

# Analyse de Survie

Thibault ALLART

CNAM

2017

- 1 Généralités sur l'analyse de survie
  - Introduction
  - Définitions
  - Censure
  - Propriétés des fonctions de survie
  
- 2 Estimation et comparaison des courbes de survie
  - L'estimateur de Kaplan-Meier
  - L'estimateur de Nelson-Aalen

On s'intéresse au temps de réalisation d'un évènement

- Durée de vie d'un patient atteint d'un cancer.
- Durée d'un composant électronique avant une panne.
- Durée d'utilisation d'un service par des clients.

L'évènement d'intérêt peut être censuré.

- Mortalité : Quelle est l'espérance de vie d'une population ?
- Durée de vie d'un appareil électronique
- Peut on démontrer l'efficacité d'un traitement médical ?
- Comment améliorer la durée d'abonnement de mes clients ?

# Les deux méthodes Statistique les plus cités

## Nonparametric estimation from incomplete observations

EL Kaplan, P Meier - Journal of the American statistical association, 1958 - Taylor & Francis

Abstract In lifetesting, medical follow-up, and other fields the observation of the time of occurrence of the event of interest (called a death) may be prevented for some of the items of the sample by the previous occurrence of some other event (called a loss). Losses may be

Cité 48411 fois [Autres articles](#) [Web of Science: 42852](#) [Citer](#) [Enregistrer](#) [Plus](#)

## Regression Models and Life-Tables - JStor

<https://www.jstor.org/stable/2985181> - Traduire cette page

de DR Cox - 1972 - Cité 42495 fois [Autres articles](#)

1972] 187. Regression Models and Life-Tables. BY D. R. Cox. Imperial College, London. [Read before the ROYAL STATISTICAL SOCIETY, at a meeting ...

On distingue l'évènement d'intérêt

- Décès du patient après l'apparition du cancer
- Panne du composant
- Fin de contrat du client

de la variable à expliquer durée avant l'apparition de l'évènement

- temps écoulé avant le décès
- temps écoulé avant la panne
- temps écoulé avant la fin du contrat

## Définition : Durée de Vie ou Survie

La durée de survie désigne le temps qui s'écoule depuis un instant initial(début du traitement, diagnostic, ...), jusqu'à la survenue d'un évènement d'intérêt final (décès du patient, rechute, rémission, guérison, ...).

La variable étudiée est appelée **durée de vie  $T$** .

On dit que le patient survit au temps  $t$  si, à cet instant, l'évènement d'intérêt final n'a pas encore eu lieu.

$T$  est une variable aléatoire positive continue.

Supposons que l'étude soit un essai clinique portant sur deux groupes de patients, recevant 2 types de traitements. Deux questions importantes se posent aux médecins :

- ❶ L'un des deux traitements est-il plus efficace que l'autre en terme d'amélioration de la survie des patients ?
- ❷ Peut-on mettre en évidence des **facteurs pronostiques** qui améliorent/détériorient la survie ?

Exemple : âge, sexe, tabagisme, antécédents familiaux, ...



Pour répondre à ces questions, on peut :

- ❶ Mettre en place des méthodes statistiques qui vont permettre de comparer les deux groupes de patients qui reçoivent les deux types de traitement.
- ❷ Proposer un modèle qui relie la durée de survie des patients à des variables explicatives et mettre en évidence des facteurs pronostiques.

La durée de survie n'est pas toujours complètement observée. Pour certains individus, l'évènement d'intérêt n'est pas observé.

## Definition : Censure

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

Exemples :

- Un patient peut être perdu de vue (déménagement, ...)
- un évènement peut survenir et entraîner la sortie de l'étude : le patient peut décéder d'une autre cause que de la maladie étudiées.
- L'étude s'arrête alors que des individus sont encore vivants

# Les différents types de 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 :

- 1 censure de type I : fixée
- 2 censure de type II : attente
- 3 censure de type III : aléatoire

Pour chaque individu  $i \in 1, \dots, n$ , on note :

- $T_i^*$  le temps de survie (pas toujours observé)
- $C_i$  le temps de censure
- $\delta_i$  l'indicateur de censure (1 observé, 0 censuré)

En pratique on observe  $T_i = \min(T_i^*, C_i)$  et  $\delta_i$

En survie, les observations sont donc  $(T_1, \delta_1), \dots, (T_n, \delta_n)$

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

La censure de type I correspond à une censure à date fixe.

$$C_i = C \quad \forall i \in 1, \dots, n$$

Exemple :

- L'étude prend fin le 10 mars.
- On dispose des données clients jusqu'à aujourd'hui.

Voir dessin au tableau

Dans le cas de la censure de type II, on observe les individus jusqu'à ce que  $r$  d'entre eux aient vu l'évènement d'intérêt se produire.

On observe  $T_1 < T_2 < \dots < T_r$

Les temps de censure  $C_i$  sont aléatoires et indépendants des temps d'évènement  $T_i$ .

Voir dessin au tableau

Dans la suite du cours, on considère que la censure est aléatoire.

De plus on suppose que la censure est indépendante du temps de survie (censure non informative). Mais si la censure est due à l'arrêt du traitement, l'hypothèse d'indépendance n'est pas valide.



