– Texte –
Estimateur de Kaplan-Meier

1 Modèles de durées de vie

Les modèles de durées sont utilisés lorsque le phénomène étudié est la durée qui s'écoule entre deux évènements: durée de vie d'un individu, de fonctionnement d'une ampoule électrique, durée entre l'apparition d'un symptôme et la guérison, entre l'obtention d'un prêt et la fin de son remboursement, durée d'un épisode de chômage, ... On parle plus généralement de *durées de vie*, quel que soit le phénomène étudié. Cette terminologie est issue de l'histoire du développement de cette branche de la statistique, principalement lié à des applications médicales.

Aujourd'hui, les domaines d'application de l'analyse statistique des durées de vie sont nombreux et variés. Citons par exemple la médecine (biostatistique, épidémiologie, ...), l'économie (analyse du marché du travail, ...), la démographie (espérance de vie, ...), la fiabilité (durée de fonctionnement de composants industriels, ...).

Les raisons qui ont conduit à l'émergence d'une branche particulière de la statistique, consacrée à l'étude des phénomènes de durée, sont nombreuses, et liées aux caractéristiques des données traitées. Tous d'abord, les durées de vie sont positives, et présentent le plus souvent des coefficients d'asymétrie négatifs. Les modèles statistiques utilisés doivent en tenir compte. Ensuite, les données recueillies sont rarement complètes, mais plutôt censurées, ou tronquées, ce qui complique sérieusement l'inférence statistique. Citons également la présence fréquente, dans l'étude des phénomènes de durée, de variables explicatives de la survie (quantité d'un certain microbe dans l'organisme d'un patient malade, revenu mensuel des clients d'une banque, ...), ce qui a conduit au développement de modèles de régression particuliers.

2 Un problème de données manquantes: la censure

Une durée est dite *censurée* lorsque l'on dispose seulement d'une information partielle sur cette durée. Un exemple permet de mieux comprendre la notion de censure.

Exemple 2.1. On souhaite étudier l'âge d'apprentissage de la lecture chez les enfants. La durée d'intérêt est l'âge d'acquisition de la lecture (notons-la T). On suit une classe d'élèves de première année de cours primaire. Trois cas peuvent se produire. Si un élève sait déjà lire en entrant au cours primaire, T n'est pas observée: on sait seulement que T est inférieure à l'âge de cet élève à l'entrée au cours primaire. On dit que T est censurée à gauche. Si un élève achève sa première année de cours primaire sans savoir lire, on n'observe pas non plus T. On sait seulement que T est supérieure à l'âge de cet élève à la fin de la première année: T est censurée à droite. Si l'élève apprend à lire en cours d'année, la durée T est bien observée et la donnée est complète.

Définition 2.2. *Soit T une variable aléatoire positive. Soit C une autre variable aléatoire positive. T est dite censurée aléatoirement à droite (respectivement à gauche) si au lieu de T, on observe le couple (X, Δ), avec $X = \min(T, C)$ et $\Delta = 1_{\{T \leq C\}}$ (respectivement $X = \max(T, C)$ et $\Delta = 1_{\{T \leq C\}}$).*

Pour chaque individu, on observe donc une durée X, et on connaît la nature de cette durée,

c'est-à-dire que l'on sait si la durée observée est complète ($\Delta = 1$) ou censurée ($\Delta = 0$).

Remarque 2.3. Lorsque les censures à droite et à gauche se combinent, on parle de *censure par intervalle.* Les variables de censure peuvent éventuellement être constantes, on parle alors de *censure fixe.*

Il existe de nombreux autres types de censure. Nous ne considérerons ici que la censure aléatoire à droite, la plus fréquente dans les applications.

La censure pose un problème d'identifiabilité de la loi de T. En effet, nous ne disposons que d'observations du couple (X, Δ) pour estimer cette loi. Même si nous connaissions parfaitement la loi de (X, Δ), serait-il possible de déterminer de manière unique la loi de T? La proposition suivante répond par l'affirmative, sous des conditions souvent rencontrées en pratique. Notons supp(T) le support de la loi de T et $\text{ssupp}(T) = \sup\{t | t \in \text{supp(T)}\}$.

