Analyse de survie : Chapitre 1/2

M. Clertant [1]

Statistiques biomédicales

1. clertant@math.univ-paris13.fr, Lab. Analyse, Géométrie et Applications

Plan

Qu'est-ce que c'est ?

Analyse de survie : Étude de la durée de survie, ou durée avant la survenue d'un événement (time-to-event).

Passage irréversible entre deux états ; terminologie courante : vivant/mort

Événements possibles :

- décés, survenue d'une maladie ou d'une guérison (médecine),
- panne d'une machine (fiabilité),
- survenue d'un sinistre (actuariat),
- faillite d'entreprise (économie) ...

Domaine biomédicale : étude de cohortes, essais thérapeutiques, données longitudinales.

De quoi parle-t-on ?

Dans une étude de survie, on parle de :

Date d'origine : date d'entrée dans l'étude de l'individu (survenue de la maladie, date d'un traitement, mise en service d'un appareil ...)

Date de point : date de l'arrêt de l'étude, les informations sur les individus inclus à l'étude ne sont plus prises en compte à partir de cette date.

date des dernières nouvelles : date des plus récentes informations sur un individu (étude séquentielle).

D'un point de vue statistique :

- Estimation de la distribution de la durée de survie,
- Étude de cofacteurs influençant cette distribution (sous-population, test, régression).

Distribution de la durée de survie (v.a. continue)

Soit X la v.a. durée de survie ($X > 0$).

La distribution de X est caractérisée par 5 fonctions se déduisant l'une de l'autre :

1. Fonction de répartition (c.d.f.)

$$F(t) = \mathbb{P}(X \leq t), t \geq 0.$$

2. Fonction de survie (survival function)

$$S(t) = \mathbb{P}(X > t) = 1 - F(t).$$

Temps continu : $S(t) = 1 - F(t) = \mathbb{P}(X > t) = \mathbb{P}(X \geq t)$, $(\mathbb{P}(X = t) = 0)$.

Temps discret (heure, jour, mois, année) :
$1 - S^{-}(t) = F^{-}(t) = \mathbb{P}(X < t)$ et $1 - S^{+}(t) = F^{+}(t) = \mathbb{P}(X \leq t)$

Distribution de la durée de survie (v.a. continue)

3. Densité de probabilité f

Soit f, telle que $f(t) \geq 0$ $(\forall t \geq 0)$ et

$$F(t) = \int_0^t f(u)du \quad \text{et} \quad S(t) = \int_t^{+\infty} f(u)du$$

Si F (et S) admet une dérivée :

$$f(t) = \lim_{h \to 0} \frac{\mathbb{P}(t \leq X \leq t+h)}{h} = F'(t) = -S'(t).$$

$f(t)$ représente la probabilité de mourir dans un petit intervalle de temps après t.

4. Risque instantané λ

La probabilité de mourir dans un petit intervalle de temps après t sachant que l'on a survécu jusqu'au temps t.

$$\lambda(t) = \lim_{h \to 0} \frac{\mathbb{P}(t \leq X \leq t+h \mid X \geq t)}{h} = \frac{f(t)}{S(t)} = -[\log(S(t)]'$$

5. Taux de hasard cumulé Λ

$$\Lambda(t) = \int_0^t \lambda(u)du = -\log(S(t)).$$

On en déduit que :

$$S(t) = \exp(-\Lambda(t)) = \exp(-\int_0^t \lambda(u)du),$$

et

$$f(t) = \lambda(t)\exp(-\int_0^t \lambda(u)du)$$

Exemples paramétriques

Loi exponentielle : $X \sim \mathcal{E}(\lambda)$, avec $\lambda > 0$, et la densité f est :

$$f(t) = \lambda \exp(-\lambda t) \ , \ \ t \in \mathbb{R}^+.$$

Loi de Weibull : $X \sim \mathcal{W}(\lambda, \alpha)$, avec $\lambda > 0$, et la densité f est :

$$f(t) = \alpha \lambda^\alpha t^{\alpha-1} \exp(-(\lambda t)^\alpha) \ , \ \ t \in \mathbb{R}^+.$$

Loi gamma : $X \sim \Gamma(k, \theta)$, avec $k, \ \theta > 0$, et la densité f est :

$$f(t) = \frac{t^{k-1} \exp(-t/\theta)}{\Gamma(k)\theta^k} \ , \ \ t \in \mathbb{R}^+,$$

où Γ est la fonction gamma d'Euler.

