

Le cadre de l'estimation non-paramétrique repose sur l'estimation de quantités d'intérêt sans hypothèse de modélisation a priori (en tout cas pas d'hyp. de modèles paramétriques). Nous verrons ici comment estimer la fonction de survie, et le taux de hazard.

Dans la suite du chapitre, nous nous plaçons dans le cadre d'une censure à droite aléatoire (type I).

I - Estimateur de Kaplan-Meier

① - Estimer la fonction de survie:

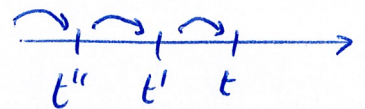
L'estimateur de Kaplan-Meier est un estimateur non-paramétrique de la fonction de survie. Il provient d'une idée simple et intuitive: survivre après une durée t revient à être en vie juste avant t et ne pas mourir en t . Ainsi, avec $t'' < t' < t$,

$$P(T > t) = P(T > t', T > t)$$

$$= P(T > t | T > t') P(T > t')$$

$$= P(T > t | T > t') P(T > t', T > t'')$$

$$= P(T > t | T > t') P(T > t' | T > t'') P(T > t'') -$$



En considérant des temps d'événements (décès et censure) distincts, noté $X_{(i)}$ ($i=1, \dots, n$) triés par ordre croissant, on obtient:

$$P(T > X_{(j)}) = \prod_{k=1}^j P(T > X_{(k)} | T > X_{(k-1)}) \quad , \quad \text{avec } X_{(0)} = 0$$

→ Soient les notations:

- Y_i l'exposition : c'est le nombre d'individus à risque de subir l'événement juste avant le temps $X_{(i)}$.
- d_i le nombre de décès en $X_{(i)}$.

⇒ On peut déduire la probabilité de mourir dans l'intervalle $]X_{(i-1)}; X_{(i)}]$ sachant que l'on était vivant en $X_{(i-1)}$: $p_i = P(T \leq X_{(i)} | T > X_{(i-1)})$.

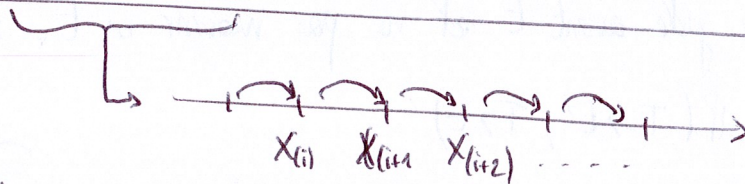
Cette quantité s'estime naturellement par $\hat{p}_i = \frac{d_i}{Y_i}$.

→ En pratique, puisqu'on considère distincts les temps d'événement, on a:

- $d_i = 0$ si l'individu i est censuré en $X_{(i)}$ ($\delta_i = 0$)
- $d_i = 1$ " " " est décédé en $X_{(i)}$ ($\delta_i = 1$). [On observe l'événement]

Définition: L'estimateur de Kaplan-Meier s'exprime alors comme suit:

$$\hat{S}(H) = \prod_{\substack{i=1 \\ X_{(i)} \leq t}}^n \left(1 - \frac{\frac{d_i}{Y_i}}{p_i} \right) = \prod_{i: X_{(i)} \leq t} \left(1 - \frac{d_i}{n - (i-1)} \right) = \prod_{i: X_{(i)} \leq t} \left(\frac{n-i}{n-i+1} \right)^{\delta_i}$$

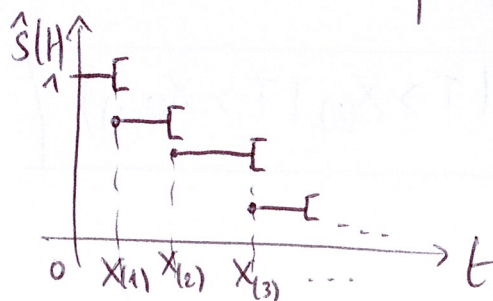


→ Cet estimateur est aussi appelé estimateur Produit Limite (c'est la limite d'un produit).

→ On peut montrer que cet estimateur est un estimateur du maximum de vraisemblance!
⇒ il possède donc d'excellentes propriétés...

→ On peut avoir un estimateur KM dans le cas de la censure, mais pas avec la censure par intervalle (car on ne connaît pas les temps d'événement).

→ Illustration:



$\hat{S}(H)$ est décroissante.
en escalier
continue à droite.