Proposition 2.4. *Si C est une censure aléatoire à droite, si T et C sont indépendantes et si ssupp$(T) \leq$ ssupp(C), alors la loi de T est identifiable à partir de la loi de (X, Δ).*

Dans la suite de ce texte, on supposera que l'on observe n copies indépendantes (X_i, Δ_i) $(i = 1, \ldots, n)$ du couple (X, Δ), où $X = \min(T, C)$ et C est une censure aléatoire à droite.

3 Fonction de survie

Soit T une variable aléatoire positive représentant une durée (de vie, par exemple. Ainsi, nous parlerons dans la suite de "décès" pour désigner l'évènement d'intérêt). Un des problèmes rencontrés en analyse des durées de vie consiste à estimer la *fonction de survie* de T, définie comme

$$S_T(t) = 1 - F_T(t) = \mathbb{P}(T > t),$$

où F_T désigne la fonction de répartition de T. Soit $T_1, \ldots, T_n$ un échantillon i.i.d. de durées, pour l'instant non censurées, de fonction de survie S_T. Un estimateur naturel de S_T est la fonction de survie empirique:

$$\mathbb{S}_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{T_i > t\}}.$$

Comme la fonction de répartition empirique ($\mathbb{F}_n = 1 - \mathbb{S}_n$), la survie empirique possède de nombreuses propriétés: citons par exemple la convergence p.s. uniforme sur $\mathbb{R}$ (théorème de Glivenko-Cantelli), la convergence en loi du processus empirique associé vers un pont Brownien (théorème de Donsker).

Intéressons-nous maintenant au cas d'une censure aléatoire à droite, on observe n copies indépendantes $X_i = \min(T_i, C_i)$ et $\Delta_i = 1_{\{T_i \leq C_i\}}$ $(i = 1, \ldots, n)$ du couple (X, Δ). Une première idée consiste à estimer S_T en utilisant seulement les durées non censurées, en posant

$$\hat{S}_n(t) = \frac{1}{\sum_{i=1}^{n} \Delta_i} \sum_{i=1}^{n} 1_{\{T_i > t, \Delta_i = 1\}}.$$

On montre cependant que

$$\hat{S}_n(t) \xrightarrow{p.s.} \int_t^{\infty} f_T(u) S_C(u)\, du \neq S_T(t), \tag{1}$$

où f_T et S_C désignent respectivement la densité de T et la fonction de survie de C. Ceci conduit donc à proposer un estimateur de S_T qui prenne en compte les censures. Pour cela, notons qu'un

individu censuré avant l'instant t peut être encore vivant en t même s'il n'est plus observé. Il semble donc raisonnable de proposer un estimateur $\hat{S}$ de S_T en posant:

$$\hat{S}(t) = \frac{1}{n}\left[\sum_{i=1}^{n} 1_{\{X_i > t\}} + \sum_{i=1}^{n} 1_{\{X_i \leq t, \Delta_i = 0\}} \frac{\hat{S}(t)}{\hat{S}(X_i)}\right]. \tag{2}$$

Cette expression repose sur l'idée que les sujets survivants en t sont:

1. ceux qui ne sont ni morts ni censurés avant t, et dont l'effectif est donné par $\sum_{i=1}^{n} 1_{\{X_i > t\}}$,
2. et ceux qui ayant été censurés en $X_i \leq t$, survivent au-delà de t, avec la probabilité $\hat{S}(t)/\hat{S}(X_i)$ qui pondère chacun d'entre eux.

Un estimateur qui vérifie (2) est dit *cohérent*. Néanmoins, si l'interprétation de (2) peut paraître raisonnable, nous ne pouvons en déduire d'expression explicite pour $\hat{S}(t)$. Un autre argument va donc être utilisé pour construire un estimateur de S_T. Il est exposé dans la partie suivante, qui présente l'estimateur de Kaplan-Meier, estimateur le plus utilisé de la fonction de survie.

