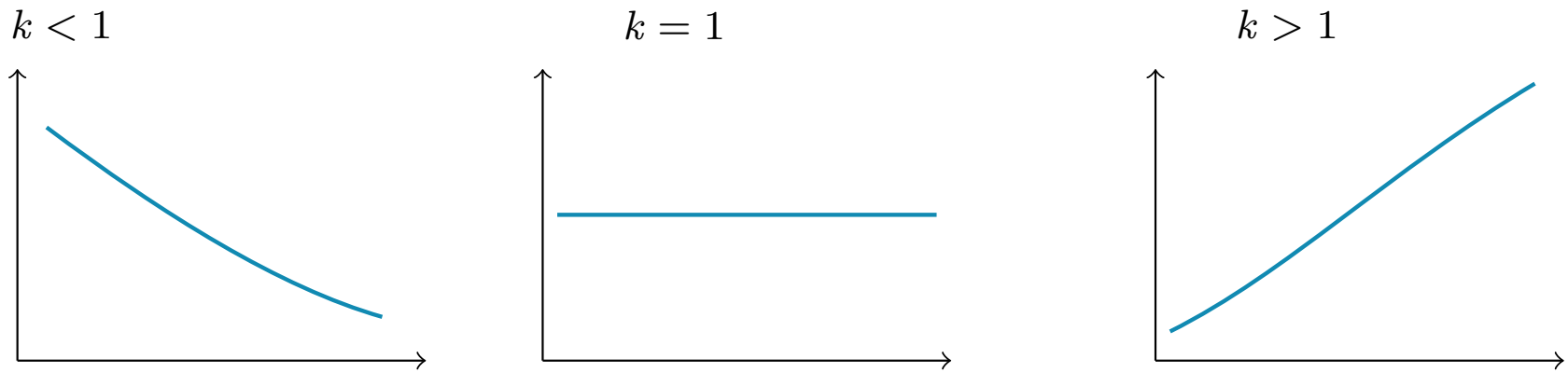
$$H(t) = \lambda t^k, \qquad S(t) = e^{-\lambda t^k}.$$

Interprétation du paramètre de forme k

- $\Rightarrow$ $k < 1$: hasard **décroissant** $\rightarrow$ sélection des robustes, apprentissage.
- $\Rightarrow$ $k = 1$: hasard **constant** $\rightarrow$ cas particulier **exponentiel**.
- $\Rightarrow$ $k > 1$: hasard **croissant** $\rightarrow$ usure, vieillissement, découragement.

Atout majeur : première loi simple offrant une **durée-dépendance flexible** (décroissante, constante, ou croissante).

Weibull : intuition sur la forme du hasard (2/4)



Weibull = la loi qui permet de passer d'un risque décroissant → constant → croissant.

Weibull : comment sont estimés λ et k ? (3/4)

Objectif : trouver les valeurs de (λ, k)où le modèle est *compatible* avec les observations.

Principe : maximum de vraisemblance

⇒ chaque individu contribue par : – la **densité** s'il a eu l'événement ;
– la **survie** s'il est censuré ;

⇒ on choisit (λ, k) qui maximisent la vraisemblance globale.

Aspects pratiques :

⇒ pas de formule explicite : estimation par **optimisation numérique** ;

Interprétation :

⇒ $\hat{k}$ décrit la **forme du risque** (croissant, constant, décroissant) ;

⇒ $\hat{\lambda}$ fixe **l'échelle temporelle**.

Weibull : usages et propriétés (4/4)

Applications :

⇒ durée de chômage (découragement/usure) ;

⇒ fiabilité matériel ;

⇒ survie médicale ;

Atouts :

⇒ très souple, estimation simple sous censure ;

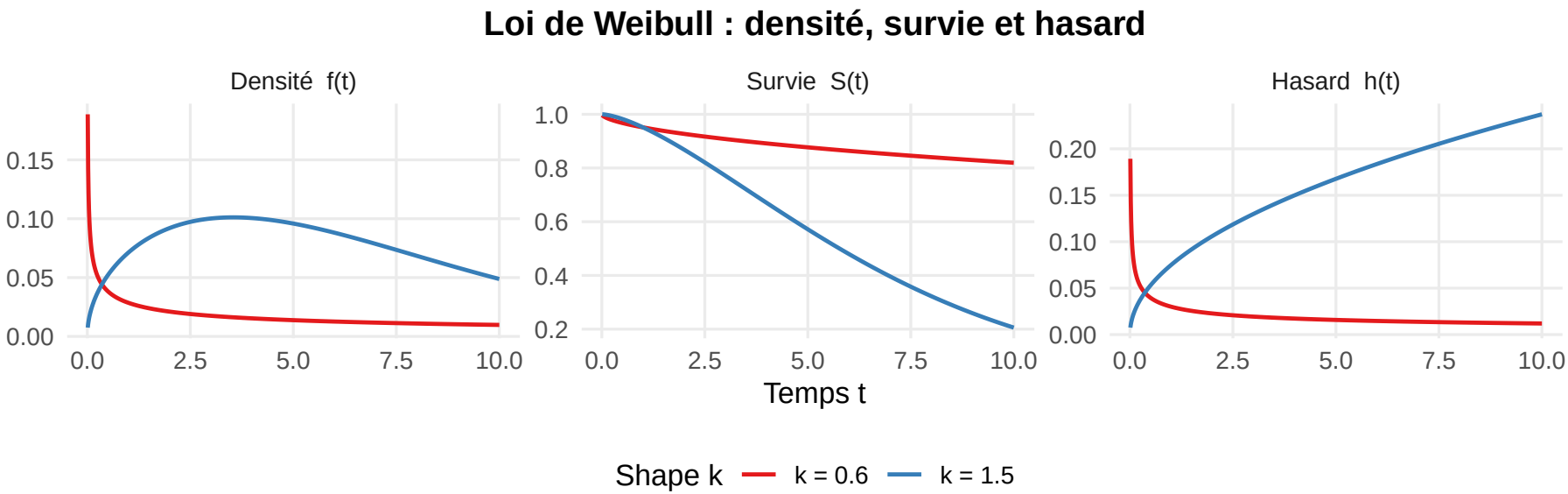
⇒ compatible **PH** ou **AFT**.

Limites :

⇒ ne capture pas les hasards en cloche ;

⇒ pas adapté aux phénomènes en cycles.

Weibull : densité, survie, hasard



Log-logistique : hasard en cloche (1/2)

Formes associées (idée générale) :

$$S(t) = \frac{1}{1 + (t/\lambda)^k}, \qquad h(t) = \frac{k}{\lambda} \frac{(t/\lambda)^{k-1}}{1 + (t/\lambda)^k}.$$

Propriété fondamentale : hasard non monotone

$$\text{augmente} \longrightarrow \text{atteint un pic} \longrightarrow \text{diminue.}$$

Le risque est faible au début, croît fortement, puis ralentit.

Interprétations typiques :

- ⇒ phénomènes avec phases successives : entrée → diffusion → saturation ;
- ⇒ processus sociaux avec “effet de mode” ou contagion ;
- ⇒ hétérogénéité forte dans la population.

Log-logistique : propriétés et usages (2/2)

Caractéristiques simples :

⇒ Médiane : $m_{0.5} = \lambda$ (interprétation directe) ;

⇒ Espérance finie uniquement si $k > 1$.

Signature du modèle :

⇒ capture très bien les **hasards en cloche** ;

⇒ structure **AFT naturelle** (pas de modèle PH possible) ;

⇒ queues lourdes : attention aux moments (moyenne parfois infinie).

