MATH 60604
Modélisation statistique
§ 7a - Concepts d'analyse de survie

Léo Belzile

HEC Montréal
Département de sciences de la décision

Données de survie

- Dans le cadre de l'analyse de survie, on s'intéresse au temps jusqu'à ce qu'un événement survienne, une variable réponse positive.
- Soit T_i le temps de survie pour le sujet i ($i = 1, \ldots, n$).
- Le temps de survie avant que l'événement d'intérêt survienne.
  - Généralement continue, mais mesurée à des moments précis (données discrètes).
  - Le temps de survie est aussi appelé temps de défaillance.
- Les données de survie sont particulières et il est essentiel d'avoir une bonne compréhension des méchanismes générateurs des données afin d'effectuer une analyse appropriée.

Exemples de données de survie

Parmi les exemples,

- le temps avant le décès d'un patient à qui on a diagnostiqué un cancer
- le temps avant qu'un hôpital donne congé à un patient
- le temps d'attente avant qu'un client soit servi dans un restaurant
- le temps précédant l'annulation d'un abonnement à un gym
- le temps avant qu'une ampoule brûle

Censure

Une des plus grandes difficultés avec l'analyse de survie est qu'**on n'observe pas nécessairement tous les événements**. Cela pourrait être parce que

- le sujet survit après la fin de la période d'étude,
  - T représente le temps de survie d'un individu après avoir été diagnostiqué avec un cancer
    - il se peut que le sujet soit toujours vivant à la fin de l'étude.
- le sujet quitte l'étude avant la fin,
  - T représente le temps que ça prend à un étudiant pour compléter leur programme
    - il se peut que certains étudiants abandonnent leur programme.
- un événement concurrentse produit, ce qui rend l'événement d'intérêt impossible.
  - T représente le nombre d'années de service avant la retraite pour des employées d'un entreprise
    - il se peut qu'un sujet décède avant de prendre sa retraite.

Censure

- Il existe plusieurs types de censure
  - **censure à droite**: on sait que l'événement se produit après un certain temps t, c'est-à-dire $T_i \geq t$.
  - **censure à gauche**: l'événement d'intérêt survient avant qu'on observe l'individu. C'est-à-dire, tout ce qu'on sait est que $T < t$.
  - **censure par intervalle**: l'événement d'intérêt survient à un point inconnu dans un intervalle de temps, mais on ne sait pas quand exactement. C'est-à-dire, tout ce qu'on sait est que $T \in [t_1, t_2]$
- Avec la censure, l'idée générale est qu'on ne sait la valeur précise de T, mais on a tout de même de l'information concernant un intervalle dans laquelle T peut tomber.
  - par ex: $T > t$, ou $T < t$, ou $T \in [t_1, t_2]$

Examples de censure

- Supposons qu'un chercheur s'intéresse à l'âge à partir duquel les enfants sont capables d'écrire leur prénom.
- T est le temps (en années) jusqu'à ce qu'un enfant puisse écrire son nom.
- Le chercheur suit un groupe d'enfants dans une classe de maternelle tout au long de l'année scolaire.
  - Lorsque le chercheur arrive, certains enfants sont déjà capables d'écrire leur nom: T est **censuré à gauche**.
  - Certains enfants apprennent à écrire leur prénom pendant les vacances de Noël: T est **censuré par intervalle**.
  - Certains enfants ne savent toujours pas comment écrire leur nom rendu à la fin de l'année scolaire: T est **censuré à droite**.

Censure non-informative

- Nous supposons généralement que la censure est non-informative.
  - C'est-à-dire, la censure est non-informative de l'événement; le temps de censure est indépendant du temps de survie.
  - Autrement dit, le temps de censure ne nous donne aucune information sur ce que pourrait être le temps de survie.
- Un exemple de censure **informative**:
  - Supposons qu'un groupe de patients en phase terminale d'une maladie suit un traitement expérimental, et que ce traitement peut avoir des effets secondaires nocifs. Donc, pour des raisons éthiques, les patients qui deviennent très malades sont retirés de l'étude. Les patients qui sont retirés de l'étude auront un temps T qui est censuré à droite. Cependant, ces patients qui abandonnent l'étude sont probablement en moins bonne santé et risquent davantage de mourir plus tôt.

