

Modèles de durée

Théorie et applications

Etienne Dagorn

Université de Lille - LEM

Qui suis-je ?

- ⇒ **Nouveau maître de conférences** à l'Université de Lille ;
- ⇒ **Thèmes de recherche** : économie expérimentale, économie de l'éducation ;
- ⇒ **Parcours académique** :
 - Doctorat en économie à l'Université de Rennes 1 ;
 - Postdoctorat à l'INED (Institut national d'études démographiques).
- ⇒ **Expériences d'enseignement** :
 - Méthodes quantitatives, économétrie appliquée ;
 - Économie de l'éducation, politiques publiques.

Objectifs du cours

À l'issue du cours, vous serez en mesure de :

- ⇒ Comprendre les notions fondamentales liées à l'analyse de durée :
 - fonction de survie, fonction de risque, censure, troncature.
- ⇒ Identifier les principaux modèles :
 - modèles non paramétriques (Kaplan-Meier), semi-paramétriques (Cox), paramétriques (Weibull, etc.).
- ⇒ Estimer et interpréter un modèle de durée sur données réelles.
- ⇒ Mobiliser ces outils dans des contextes appliqués (économie du travail, santé, finance...).
- ⇒ Utiliser R pour produire des analyses reproductibles.

Organisation du cours

⇒ **Format** : 4 séances de 3h (cours + TP sur logiciel).

- Trois cours théoriques ;
- Dernier cours sur marchine ;

⇒ **Évaluation** :

- Produire un rendu de 5 pages

⇒ Comment échanger ?

- etienne.dagorn@univ-lille.fr
- Formule de politesse par mail ;
- Réponse aux mails le lundi matin **uniquement** ;

Séance 1 : Introduction

Compétences visées aujourd'hui

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

- ⇒ **Identifier** la censure/troncature dans une base réelle.
- ⇒ **Expliquer** les 5 fonctions clés (F, S, f, h, H) et leurs liens.
 - Répartition ($F(t)$) ; Survie ($S(t)$) ; Densité ($f(t)$) ; hasard ($h(t)$) ; Hasard cumulé $H(t)$
- ⇒ **Diagnostiquer** un biais sur une moyenne naïve sous censure.
- ⇒ **Construire** (T, δ) à partir d'une frise temporelle.

Questions de recherche

- ❶ **Combien de temps dure un épisode** (ex. : chômage, traitement, contrat) ?
- ❷ **Quels facteurs influencent cette durée** (âge, formation, traitement, etc.) ?
- ❸ **Tous les individus ont-ils le même risque** de sortir d'un état donné ?

Questions de recherche

- ❶ Combien de temps dure un épisode (ex. : chômage, traitement, contrat) ?
- ❷ Quels facteurs influencent cette durée (âge, formation, traitement, etc.) ?
- ❸ Tous les individus ont-ils le même risque de sortir d'un état donné ?

Termes clés

- ⇒ La **durée** mesure le **temps d'attente** jusqu'à un **événement** (sortie, rechute, embauche...).
- ⇒ **Observation complète** : l'événement est observé pendant la période d'étude.
- ⇒ **Censure à droite** : l'événement n'a pas encore eu lieu à la fin de l'observation.

Où rencontre-t-on des *durées*? (exemples concrets)

Travail

- ⇒ Durée de chômage → événement
= *emploi*
- ⇒ *Censure à droite* : encore au chômage à la fin d'étude

Santé

- ⇒ Temps jusqu'à récurrence → *rechute*
- ⇒ *Perte de vue* = censure aléatoire

Industrie

- ⇒ Temps jusqu'à panne →
défaillance
- ⇒ Maintenance préventive =
observation interrompue

Marketing

- ⇒ Temps jusqu'à la résiliation
- ⇒ Clients encore actifs = censurés

Finance

- ⇒ Temps jusqu'au *défaut*
- ⇒ Portefeuille encore sain = censure
déterministe (fin d'horizon)

Point commun : on observe des **temps d'attente** avec **censure** ; on veut **décrire, comparer, expliquer**.

Objectif : Estimer la durée moyenne de chômage de 5 individus suivis pendant 2 ans (de janvier 2020 à janvier 2022).

Individu	Sortie du chômage ?	Durée (mois)	Censuré ?
A	Oui (emploi)	6	Non
B	Oui (formation)	12	Non
C	Non (toujours au chômage)	24	Oui
D	Oui (emploi)	15	Non
E	Non	24	Oui

Question 1 : Quelle est la durée moyenne observée si on exclut les censurés ?

Objectif : Estimer la durée moyenne de chômage de 5 individus suivis pendant 2 ans (de janvier 2020 à janvier 2022).

Individu	Sortie du chômage ?	Durée (mois)	Censuré ?
A	Oui (emploi)	6	Non
B	Oui (formation)	12	Non
C	Non (toujours au chômage)	24	Oui
D	Oui (emploi)	15	Non
E	Non	24	Oui

Question 1 : Quelle est la durée moyenne observée si on exclut les censurés ?

$$\frac{6+12+15}{3} = 11 \text{ mois}$$

Objectif : Estimer la durée moyenne de chômage de 5 individus suivis pendant 2 ans (de janvier 2020 à janvier 2022).

Individu	Sortie du chômage ?	Durée (mois)	Censuré ?
A	Oui (emploi)	6	Non
B	Oui (formation)	12	Non
C	Non (toujours au chômage)	24	Oui
D	Oui (emploi)	15	Non
E	Non	24	Oui

Question 1 : Quelle est la durée moyenne observée si on exclut les censurés ?

$$\frac{6+12+15}{3} = 11 \text{ mois}$$

Question 2 : Ce calcul est-il juste ?

Objectif : Estimer la durée moyenne de chômage de 5 individus suivis pendant 2 ans (de janvier 2020 à janvier 2022).

Individu	Sortie du chômage ?	Durée (mois)	Censuré ?
A	Oui (emploi)	6	Non
B	Oui (formation)	12	Non
C	Non (toujours au chômage)	24	Oui
D	Oui (emploi)	15	Non
E	Non	24	Oui

Question 1 : Quelle est la durée moyenne observée si on exclut les censurés ?

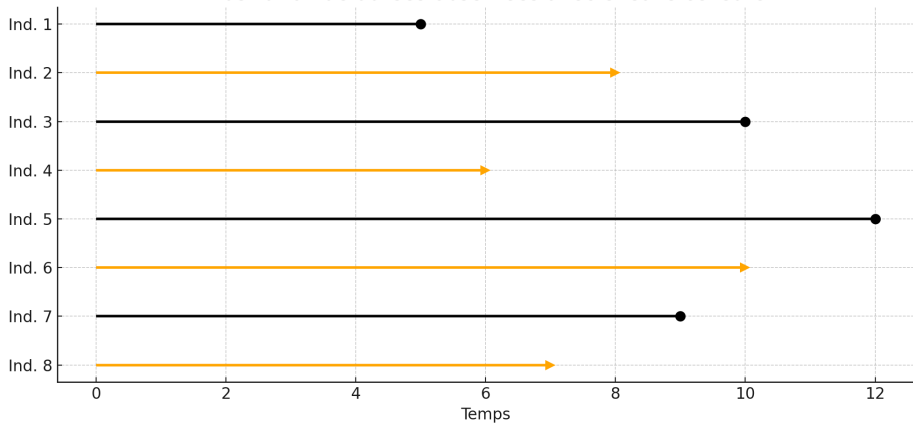
$$\frac{6+12+15}{3} = 11 \text{ mois}$$

Question 2 : Ce calcul est-il juste ?