Exercice : On suppose que X suit la loi de Weibull $\mathcal{W}(\lambda, \alpha)$.
1. Calculer :

- la fonction de survie,
- la fonction de répartition,
- le risque instantané,
- le taux de hasard cumulé.

2. Soit la durée de survie Y dont le risque instantané est :

$$\lambda(t) = \exp(a + bt).$$

Ceci est le modèle de Gompertz-Makeham (très bon ajustement pour la mortalité des adultes dans les pays développées).
Calculer la fonction de survie et la densité de ce modèle.

Exercice : Montrer que :

$$\mathbb{E}(X) = \int_0^{+\infty} S(t)dt,$$

et

$$Var(X) = 2\int_0^{+\infty} t \times S(t)dt - \left(\mathbb{E}(X)\right)^2$$

Distribution de la durée de survie (v.a. discrète)

X suit une distribution discrète sur l'ens. ordonné $\{t_1, t_2, \ldots\}$:

$$\mathbb{P}(X = t_i) = p_i$$

Fonction de survie et fonction de répartition

Pour $t \in \mathbb{R}^+$,

$$S(t) = \sum_{i:t_i > t} p_i \quad \text{et} \quad F(t) = \sum_{i:t_i \leq t} p_i$$

Risque instantané λ (aussi appelé taux de hasard)

Pour $i \in \mathbb{N}^\star$,

$$\lambda_i = \lambda(t_i) = \lim_{h \to 0} \mathbb{P}(t_i \leq X \leq t_i + h \mid X \geq t_i) = \frac{p_i}{S(t_{i-1})} = \frac{p_i}{\sum_{k:t_k \geq t_i} p_k}$$

Taux de hasard cumulé Λ

Pour $i \in \mathbb{N}^\star$,

$$\Lambda(t) = \sum_{i:t_i \leq t} \lambda_i$$

Exercice : Dans le cas discret, montrer que, pour $t \in [t_i, t_{i+1}[$, on a :

$$S(t) = \prod_{k=1}^{i} (1 - \lambda_k).$$

Exercice : On suppose que X est une v.a. discrète sur $\{t_1, t_2, \ldots\}$ telle que :

$$\mathbb{P}(X = t_k) = p_k = p(1-p)^{k-1},$$

pour $p \in [0, 1]$ (loi géométrique). Calculer la fonction de survie, le taux de hasard (risque instantané), le taux de hasard cumulé.

-

Plan

Censure

Temps de survie (survival time or time-to-event)

Il s'agit du temps écoulé entre un point d'entrée (début d'un certain état) et la survenue d'un événement (point de sortie).

Censure (censoring)

On parle de censure lorsque le point d'entrée et/ou l'événement n'est pas observé.

Exemples :

- point d'entrée : infection par une maladie ; événement : hospitalisation ;
- point d'entrée : traitement du patient ; événement : rémission (ou rechute, selon les études) ;
- point d'entrée : perte d'emploi ; événement : reprise d'une activité rémunératrice.

Censure à droite

Elle consiste à ne pas observer l'événement (point de sortie) pour certains sujets.

