Analyse de Survie

Professeur Abdellatif El Afia

Motivation

Analyse de survie: (X,E,T)

L'analyse de survie est une branche des statistique qui cherche à modéliser le temps **T** jusqu'à l'occurrence d'un événement précis **E**, ayant données des variables explicatives **X**.

Le temps T : est une variable aléatoire positive mesurant le temps t depuis une exposition jusqu'à l'événement **E**.

L'événement E : est le passage irréversible entre deux états. Il peut être la mort d'un sujet **X** (Biostatistiques), la défaillance d'une pièce (Industrie), trouver un emploi pour un jeune diplômé (Economie), divorce (Social)…

Les variables explicatives X : sont des vecteurs de données caractérisant chaque individu ayant ou pas un impact sur la survie.

Exemple 1

Une étude s'intéressant au temps de survie des patients après une transplantation d'un organe.

Exposition :

Evénement :

T :

Variables X :

Exemple 1

Une étude s'intéressant au temps de survie des patients après une transplantation d'un organe.
Exposition : la transplantation
Evénement : la mort du patient
T : le temps entre la transplantation et la mort du patient
Variables X : l'âge, le groupe sanguin, antécédant médicaux, ….

Exemple 2

Une étude de performance de pièces industrielles pour prévoir un entretient ou un remplacement.

Exposition :

Evénement :

T :

Variables X :

Exemple 2

Une étude de performance de pièces industrielles pour prévoir un entretient ou un remplacement.
Exposition : Mise en service.
Evénement : défaillance.
T : temps entre la mise en service et la défaillance.
Variables X : température, pression, matière d'origine, marque….

Exemple 3

Une étude sociale concernant le divorce.
Exposition :
Evénement :
T :
Variables X :

Exemple 3

Une étude sociale concernant le divorce.
Exposition : Mariage.
Evénement : Divorce.
T : temps entre les deux.
Variables X : type de mariage, nombre d'enfant, travail, grand parents….

Exemple 4

Une étude de temps de chaumage de jeunes diplômés.
Exposition :
Evénement :
T :
Variables X :

Exemple 4

Une étude de temps de chaumage de jeunes diplômés.
Exposition : obtention du diplôme.
Evénement : trouver un travail.
T : temps de chaumage.
Variables X : âge, spécialité, stage, …

Exemple 5

Le temps de résiliation d'un abonnement.
Exposition :
Evénement :
T :
Variables X :

Exemple 5

Le temps de résiliation d'un abonnement.
Exposition : le début de l'abonnement.
Evénement : la résiliation.
T : le temps d'abonnement.
Variables X : revenu, prix, concurrence….

Exemple 6

Temps d'une publicité.
Exposition :
Evénement :
T :
Variables X :

Exemple 6

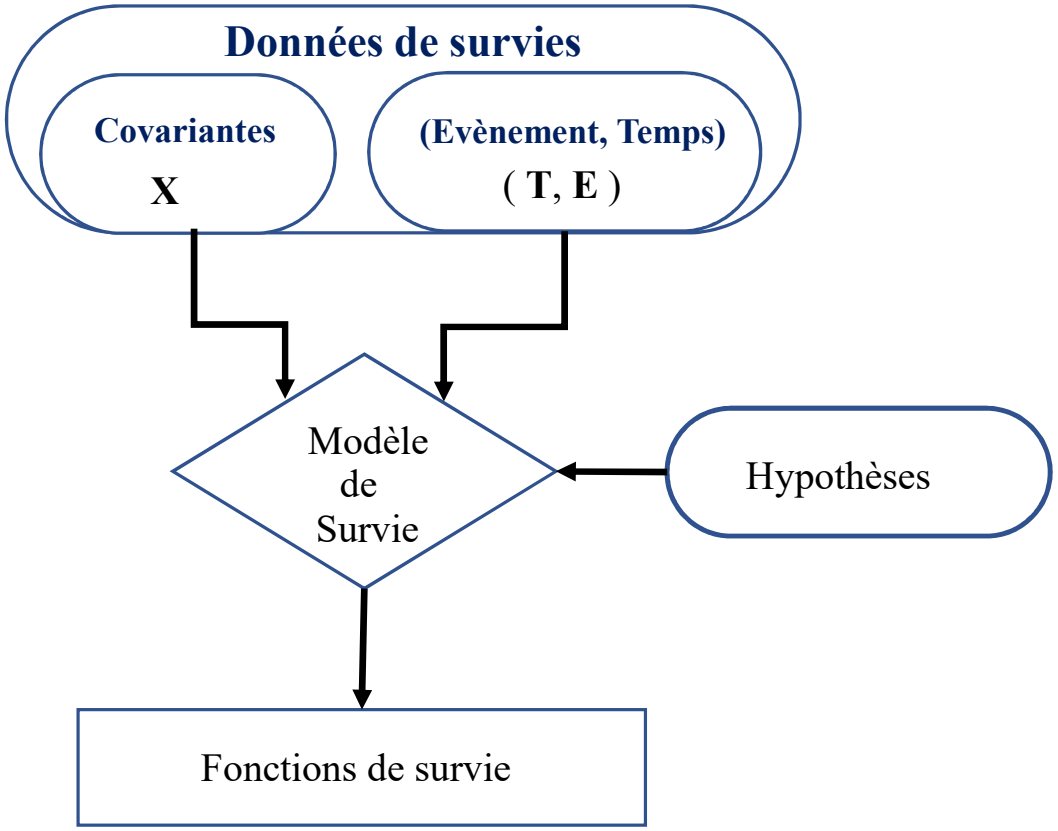
Temps d'une publicité.
Exposition : début d'une publicité.
Evénement : skip.
T : temps entre les deux.
Variables X : historique, informations personnelles Ads ID, sujet de la publicité…