Non : on exclut les chômeurs de longue durée (C et E), ce qui **biaisera l'estimation vers le bas**.

Moralité : Ne pas tenir compte de la censure revient à ignorer les individus les plus longs à sortir du chômage.

Illustration de durées observées avec et sans censure



Pourquoi pas une régression linéaire (OLS) ?

Problèmes structurels pour les durées

- ⇒ **Censure** : beaucoup de T sont tronqués à droite ⇒ *biais de sélection* si on les exclut.
- ⇒ **Support $\mathbb{R}_+$, asymétrie, queues** : OLS suppose erreurs symétriques ; $\log(T)$ ne règle ni la censure ni le *risk set*.
- ⇒ **Hétéroscédasticité forte & non-normalité** : variances croissantes avec t ; IC douteux.
- ⇒ **Temps au coeur du mécanisme** : le *rythme* de sortie (hazard) évolue ; OLS ignore le *timing*.

Alternatives adaptées : **Kaplan-Meier** (décrire), **Cox PH** (comparer via *hazard ratios*), **AFT** (effet = *accélération* du temps).

Bénéfices : gèrent la **censure**, respectent le support, interprétations claires (médiane, RMST, HR).

Pourquoi des modèles de *durée*? (intuition)

- ⇒ On n'observe pas toujours la fin : **censure** (certains n'ont pas encore l'événement).
- ⇒ Le **rythme d'arrivée** de l'événement *change dans le temps*.
- ⇒ On veut comparer des groupes / effets d'un facteur *en tenant compte du temps*.

Image mentale : une **cohorte qui s'effrite** au fil du temps. Les bons modèles *organisent* cet effritement pour décrire, comparer, expliquer.

Hypothèses transversales à garder en tête

- ⇒ **Censure non-informative** (conditionnellement aux covariables) : la probabilité d'être censuré ne dépend pas du temps vrai restant.
- ⇒ **Ensemble à risque** : au temps t , seules les personnes encore “exposées” doivent contribuer aux calculs (risk set).
- ⇒ **Origine du temps cohérente** : toutes les durées sont mesurées depuis le même événement initial défini.
- ⇒ **Unité et calendrier** : homogénéiser jours/mois, gérer les entrées retardées (troncature à gauche).

Pourquoi maintenant ? Ces hypothèses guident *tous* les choix de nettoyage et d'estimation que vous ferez dès la séance 2.

Trois questions $\Rightarrow$ trois familles

- ① **Décrire** “combien de temps ça dure?”
 $\Rightarrow$ **Kaplan–Meier** : photographie *non paramétrique* de la survie.
- ② **Comparer/Expliquer** par un facteur X en termes de *risque instantané*
 $\Rightarrow$ **Cox (PH)** : X **multiplie** un *métronome de risque* commun.
- ③ **Expliquer** en termes d'*accélération du temps*
 $\Rightarrow$ **AFT** : X **étire/comprime** la *règle du temps*.

Notations & symboles

T	Durée (positive), temps jusqu'à l'événement
δ	Indicateur d'événement observé (1 si évènement, 0 si censuré)
$F(t)$	Fonction de répartition $\Pr(T \leq t)$
$S(t)$	Survie $\Pr(T > t) = 1 - F(t)$
$f(t)$	Densité (si F est lisse), $f(t) = F'(t) = -S'(t)$
$h(t)$	Hasard (risque instantané conditionnel) $= f(t)/S(t)$
$H(t)$	Hasard cumulé $= \int_0^t h(u) du$, avec $S(t) = e^{-H(t)}$
t_0	Origine du temps (ex. inscription, début suivi)

Règle d'or : préciser *l'origine* t_0 , *l'unité de temps* (jours/mois) et *le mécanisme d'observation* (censure/troncature) avant toute estimation.

Définitions

Répartition et densité

Survie

Hasard

Pourquoi modéliser les durées ?

- ⇒ Les durées ont des caractéristiques spécifiques :
- **Censurées** : certaines durées ne sont pas complètement observées ;
 - **Tronquées** : certaines durées ne sont jamais observables.

Pourquoi modéliser les durées ?

- ⇒ Les durées ont des caractéristiques spécifiques :
- **Censurées** : certaines durées ne sont pas complètement observées ;
 - **Tronquées** : certaines durées ne sont jamais observables.
- ⇒ Modéliser ces durées permet de :
- **Décrire** les comportements temporels (sortie d'un état, survie) ;
 - **Comparer** des groupes (hommes/femmes, traités/non traités) ;
 - **Évaluer** l'impact de politiques ou d'événements.

Pourquoi modéliser les durées ?

- ⇒ Les durées ont des caractéristiques spécifiques :
 - **Censurées** : certaines durées ne sont pas complètement observées ;
 - **Tronquées** : certaines durées ne sont jamais observables.
- ⇒ Modéliser ces durées permet de :
 - **Décrire** les comportements temporels (sortie d'un état, survie) ;
 - **Comparer** des groupes (hommes/femmes, traités/non traités) ;
 - **Évaluer** l'impact de politiques ou d'événements.
- ⇒ À noter :
 - **durée totale jusqu'à l'événement** (temps d'attente)
 - **durée résiduelle** : temps restant conditionnellement à la survie jusqu'à un moment donné.

Définitions

Répartition et densité

Survie

Hasard

Variables de durée : définitions et spécificités

- ⇒ Les modèles de durée ont été développés pour analyser la **durée de vie**, mais s'appliquent à de nombreux domaines.
- ⇒ Les variables de durée sont :
 - **Strictement positives** (pas de valeur négative ou nulle) ;
 - **Souvent censurées**, c'est-à-dire partiellement observées.
- ⇒ Types de censure :
 - **À droite** : la durée n'est pas terminée à la fin de l'étude ;
 - **À gauche** : le processus a déjà commencé à l'entrée dans l'échantillon.

Variables de durée : définitions et spécificités

- ⇒ Les modèles de durée ont été développés pour analyser la **durée de vie**, mais s'appliquent à de nombreux domaines.
- ⇒ Les variables de durée sont :
 - **Strictement positives** (pas de valeur négative ou nulle) ;
 - **Souvent censurées**, c'est-à-dire partiellement observées.
- ⇒ Types de censure :
 - **À droite** : la durée n'est pas terminée à la fin de l'étude ;
 - **À gauche** : le processus a déjà commencé à l'entrée dans l'échantillon.
- ⇒ Ne pas tenir compte des observations censurées introduit un **biais de sélection** :
 - Exemple : ignorer les chômeurs de longue durée fausse l'analyse.
- ⇒ **Durée vs. temps d'attente** :
 - Une longue attente ne signifie pas qu'on est "près de la sortie" ;
 - Cela dépend du processus sous-jacent (mémoire ou non).