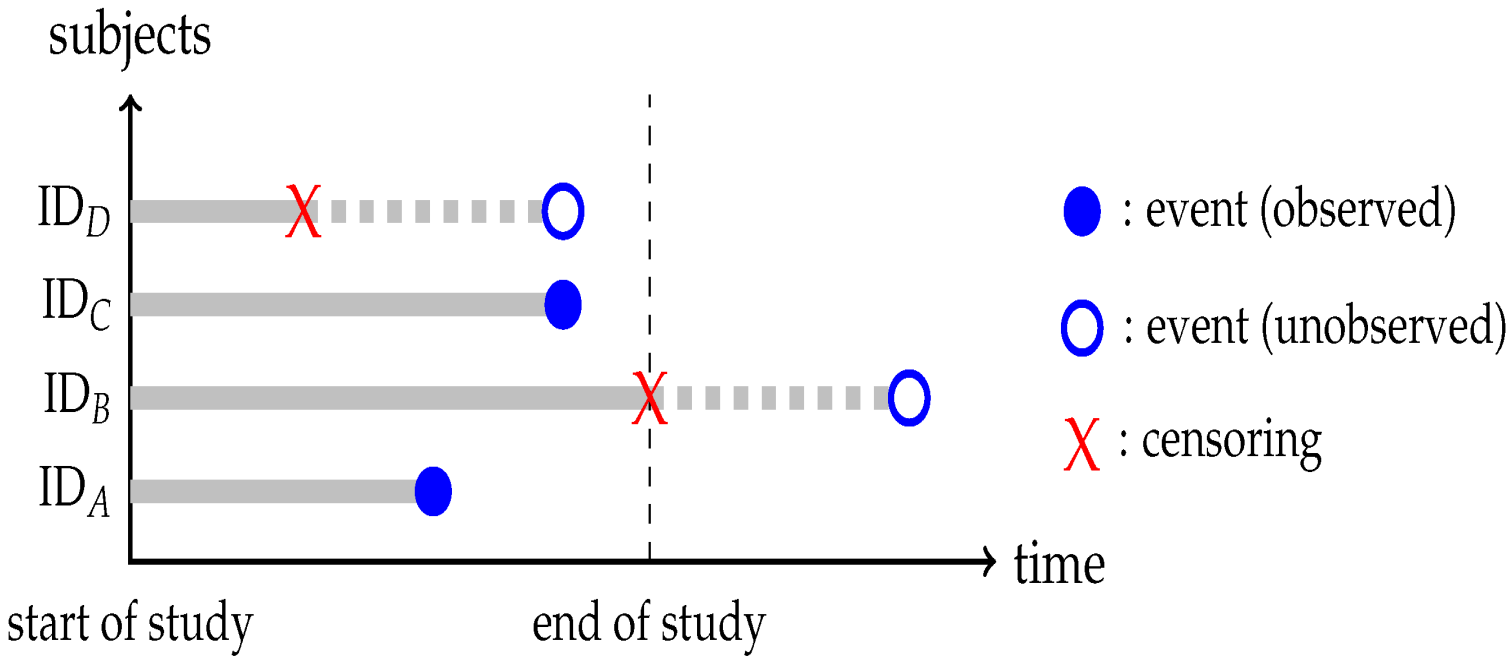

FIGURE – Censure à droite, 4 sujets ; *Introduction to Survival Analysis in Practice*, Frank Emmert-Streib and Matthias Dehmer

Censure à droite

$$T = X \wedge C = \min(X, C)$$

Censure type I : Tous les sujets commencent au même moment et la fin C est fixée (expérience en labo.) :

$$T_i = X_i \wedge C$$

Censure type II : L'étude prend fin après l'observation d'un nombre k fixé d'évts (expérience en labo.) :$X_{(1)}, \ldots, X_{(n)}$ les temps d'évts ordonnés

$$\text{Pour } 1 \leq i \leq k,\ T_{(i)} = X_{(i)},\ \text{et pour } k \leq i \leq n,\ T_{(i)} = X_{(k)}.$$

Censure type III : L'évt pour certains sujets n'est pas observé pour des raisons aléatoires (perte de vue, changement de traitement, fin de l'étude) :

$$T_i = X_i \wedge C_i$$

Autres censures

Censure à gauche : certains individus ont pu subir l'évt avant l'arrivée de l'observateur :

$$T = X \vee C.$$

Censure par intervalle : les individus ne sont pas observés durant des intervalles de temps. En pratique : $C =$ temps de la dernière visite avant l'évt (censure à droite)

Troncature : Non-observation d'individus lié à l'échantillonnage. Exemple : les individus morts avant le début de l'étude ne font pas partie de l'étude.
Problème "difficile" en l'absence d'information sur l'échantillonnage.

Censure à droite indépendante de X

C'est le cas étudié dans la suite de ce cours.

C, la censure aléatoire et X, la durée de survie. On observe :

$$T = X \wedge C \text{ avec } X \perp\!\!\!\perp C \text{ (indépendance),}$$

et aussi δ, la v.a. indiquant si l'événement a été observé ou non (oui : 1, non : 0) :

$$\delta = \mathbb{1}_{\{X \leq C\}}.$$

En temps continue : X et C ont pour densité f et g et pour fonction de survie S et $\overline{G}$.
On a :

$$\mathbb{P}(T \in [t, t+dt], \delta = 1) = \frac{d\mathbb{P}(T \leq t, \delta = 1)}{dt} = f(t)\overline{G}(t)$$

Exercice : 1. Montrer l'égalité ci-dessus.
2. Déterminer une expression équivalente pour $\mathbb{P}(T \in [t, t+dt], \delta = 0)$.

Vraisemblance (Censure dr. ind.)

On observe n individus : $(T_1, \delta_1), \ldots, (T_n, \delta_n)$.