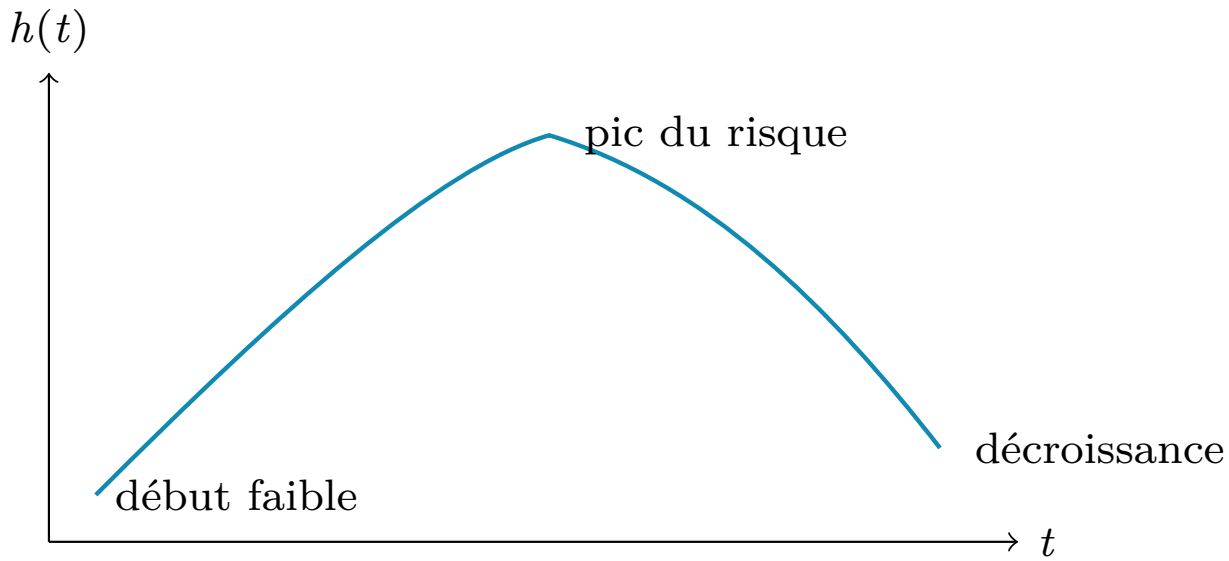
Applications classiques :

⇒ diffusion d'innovations ou d'idées ;

⇒ processus sociaux avec phases d'adoption puis stabilisation ;

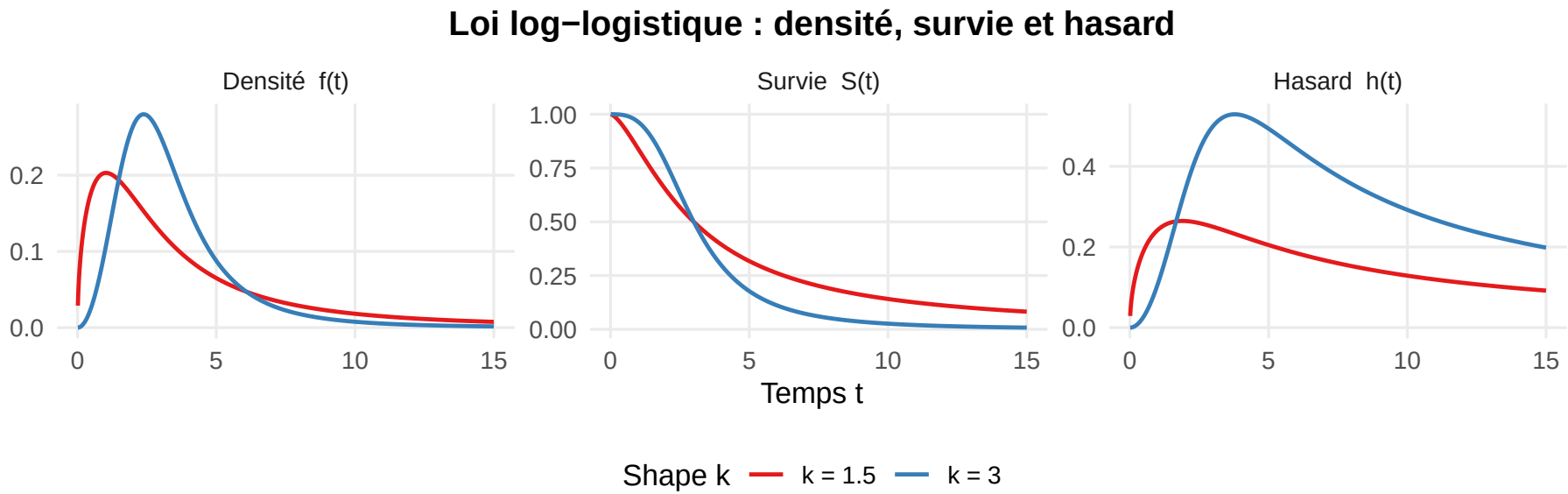
⇒ dynamiques où l'hétérogénéité joue un rôle majeur.

Log-logistique : forme du hasard



Lecture : au début, peu de sorties (risque faible) → le risque augmente jusqu'à un **pic** → puis diminue (saturation, épuisement du potentiel).

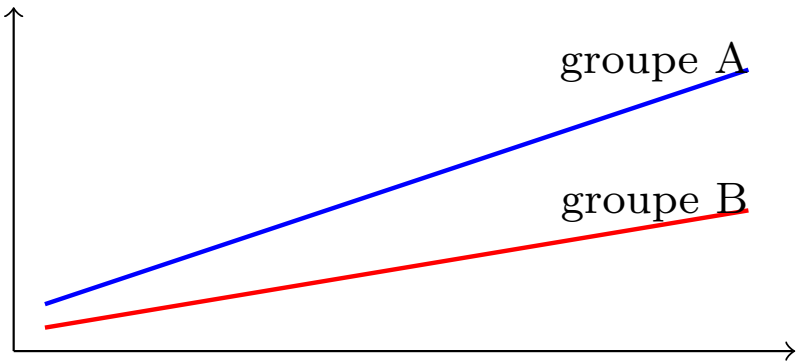
log-logistic : densité, survie, hasard



Risques proportionnels : intuition

Idée intuitive : Dans certains contextes, deux groupes peuvent avoir des profils de risque qui gardent **la même forme au cours du temps**. Seul le "niveau" du risque change.

Exemple visuel : deux courbes en parallèle



Même forme → une simple différence de niveau.

Signification : - le groupe A a toujours un risque plus élevé que B ; - **mais la différence reste stable dans le temps** ; - les deux groupes évoluent de manière "parallèle".

Cette situation apparaît dans certains modèles, mais pas tous.

Log-logistique : pourquoi les risques ne restent jamais parallèles ?

Rappel : la loi log-logistique produit un **hasard en cloche** :

faible au début → augmente → puis diminue.

Conséquence immédiate : Deux groupes log-logistiques ne peuvent pas avoir des courbes de risque gardant la **même forme** dans le temps.
Même si les groupes diffèrent seulement par un changement d'échelle (λ_1 vs λ_2) :

- ⇒ leurs courbes montent,
- ⇒ atteignent leur pic à des moments différents,
- ⇒ puis redescendent à des vitesses différentes.

Conclusion intuitive : Avec une loi log-logistique, les **formes de risque changent dans le temps**. Deux groupes ne peuvent donc **jamais évoluer en parallèle**. Le rapport de leurs risques **ne reste pas constant**.

Applications typiques des lois paramétriques

Quel modèle pour quel phénomène ?

Chaque loi encode une **géométrie du hasard** → **mécanismes** différents

⇒ **Exponentielle (hasard constant)**

- pannes électroniques sans usure (“memoryless failures”) ;
- phénomènes sans durée-dépendance : *risque constant dans le temps.*

⇒ **Weibull (hasard croissant ou décroissant)**

- $k > 1$ (croissant) : mortalité adulte, fatigue de matériaux, usure, découragement en chômage ;
- $k < 1$ (décroissant) : sélection des robustes, guérison progressive, apprentissage ;

⇒ **Log-logistique (hasard en cloche)**

- diffusion d’innovations (phases d’adoption puis saturation) ;
- délais de décision avec forte hétérogénéité.

Exemples : quelle loi pour quel phénomène ?

Exemple 1 : durée de chômage des jeunes diplômés

⇒ Forte sortie au début (les plus employables), puis ralentissement.

⇒ **Hypothèse plausible :** Weibull avec $k < 1$ (sélection des plus robustes)

Exemple 2 : adoption d'une nouvelle technologie

⇒ Début lent, puis accélération, puis saturation.

⇒ **Hypothèse plausible :** log-logistique (hasard en cloche).

Exemple 3 : panne d'un composant électronique simple

⇒ Pannes indépendantes du temps d'utilisation.

⇒ **Hypothèse plausible :** exponentielle (hasard constant).

Réflexe à développer : partir du *mécanisme* imaginé dans la vraie vie, puis choisir la loi dont la forme de $h(t)$ lui ressemble.

Synthèse : trois lois paramétriques de base

Réflexe : choisir la loi dont la forme de $h(t)$ colle le mieux au mécanisme économique.