Durées : domaine & codage

Domaine & codage

- ⇒ **Support** : $T \geq 0$ (souvent $T > 0$).
- ⇒ **Échelle** : **continue** (heures/jours) *ou* **discrète** (mois/trimestres).
- ⇒ **Origine du temps** : fixer t_0 (ex. inscription au chômage).
- ⇒ **Unité** : explicite et **cohérente** avec les covariables.

Exemples : durée chômage (mois), temps jusqu'à panne (heures), réadmission hôpital (jours).

Fonctions utilisées pour décrire les durées

⇒ Une durée est décrite par sa **fonction de répartition** :

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

⇒ En pratique, on mobilise des objets plus adaptés aux durées :

- **Fonction de survie** $S(t) = \Pr(T > t) = 1 - F(t)$: probabilité d'être encore en vie/à risque " à t .
- **Densité** $f(t) = \frac{dF}{dt} = -\frac{dS}{dt}$: *vitesse* d'arrivée des événements autour de t .
- **Fonction de risque (hasard)** $h(t) = \frac{f(t)}{S(t)}$: *intensité instantanée* de sortie conditionnelle à survivre jusqu'à t .
- **Hasard cumulé** $H(t) = \int_0^t h(u) du$, avec le lien clé $S(t) = e^{-H(t)}$.

⇒ Ces notions sont équivalentes aux concepts probabilistes classiques mais plus **interprétables** pour des durées (ex. démographie : *taux de mortalité, probabilité de survie, espérance de vie*).

Censure, troncature & implications

Schéma d'observation

Censure = l'instant exact de l'événement n'est **pas entièrement observé**.

- ⇒ **Censure à droite** : l'étude se termine avant l'événement ⇒ on sait seulement que $T > t_{\text{fin}}$.
- ⇒ **Censure à gauche** : l'événement a eu lieu avant l'entrée ⇒ on sait seulement que $T < t_{\text{entrée}}$.
- ⇒ **Censure par intervalle** : on sait que l'événement survient dans $[a, b]$.

Troncature = inclusion conditionnelle dans l'échantillon.

- ⇒ **Troncature à gauche (entrée retardée)** : on n'observe un individu que s'il vérifie $T > t_{\text{entrée}}$.

Cas typiques & troncature (entrée retardée)

Cas d'usage (censure) :

- ⇒ **À droite** : fin d'étude à date fixe $\Rightarrow \delta = 0$ si l'événement n'est pas encore survenu.
- ⇒ **Aléatoire** : *perte de vue* avant la fin $\Rightarrow \delta = 0$ (même traitement côté estimation).
- ⇒ **À gauche / par intervalle** : événement avant l'entrée / connu seulement dans $[a, b]$.
→ avant l'entrée dans l'observation, mais on ne connaît pas exactement quand

Troncature à gauche (entrée retardée) :

- ⇒ Inclusion conditionnelle : on n'observe que les individus tels que $T > t_{\text{entrée}}$.
- ⇒ À déclarer explicitement dans le logiciel (ex. entrée retardée / left-truncation).

Exemple : ignorer la censure $\Rightarrow$ moyenne biaisée

Contexte : suivi de 6 personnes pendant 24 mois, événement = *retour à l'emploi*.

Individu	Sortie ?	Temps (mois)	δ	Commentaire
A	Oui	6	1	Événement observé
B	Oui	12	1	Événement observé
C	Non	24	0	Censuré à droite (tjr au chômage fin d'étude)
D	Oui	15	1	Événement observé
E	Non	24	0	Censuré à droite
F	?	< 3	0	<i>Censuré à gauche</i> (déjà sorti avant entrée dans l'étude)

Naïf (faux) : moyenne sur les seuls non-censurés $\Rightarrow (6 + 12 + 15)/3 = 11$ mois.

Problème : on **exclut** les plus longues durées (C,E) et on **ignore** F (censure à gauche) $\Rightarrow$ **sous-estimation**.

Moralité : utiliser des méthodes qui exploitent *toute* l'information disponible au lieu de jeter les censurés.

Du calendrier au couple (T, δ)

Objectif : transformer des dates (début, fin, événement) en un temps observé T et un indicateur δ utilisables par les modèles de durée.

Notation minimale (par individu i)

⇒ **Origine (référence)** : t_{0i} (ex. date d'inscription au chômage).

⇒ **Entrée à risque (éventuelle)** : a_i (*entrée retardée / left truncation*).

⇒ **Vrai temps d'événement** : T_i^* (inobservable si censuré).

⇒ **Temps de censure (fin de suivi)** : C_i .

$$\underbrace{Y_i}_{\text{temps observé (calendrier)}} = \min(T_i^*, C_i), \quad \underbrace{\delta_i}_{\text{statut}} = \mathbf{1}\{T_i^* \leq C_i\}.$$

$$\underbrace{T_i}_{\text{durée analysée}} = Y_i - \max(t_{0i}, a_i) \quad (\text{dans une unité cohérente : jours/mois}).$$

Censure - exemple (apprentissage de la marche)

- ⇒ Étude de l'âge d'acquisition de la marche (durée = naissance → premiers pas).
- ⇒ Données observées entre 10 et 16 mois.

Trois cas dans l'échantillon

- ① Marche avant 12 mois ⇒ censure à gauche (fixe).
- ② Apprentissage entre 10 et 16 mois ⇒ donnée complète.
- ③ ≥ 16 mois sans marche ou perdu de vue ⇒ censure à droite.

Censure à droite - précisions

Censure à droite : l'événement n'est pas observé pendant la fenêtre de suivi.

- ⇒ **Censure déterministe** : fin d'observation à 16 mois.
⇒ l'enfant ne marche pas encore à 16 mois, il marchera *après*.
- ⇒ **Censure aléatoire** : suivi interrompu à 12 mois pour certains enfants.
⇒ on *ignore* s'ils ont marché entre 12 et 16 mois.

Point-clé : traiter correctement la censure à droite évite de **sous-estimer** l'âge médian d'acquisition.

Censure à droite - autre exemple

Données longitudinales Céreq *Génération* (parcours pro jusqu'à l'interview).

- ⇒ Les individus décrivent les épisodes (emploi, chômage...) jusqu'à la date d'enquête.
- ⇒ *Génération 98* : en **printemps 2005**, on sait qu'un individu est au chômage, mais on **ne connaît pas sa date de sortie** de chômage ⇒ **censure à droite**.

Conséquence : si on exclut ces observations censurées, on **sous-estime** la durée moyenne passée au chômage.

Hypothèse clé : censure non-informative

Définition : conditionnellement aux covariables, le mécanisme de censure est **indépendant** de la durée vraie T .