Objectif

- **L'objectif** est d'utiliser les méthodes d'analyse de survie afin de :

  - Estimé la distribution de temps de survie (Fonctions de survie S(t), fonction de densité $\boldsymbol{f(t)}$, fonction de répartition $\boldsymbol{F(t)}$, fonction de risque $\boldsymbol{\lambda(t)}$, fonction de risque cumulé $\boldsymbol{\Lambda(t)}$).
  - Comparer la fonction de survie de deux ou plusieurs groupes.
  - Analyser la manière dont des variables explicatives **X** influencent les fonctions de survie.

Modèle de survie



Fonction de survie :

$S(t) = P(T \geq t) = 1 - F(t),\ t > 0$

Fonction de densité :

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

Fonction de risque :

$$\lambda(t)\Delta t \approx P(t \leq T < t + \Delta t | T \geq t)$$

Fonction de risque Cumulative :

$$\Lambda(t) = \int_0^t \lambda(x)dx\ ,\ t \geq 0$$

Données de survie

La censure : La durée T est dite censurée si la durée n'est pas intégralement observée.

La troncature : a lieu si l'observation de la variable d'intérêt T n'a lieu que conditionnellement à un événement B.

Données manquantes : certaines valeurs de données sont manquantes soit pour les variables explicatives X ou le statut de l'événement E ou encore le temps T

La censure

- Pour chaque individu $i \in 1, \ldots, n$, on note :
  - T_i^* le temps de survie (pas toujours observé).
  - C_i le temps de censure.
  - δ_i l'indicateur de censure (1 observé, 0 censuré).

- En pratique, on observe $T_i = \min(T_i^*, C_i)$ et δ_i.
- Les observations de survie sont donc : $(T_1, \delta_1), (T_2, \delta_2), \ldots, (T_n, \delta_n)$

$$(T_i, \delta_i) = \begin{cases} (T_i, 1) & si \quad C_i \geq T_i^* \ (non\ censuré) \\ (T_i, 0) & si \quad C_i < T_i^* \ (censuré) \end{cases}$$

La censure

- La censure est dite indépendante si elle n'apporte pas d'information sur la durée de survie.
- Du fait de la censure, on ne peut pas utiliser les méthodes statistiques classiques (t-test, régression linéaire).
- On ne peut même pas calculer de moyenne.
- Les différents types de censure :

  censure de type I : fixée

  censure de type II : attente

  censure de type III : aléatoire

La censure à droite :

T correspond à l'Age en années de l'individu au moment du décès, l'événement est censuré à droite à l'Age 60 si tout ce que nous pouvons savoir est que T>60.