Loi	**Survie** $S(t)$	**Hasard** $h(t)$	**PH / AFT ?**	**Interprétation / usages**
Exponentielle	$S(t) = e^{-\lambda t}$	$h(t) = \lambda$ (constant)	PH & AFT possibles	phénomènes sans durée-dépendance.
Weibull	$S(t) = e^{-\lambda t^k}$	$h(t) = \lambda k t^{k-1}$, $k < 1$: décroissant, $k = 1$: constant, $k > 1$: croissant	PH *ou* AFT	usure ou sélection des robustes.
Log-logistique	$S(t) = \dfrac{1}{1 + (t/\lambda)^k}$	$h(t) = \dfrac{k}{\lambda} \dfrac{(t/\lambda)^{k-1}}{1 + (t/\lambda)^k}$ (hasard en cloche)	AFT seulement	phases d'adoption puis saturation.

Motivation : au-delà d'une loi commune pour tous

Jusqu'ici : une seule loi pour toutes les durées. **En réalité :** la durée dépend d'individus, de contextes, de choix.

Exemples de facteurs :

- ⇒ âge, genre, éducation ;
- ⇒ traitement / contrôle ;
- ⇒ santé, expérience, capital humain ;
- ⇒ exposition au risque, secteur, localisation.

Question centrale : *comment une covariable déforme-t-elle la distribution des durées ?*

Deux visions dominantes :

- ⇒ **PH** : changer l'**intensité** du risque ;
- ⇒ **AFT** : changer l'**vitesse** du temps.

→s concurrentes : PH vs AFT

Vision 1 : Proportional Hazards (PH)

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

⇒ les covariables **modifient le niveau du risque** ;

⇒ la **forme temporelle reste identique** pour tous les groupes.

Vision 2 : Accelerated Failure Time (AFT)

$$T = T_0 \cdot \exp(-X\gamma)$$

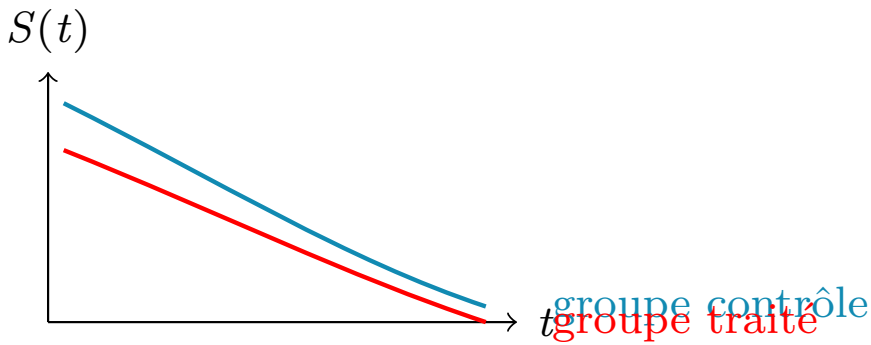
⇒ les covariables **accélèrent ou ralentissent le temps** ;

⇒ l'événement survient **plus tôt ou plus tard**.

Métaphores : PH = déplacer la courbe **verticalement**. AFT = étirer/-compresser la courbe **horizontalement**.

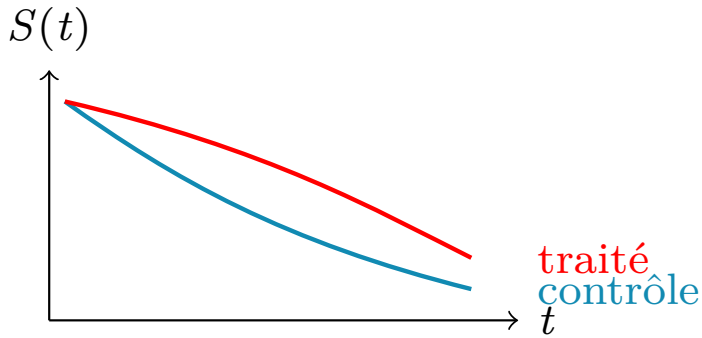
PH vs AFT : comparaison visuelle simplifiée

À gauche : vision PH (risques proportionnels)



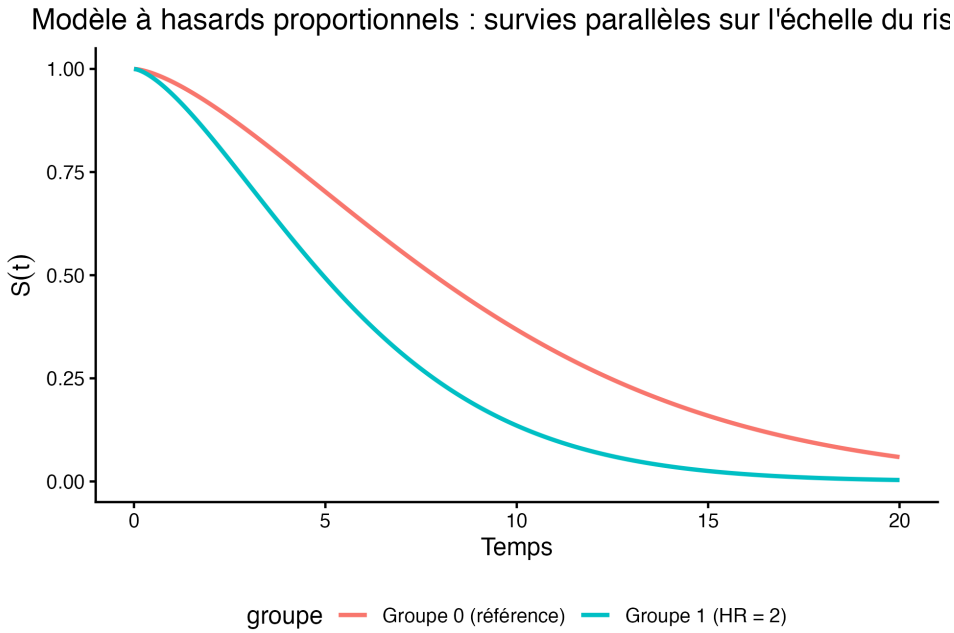
même forme, seulement un décalage vertical

À droite : vision AFT (temps accéléré)



même forme, mais ”temps étiré”

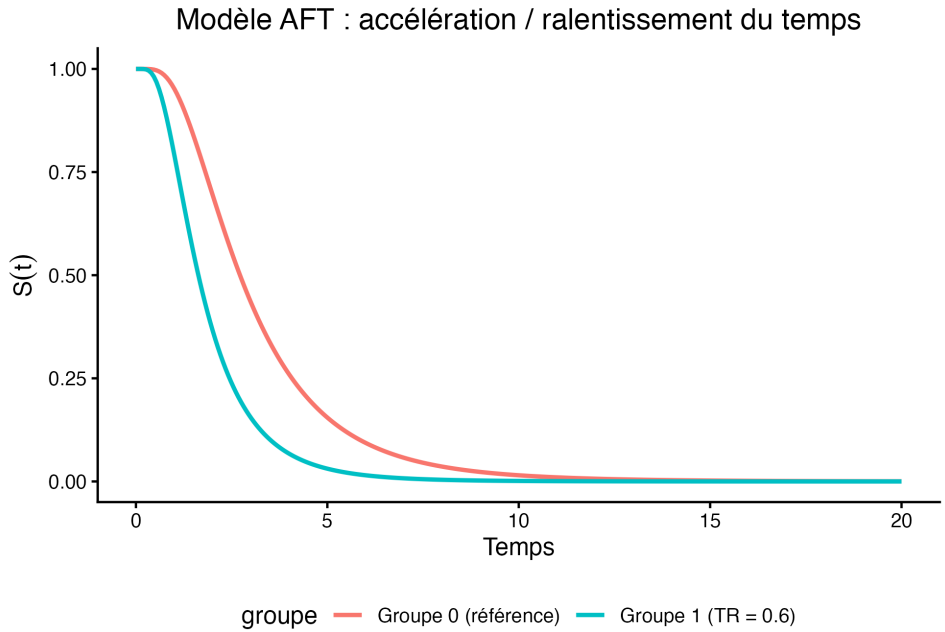
Illustration : effet d'une variable en modèle PH



Signature des PH :

- ⇒ Les courbes de risque sont **proportionnelles** : ratio constant.
- ⇒ Les courbes de survie sur échelle log(-log) sont **parallèles**.
- ⇒ Interprétation directe : **hazard ratio** $= e^{\beta}$.

Illustration : effet d'une variable en modèle AFT



Signature des AFT :

⇒ Les courbes de survie sont "étirées" ou "compressées".

⇒ Pas de proportionalité des hasards.

⇒ Interprétation directe : **time ratio** $= e^{-\gamma}$.