4 Estimateur de Kaplan-Meier de la fonction de survie

On dispose de n répliques indépendantes (X_i, Δ_i) où $X_i = \min(T_i, C_i)$ et $\Delta_i = 1_{\{T_i \leq C_i\}}$. Notons $X_{(1)} < X_{(2)} < \ldots < X_{(L)}$ $(L \leq n)$ les L instants distincts d'évènements (censurés ou non) parmi $X_1, \ldots, X_n$. Notons également $M(t)$ le nombre de "décès" en t et $R(t)$ le nombre de *sujets à risque* en t (c'est-à-dire ni morts ni censurés juste avant t: $R(t) = \sum_{i=1}^{n} 1_{\{X_i \geq t\}}$).

Définition 4.1. *Soit $t > 0$. L'estimateur de Kaplan-Meier $\hat{S}_{KM}(t)$ de $S_T(t)$ est défini par*

$$\hat{S}_{KM}(t) = \prod_{i=1}^{L} \left(1 - \frac{M(X_{(i)})}{R(X_{(i)})}\right)^{1_{\{X_{(i)} \leq t\}}}.$$

La construction de l'estimateur de Kaplan-Meier repose sur la décomposition suivante de S_T:

$$\begin{aligned} S_T(X_{(i)}) &= \mathbb{P}(T > X_{(i)}) \\ &= \mathbb{P}(T > X_{(i)} | T > X_{(i-1)}) \cdot P(T > X_{(i-1)}) \\ &= \ldots \\ &= \mathbb{P}(T > X_{(i)} | T \geq X_{(i)}) \cdots P(T > X_{(1)} | T \geq X_{(1)}) \cdot P(T \geq X_{(1)}). \end{aligned}$$

Il est naturel d'estimer $\mathbb{P}(T > X_{(i)} | T \geq X_{(i)})$ par le rapport entre le nombre de sujets non décédés en $X_{(i)}$ et le nombre de sujets à risque en $X_{(i)}$, soit $(R(X_{(i)}) - M(X_{(i)}))/R(X_{(i)})$.

L'estimateur ainsi obtenu est une fonction càdlàg en escalier, les sauts ayant lieu aux seuls instants $X_{(i)}$ où se produit au moins un décès. En effet, si en $X_{(i)}$ ne se produisent que des censures, le facteur correspondant dans le produit vaut 1.

Remarque 4.2. S'il n'y a pas d'ex-aequo, $L = n$. Ordonnons alors les observations X_i $(i = 1, \ldots, n)$ par ordre croissant, on obtient la statistique d'ordre $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$. L'estimateur de Kaplan-Meier se réécrit

$$\begin{aligned} \hat{S}_{KM}(t) &= \prod_{i=1}^{n} \left(1 - \frac{M(X_{(i)})}{R(X_{(i)})}\right)^{1_{\{X_{(i)} \leq t\}}} \\ &= \prod_{\substack{i=1 \\ X_{(i)} \leq t}}^{n} \left(1 - \frac{1}{n-i+1}\right)^{\Delta_{(i)}}. \end{aligned} \tag{3}$$

Remarque 4.3. Si $t > X_{(n)}$, $\hat{S}_{KM}(t) = 0$ si $\Delta_{(n)} = 1$ et $\hat{S}_{KM}(t) > 0$ si $\Delta_{(n)} = 0$.

5 Propriétés de $\hat{S}_{KM}$

On montre que:

Proposition 5.1. *L'estimateur de Kaplan-Meier est l'unique estimateur cohérent de la fonction de survie S_T.*

La proposition suivante caractérise le biais de l'estimateur de Kaplan-Meier:

Proposition 5.2. *Si $S_T(t) > 0$,*

$$0 \leq E[\hat{S}_{KM}(t) - S_T(t)] \leq \mathbb{P}(T \leq t) \cdot [1 - \mathbb{P}(X \geq t)]^n.$$

Notons S_C la fonction de survie de C et $\tau_X = \inf\{x \geq 0 | S_T(x).S_C(x) = 0\}$. Le théorème suivant établit deux résultats asymptotiques pour l'estimateur de Kaplan-Meier dans le cas de la censure aléatoire à droite:

Théorème 5.3.