⇒ les censurés ne sont pas *systématiquement* plus rapides/lents que les non-censurés (à X donnés).

Exemples

- ⇒ **Non-informative (OK)** : fin administrative de l'étude au 31/12/2024.
- ⇒ **Informative (problème)** : patients très malades quittent l'essai ⇒ censure liée au risque.

Signaux faibles (diagnostic empirique)

- ⇒ Comparer **profils X** des censurés vs non-censurés (écarts substantiels?).
- ⇒ Taux de censure très **différent par groupe** (ex. traitement vs contrôle).
- ⇒ Sorties de suivi liées à l'issue (pertes de vue post-événement probable).

Fenêtre d'observation & censure déterministe (contrats/abonnements)

Fenêtre : janvier 2000 $\rightarrow$ janvier 2004. Entrées à des **dates calendaires différentes**.

- $\Rightarrow$ Client A : souscrit **janv. 2001**, résilie **janv. 2003** $\Rightarrow$ **durée non censurée** $T = 25$ mois.
- $\Rightarrow$ Client B : souscrit **janv. 2002**, toujours client **janv. 2004** $\Rightarrow$ **censure déterministe** $T = 25^+$ mois.

Notation utile : $T = 25^+$ signale une **borne inférieure** (durée minimale observée).

Caractéristiques des durées & ensemble à risque

On étudie :

- ⇒ la **durée** passée dans un état avant la réalisation (ou non) d'un événement ;
- ⇒ la **probabilité de transition** d'une situation à une autre.

Condition essentielle

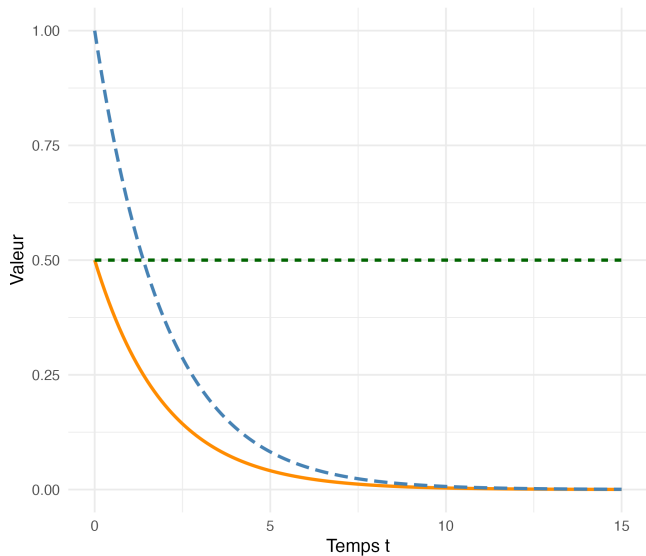
Tous les individus de la base doivent être **exposés au risque** de l'événement étudié (**ensemble à risque**).

Exemples :

- ⇒ **Durée/sortie du chômage** : ne considérer que des chômeurs (ou ex-chômeurs).
- ⇒ **Survie** : individus **vivants** au début de la période d'observation.

Fonctions de durée : Densité, Survie, Hasard

Fonction — Densité — Hasard — Survie



Définitions

Répartition et densité

Survie

Hasard

Durées : 5 objets (définitions & intuitions)

⇒ **Répartition** $F(t) = \Pr(T \leq t)$

Intuition : part ayant déjà eu l'événement avant t .

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

Intuition : probabilité d'être encore "en vie" à t .

⇒ **Densité** $f(t) = F'(t) = -S'(t)$ (si F différentiable)

Intuition : vitesse d'arrivée d'événements autour de t .

⇒ **Hasard (hazard)** $h(t) = \frac{f(t)}{S(t)}$

Intuition : rythme instantané *conditionnel* à être encore à risque.

⇒ **Hasard cumulé** $H(t) = \int_0^t h(u) du$

Intuition : accumulation du risque jusqu'à t .

Repères visuels : $F \uparrow (0 \Rightarrow 1)$, $S \downarrow (1 \Rightarrow 0)$. Pic de $f \Rightarrow$ pente forte de $F \Rightarrow$ chute rapide de S .

Durées : carte mentale des 5 objets

Objet	Formule	Question à laquelle il répond
Répartition	$F(t) = \Pr(T \leq t)$	“Quelle % part a déjà eu l'événement avant t ?”
Survie	$S(t) = \Pr(T > t) = 1 - F(t)$	“Quelle % part n'a <i>pas</i> encore eu l'événement à t ?”
Densité	$f(t) = F'(t) = -S'(t)$	“À quelle <i>vitesse</i> arrivent des évènements autour de t ?”
Hasard	$h(t) = \frac{f(t)}{S(t)}$	“Risque <i>instantané</i> de sortie à t <i>conditionnel</i> à être encore à risque.”
Hasard cumulé	$H(t) = \int_0^t h(u) du$	“Accumulation du risque jusqu'à t ; $S(t) = e^{-H(t)}$.”

À retenir : $F \uparrow$ de $0 \rightarrow 1$; $S \downarrow$ de $1 \rightarrow 0$; f et h ont une *unité de temps*, F, S, H sont sans unité.

Exemple chiffré : hasard **constant** $\lambda = 0,05$ par mois

Modèle (exponentielle, mémoire nulle) : $h(t) \equiv \lambda = 0,05 \text{ mois}^{-1}$.

$$H(t) = \lambda t = 0,05 t, \quad S(t) = e^{-\lambda t} = e^{-0,05t}, \quad f(t) = \lambda e^{-\lambda t}.$$

Repères numériques

⇒ **Survie à 6 mois** : $S(6) = e^{-0,30} \approx 0,74 \Rightarrow 74\%$ encore “en vie”.

⇒ **Espérance** : $\mathbb{E}[T] = \frac{1}{\lambda} = 20 \text{ mois}$.

Probabilité de sortie *conditionnelle* depuis t sur un intervalle Δt

$$\Pr(t < T \leq t + \Delta t \mid T > t) = 1 - e^{-\lambda \Delta t} \approx \underbrace{\lambda \Delta t}_{\text{approx. si } \Delta t \text{ petit}}.$$

⇒ **1 mois** : exact = $1 - e^{-0,05} \approx 4,88\%$ vs approx = 5%.

⇒ **3 mois** : exact = $1 - e^{-0,15} \approx 13,9\%$ (l'approx = 15%).

Mémoire nulle : $S(t+s \mid T > t) = S(s)$ pour tout s ; “ce qu’il reste à attendre” ne dépend pas de l’âge t .

Temps discret : hazard, données person-period

Définitions (temps discret : mois/trimestres)

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

Données *person-period* :

$Y_{it} = \mathbf{1}\{T_i = t\}$ observée si $T_i \geq t$ (une ligne par individu i et période t).