Modèle AFT : comprendre le *time ratio*

Équation clé du modèle AFT :

$$T = T_0 \exp(-X\gamma)$$

$$\text{time ratio} = \exp(-\gamma)$$

⇒ > 1 : **durée étirée** → événement plus tard ;

⇒ < 1 : **durée compressée** → événement plus tôt.

Intuition :

⇒ PH = agit sur le **risque** (décalage vertical) ;

⇒ AFT = agit sur le **temps** (étirement horizontal).

Lecture : Le time ratio indique “de combien” la vidéo du temps est accélérée ou ralentie.

Une même loi peut être PH, AFT... ou aucun des deux

Idée générale : Selon la forme du hasard, une loi paramétrique peut s'écrire :

- ⇒ comme un **modèle PH** (effet vertical sur le risque),
- ⇒ ou comme un **modèle AFT** (effet horizontal sur le temps),
- ⇒ ou comme **ni l'un ni l'autre**.

Exponentielle : PH & AFT Hasard constant → les deux représentations coïncident.

Weibull : PH *et* AFT Hasard monotone (constant).

Log-normal : AFT seulement Survie en "S" asymétrique → impossible d'obtenir deux courbes proportionnelles.

Log-logistique : AFT seulement Hasard en cloche → les courbes ne peuvent jamais évoluer "en parallèle".

Message clé : la **forme du hasard** détermine si PH, AFT, ou les deux sont possibles.

PH ou AFT ? Résumé visuel

Loi	**Forme du hasard**	**PH / AFT ?**
Exponentielle	constant	PH & AFT
Weibull	monotone	PH & AFT
Log-normal	non monotone	AFT seulement
Log-logistique	en cloche	AFT seulement

PH = courbes "en parallèle". AFT = même forme, mais étirées/compressées dans le temps.

Quand utiliser PH ? Quand utiliser AFT ?

Choisir PH si :

⇒ l'effet des covariables déplace **verticalement le hasard** ;

⇒ la courbe $\log(-\log S(t))$ est **parallèle entre groupes** ;

⇒ interprétation en **hazard ratio** = centrale.

Choisir AFT si :

⇒ les groupes semblent "aller plus vite / lentement" ;

⇒ les courbes de survie semblent "étirées" ;

⇒ interprétation en **time ratio** plus intuitive.

En pratique : Tester PH → si non respecté → préférer AFT (log-normal, log-logistique, Weibull).

Quiz : PH ou AFT ?

Situation 1. On estime un modèle et on trouve :

$$\hat{\beta} = 0{,}5, \quad e^{\hat{\beta}} \approx 1{,}65.$$

On dit : « le traitement multiplie le **risque instantané** par 1,65 à tout instant ».

Situation 2. On estime un modèle et on trouve :

$$\hat{\gamma} = -0{,}3, \quad e^{-\hat{\gamma}} \approx 1{,}35.$$

On dit : « la durée médiane avant l'événement est **35 % plus longue** pour les traités ».

Question 1. Associez chaque situation au bon cadre :

1. PH (hazards proportionnels)
2. AFT (Accelerated Failure Time)

Question 2. Dans quel cadre (PH ou AFT) l'interprétation suivante est-elle naturelle ? « Le traitement ralentit le temps jusqu'à l'événement ».

Correction du quiz : PH ou AFT ?

Situation 1.

$$e^{\hat{\beta}} \approx 1{,}65 \Rightarrow \text{hazard ratio}$$

⇒ **Cadre : PH**.

⇒ Interprétation : à tout instant t, le traitement augmente le **risque instantané** d'un facteur 1,65.

Situation 2.

$$e^{-\hat{\gamma}} \approx 1{,}35 \Rightarrow \text{time ratio}$$

⇒ **Cadre : AFT**.

⇒ Durée médiane est 35 % plus longue → le temps est **ralenti**.

Question 2 : « Le traitement ralentit le temps jusqu'à l'événement » ⇒ **AFT**.

À retenir : PH ⇒ *hazard ratio* ; AFT ⇒ *time ratio* (accélération / ralentissement du temps).

Feuille de route : modèles paramétriques avec covariables

Maintenant que nous avons :

- ⇒ les lois de base (exponentielle, Weibull, log-logistique),
- ⇒ la philosophie des modèles PH et AFT,
- ⇒ les signatures visuelles pour diagnostiquer l'un ou l'autre,

Nous allons estimer :

1. le modèle exponentiel PH / AFT,
2. le modèle Weibull PH,
3. le modèle Weibull AFT,
4. le modèle log-logistique AFT,
5. comparaison des ajustements.

Objectif final : savoir choisir, estimer et interpréter un modèle de durée paramétrique.

Rappels

Motivation

Lois de base

Extensions

Panorama & extensions

Vraisemblance sous censure : comprendre ce que l'on observe (1/2)

Pour chaque individu i, on observe :

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

⇒ **Cas 1 : événement observé** ($\delta_i = 1$) On connaît exactement la durée : $T_i = t_i$.

⇒ **Cas 2 : censure** ($\delta_i = 0$) On sait seulement que l'événement **n'est pas encore arrivé** :

$$T_i > t_i$$

Idée simple : Pour un censuré, on n'a pas la durée exacte... mais on a une **information partielle précieuse** : "il a survécu jusqu'à t_i".

Comment construire la vraisemblance avec censure ?

Étape 1 : Identifier ce qu'on observe pour chaque individu

⇒ Si l'événement est observé ($\delta_i = 1$) : on connaît la **durée exacte** : $T_i = t_i$.

⇒ S'il est censuré ($\delta_i = 0$) : seulement une **borne inférieure** : $T_i > t_i$.

Étape 2 : Traduire cette information en probabilité

⇒ **Événement** : probabilité d'observer exactement $t_i \Rightarrow f(t_i)$ (densité)

⇒ **Censure** : probabilité d'être encore en vie à $t_i \Rightarrow S(t_i)$ (survie)

Étape 3 : Une formule unique pour les deux cas

$$L_i(\theta) = [f(t_i \mid \theta)]^{\delta_i} [S(t_i \mid \theta)]^{1-\delta_i}$$

⇒ si $\delta_i = 1$: $f^1 S^0 = f(t_i)$

⇒ si $\delta_i = 0$: $f^0 S^1 = S(t_i)$

Idée clé : La censure ne retire pas d'individus : elle change seulement **la façon dont ils contribuent** à la vraisemblance.

Vraisemblance sous censure : contributions individuelles (2/2)

Chaque individu contribue selon ce que l'on observe :

- ⇒ **Événement observé** ($\delta_i = 1$) → contribution par la **densité** : $f(t_i)$.
- ⇒ **Censuré** ($\delta_i = 0$) → contribution par la **survie** : $S(t_i)$.

On regroupe les deux cas :

$$L_i(\theta) = \left[f(t_i \mid \theta)\right]^{\delta_i} \left[S(t_i \mid \theta)\right]^{1-\delta_i}$$

Et la log-vraisemblance totale :

$$\ell(\theta) = \sum_{i=1}^{n} \left[\delta_i \log f(t_i \mid \theta) + (1 - \delta_i) \log S(t_i \mid \theta)\right].$$

Message clé : La censure **modifie la contribution** dans la vraisemblance, mais **aucune observation n'est supprimée**.

Exemple simple : contribution à la vraisemblance

individu jusqu'à $t_i = 5$. Que peut-on dire sur sa durée réelle T_i ?

Cas 1 : événement observé ($\delta_i = 1$)

⇒ On connaît la **valeur exacte** : $T_i = 5$.

⇒ La contribution est donc la **densité** au point 5 :

$$L_i = f(5 \mid \theta)$$

Cas 2 : censure ($\delta_i = 0$)

⇒ On sait seulement que l'événement n'est pas encore arrivé :

$$T_i > 5$$

⇒ Contribution = **probabilité de survivre au-delà de 5** :

$$L_i = S(5 \mid \theta)$$

Événement = durée exacte → *densité.* **Censure** = durée minimale → *survie.*

Exponentielle : intuition de la log-vraisemblance (1/2)

Rappel : modèle exponentiel Hasard constant $\lambda \Rightarrow$ durée moyenne $= 1/\lambda$.

Contribution d'un individu

- $\Rightarrow$ **Événement** $(\delta_i = 1) : f(t_i) = \lambda e^{-\lambda t_i}$ (on connaît la durée exacte).
- $\Rightarrow$ **Censure** $(\delta_i = 0) : S(t_i) = e^{-\lambda t_i}$ (on sait seulement $T_i > t_i$).