1. *Si S_T et S_C n'ont pas de discontinuités en commun, on a pour tout $\tau < \tau_X$:*

$$\sup_{0 \leq t \leq \tau} |\hat{S}_{KM}(t) - S_T(t)| \xrightarrow{p.s.} 0.$$

2. *En tout point $t \in [0, \tau]$:*

$$\sqrt{n}(\hat{S}_{KM}(t) - S_T(t)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V^2(t)), \text{ avec } V^2(t) = -S_T^2(t) \int_0^t \frac{S_T(du)}{S_T^2(u) S_C(u)}. \tag{4}$$

L'estimateur usuel de la variance $V^2(t)$ est donné par la

Définition 5.4. (Estimateur de Greenwood) *Reprenant les notations de la définition 4.1, l'estimateur de Greenwood de $V^2(t)$ est défini comme*

$$\hat{V}_n^2(t) = n\hat{\sigma}_n^2(t), \text{ où } \hat{\sigma}_n^2(t) = \hat{S}_{KM}^2(t) \cdot \sum_{i: X_{(i)} \leq t} \frac{M(X_{(i)})}{R(X_{(i)}) \cdot [R(X_{(i)}) - M(X_{(i)})]}. \tag{5}$$

On peut montrer que $\hat{V}_n^2(t)$ converge p.s. vers $V^2(t)$. Les résultats (4) et (5) permettent de déterminer un intervalle de confiance de niveau asymptotique $1 - \alpha$ $(0 < \alpha < 1)$ pour $S_T(t)$:

$$\left[\hat{S}_{KM}(t) - u_{1-\frac{\alpha}{2}} \cdot \hat{\sigma}_n(t), \hat{S}_{KM}(t) + u_{1-\frac{\alpha}{2}} \cdot \hat{\sigma}_n(t)\right], \tag{6}$$

où $u_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi $\mathcal{N}(0, 1)$.

Remarque 5.5. Notons que si l'on souhaite obtenir une *bande de confiance* de niveau asymptotique donné pour S_T, nous avons besoin du résultat suivant. Pour tout $\tau < \tau_X$, notons $D[0, \tau]$ l'espace des fonctions càdlàg sur $[0, \tau]$. Si S_T est continue sur $[0, \tau]$, alors

$$\sqrt{n}(\hat{S}_{KM} - S_T) \Longrightarrow Z \quad \text{dans } D[0, \tau],$$

où Z est un processus gaussien centré, de fonction de covariance

$$\text{cov}(Z(s), Z(t)) = -S_T(s) S_T(t) \int_0^{\min(s,t)} \frac{S_T(du)}{S_T^2(u) S_C(u)}.$$

Remarque 5.6. Les réalisations de l'intervalle (6) peuvent être supérieures à 1 ou inférieures à 0. Une solution consiste à déterminer la loi asymptotique de $\ln(-\ln \hat{S}_{KM}(t))$, puis à en déduire un intervalle de confiance pour $S_T(t)$.

6 Suggestions

Quelques suggestions pour traiter ce sujet:

1. Démontrer le résultat (1). On pourra aussi illustrer par la simulation le fait que $\hat{S}_n$ n'est pas un estimateur satisfaisant de S_T.
2. Expliquer la définition 4.1.
3. Démontrer le résultat (3).
4. Montrer qu'en l'absence de censures, l'expression (3) se ramène à la fonction de survie empirique.
5. Illustrer par la simulation la convergence de $\hat{S}_{KM}(t)$.
6. Montrer qu'en l'absence de censures, $V^2(t)$ se ramène à $F_T(t)(1 - F_T(t))$. En supposant par exemple que $T \sim \mathcal{E}(\lambda)$ et $C \sim \mathcal{E}(\gamma)$ (où $\mathcal{E}(\lambda)$ désigne la loi exponentielle de paramètre $\lambda > 0$), on pourra représenter graphiquement le rapport de la variance asymptotique de $\sqrt{n}(\hat{S}_{KM}(t) - S_T(t))$ en présence de censures et de la variance asymptotique de $\sqrt{n}(\hat{S}_{KM}(t) - S_T(t))$ en l'absence de censures.
7. Démontrer le résultat (6). Déterminer ensuite un intervalle de confiance de niveau asymptotique $1 - \alpha$ pour $S_T(t)$ en utilisant la remarque 5.6.