Freireich (1963) à observé la durée de rémission (en semaines) de patients atteints de leucémie aiguë, traités soit par placebo soit par 6-mercaptopurine (6-MP)

6-MP	6	6	6	6 <sup>+</sup>	7	9 <sup>+</sup>	10	10 <sup>+</sup>	11 <sup>+</sup>	13		
	16	17 <sup>+</sup>	19 <sup>+</sup>	20 <sup>+</sup>	22	23	25 <sup>+</sup>	32 <sup>+</sup>				
	32 <sup>+</sup>	34 <sup>+</sup>	35 <sup>+</sup>									
Placebo	1	1	2	2	3	4	4	5	5	8	8	8
	8	11	11	12	12	15	17	22	23			

Le

signe + correspond à des patients qui ont quitté l'étude à la date considérée. Pour ces individus, les durées sont donc censurées.

Les données censurées demandent un traitement particulier. Si on enlève les données censurées → perte d'information.

Dans l'exemple précédent, si on enlève les données censurées, on ne tient pas compte des durées de rémission les plus longues et on sous-évalue l'effet du traitement 6-MP.

## 5 fonctions caractéristique

La loi du temps de survie  $T$  est décrite par 5 fonctions :

- La fonction de survie  $S(t)$
- La fonction de répartition  $F(t)$
- La densité  $f(t)$
- Le taux de risque instantané  $\lambda(t)$
- Le taux de risque cumulé  $H(t)$

A partir de l'une d'elles, on peut déduire toute les autres.

La fonction de survie  $S(t)$  représente la probabilité de survivre au moins jusqu'au temps  $t$ .

$$S(t) = \mathbb{P}(T \geq t)$$

## propriétés

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

La fonction de répartition désigne la probabilité que l'évènement d'intérêt ait lieu avant  $t$

$$F(t) = \mathbb{P}(T \leq t)$$

## propriétés

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

La densité désigne la probabilité que l'évènement d'intérêt ait lieu après  $t$ , dans un petit intervalle de temps.

$$f(t) = \lim_{dt \rightarrow 0} \frac{\mathbb{P}(t < T \leq t + dt)}{dt}$$

## propriétés

- $F(t) = \int_0^t f(s)ds$
- $f(t) = F'(t)$

Le taux de risque instantané  $h(t)$ , aussi noté  $\lambda(t)$ , est la probabilité qu'un évènement survienne dans un petit intervalle de temps après  $t$ , sachant qu'il n'a pas eu lieu avant  $t$ .

$$h(t) = \lim_{dt \rightarrow 0} \frac{\mathbb{P}(t \leq T < t + dt | T \geq t)}{dt}$$

Remarques :

$h(t)$  est lié à une unité de temps. Si  $t$  est en heure, alors  $h(t)$  mesure le risque qu'un évènement survienne dans l'heure. Ce n'est pas une densité donc son intégrale ne vaut pas nécessairement 1.

Le taux de risque cumulé est défini par :

$$H(t) = \int_0^t h(s) ds$$



## Propriétés

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

Ces 5 fonctions caractérisent la loi de T.

Elles sont inconnues.

On va chercher à les estimer à partir des observations  $(X_i, \delta_i)$ .

Pour répondre à la question 1 (**comparaison de traitements**) : on va travailler avec  $S(t)$ , que l'on cherchera à estimer.

Pour répondre à la question 2 (**facteurs pronostiques**) : on va travailler avec le taux de risque instantané  $\lambda(t)$

# Exercice : loi exponentielle

Calculez les 5 fonctions associés à la loi exponentielle, sachant que le taux de risque instantané est constant :  $\lambda(t) = \lambda$

## Loi exponentielle

- $S(t) = e^{-\lambda t}$
- $F(t) = 1 - e^{-\lambda t}$
- $f(t) = \lambda e^{-\lambda t}$
- $\lambda(t) = \lambda$
- $H(t) = \lambda t$

Nous verrons d'autres lois (Weibull, log-Normale) dans le chapitre sur les modèles paramétriques.

La moyenne :

$$\mathbb{E}(T) = \int_0^{\infty} S(t)dt$$

Le moment d'ordre 2 :

$$\mathbb{E}(T^2) = 2 \int_0^{\infty} tS(t)dt$$

La variance :

$$\text{Var}(T) = 2 \int_0^{\infty} tS(t)dt - \left( \int_0^{\infty} S(t)dt \right)^2$$

En pratique, on peut rarement estimer la durée moyenne de survie.

$$\mathbb{E}(t) = \int_0^{\infty} S(u) du$$

On considère alors la moyenne restreinte, i.e le temps moyen de survie sur un intervalle de temps  $[0, t]$

$$\mu(t) = \int_0^t S(u) du$$

Que l'on peut estimer par

$$\hat{\mu}(t) = \int_0^t \hat{S}(u) du$$

La fonction quantile de la durée de survie est définie par

$$Q(p) = \inf\{t : 1 - S(t) \geq p\} = \inf\{t : F(t) \geq p\}, \quad 0 < p < 1$$

Si  $F$  est strictement croissante et continue, alors :

$$Q(p) = F^{-1}(p) = S^{-1}(1 - p), \quad 0 < p < 1$$

A partir des données de l'INSEE

<http://www.ined.fr/fr/tout-savoir-population/chiffres/france/mortalite-cause-deces/table-mortalite/>

- Affichez la courbe de survie
- Quelle est la probabilité de survivre au moins jusqu'à  $t$  ?
- Calculez le temps de vie moyen et médian.
- Quelle est la probabilité de survivre au moins jusqu'à  $t_2$ , sachant que l'on a survécu jusqu'à  $t_1$  ?
- Quelle est la probabilité de mourir aujourd'hui ?  
Calculer  $\lambda(t)$