Idée clé :

- $\Rightarrow$ tous les individus apportent un facteur $e^{-\lambda t_i}$;
- $\Rightarrow$ seuls les événements apportent un facteur supplémentaire λ.

Ce qui compte au final : nombre d'événements D et temps total observé $T = \sum t_i$.

Exponentielle : log-vraisemblance et estimateur (2/2)

Log-vraisemblance individuelle

$$\ell_i = \delta_i(\log\lambda - \lambda t_i) + (1-\delta_i)(-\lambda t_i).$$

Lecture : $-\lambda t_i$ apparaît pour tout le monde ; $\log\lambda$ seulement pour les événements.

En sommant :

$$\ell(\lambda) = D\log\lambda - \lambda T \qquad \text{avec } D = \sum_i \delta_i, \quad T = \sum_i t_i.$$

MLE :

$$\hat{\lambda} = \frac{D}{T}.$$

$$\hat{\lambda} = \frac{\text{événements}}{\text{temps exposé}} = \text{un } \textbf{taux d'incidence}.$$

Weibull : comprendre la vraisemblance (1/2)

Rappel : Weibull (λ, k)

$$f(t) = \lambda k t^{k-1} e^{-\lambda t^k}, \qquad S(t) = e^{-\lambda t^k}.$$

Rôle des paramètres :

- $\Rightarrow$ λ : **niveau** du risque (scale) ;
- $\Rightarrow$ k : **forme** du risque (croissant, constant, décroissant).

Contributions à la vraisemblance :

- $\Rightarrow$ **Événement** : probabilité d'observer $t_i \Rightarrow f(t_i)$;
- $\Rightarrow$ **Censure** : probabilité que $T_i > t_i \Rightarrow S(t_i)$.

Weibull = exponentielle "généralisée" : λ contrôle le niveau, k contrôle la **forme du hasard**.

Weibull : log-vraisemblance (structure) (2/2)

Log-vraisemblance individuelle

$$\ell_i = \delta_i \log f(t_i) + (1 - \delta_i) \log S(t_i).$$

En remplaçant f et S du Weibull :

$$\ell(\lambda, k) = \sum_i \Big[\delta_i \big(\log(\lambda k) + (k-1) \log t_i \big) - \lambda t_i^k \Big].$$

Interprétation des termes :

- $\Rightarrow$ $\log(\lambda k)$: contribution des événements au niveau du risque ;
- $\Rightarrow$ $(k-1) \log t_i$: effet de la **forme du hasard** (croissant/décroissant) ;
- $\Rightarrow$ $-\lambda t_i^k$: **terme de survie**, présent pour tous (événements + censures).

Estimation :

- $\Rightarrow$ Pas de solution analytique pour $(\hat{\lambda}, \hat{k}) \rightarrow$ Maximisation de $\ell(\lambda, k)$

densité pour les événements, survie pour les censures.

Sélection de modèle : expo vs Weibull (1/2)

Lien fondamental : L'exponentielle est un **cas particulier** du Weibull :

$$\text{Weibull}(\lambda, k) \qquad \text{et} \qquad \text{Expo}(\lambda) = \text{Weibull}(\lambda, 1).$$

Que signifie le paramètre k ?

- $\Rightarrow$ $k < 1$: le risque **diminue** avec le temps (apprentissage, sélection) ;
- $\Rightarrow$ $k = 1$: **risque constant** (processus "sans mémoire") ;
- $\Rightarrow$ $k > 1$: le risque **augmente** (vieillissement, usure).

Interprétation du test :

- $\Rightarrow$ si les données sont compatibles avec $k = 1 \Rightarrow$ modèle exponentiel suffisant ;
- $\Rightarrow$ si $k \neq 1 \Rightarrow$ il existe une **forme du hasard non constante** $\Rightarrow$ Weibull nécessaire.

Question pratique : Vos données suggèrent-elles que le **risque change au fil du temps** (croît-il ? décroît-il ?) ou restent-elles compatibles avec un risque constant ?

Sélection de modèle : expo vs Weibull (2/2)

Hypothèses :

$$H_0 : k = 1 \qquad \text{(hasard constant)}$$
$$H_1 : k \neq 1 \qquad \text{(hasard non constant)}$$

Statistique du rapport de vraisemblance :

$$LR = 2\big(\ell_{\text{Weibull}} - \ell_{\text{Expo}}\big), \qquad LR \overset{H_0}{\sim} \chi^2(1).$$

Interprétation :

- ⇒ **LR petit** (p-value grande) ⇒ les données ne suggèrent pas de changement du risque. ⇒ **exponentielle acceptable**.
- ⇒ **LR grand** (p-value faible) ⇒ le risque évolue avec le temps. ⇒ **préférer Weibull**.

En clair : Ce test vérifie si le **hasard est constant** ou non.

Exemple numérique : test expo vs Weibull

Supposons :

$$\ell_{\text{Expo}} = -520.3, \qquad \ell_{\text{Weibull}} = -510.1$$

$$LR = 2(\ell_W - \ell_E) = 20.4$$

p-value $\approx 0.00001 \Rightarrow$ on **rejette l'exponentielle**. Le hasard n'est **pas constant**.

Pourquoi AIC et BIC ? (Sélection de modèle 1/2)

Problème : Le test du rapport de vraisemblance (LR) ne compare que des **modèles imbriqués** (ex : exponential $\subset$ Weibull).

Mais en pratique, on veut souvent comparer :

- $\Rightarrow$ Weibull vs log-normal ;
- $\Rightarrow$ log-logistique vs exponentielle ;
- $\Rightarrow$ plusieurs modèles AFT ou PH entre eux ;
- $\Rightarrow$ modèles avec covariables vs modèles sans covariables.

Besoin : un outil qui compare **l'ajustement** du modèle et **la pénalité de complexité**, même pour des modèles totalement différents.

Solution : AIC et BIC Deux critères simples pour choisir le *meilleur modèle* parmi plusieurs, même s'ils ne sont pas imbriqués.

Sélection de modèle : AIC et BIC (2/2)

Définitions :

$$AIC = -2\ell + 2p, \qquad BIC = -2\ell + p \log n.$$

⇒ ℓ : log-vraisemblance maximisée (qualité d'ajustement).

⇒ p : nombre de paramètres (complexité du modèle).

⇒ n : taille de l'échantillon.

Interprétation :

⇒ Le terme -2ℓ récompense **l'ajustement** (plus petit = meilleure fit).

⇒ Les termes $2p$ ou $p \log n$ pénalisent **la complexité**.

⇒ **Modèle préféré = celui avec le critère le plus petit**.

Sélection de modèle : AIC et BIC (2/2)

Définitions :

$$AIC = -2\ell + 2p, \qquad BIC = -2\ell + p \log n.$$

⇒ ℓ : log-vraisemblance maximisée (qualité d'ajustement).

⇒ p : nombre de paramètres (complexité du modèle).

⇒ n : taille de l'échantillon.

AIC : favorise l'ajustement (moins sévère). **BIC** : pénalise davantage les modèles complexes (critère plus conservateur).

AIC/BIC permettent de comparer **des modèles non imbriqués**, sur les **mêmes données**, quel que soit le cadre (PH, AFT, paramétriques...).

Exemple numérique : comparaison AIC/BIC

Trois modèles ajustés sur les mêmes données :

Modèle	p (param.)	AIC	BIC
Exponentielle	1	1042	1046
Weibull	2	1026	1034
Log-logistique	2	1022	1030

Lecture du tableau :

- ⇒ L'exponentielle a l'AIC/BIC les plus **grands** → mauvais ajustement.
- ⇒ Weibull améliore nettement l'ajustement (AIC/BIC plus faibles).
- ⇒ Log-logistique est encore meilleur : **gains supplémentaires** malgré la même complexité que Weibull.

Plus petit = meilleur compromis ajustement / complexité. Ici : **log-logistique** gagne selon AIC et BIC.

Diagnostics visuels pour choisir une loi

1. KM vs modèle théorique

⇒ Superposer la courbe de survie paramétrique à la KM.

⇒ Chercher des écarts systémiques (début / fin de support).

2. Graphique $\log(-\log \hat{S}(t))$

⇒ Linéaire en t^{α} ⇒ Weibull plausible.

⇒ Courbure marquée ⇒ loi plus flexible à envisager.

3. Nelson-Aalen vs $H(t)$ **théorique**

⇒ Vérifier la forme du **risque cumulé**.

⇒ Segments quasi-linéaires vs convexes/concaves.

Un bon fit visuel ne suffit pas : il faut aussi vérifier la **cohérence de** $h(t)$ avec l'histoire économique.