Remarques:

(2)

- S'il y a des ex-aequo pour les temps d'événement, on compte le nombre de décès correspondants. Par exemple, s'il y a d_i décès au temps $T_{(i)}$ (avec $d_i > 1$),

$$\text{on a } \hat{S}(t) = \prod_{\substack{i=1 \\ T_{(i)} \leq t}}^n \left(1 - \frac{d_i}{Y_i}\right)$$

- Lorsqu'on observe un échantillon iid de durées non-censurées $(T_i)_{i=1, \dots, n}$, on utilise naturellement la fonction de survie empirique: $S_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{T_i > t\}}$.
→ S'il y a censure, cet estimateur devient biaisé et a tendance à sous-estimer la survie.
→ Sans censure, ce dernier estimateur a d'excellentes propriétés, héritées de celles de la fonction de répartition empirique (convergence p.s. grâce à Glivenko-Cantelli).
- $\hat{S}(t)$ est un estimateur, donc une fonction des réalisations X_i , donc une v.e.!

②. Variance de l'estimateur de Kaplan-Meier.

Le but ici est de proposer un estimateur de la variance de l'estimateur Kaplan-Meier de la fonction de survie. Pour commencer, on partira de la propriété de normalité asymptotique de l'estimateur Kaplan-Meier.

Théorème: En tout point de continuité de S , $t_0 \in [0, \tau]$ et $S(t_0^-) > 0$,

$$\sqrt{n} (\hat{S}(t_0) - S(t_0)) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, V^2(t_0)), \text{ avec}$$

$$V^2(t_0) = -S^2(t_0) \int_0^{t_0} \frac{S(du)}{S^2(u)G(u)}, \text{ où } G(u) \text{ est la fonction de survie de la censure } C.$$

→ Asymptotiquement, $E[\hat{S}(t_0)] = S(t_0) \rightarrow$ estimateur sans biais...

On peut montrer par ailleurs que la variance $V^2(t_0)$ de l'estimateur Kaplan-Meier peut se récrire en fonction des quantités suivantes (ce qui sera ensuite utile pour passer à la version empirique):

- $H(t) = P(X > t) = P(\min(T, C) > t) = P(T > t, C > t) \stackrel{T \perp C}{=} P(T > t) P(C > t) = S(t) G(t)$
- $H_1(t) = P(X > t, \delta = 1) = P(\min(T, C) > t, T \leq C) = P(T > t, C > t, T \leq C)$
 $= P(T > t, T \leq C) = E[P(C > T) \mathbb{1}_{\{T > t\}}] = E[G(T^-) \mathbb{1}_{\{T > t\}}]$
 $= \int_t^{\infty} G(u^-) f(u) du = - \int_t^{\infty} G(u^-) S(du) \quad \text{Ainsi, } H_1(dt) = G(t^-) S(dt).$
 $\underbrace{f(u)}_{f(u) = -\frac{dS(u)}{du}}$

On obtient donc:

$$V^2(t_0) = -S^2(t_0) \int_0^{t_0} \frac{S(du)}{S^2(u) G(u)} = -S^2(t_0) \int_0^{t_0} \frac{\frac{H_1(du)}{G(u^-)}}{\left(\frac{H(u)}{G(u)}\right)^2 G(u)} = -S^2(t_0) \int_0^{t_0} \frac{H_1(du)}{G(u^-) \frac{H^2(u)}{G(u)}}$$

avec $H(t) = S(t) G(t) \Rightarrow S(t) = H(t) / G(t)$

d'où $V^2(t_0) = -S^2(t_0) \int_0^{t_0} \frac{H_1(du)}{G(u^-) H(u) S(u)}$

Il reste à remplacer H et H_1 par leurs équivalents empiriques (avec X et δ observés!):

- $\hat{H}(u) = \widehat{P(X > u)} = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i > u\}}$

- $\hat{H}_1(u) = \widehat{P(X > u, \delta = 1)} = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i > u, \delta_i = 1\}}$,

et on obtient: $\hat{V}^2(t) = -\hat{S}^2(t) \int_0^t \frac{\hat{H}_1(du)}{\hat{H}(u) \hat{H}(u^-)}$

Par le théorème, $\widehat{\text{Var}}(\hat{S}(t)) = \frac{1}{n} \hat{V}^2(t)$.