Modèle binomial (lien compl. log-log) :

$$\Pr(Y_{it} = 1 \mid Y_{i,t-1} = 0, X_{it}) = 1 - \exp\{-\exp(\alpha_t + X_{it}\beta)\}.$$

$$\underbrace{\log(-\log(1 - p_{it}))}_{\text{lien cloglog}} = \alpha_t + X_{it}\beta, \quad p_{it} \equiv \Pr(Y_{it} = 1 \mid \cdot)$$

Lecture de β : à intervalle donné t , e^β multiplie le *hazard discret* (probabilité conditionnelle d'évènement à t).

Pièges fréquents & mini-diagnostics

- ⇒ Confondre f (densité, peut > 1) et probabilité $\Rightarrow$ toujours interpréter via $f(t)\Delta t$.
- ⇒ Lire h comme une proba brute $\Rightarrow$ rappeler *conditionnel à survivre jusqu'à t* .
- ⇒ Oublier l'unité (jours/mois) $\Rightarrow$ incohérences de h et de la médiane.
- ⇒ Ignorer censure/troncature $\Rightarrow$ biais systématique des moyennes.

Auto-check : mes échelles sont cohérentes ? $S(0) = 1$? $F(\infty) = 1$? $S = e^{-H}$?

Fonction de répartition F

Considérons une durée $T > 0$. Sa distribution se décrit par :

$$F(t) = \Pr(T \leq t), \quad t \in \mathbb{R}_+, \quad 0 \leq F(t) \leq 1.$$

Définition : la fonction de répartition d'une variable aléatoire T est la fonction qui, à tout t , associe la probabilité que T prenne une valeur *inférieure ou égale* à t (probabilité cumulée).

⇒ **Propriétés** : F est **croissante**, $F(0) = \Pr(T \leq 0) \approx 0$,
 $\lim_{t \rightarrow \infty} F(t) = 1$.

⇒ **Intuition** : $F(t)$ = *part de la cohorte ayant déjà connu l'événement avant t* .

⇒ **Liens essentiels** :

$$S(t) = 1 - F(t) \quad (\text{survie}), \quad f(t) = F'(t) = -S'(t) \quad (\text{densité, si } F \text{ est lisse}).$$

À lire sur un graphe : F qui grimpe vite $\Rightarrow$ beaucoup d'événements précoces ; plateaux $\Rightarrow$ peu d'événements.

Densité et lien avec F

La **densité** (si elle existe) est :

$$f(t) = \frac{dF(t)}{dt} = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \Pr(t < T \leq t + \Delta t), \quad f(t) \geq 0.$$

⇒ *Intuition* : **vitesse** à laquelle surviennent des événements autour de t .

⇒ Ce n'est pas une probabilité (peut être > 1) mais une **densité** :
 $\Pr(t < T \leq t + \Delta t) \approx f(t) \Delta t.$

⇒ **Lien clé** : $F(t) = \int_0^t f(x) dx.$

Pense-bête : grande $f(t) \Rightarrow$ beaucoup d'événements autour de t .

Répartition F & densité f : lire les deux

$$F(t) = \Pr(T \leq t), \quad 0 \leq F(t) \leq 1, \quad F \uparrow \qquad f(t) = F'(t) = -S'(t) \geq 0$$

⇒ **Lire** F : à l'instant t , $F(t)$ = *fraction déjà "arrivée"*. Ex. : $F(6) = 0,4 \Rightarrow$ 40% ont vécu l'événement avant 6.

⇒ **Lire** f : *vitesse d'arrivées* autour de t (taux instantané non normalisé).

⇒ **Aire** : $F(t) = \int_0^t f(x) dx$.

⇒ **Lien avec la survie** : $S(t) = 1 - F(t)$ et $f(t) = -S'(t)$.

Guides visuels

Sur un même graphique, tracer $F(t)$ (courbe montante) et $f(t)$ (forme/pics).

⇒ **Pic de f** ⇒ **pente forte de F** ⇒ **chute rapide de S** .

Temps *discret* (mois, trimestres) - Notions de base

Cadre : $T \in \{1, 2, 3, \dots\}$ (ex. durée de chômage en mois).

Ce qu'on mesure

$$p(t) = \Pr(T = t) \quad (\text{proba de sortir à } t)$$

$$F(t) = \Pr(T \leq t) = \sum_{k=1}^t p(k), \quad S(t) = 1 - F(t) = \Pr(T > t)$$

Hasard discret (probabilité conditionnelle de sortie à t)

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

Relations clés :

$$S(t) = S(t-1)(1 - h(t)), \quad p(t) = S(t-1)h(t), \quad S(t) = \prod_{k=1}^t (1 - h(k)).$$

Exercice minute : lire ***F*** et ***f***

On observe à $t = 6$: $F(6) = 0,40$ et un pic de $f(t)$ autour de $t = 6-7$.

- ❶ Quelle part a déjà eu l'événement avant 6 ?
- ❷ Que dit le pic de f à 6-7 ?
- ❸ Vrai/Faux : si f est élevé à 6, alors F est plat à 6.

Exercice minute : lire ***F*** et ***f***

On observe à $t = 6$: $F(6) = 0,40$ et un pic de $f(t)$ autour de $t = 6-7$.

- ❶ Quelle part a déjà eu l'événement avant 6 ?
(40%)
- ❷ Que dit le pic de f à 6-7 ?
- ❸ Vrai/Faux : si f est élevé à 6, alors F est plat à 6.

Exercice minute : lire ***F*** et ***f***

On observe à $t = 6$: $F(6) = 0,40$ et un pic de $f(t)$ autour de $t = 6-7$.

- ❶ Quelle part a déjà eu l'événement avant 6 ?
- ❷ Que dit le pic de f à 6-7 ?
(*Nombreux événements qui arrivent à ce moment*)
- ❸ Vrai/Faux : si f est élevé à 6, alors F est plat à 6.

Exercice minute : lire ***F*** et ***f***

On observe à $t = 6$: $F(6) = 0,40$ et un pic de $f(t)$ autour de $t = 6-7$.

- ❶ Quelle part a déjà eu l'événement avant 6 ?
- ❷ Que dit le pic de f à 6-7 ?
- ❸ Vrai/Faux : si f est élevé à 6, alors F est plat à 6.
(*Faux : pente forte*)

Définitions

Répartition et densité

Survie

Hasard

Fonction de survie : définition & intuition

Définition (continu) : probabilité de *ne pas* avoir encore eu l'événement à t :

$$S(t) = \Pr(T > t) = 1 - F(t), \quad t \geq 0.$$

Intuition :

- ⇒ $S(t)$ mesure la **part encore “en vie”** à l'instant t ;
- ⇒ complémentaire de la répartition $F(t) = \Pr(T \leq t)$ (on utilise S car c'est plus naturel à lire pour des durées/survies).

Propriétés de base :

$$S(0) = 1, \quad S(t) \downarrow, \quad \lim_{t \rightarrow \infty} S(t) = 0.$$

À retenir : $S(t)$ est une *probabilité cumulée de “survivre”*, donc décroissante et bornée dans $[0, 1]$.

Sur un petit intervalle : lecture locale de S et h

Pour Δt petit et conditionnellement à $T > t$:

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

Idée : $h(t)$ est un *taux instantané* (1/unité temps) ; la probabilité locale est ce taux $\times$ la longueur de l'intervalle.