Durée-dépendance ou hétérogénéité latente ? (1/2)

Observation courante en données de durée : **un hazard décroissant** (risque plus élevé au début, puis plus faible).

Deux interprétations possibles :

⇒ **Durée-dépendance réelle** Le risque change réellement avec le temps :

- apprentissage (chômage) ;
- découragement (prospection) ;
- épuisement d'opportunités.

⇒ **Hétérogénéité non observée / sélection** Les individus sont différents :

- ceux avec risque élevé sortent très tôt ;
- ceux qui restent sont les "robustes".

Message clé : Un hazard décroissant observé *peut être dû à la composition du groupe*, pas à un changement du risque individuel.

Pourquoi un hazard peut *sembler* décroissant ? (2/2)

Exemple : deux types d'individus

⇒ 50% "**fragiles**" : hazard $h = 0.5$ (risque élevé)

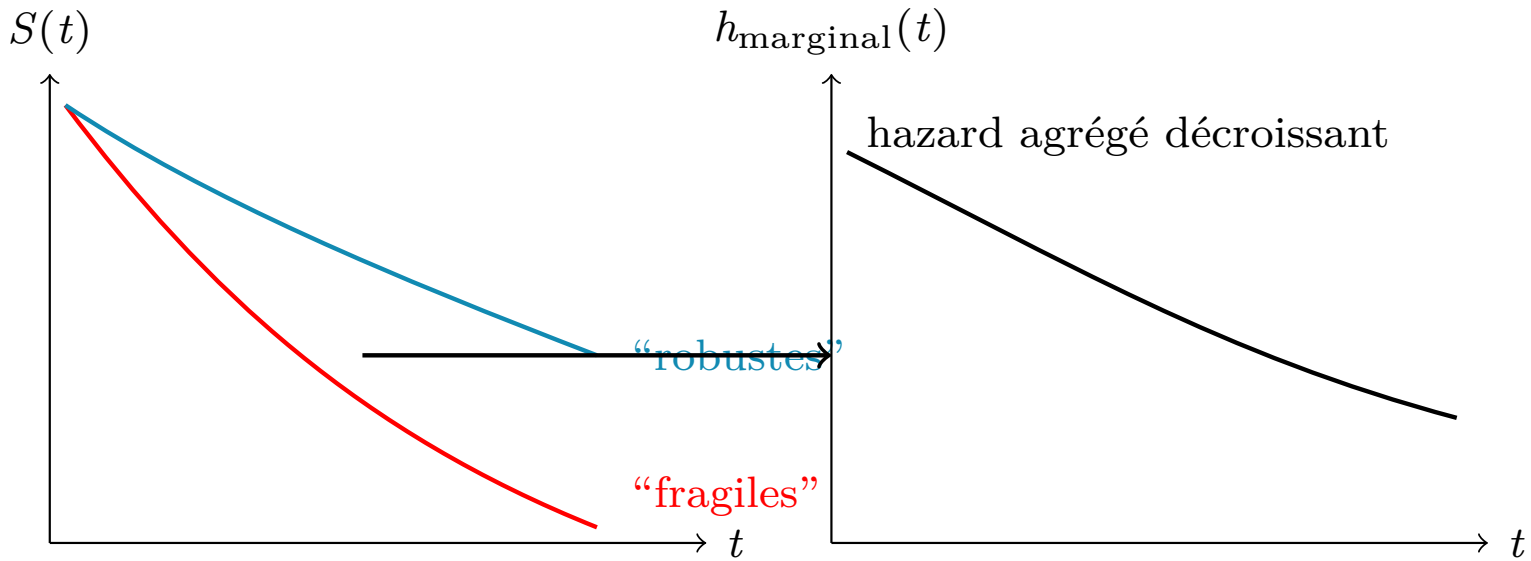
⇒ 50% "**robustes**" : hazard $h = 0.1$ (risque faible)

Ce qui se passe dans le temps :

1. Au début, beaucoup de "fragiles" sortent très vite.
2. Après un moment, il ne reste principalement que les "robustes".
3. Le hazard **observé** dans l'échantillon diminue... même si le hazard **individuel** est constant !

Hazard moyen décroissant ≠ hazard individuel décroissant. La baisse peut venir d'une **sélection progressive** des survivants.

Frailty : sélection des fragiles vs robustes



Même si chaque individu a un **hasard constant**, la **moyenne** peut être décroissante car les plus fragiles sortent plus tôt. C'est l'effet **frailty**.

Pourquoi un mélange d'individus crée un hazard décroissant ?

Idée centrale : Les individus n'ont pas tous le même risque **initial**. Certains sont fragiles (risque élevé), d'autres robustes (risque faible).

Ce qui se passe dans le temps :

1. Les individus **fragiles** ont plus de chances de connaître l'événement tôt.
2. Ils quittent donc l'échantillon rapidement.
3. Plus le temps passe, plus la population restante est composée de **survivants robustes**.
4. Le **risque moyen observé** baisse - même si le risque individuel ne change pas.

Un hazard décroissant peut venir d'une **sélection naturelle des survivants**, pas d'un changement du risque individuel.

Frailty Gamma : l'intuition avant les maths

Idée centrale : Les individus n'ont pas tous le même risque initial. Leur " fragilité " est un facteur multiplicatif noté v.

$$h(t \mid v) = v\lambda$$

Conditionnellement à v : chaque individu suit une loi exponentielle (hasard constant).

$$S(t \mid v) = e^{-v\lambda t}$$

Mais l'échantillon contient un **mélange** d'individus :

- $\Rightarrow$ les plus " fragiles " (v élevé) sortent vite ;
- $\Rightarrow$ les plus " robustes " (v faible) restent plus longtemps.

Survie observée = **moyenne** des survies individuelles. C'est ce mélange qui modifie la forme de $S(t)$.

Frailty Gamma : l'essentiel mathématique

1. Survie conditionnelle :

$$S(t \mid v) = e^{-v\lambda t}$$

2. Survie marginale = moyenne sur la distribution de v :

$$S(t) = \mathbb{E}_v\left[e^{-v\lambda t}\right]$$

3. Choix du Gamma : Si $v \sim \Gamma(\alpha, \alpha)$, alors

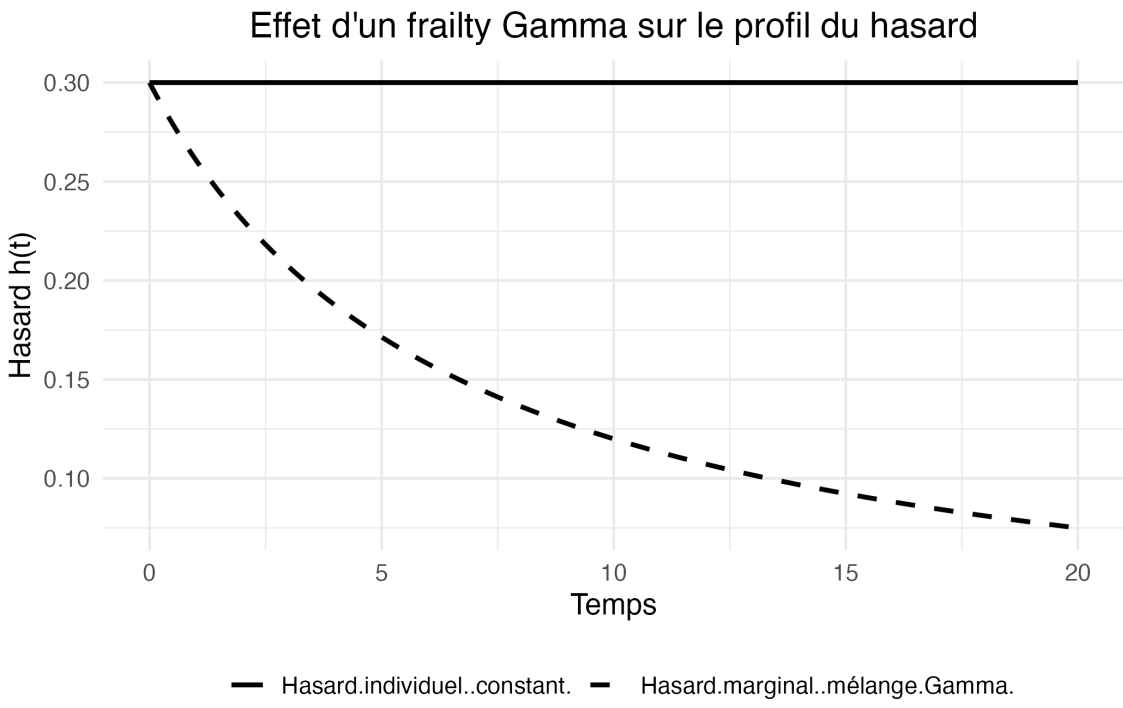
$$\mathbb{E}(e^{-vx}) = \left(1 + \frac{x}{\alpha}\right)^{-\alpha}.$$

En posant $x = \lambda t$:

$$S(t) = \left(1 + \frac{\lambda t}{\alpha}\right)^{-\alpha}.$$

Pourquoi le Gamma ? Parce qu'il donne une forme **fermée, simple et interprétable** pour $S(t)$.