Censure à gauche :

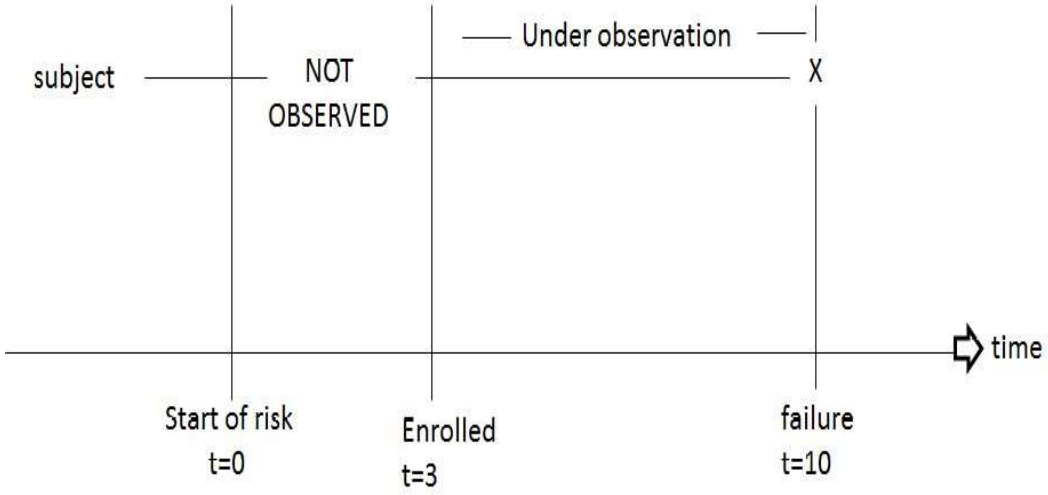
une étude de ménarche (le début de la menstruation), si nous prenons un échantillant de filles de 12 ans et que certaines ont déjà commencé à avoir leurs règles sans pouvoir en déterminer le temps de la première apparition, l'Age de la ménarche chez elles est censuré à 12 ans.

Censure bornée :

un échantillant d'individus suivis pour infection de AIDS, des analyses sont réalisées chaque trimestre depuis l'admission. Si un individu s'avère séro-positif lors du troisième test, son temps à événement est censuré par intervalle entre la valeur 3 et 6 (en mois).

Troncature

Un intervalle de temps durant lequel le sujet n'a pas été suivit mais n'a pas eu l'événement non plus est appelé une troncature.



Troncature à gauche :

Seules les personnes qui survivent au stade initial de l'infarctus du myocarde et arrivent à l'hôpital seront incluses dans l'étude. Si un individu a été admis à l'hôpital il est ajouté à l'étude avec comme temps d'origine le temps de l'infarctus. Pour différents patients, cela peut se produire à des moments différents, mais ces patients ne seront jamais admis à l'étude s'ils meurent avant d'arriver à l'hôpital.

Troncature à droite :

La troncature à droite se produit quand l'information concernant un sujet n'est obtenue que lorsqu'il vit l'événement. Les sujets qui survivent après la fin de l'étude ne sont pas diagnostiqués et ne sont donc pas inclus dans l'échantillon de l'étude, ce qui donne un échantillon biaisé en faveur des sujets avec des temps de survie plus courts.

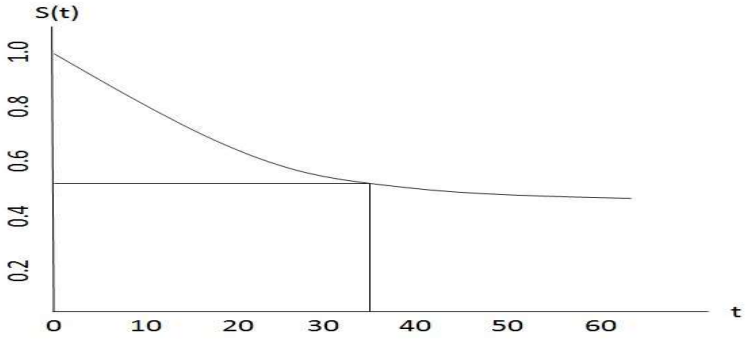
Troncature d'intervalle :

La troncature d'intervalle n'est qu'une adoption de la troncature à gauche où un individu entre dans l'étude au temps zéro mais disparaît pendant un certain temps puis revient à l'étude en générant un écart entre les observations. Le problème est que cette personne aurait pu mourir lorsqu'elle a disparu.