Lire des *nombres* sur $S(t)$: médiane, quantiles, RMST

Quantiles : $q_p = \inf\{t : S(t) \leq 1 - p\}$ ($p = 0,5$ pour la médiane).

RMST : restricted Mean Survival Time

RMST à horizon τ : temps moyen *jusqu'à* τ :

$$\text{RMST}(\tau) = \int_0^{\tau} S(t) dt \quad (\text{discret : } \sum_{t=1}^{\tau} S(t)).$$

- ⇒ La RMST compare des groupes même quand la médiane n'est pas atteinte.
- ⇒ Interprétation ⇒ “ heures/mois de vie/état *gagnés* avant τ ”.

Survie, densité & hazard : les liens utiles

Continu :

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt} \geq 0, \quad h(t) = \frac{f(t)}{S(t)},$$

$$S(t) = e^{-H(t)} \text{ avec } H(t) = \int_0^t h(u) du.$$

Lecture : f est la **vitesse d'arrivées**, h le **rythme instantané conditionnel** parmi ceux encore à risque.

Temps discret (mois/trimestres) :

$$h(t) = \Pr(T = t \mid T \geq t), \quad S(t) = \prod_{k=1}^t (1 - h(k)),$$

$$p(t) = \Pr(T = t) = S(t-1) h(t).$$

Repères rapides : f grand $\Rightarrow S$ chute vite ; h constant $\Rightarrow S(t) = e^{-\lambda t}$ (mémoire nulle).

Fonction de survie

Lien avec la densité :

$$f(t) = \frac{dF(t)}{dt} = \frac{d}{dt}(1 - S(t)) = -\frac{dS(t)}{dt}$$

Interprétation :

- ❶ la densité $f(t)$ correspond à la **vitesse à laquelle la fonction de survie décroît**.
- ❷ $\Rightarrow$ **probabilité que l'événement ait lieu autour de t** .
- ❸ $\Rightarrow$ forte densité implique une chute rapide de $S(t)$: beaucoup d'événements surviennent à ce moment.

Survie conditionnelle : remettre l'horloge à t_0

Pourquoi s'y intéresser ?

Question pratique : “ Combien de temps *reste-t-il* à attendre sachant qu'on a déjà survécu jusqu'à t_0 ? ”

Définition (conditionnement à avoir tenu jusqu'à t_0) :

$$S(t \mid T > t_0) = \Pr(T > t \mid T > t_0) = \frac{S(t)}{S(t_0)}, \quad t \geq t_0.$$

Intuition pédagogique : on *recalibre* la courbe de survie en partant de t_0 :

- ⇒ on “ remet l'horloge à t_0 ” ;
- ⇒ la probabilité d'être encore en vie à t étant donné la survie jusqu'à t_0 est la survie *relative* $S(t)/S(t_0)$.

À lire sur un graphe : prenez la courbe S , repérez $S(t_0)$, puis renormalisez la suite de la courbe en la divisant par $S(t_0)$.

Durée résiduelle espérée

Définition (temps restant moyen à partir de t_0) :

$$\mathbb{E}[T - t_0 \mid T > t_0] = \int_{t_0}^{\infty} \frac{S(u)}{S(t_0)} du \quad (\text{cont.})$$

$$\mathbb{E}[T - t_0 \mid T > t_0] = \sum_{u=t_0+1}^{\infty} \frac{S(u)}{S(t_0)} \quad (\text{dis.}).$$

Lecture : c'est l'aire sous la courbe de survie *conditionnelle* (ou la somme de ses “ barres ” en discret).

Cas repère : mémoire nulle (exponentielle)

Si $S(t) = e^{-\lambda t}$:

$$S(t \mid T > t_0) = e^{-\lambda(t-t_0)}, \quad \mathbb{E}[T - t_0 \mid T > t_0] = \frac{1}{\lambda} \quad (\text{indépendant de } t_0).$$

Repère mental : avec une loi exponentielle, “ ce qu'il reste à attendre ” ne dépend pas de l'âge courant.

Lire une courbe de survie (pratique)

- ⇒ $S(t)$ = **proba d'être encore "en vie"** à t ; les **marches** surviennent aux événements;
- ⇒ **plateaux** = périodes sans événements (les censures ne font pas chuter S).
- ⇒ **Médiane de survie** : première abscisse où $S(t) \leq 0,5$.
- ⇒ **N à risque** (sous l'axe) : quand N devient faible, l'incertitude augmente.

Pièges courants :

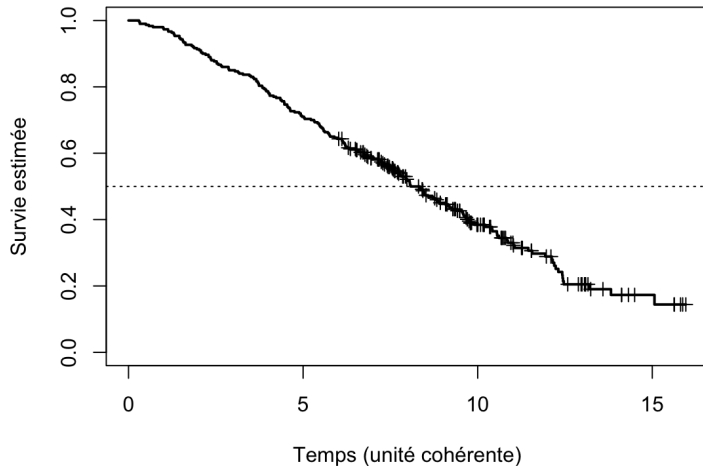
- ❶ lire S comme une fréquence brute (sans tenir compte des censures)
- ❷ comparer des courbes sans regarder N à risque
- ❸ oublier l'échelle du temps (jours vs mois)

Définitions
oooooooooooooooo

Répartition et densité
oooooooooooo

Survie
oooooooooooo●ooo

Hasard
oooooooooooooooo



Quiz express - (1/3)

Q1 - Vrai/Faux : Une variable de durée peut prendre des valeurs négatives.

Q2 - Interprétation : Sur un graphe, une *densité* $f(t)$ “ plate ” (quasi constante) signifie :

- (A) qu'il y a très peu d'événements au début puis beaucoup à la fin ;
- (B) qu'en moyenne la *vitesse d'arrivée* des événements varie peu avec t ;
- (C) que $F(t)$ décroît de façon à peu près linéaire ;
- (D) que $S(t)$ est constant dans le temps.

Quiz express - (1/3)

Q1 - Vrai/Faux : Une variable de durée peut prendre des valeurs négatives.

Q2 - Interprétation : Sur un graphe, une *densité* $f(t)$ “ plate ” (quasi constante) signifie :

- (A) qu'il y a très peu d'événements au début puis beaucoup à la fin ;
- (B) qu'en moyenne la *vitesse d'arrivée* des événements varie peu avec t ;
- (C) que $F(t)$ décroît de façon à peu près linéaire ;
- (D) que $S(t)$ est constant dans le temps.