Frailty : un effet simple à visualiser



⇒ Hasard individuel : constant.

⇒ Hasard agrégé : décroissant.

La sélection suffit à créer une **fausse** durée-dépendance.

Pourquoi intégrer des covariables dépendantes du temps ?

Dans beaucoup de situations, ce qui influence la durée change lui-même avec le temps :

- ⇒ un traitement commence (ou s'arrête) en cours de suivi ;
- ⇒ l'état de santé évolue ;
- ⇒ un individu suit une formation pendant le chômage ;
- ⇒ un revenu, un statut familial, une exposition au risque se modifient.

Limite des modèles "simples" : Ils supposent que les covariables restent **fixes** pendant toute la durée.

Idée clé : représenter le parcours d'un individu comme une suite d'intervalles où les covariables sont **constants**. Chaque changement d'état → un **nouvel épisode**.

Idée intuitive : découper le temps en épisodes

Principe : À chaque fois qu'une covariable change, on coupe le suivi en un nouvel intervalle.

Exemple :

Début	Fin	Événement ?	Statut du traitement
0	6	0	0 (non traité)
6	10	1	1 (traité)

Lecture :

⇒ L'individu est "non traité" de 0 à 6.

⇒ À 6, il commence le traitement → nouveau segment.

⇒ L'événement arrive à 10 pendant le dernier segment.

Cette structure "(début, fin, événement)" permet d'utiliser Cox, Weibull, log-logistique, etc. Les logiciels se chargent du reste.

Idée clé : découper le temps en épisodes (“counting process”)

Principe : Chaque fois qu’une covariable change, on crée un **nouvel intervalle de temps**. L’individu est donc représenté par plusieurs “ épisodes ” successifs :

Début	Fin	Événement ?	Statut $X(t)$
0	6	0	0 (non traité)
6	10	1	1 (traité)

Lecture de l’exemple :

- ⇒ de 0 à 6 : l’individu n’est pas traité, pas d’événement ;
- ⇒ à 6 : le statut change → nouvel épisode ;
- ⇒ de 6 à 10 : traité, et l’événement survient à la fin.

L’idée centrale : **actualiser** $X(t)$ **dans le temps** en découpant la trajectoire en épisodes successifs.

Pourquoi découper en épisodes (tstart-tstop) ?

Exemple concret : Un individu commence un traitement au temps $t = 6$.

$$\underbrace{[0, 6]}_{X(t)=0 \text{ : non traité}} \quad \Rightarrow \quad \underbrace{[6, 10]}_{X(t)=1 \text{ : traité}}$$

Problème : Les modèles de durée (Cox, Weibull, etc.) supposent que les covariables **restent constantes** pendant l'intervalle analysé.

Solution : découper le suivi en épisodes

- $\Rightarrow$ un épisode = une période où $X(t)$ ne change pas ;
- $\Rightarrow$ chaque changement de statut crée un nouvel intervalle ;
- $\Rightarrow$ l'événement appartient à l'épisode où il se produit.

Découper permet au modèle d'utiliser la **bonne valeur de la covariable** au **bon moment**.

Hazards cause-spécifiques : l'idée

Contexte : Plusieurs types possibles d'événement (ex. démission, licenciement).

Définition : pour une cause k

$$h_k(t) = \lim_{\Delta \to 0} \frac{\Pr(t \leq T < t + \Delta,\ \text{cause} = k \mid T \geq t)}{\Delta}$$

Interprétation :

- $\Rightarrow$ risque instantané de sortir **par la cause** k ;
- $\Rightarrow$ en traitant les autres causes comme des **censures**.

On étudie l'évolution du **mécanisme** k en lui-même.

Comment utiliser les hazards cause-spécifiques ?

Estimation :

⇒ créer un jeu de données par cause ;

⇒ cause k = événement ;

⇒ autres causes = censures ;

⇒ ajuster Cox ou un modèle paramétrique sur chaque cause.

Ce que cela dit :

⇒ comment les covariables influencent **chaque mécanisme** (ex. effets différents sur démission / licenciement).

Ne décrit pas la probabilité finale de sortir par $k \rightarrow$ il faut la **cumulative incidence** pour cela.

Modèle de Fine-Gray : l'idée

But : modéliser directement la **probabilité cumulée** d'un événement de cause k :

$$\mathrm{CIF}_k(t) = \Pr(T \leq t,\ \text{cause} = k).$$

Pourquoi ? Parce que la CIF répond à une question centrale :

$\Rightarrow$ " Quelle est la **probabilité** que l'événement k se produise avant t ? "

Idée Fine-Gray : modifier la définition du " risque " pour que la CIF soit directement reliée au modèle.

> Au lieu de modéliser le **hazard** d'une cause, on modélise la **CIF** elle-même.

Fine-Gray : intuition du ń subdistribution hazard ż

Pour relier directement les covariables à la **CIF**, Fine-Gray définit un hazard modifié :

$$\tilde{h}_k(t) = \lim_{\Delta \to 0} \frac{\Pr(t \leq T < t + \Delta,\ \text{cause} = k \mid \text{encore “à risque”})}{\Delta}$$

Différence clé : Les individus sortis par une autre cause **restent dans le dénominateur**.

ils ne disparaissent pas du “risk set”

⇒ Cela garde la probabilité cumulée dans la dynamique du modèle.

⇒ On obtient directement des effets sur la **CIF**, pas seulement sur les hazards.

À retenir : Cause-spécifique → effets sur les *hazards*. Fine-Gray → effets sur les *probabilités cumulées (CIF)*.

Cause-spécifique vs Fine–Gray : deux questions différentes

Modèle cause-spécifique (Cox appliqué à chaque cause)

⇒ On étudie le **mécanisme instantané** de la cause k.

⇒ Question :

" Comment X modifie **le hazard de la cause** k ? "

⇒ Les autres causes sont traitées comme des censures.

Modèle de Fine–Gray

⇒ On étudie directement la **probabilité cumulée** d'observer la cause k.

⇒ Question :

" Comment X modifie la **CIF** de la cause k ? "

⇒ Les autres causes restent dans le risk set (subdistribution hazard).

Résumé intuitif : Cause-spécifique → *comment ça se produit.* Fine–Gray → *avec quelle probabilité cela arrivera au final.*

Risques concurrents : pièges et intuition

Situation typique : un individu peut connaître plusieurs issues mutuellement exclusives. Exemples :

- ⇒ chômage : emploi / inactivité / formation ;
- ⇒ entreprise : faillite / rachat / sortie du marché ;
- ⇒ étudiant : diplomation / abandon / changement d'établissement.

Piège majeur : Kaplan-Meier suppose que la censure est **non-informative**.

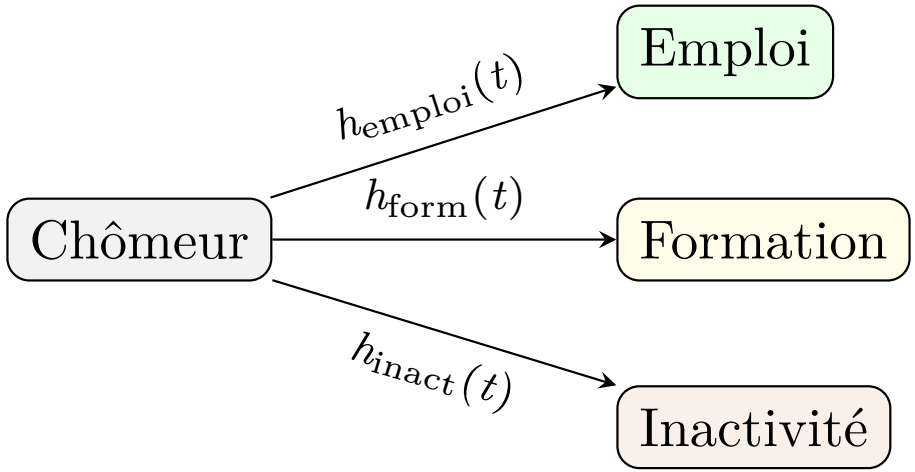
Mais ici :être "censuré" par une autre cause = événement informatif.