Type de censure à droite non-informative

On distingue entre plusieurs formes de censure à droite qui est non-informative:

- censure de type 1: la collecte de donnée prend fin au temps C; toute observation résiduelle est censurée à droite.
- censure de type 2: on collecte des données jusqu'à un nombre prédéterminé k d'événements.
- **censure aléatoire**: le temps de survie observé est $T_i = \min\{T_i^0, C_i\}$, où la durée de survie T_i^0 et le temps de censure C_i sont des variables aléatoires **indépendantes**.

Troncation

Dans certaines études, on collecte les données seulement pendant un créneau prédéterminé $[a, b]$.

- La trajectoire du temps est **tronquée à gauche** si le temps de survie excède zéro au temps a

Par exemple, lors d'une enquête sur le chômage, on considère toutes les personnes qui sont inscrites au chômage entre janvier et mars.

- Certaines personnes ont perdu leur emploi depuis plusieurs mois lors du début de l'enquête (troncation à gauche).
- Si la personne est toujours en recherche d'emploi à b, elle sera censurée à droite (censure à droite de type 1).

Diagramme de Lexis

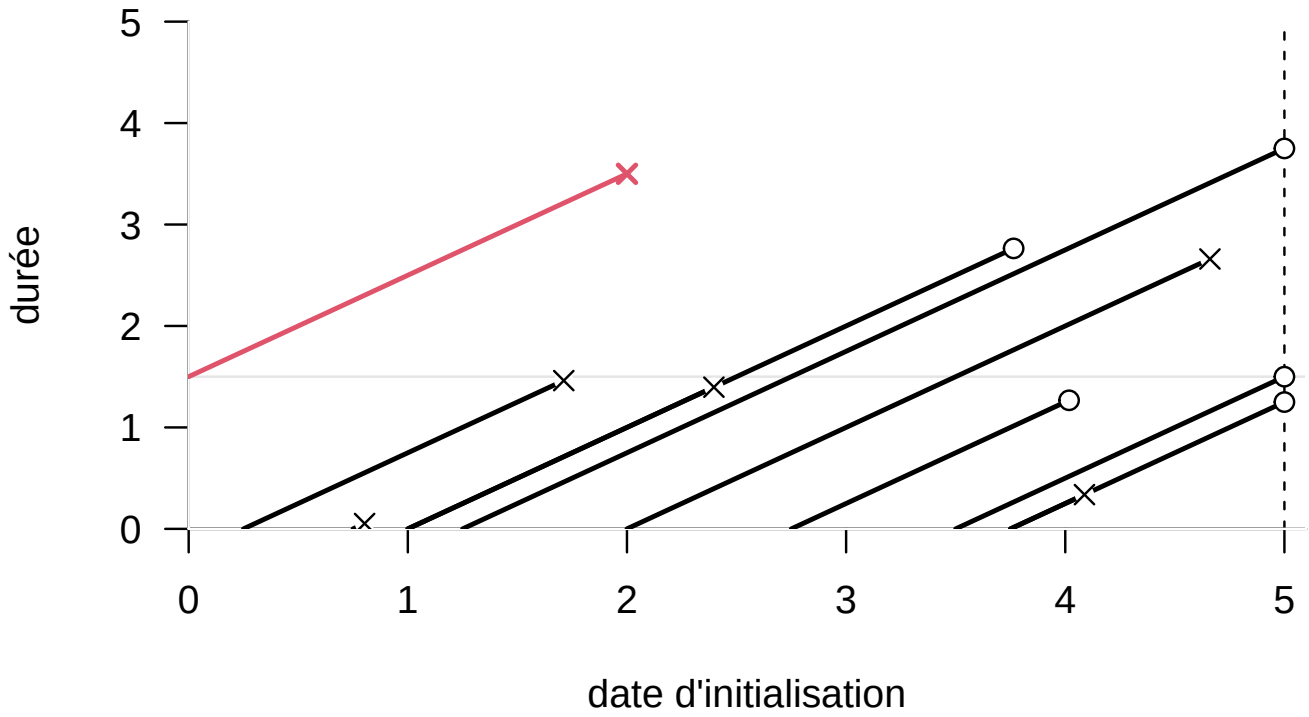

Diagramme de Lexis représentant les trajectoires temporelles. Les x dénotent les temps de défaillance observés, tandis que les ○ indiquent les valeurs censurées. Les temps de survie résiduels au temps 5 sont censurés. La trajectoire en rouge dénote un individu dont le temps de survie est tronqué à gauche.