Le modèle f de la durée de survie X dépend d'un paramètre θ.

Dans le cas continue la vraisemblance pour l'individu i est :

$$\begin{aligned}\mathcal{L}_{(T_i,\delta_i)}(\theta) &= \mathbb{P}(T_i \in [t_i, t_i + dt], \delta_i = 1 \mid \theta)^{\delta_i} \times \mathbb{P}(T_i \in [t_i, t_i + dt], \delta_i = 0 \mid \theta)^{1-\delta_i} \\ &= [f(t_i;\theta)\overline{G}(t_i)]^{\delta_i} \times [g(t_i)S(t_i;\theta)]^{1-\delta_i}\end{aligned}$$

Ainsi, la vraisemblance est :

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} \mathcal{L}_{(T_i,\delta_i)}(\theta) = \prod_{i=1}^{n} f(t_i;\theta)^{\delta_i} S(t_i;\theta)^{1-\delta_i} \times \prod_{i=1}^{n} G(t_i)^{\delta_i} g(t_i)^{1-\delta_i}$$

La partie utile de la vraisemblance est :

$$\mathcal{L}_n(\theta) \propto \prod_{i=1}^{n} f(t_i;\theta)^{\delta_i} S(t_i;\theta)^{1-\delta_i}.$$

Remarques sur la vraisemblance

Vraisemblance naïve

Sur le sous échantillon des données non-censurées, la vraisemblance est :

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(t_i;\theta)^{\delta_i}.$$

Le maximum de vrais. a une plus grande variance et est asymptotiquement biaisé.

Censure non-aléatoire (vraisemblance identique)

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(t_i;\theta)^{\delta_i} S(t_i;\theta)^{1-\delta_i}.$$

Censure à droite de type II

On observe k évts avant arrêt de l'étude :

$$\mathcal{L}_n(\theta) = \binom{n}{k} \times S(t_k;\theta)^{n-k} \times \prod_{i=1}^{k} f(t_i;\theta).$$

Exercice : On suppose que :

- $X \sim \mathcal{E}(\theta)$,
- $C \sim \mathcal{W}(\lambda, \alpha)$.
- La durée de survie est indépendante de la censure.

On observe les données $(t_1, \delta_1), \ldots, (t_n, \delta_n)$. Déterminer le maximum de vraisemblance.

Exercice : Dans un étude (Klein and Moeschberger, 2005), 191 lycéens ont répondu à la question "quand avez-vous consommé pour la première fois de la marijuana ?".

Les réponses furent de trois types : 1. "Jamais consommé" 2. "Ma première fois était au mois de ..." 3."J'en ai consommé, mais je ne me rappelle pas la première fois. "

Déterminer le type de censure pour chaque réponse (si censure il y a).

Exercice : On observe pour 462 résidents d'une maison de retraite (Klein and Moeschberger, 2005) :

- Mort (0 ou 1),
- date d'entrée dans la maison de retraite,
- date de mort ou de retour à la maison.

On cherche à comprendre la durée de survie des personnes à la retraite. Que pensez-vous de ces données ?

Plan

Cas simple : pas de censure

On observe un échantillon i.i.d : $x_1 < x_2 < \ldots < x_n$.

Soit U la variable aléatoire tirant uniformément une valeur dans l'échantillon. On a :

$$\mathbb{E}(U) = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ (moyenne empirique).}$$

La fonction de survie de U est :

$$S_n(t) = \mathbb{P}(U > t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i > t\}}.$$

C'est un estimateur convergent p.s. vers la la fonction de survie de X ; convergence en loi vers un pont brownien.

Exercice : Au slide 12, on a montré que, dans le cas discret, pour $t \in [t_i, t_{i+1}[$, on a :

$$S(t) = \prod_{k=1}^{i}(1 - \lambda_k).$$

Proposer une autre manière d'obtenir $S_n(t)$.

Avec censure à droite indépendante de X

On observe un échantillon i.i.d : $(t_1, \delta_1), (t_2, \delta_2), \ldots, (t_n, \delta_n)$, avec $t_1 < t_2 < \ldots < t_n$

Les estimateurs :

$$S_{n,1}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{t_i > t\}} \text{ et } S_{n,2}(t) = \frac{1}{n}\sum_{i=1}^{n} (\mathbb{1}_{\{t_i > t\}})^{\delta_i}$$

ne convergent pas vers $S(t)$ (fct de survie de X) ; ils sont asymptotiquement biaisés.