Fonctions de survie

Fonction de survie :

$$S(t) = P(T \geq t) = 1 - F(t),\ t > 0$$



Fonction de répartition :

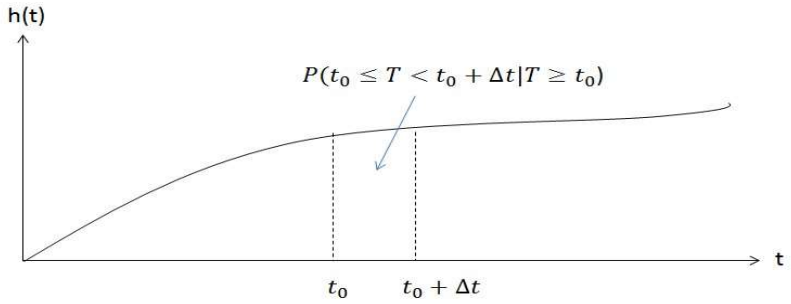
$$F(t) = 1 - S(t)$$

Fonction de densité :

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

Fonction de risque :

$$\boldsymbol{\lambda(t)} = h(x) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \,/\, T \geq t)}{\Delta t}$$



Fonction de risque Cumulative :

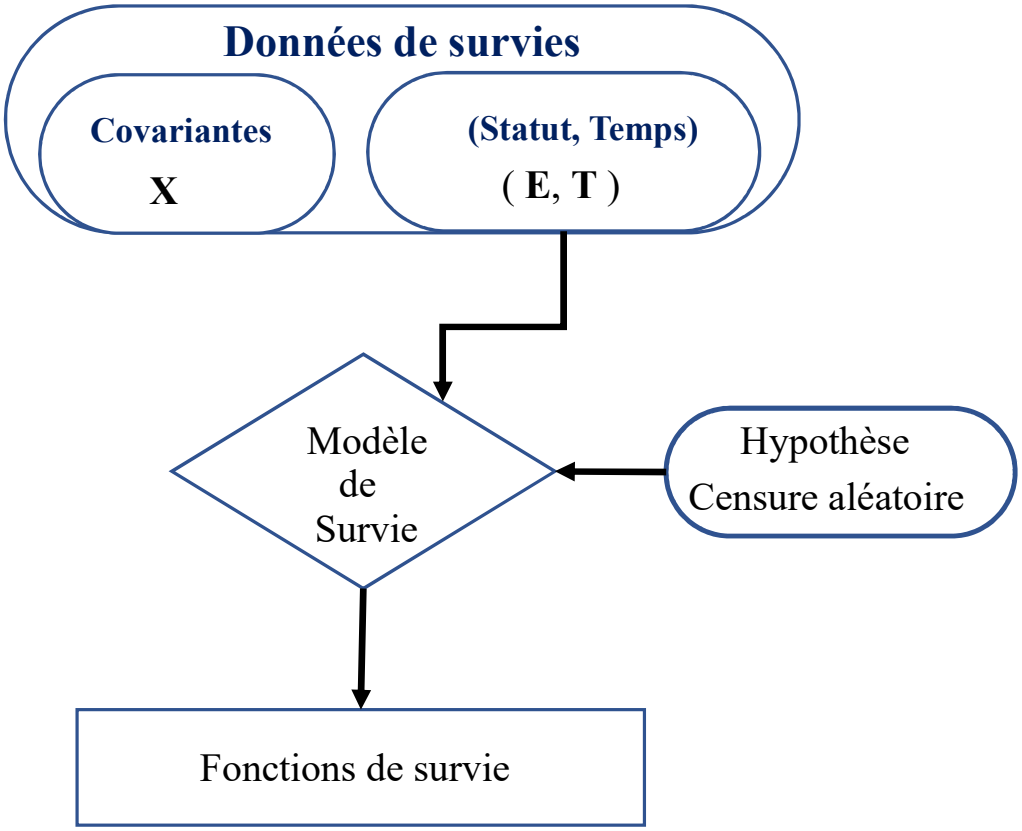
$$\boldsymbol{H(t)} = \int_0^t h(x)dx \,, t \geq 0$$

Propriétés :

- $\boldsymbol{\lambda}(\boldsymbol{t}) = H'(t) = \frac{-S\prime(t)}{S(t)} = \frac{f(t)}{S(t)}$
- $H(t) = -\ln\big(S(t)\big)$
- $S(t) = e^{-H(t)} = e^{-\int_0^t \boldsymbol{\lambda}(\boldsymbol{s})\boldsymbol{ds}}$

- Ces 5 fonctions caractérisent la loi de T.
- Elles sont inconnues.
- On va chercher à les estimer à partir des observations (X_i, δ_i).