Conséquence :

- ⇒ KM **surestime** la survie spécifique à une cause.
- ⇒ KM répond à "probabilité de **n'avoir aucun événement**", pas à "probabilité de **ne pas avoir l'événement** k".

Message essentiel : Avec plusieurs causes, la survie et le risque doivent être définis **cause par cause**.

Exemple de risques concurrents : sortie du chômage



⇒ Trois **hazards cause-spécifiques** : emploi, formation, inactivité.

⇒ Un individu peut quitter l'état "chômeur" par **une seule** de ces issues.

Pour étudier l'effet d'une politique, on peut :

⇒ soit modéliser chaque hazard $h_k(t)$ séparément (cause-spécifique),

⇒ soit modéliser directement la **probabilité cumulée** d'accès à l'emploi (Fine–Gray).

Incidence cumulée (CIF) : interprétation économique

$$\mathrm{CIF}_k(t) = \Pr(T \leq t,\ \text{cause} = k)$$

Lecture : probabilité que l'individu ait connu l'issue k avant le temps t.

Distinction cruciale :

⇒ **CIF** = probabilités réelles observées dans une population.

⇒ **Hazard cause-spécifique** = *mécanismes instantanés.*

En présence de concurrence :

$$\mathrm{CIF}_k(t) = \int_0^t S(u^-)\, h_k(u)\, du.$$

⇒ h_k détermine la vitesse vers la cause k.

⇒ $S(u)$ capture la compétition des autres causes.

Donc : Analyser h_k = comprendre les **forces** ; Analyser CIF = comprendre les **probabilités finales** observées.

Rappels

Motivation

Lois de base

Extensions

Panorama & extensions

Quel modèle retenir ?

Avant de choisir un modèle : regarder les données. La forme de $S(t)$, $h(t)$, et la présence de queues lourdes ou de durée-dépendance guident le choix.

Quelques repères pratiques :

⇒ **Weibull**

- flexible : hasard monotone (croissant ou décroissant) ;
- compatible **PH** *et* **AFT** ;
- très bon point de départ.

⇒ **Log-logistique**

- très utile si $S(t)$ a une **queue lourde** ;
- hazard en **cloche** (non monotone) ;
- souvent gagnant en **AIC/BIC**.

⇒ **KM / NA**

- indispensable pour **visualiser** la structure des données ;
- permet de repérer monotonicité, points de rupture, hétérogénéité.

Réflexe pratique :

Extensions : les 5 idées à retenir

1. Vraisemblance sous censure

$$\text{événement} \rightarrow f(t) \qquad \text{censure} \rightarrow S(t)$$

2. Choix de modèle

$$\text{Imbriqués : LR-test} \qquad \text{Non imbriqués : AIC/BIC}$$

3. Frailty

$$h(t)\downarrow \text{ observé} \not\Rightarrow h(t)\downarrow \text{ individuel}$$

4. Covariables dépendantes du temps

$$(t_{\text{start}}, t_{\text{stop}}] \text{ pour rendre } X(t) \text{ constant par épisode}$$

5. Risques concurrents

$$\text{Cause-spécifique : hazard} \qquad \text{Fine–Gray : CIF}$$

Trois cadres pour analyser les durées

1. Non-paramétrique (KM / NA)

Décrit les données, sans hypothèses.

2. Semi-paramétrique (Cox PH)

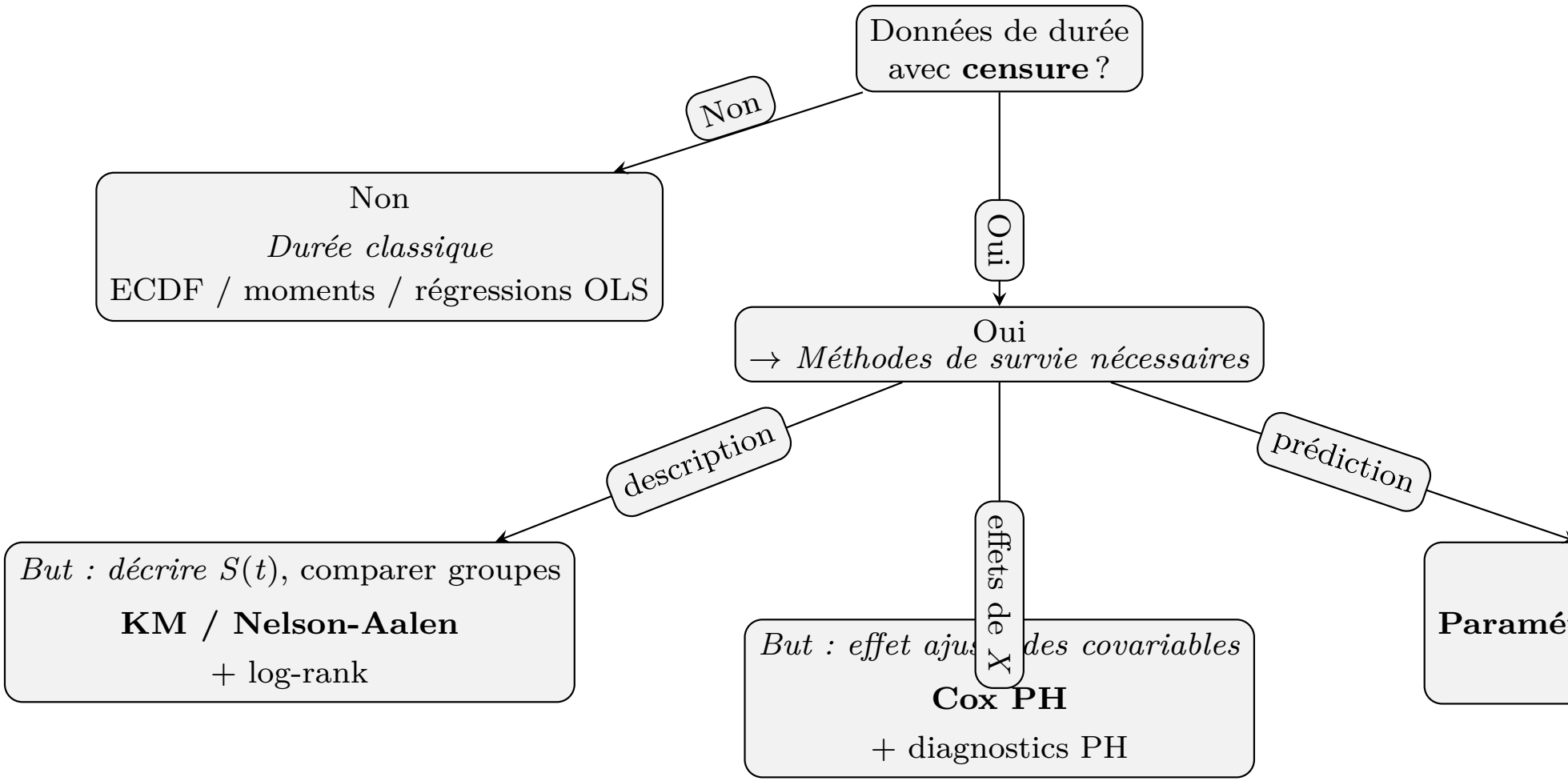
Effets de covariables ; forme de $h_0(t)$ libre.

3. Paramétrique (Expo, Weibull, Log-logistique)

Forme imposée $\Rightarrow$ interprétation + prédiction.

Résumé : Souplesse (KM) $\rightarrow$ équilibre (Cox) $\rightarrow$ prédiction (paramétrique).

Choisir le bon cadre : arbre décisionnel enrichi



Règle d'or : commencer simple (KM), tester PH, puis affiner (Cox ou paramétrique).

Checklist pratique : analyser des durées

1. Explorer

⇒ Courbes KM par groupe.

⇒ NA pour la forme du risque.

⇒ Niveau de censure.

2. Tester

⇒ Diagnostics PH (log–log parallèle ?)

⇒ Expo vs Weibull : LR-test.

⇒ Comparer les lois : AIC / BIC.

3. Choisir

⇒ **PH** si hazard ratio central.

⇒ **AFT** si time ratio plus parlant.

⇒ **Frailty** si hétérogénéité forte.

⇒ **Risques concurrents** si plusieurs issues.