Soit (U, D) la variable aléatoire tirant uniformément une observation dans l'échantillon. On cherche donc à estimer :

$$\lambda(t) = \lim_{h \to 0} \frac{\mathbb{P}(t \leq X \leq t + h,\ \delta = 1)}{\mathbb{P}(X \geq t)}$$

par

$$\hat{\lambda}(t_i) = \lim_{h \to 0} \frac{\mathbb{P}(t_i \leq U \leq t_i + h,\ D = 1)}{\mathbb{P}(U \geq t_i)}.$$

Estimateur de Kaplan-Meier

On a :

$$\mathbb{P}(t_i \leq U \leq t_i + h,\ D = 1) = \frac{\delta_i}{n}$$

et

$$\mathbb{P}(U \geq t_i) = \frac{n - i + 1}{n}.$$

Définition (Kaplan-Meier)

L'estimateur de Kaplan-Meier est, pour $t \geq t_1$:

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{\delta_i}{n - i + 1}\right)$$

et 1 sinon.

Le cas des ex-aequo :

- Pour les événements de nature différente, on considère que les obs. non censurées ont lieu avant les censurées.
- d_i, nombre de décès au temps t_i,

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n - i + 1}\right)$$

La variance de $\hat{S}(t)$

Théorème (Normalité asymptotique de Kaplan-Meier)

En tout point de continuité de S, pour tout t telle que $S(t^-) > 0$, on a :

$$\sqrt{n}\left(\hat{S}(t) - S(t)\right) \underset{n\to+\infty}{\longrightarrow} \mathcal{N}(0, V^2(t)),$$

avec

$$V^2(t) = -S^2(t)\int_0^t \frac{S(du)}{S^2(u)G(u)}.$$

De plus, et c'est une conséquence "directe", on obtient un estimateur de la variance.
Y_i, le nombre d'individu encore dans l'étude au temps t_i^-.

Définition (Estimateur de Greenwood)

$$\widehat{Var}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{i:t_i\leq t} \frac{d_i}{Y_i(Y_i - d_i)}.$$

Le risque cumulé

On rappelle que : $\Lambda(t) = \int_0^t \lambda(u)du = \int_0^t \frac{f(u)}{S(u)}du$

Définition (Estimateur de Nelson-Aalen)

$$\hat{\Lambda}(t) = \sum_{i:t_i \leq t} \frac{d_i}{Y_i}$$

Définition (Estimateur de la variance de $\hat{\Lambda}(t)$)

$$\widehat{Var}(\hat{\Lambda}(t)) = \sum_{i:t_i \leq t} \frac{d_i}{Y_i^2}.$$

Relation : $\Lambda(t) = -\log(S(t)) +$ l'estimateur de Kaplan-Meier.

Définition (Estimateur de Breslow)

$$\hat{\Lambda}_2(t) = -\sum_{i:t_i \leq t} \log\left(1 - \frac{d_i}{Y_i}\right).$$

Méthode actuarielle

Méthode identique à Kaplan-Meier. Découpage du temps fixé pour l'estimation des $\lambda(t)$ (pas les intervalles des événements) :

$$0 < \tau_1 < \tau_2 < \ldots < \tau_n < \infty$$

Dans l'intervalle $[\tau_{i-1}, \tau_i[$:

- d_i, nombre de décés,
- n_{i-1}, nombre de sujets vivant juste avant,
- c_i, nombre de sujet censurés,
- Les sujets censurés durant un intervalle de temps sont considérés à moitié à risque, comme si la censure intervenait pour tous au milieu de l'intervalle (ou comme si la censure intervenait selon une loi uniforme dans l'intervalle sans qu'on en connaisse les résultats exacts $->$ moyenne).

Exercice (actuariat) : Déterminer un estimateur $\hat{S}_2(t)$ de la fonction de survie à partir de la modélisation précédente. Proposer un estimateur de la variance basé sur celui de Greenwood.

Exercice (Harrington et Fleming) : Déterminer un estimateur $\hat{S}_3(t)$ à partir de l'estimateur de Nelson-Aalen du risque cumulé.

References


Frank Emmert-Streib and Matthias Dehmer
Introduction to Survival Analysis in Practice, Machine learning knowledge extraction, 2019.


John P Klein and Melvin L Moeschberger
Survival analysis : techniques for censored and truncated data. Springer Science & Business Media, 2005.


Kaplan, E. L. ; Meier, P.
Nonparametric estimation from incomplete observations. J. Amer. Statist. Assn. 53 :457–481, 1958.