Corrigé Q1 : Faux. Les durées sont ≥ 0 (souvent > 0).

Quiz express - (1/3)

Q1 - Vrai/Faux : Une variable de durée peut prendre des valeurs négatives.

Q2 - Interprétation : Sur un graphe, une *densité* $f(t)$ “ plate ” (quasi constante) signifie :

- (A) qu'il y a très peu d'événements au début puis beaucoup à la fin ;
- (B) qu'en moyenne la *vitesse d'arrivée* des événements varie peu avec t ;
- (C) que $F(t)$ décroît de façon à peu près linéaire ;
- (D) que $S(t)$ est constant dans le temps.

Corrigé Q1 : Faux. Les durées sont ≥ 0 (souvent > 0).

Indice Q2 : “ densité *plate* ” $\Rightarrow$ *vitesse d'arrivées* peu variable.

Quiz express - (1/3)

Q1 - Vrai/Faux : Une variable de durée peut prendre des valeurs négatives.

Q2 - Interprétation : Sur un graphe, une *densité* $f(t)$ “ plate ” (quasi constante) signifie :

- (A) qu'il y a très peu d'événements au début puis beaucoup à la fin ;
- (B) qu'en moyenne la *vitesse d'arrivée* des événements varie peu avec t ;
- (C) que $F(t)$ décroît de façon à peu près linéaire ;
- (D) que $S(t)$ est constant dans le temps.

Corrigé Q1 : Faux. Les durées sont ≥ 0 (souvent > 0).

Corrigé Q2 : (B) seulement.

(C) est faux : F *augmente* (ne décroît pas).

(D) est faux : S décroît (sauf cas trivial).

Quiz express - (2/3)

Q3 - Censure *vs* troncature : Choisir la (les) bonne(s) réponse(s).

- (A) Censure à droite : on sait que $T > t_{\text{fin}}$, mais pas la valeur exacte de T .
- (B) Troncature à gauche : on n'inclut que les individus avec $T > t_{\text{entrée}}$.
- (C) Censure à gauche : on sait seulement que $T < t_{\text{début}}$.
- (D) Troncature = censure systématique.

Quiz express - (2/3)

Q3 - Censure *vs* troncature : Choisir la (les) bonne(s) réponse(s).

- (A) Censure à droite : on sait que $T > t_{\text{fin}}$, mais pas la valeur exacte de T .
- (B) Troncature à gauche : on n'inclut que les individus avec $T > t_{\text{entrée}}$.
- (C) Censure à gauche : on sait seulement que $T < t_{\text{début}}$.
- (D) Troncature = censure systématique.

Indice : Censure = on garde l'observation (info partielle). Troncature = on n'entre jamais dans l'échantillon.

Quiz express - (2/3)

Q3 - Censure vs troncature : Choisir la (les) bonne(s) réponse(s).

- (A) Censure à droite : on sait que $T > t_{\text{fin}}$, mais pas la valeur exacte de T .
- (B) Troncature à gauche : on n'inclut que les individus avec $T > t_{\text{entrée}}$.
- (C) Censure à gauche : on sait seulement que $T < t_{\text{début}}$.
- (D) Troncature = censure systématique.

Corrigé Q3 : (A), (B), (C) vraies. (D) faux.

Pourquoi ? La troncature est un biais de **sélection à l'entrée**, pas une censure.

Quiz express - (3/3)

Q4 - Effet pratique : Si j'exclus les observations censurées à droite et je calcule la moyenne des durées restantes, le biais attendu est :

(A) vers le bas ; (B) vers le haut ; (C) nul ; (D) indéterminé.

Q5 - “ Densité > 1 ” ? : Peut-on avoir $f(t) > 1$? Si oui, pourquoi ce n'est pas un problème de probabilité ?

Q6 - Discret (mois) : Cocher la vraie relation.

(A) $h(t) = \frac{p(t)}{S(t-1)}$ (B) $S(t) = \prod_{k=1}^t (1 - h(k))$ (C) $p(t) = S(t-1)h(t)$ (D) $S(t) = 1 - \sum_{k=1}^t h(k)$

Quiz express - (3/3)

Q4 - Effet pratique : Si j'exclus les observations censurées à droite et je calcule la moyenne des durées restantes, le biais attendu est :

(A) vers le bas ; (B) vers le haut ; (C) nul ; (D) indéterminé.

Q5 - “ Densité > 1 ” ? : Peut-on avoir $f(t) > 1$? Si oui, pourquoi ce n'est pas un problème de probabilité ?

Q6 - Discret (mois) : Cocher la vraie relation.

(A) $h(t) = \frac{p(t)}{S(t-1)}$ (B) $S(t) = \prod_{k=1}^t (1 - h(k))$ (C) $p(t) = S(t-1)h(t)$ (D)
 $S(t) = 1 - \sum_{k=1}^t h(k)$

Indices : Q4 - qui retire-t-on quand on exclut les censurés ? Q5 - f est une *densité* (valeur par unité de temps). Q6 - pense “ conditionnel ” puis “ produit des survies ”.

Quiz express - (3/3)

Q4 - Effet pratique : Si j'exclus les observations censurées à droite et je calcule la moyenne des durées restantes, le biais attendu est :

(A) vers le bas ; (B) vers le haut ; (C) nul ; (D) indéterminé.

Q5 - “ Densité > 1 ” ? : Peut-on avoir $f(t) > 1$? Si oui, pourquoi ce n'est pas un problème de probabilité ?

Q6 - Discret (mois) : Cocher la vraie relation.

(A) $h(t) = \frac{p(t)}{S(t-1)}$ (B) $S(t) = \prod_{k=1}^t (1 - h(k))$ (C) $p(t) = S(t-1)h(t)$ (D)
 $S(t) = 1 - \sum_{k=1}^t h(k)$

Corrigés

Q4 : (A) biais *vers le bas* (on retire souvent des durées longues).

Q5 : Oui. f est une *densité* (1/temps), pas une proba ; la proba locale est $\Pr(t < T \leq t + \Delta t) \approx f(t)\Delta t \leq 1$.

Q6 : (A), (B), (C) vraies ; **(D)** faux.

$h(t) = \frac{p(t)}{S(t-1)}$, $S(t) = \prod_{k=1}^t (1 - h(k))$, $p(t) = S(t-1)h(t)$.

Définitions

Répartition et densité

Survie

Hasard

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

Idée clé

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

Définition compacte

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

Unités & lecture

⇒ Unités : (événements)/(individus à risque)/(unité de temps).

⇒ Lecture locale : $\Pr(t < T \leq t + \Delta t \mid T > t) \approx h(t) \Delta t$.