Modèles Non Paramétriques



- **Estimateur de Kaplan Meier (KM) :**

$$\hat{S}(t_k) = \prod_{t_k<t} S(t_{k-1})\left(1 - \frac{d_k}{n_k}\right) \qquad 1 < k < j$$

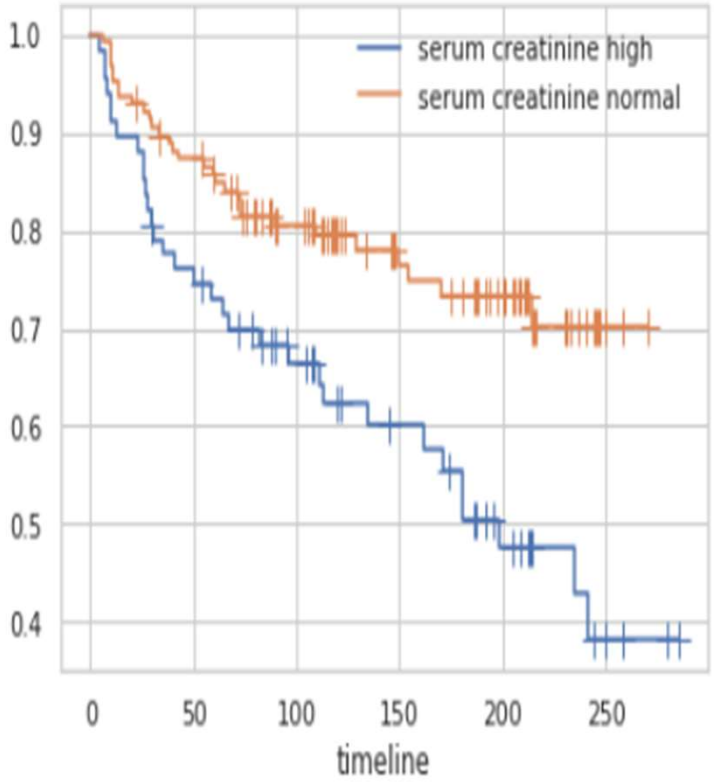
- **Estimateur de Nelson–Aalen :**

$$\tilde{H}(t) = \sum_{t_i<t} \frac{d_i}{n_i}$$

- **Tables de survie.**

Kaplan Meier

- La méthode de **Kaplan-Meier** permet d'obtenir rapidement une courbe de survie sans nécessiter que les intervalles de temps soient réguliers, contrairement à la méthode des tables actuarielles de survie (**Life tables**)
- Simple à interpréter et permet d'estimer la fonction de survie.
- Ne fourni pas de formule-Fonction et ne permet pas d'estimer le « rapport de risque » (Hazard ratio)
- Ne prend que quelques variables catégoriques X



KM formule de produit :

$$\hat{S}(t_k) = \hat{S}(t_{k-1}) \times \hat{P}(T > t_k \mid T \geq t_k)$$

$$\hat{S}(t_k) = \prod_{i=1}^{k} \hat{P}(T > t_i \mid T \geq t_i)$$

$$P(A \cap B) = P(A) \times P(B \mid A)$$

$$\hat{S}(t) = \prod_{t_i < t} (1 - \frac{d_i}{r_i})$$

d_i le nombre d'événement à l'instant t_i.
r_i le nombre d'individus à risque à l'instant t_i.

Log Rank test

- Un test chi-deux χ^2 utilisé pour tester l'hypothèse nulle:

  H_0 : "$Il\ n'y\ a\ pas\ de\ difference\ statistique\ entre\ les\ deux\ groupe$"
- Utilise la technique d'observés **O** Vs attendues **E** sur les catégories, ces dernières étant définies par chacun des évènements ordonnés pour l'ensemble des données analysées (les deux groupes).

$$Log\text{-}rank\ statistic = \frac{(O^g - E^g)^2}{Var(O^g - E^g)}$$

- Log-rank statistic est approximativement du khi-deux χ^2 avec un degré de liberté.
- Approximation : $\chi^2 = \sum_g \frac{(O^g - E^g)^2}{E^g}$
- La p-value de Log-rank détermine si H_0 est rejetée ou non.

Références

- Kartsonaki, C. (2016). Survival analysis. Diagnostic Histopathology, 22(7), 263-270
- Lee, Elisa T., and John Wang. Statistical methods for survival data analysis. Vol. 476. John Wiley & Sons, 2003
- survival-analysis-self-learning-book third edition