On peut réécrire les quantités $\hat{H}(u)$ et $\hat{H}_1(u)$ avec Y_i l'exposition juste avant le temps $X(i)$ et d_i le nombre de décès en $X(i)$: en effet, (3)

$$\bullet \hat{H}(u) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X(i) > u\}} = \frac{1}{n} (Y_i - d_i)$$

$$\bullet \hat{H}(u^-) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X(i) \geq u\}} = \frac{1}{n} Y_i$$

$$\bullet \hat{H}_1(du) = \hat{H}(u) - \hat{H}(u^-) = -\frac{1}{n} \sum \mathbb{1}_{\{X(i) \in [u, u+du], d_i = 1\}} = -\frac{d_i}{n}$$

D'où

$$\begin{aligned} \widehat{\text{Var}}(\hat{S}(H)) &= \frac{1}{n} \hat{V}^2(H) = \frac{1}{n} \left(-\hat{S}^2(H) \int_0^t \frac{\hat{H}_1(du)}{\hat{H}(u)\hat{H}(u^-)} \right) \\ &= -\hat{S}^2(H) \times \frac{1}{n} \int_0^t \frac{\hat{H}_1(du)}{\hat{H}(u)\hat{H}(u^-)} = +\hat{S}^2(H) \sum_{i: X(i) \leq t} \frac{d_i}{(Y_i - d_i) Y_i} \end{aligned}$$

On vient d'aboutir à l'estimateur de Greenwood.

→ Remarque: On obtient cet estimateur également en réalisant que nous utilisons ici la "Delta méthode": en effet, en utilisant l'approximation

$$\widehat{\text{Var}}(\log(\hat{S}(H))) \approx \sum_{i: X(i) \leq t} \frac{d_i}{Y_i(Y_i - d_i)}$$

et en appliquant la Delta-méthode selon laquelle $\text{Var}(f(z)) \approx f'(\theta(z))^2 \text{Var}(z)$,

où $\begin{cases} f = \log \\ z = \hat{S}(H) \\ \theta(z) = S(H) \end{cases}$, on obtient $\widehat{\text{Var}}(\underbrace{\log(\hat{S}(H))}_{\hat{A}(H)}) \approx \frac{1}{S(H)^2} \text{Var}(\hat{S}(H))$

Conclusion: L'estimateur de la variance de l'estimateur Kaplan-Meier est

l'estimateur de Greenwood, donné par: $\widehat{\text{Var}}(\hat{S}(H)) = \hat{S}^2(H) \sum_{i: X(i) \leq t} \frac{d_i}{(Y_i - d_i) Y_i}$

③ - Intervalle de confiance de la survie:

L'estimateur de Kaplan-Meier fournit une estimation en moyenne de la fonction de survie. Or, dans chaque intervalle temporel, la survie est une proportion.

Sous certaines conditions, on peut donc faire une approximation par une loi normale, ce qui permet d'obtenir un intervalle de confiance (IC) de la survie:

$$IC_{1-\alpha}(\hat{S}(H)) = \left[\hat{S}(H) - z_{\frac{\alpha}{2}} \sqrt{\widehat{\text{Var}}(\hat{S}(H))} ; \hat{S}(H) + z_{\frac{\alpha}{2}} \sqrt{\widehat{\text{Var}}(\hat{S}(H))} \right]$$

Typiquement, le niveau α de l'IC (en général, on prend $\alpha = 5\%$, d'où une confiance de 95% d'appartenir à cet intervalle, et ainsi $z_{1-\frac{\alpha}{2}} = z_{97,5\%}$ est le quantile à 97,5% de la loi $N(0,1)$ qui vaut 1,96).

→ Remarque: Lorsque $\hat{S}(H)$ est proche de 1 ou 0, il faut éviter d'utiliser cet IC. En effet, puisqu'il est symétrique, les bornes peuvent sortir de l'intervalle $[0,1]$ (ce qui est impossible pour une fonction de survie!).

Dans ce cas, on utilise l'IC de Rothman, donné par:

$$IC_{1-\alpha}(\hat{S}(H)) = \frac{K}{K + \frac{z_{1-\frac{\alpha}{2}}^2}{2K}} \left[\hat{S}(H) + \frac{z_{1-\frac{\alpha}{2}}^2}{2K} \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\text{Var}}(\hat{S}(H)) + \frac{z_{1-\frac{\alpha}{2}}^2}{4K^2}} \right]$$

avec $K = \hat{S}(H)(1 - \hat{S}(H)) / \widehat{\text{Var}}(\hat{S}(H))$.

II - Estimateur de Nelson-Aalen.

Cet estimateur est un estimateur du risque cumulé. Rappelons qu'il s'agit donc de $\Lambda(H) = \int_0^H \lambda(u) du$.