Mini-exemple (ordre de grandeur)

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

De $h(t)$ à $H(t)$ puis à $S(t)$

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

Ce que ça implique

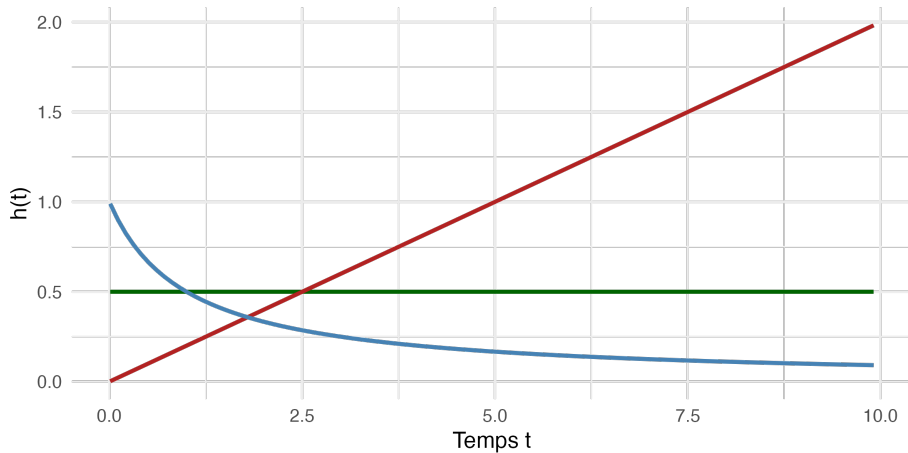
- ⇒ h constant $= \lambda \Rightarrow H(t) = \lambda t$, $S(t) = e^{-\lambda t}$ (exponentielle, *mémoire nulle*).
- ⇒ h croissant $\Rightarrow H$ accélère, S chute plus vite avec t (usure/vieillessement).
- ⇒ h décroissant $\Rightarrow$ sélection des “résistants”, queue plus lourde.

Estimation

Estimateur de Nelson-Aalen : $\hat{H}(t) = \sum_{t_j \leq t} \frac{d_j}{n_j}$, $\hat{S}(t) = \exp\{-\hat{H}(t)\}$.

Fonction de hasard : comparaison de formes

Type — Constante — Croissante — Décroissante



Hasard par morceaux (piecewise-constant)

Découper $0 = \tau_0 < \tau_1 < \dots < \tau_J$. Sur $[\tau_{j-1}, \tau_j) : h(t) = \lambda_j$.

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

Atouts :

- ⇒ Flexible pour capturer des changements de rythme (court/moyen/long terme).
- ⇒ Estimation simple, interprétation claire (taux par intervalle).
- ⇒ Pont naturel avec Poisson/binomial sur données agrégées.

Hasard en temps *discret* (mois/trimestres)

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

Modélisation **person-period** :

$$\Pr(T = t \mid T \geq t, X) = 1 - \exp\{-\exp(\alpha_t + X\beta)\} \quad (\text{cloglog}).$$

Intuition : la compl. log-log approxime un modèle à hasards proportionnels en discret.

Pratique : dummies de durée pour α_t ; censure à droite gérée naturellement.

Hasards proportionnels : intuition de l'effet covariable

Idée : une covariable X *multiplie* le hasard à tout t :

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

Lecture : un **hazard ratio** $e^\beta = 1.3$ signifie “+30% de risque *instantané* de sortie à tout t ”.

Important : la covariable *ne change pas la forme temporelle* de $h_0(t)$, seulement son échelle.

Durée-dépendance *vs* sélection des plus résistants (frailty)

Attention : un $h(t)$ **décroissant** peut venir de deux sources :

- ⇒ **Vraie durée-dépendance** : le risque baisse avec le temps pour un individu.
- ⇒ **Sélection** (frailty) : les individus à haut risque sortent tôt ; ceux qui restent sont plus “résistants”.

Conséquence : interpréter la forme de $h(t)$ *avec prudence* ; envisager des modèles à **frailty** si nécessaire.

Comparaison de formes de densité

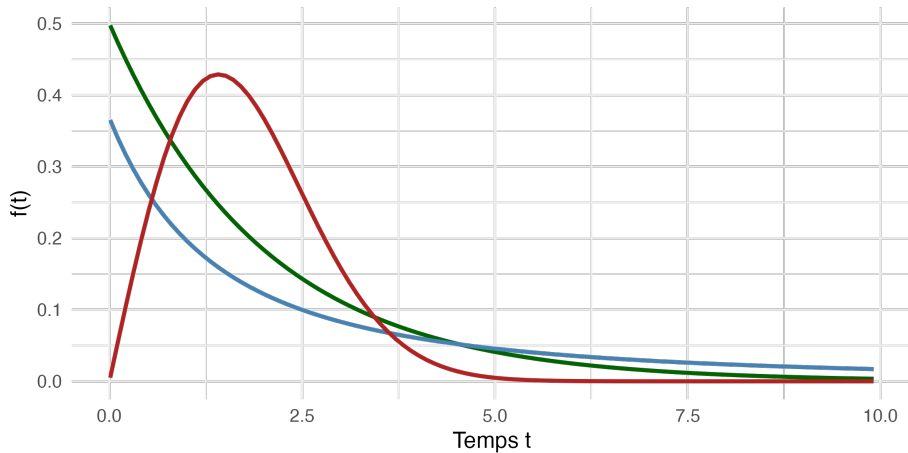
- ⇒ La densité $f(t)$ indique la **probabilité instantanée de sortie** à l'instant t .
- ⇒ Elle dépend de la forme de la distribution de durée sous-jacente :

Exemples usuels

- ⇒ **Exponentielle** : pic immédiat, décroissance rapide (durée "sans mémoire").
- ⇒ **Weibull (shape > 1)** : montée puis décroissance (vieillessement).
- ⇒ **Log-logistique** : longue traîne (certaines durées très longues).

Fonction de densité : comparaison de lois

Distribution — Exponentielle — Log.logistique — Weibull



Erreurs fréquentes & corrections

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

Ce que vous devez retenir aujourd'hui

- ⇒ Les **modèles de durée** permettent d'analyser le temps jusqu'à la survenue d'un événement (chômage, décès, faillite, etc.).
- ⇒ Les variables de durée ont des propriétés spécifiques :
 - elles sont strictement positives ;
 - elles peuvent être **censurées** (droite, gauche, intervalle).
- ⇒ Trois fonctions fondamentales décrivent les durées :
 - la fonction de densité $f(t)$,
 - la fonction de survie $S(t)$,
 - la fonction de hasard $h(t)$.
- ⇒ Ignorer la censure mène à des **biais de sélection**.

Une bonne modélisation = bien prendre en compte la censure et le comportement du risque dans le temps.

La prochaine fois...

- ⇒ Nous verrons notre premier modèle d'estimation : **Kaplan-Meier**.
 - Estimation non paramétrique de la fonction de survie.
 - Utilisable même en présence de censure à droite.
 - Comparaison de groupes : test de log-rank.
- ⇒ Introduction à l'analyse en présence de **covariables**.
 - Pourquoi on ne peut pas utiliser une régression linéaire ?
 - Limites de l'approche naïve.
- ⇒ Et bien sûr **du code R pour manipuler vos premières données de durée !**