① - Estimer le taux de hasard cumulé

④

• Si la v.a. T admet une densité, on sait qu'on a la relation suivante:

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

• Si ce n'est pas le cas, c-à-d si la FdR de T n'admet pas de dérivée en tout point de \mathbb{R}^+ , on peut toujours définir $\Lambda(t)$ en utilisant la définition de la densité de T : $\Lambda(t) = - \int_0^t \frac{S(du)}{S(u^-)}$

→ En remplaçant les quantités $H(t)$ et $H_1(t)$ introduites précédemment, on peut écrire:

$$\Lambda(t) = - \int_0^t \frac{H_1(du)}{H(u^-)}$$

Ainsi, un estimateur naturel s'obtient par $\hat{\Lambda}(t) = - \int_0^t \frac{\hat{H}_1(du)}{\hat{H}(u^-)}$

D'où

$$\hat{\Lambda}(t) = \sum_{i: X_i \leq t} \frac{\sum_{j=1}^n \mathbb{1}_{\{X_j = X_i, \delta_j = 1\}}}{\sum_{j=1}^n \mathbb{1}_{\{X_j \geq X_i\}}} = \sum_{i: X_i \leq t} \frac{d_i}{Y_i}$$

⇒ L'estimateur de Nelson-Aalen est une fonction en escalier avec un saut de taille $\frac{d_i}{Y_i}$ à chaque instant de décès.

② - Variance de cet estimateur

Par une approximation par une loi de Poisson et en utilisant la théorie des processus de comptage, on peut aboutir à l'estimateur de la variance de l'estimateur de Nelson-Aalen, donné par

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

avec d_i et Y_i le nb de décès et d'individus à risque en X_i .

III - Autres estimateurs.

① - Estimateur de Breslow

C'est un estimateur du risque cumulé, obtenu à partir de l'estimateur KM:
en effet, $\Lambda(H) = -\log(S(H))$. Ainsi,

$$\boxed{\hat{\Lambda}_2(H) = -\ln(\hat{S}(H)) = -\sum_{i: X_i \leq t} \ln\left(1 - \frac{d_i}{Y_i}\right)}$$

→ la variance de cet estimateur vaut $\widehat{\text{Var}}(\hat{\Lambda}_2(H)) = \sum_{i: X_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}$

② - Estimateur de Harrington et Fleming.

C'est le "symétrique" de l'estimateur de Breslow: il permet d'estimer la fonction de survie à partir de l'estimateur de Nelson-Aalen.

En effet, sachant que $S(H) = e^{-\Lambda(H)}$, on peut naturellement utiliser:

$$\boxed{\hat{S}_2(H) = e^{-\hat{\Lambda}(H)} = \prod_{i: X_i \leq t} e^{-\frac{d_i}{Y_i}} \approx \prod_{i: X_i \leq t} \left(1 - \frac{d_i}{Y_i}\right) \quad \left(\text{si } \frac{d_i}{Y_i} \rightarrow 0\right).}$$

→ Remarques:

- En faisant un développement limité, on retrouve ainsi l'estimateur KM.
- Avec la méthode Delta, i.e. $\text{Var}(f(z)) \approx (f'(z))^2 \text{Var}(z)$, on obtient un estimateur de la variance de cet estimateur:

$$\widehat{\text{Var}}(\hat{S}_2(H)) = (\hat{S}_2(H))^2 \widehat{\text{Var}}(\hat{\Lambda}(H))$$

$$\Rightarrow \boxed{\widehat{\text{Var}}(\hat{S}_2(H)) = e^{-2 \sum_{i: X_i \leq t} \frac{d_i}{Y_i}} \times \sum_{i: X_i \leq t} \frac{d_i}{Y_i^2}}$$

③ - Méthode actuarielle pour l'estimation de la survie.

C'est le même principe que l'estimateur KM.

Différence: au lieu de définir les intervalles de temps (sur lesquels les probas. conditionnelles sont estimées) grâce au temps d'événement, ces derniers sont fixés par l'utilisateur. En général, on considère des intervalles de même longueur: 1 mois, 1 trimestre, ...

Ainsi, on a k intervalles $[0, t_1[$, $[t_1, t_2[$, ..., $[t_{k-1}, +\infty[$ fixés a priori.

Notons:

- d_i le nb de décès dans le i^e intervalle $[t_{i-1}, t_i[$;
- n_{i-1} le nb d'indiv. vivants au temps t_{i-1} ;
- c_i le nb " censurés dans $[t_{i-1}, t_i[$;
- r_i le nb " à risque dans $[t_{i-1}, t_i[$.

• Pour simplifier les calculs, on suppose souvent que les censures sont uniformément réparties dans l'intervalle. (les indiv. ~~sont~~ censurés sont exposés en moyenne une moitié d'intervalle). \Rightarrow pour les individus à risque, leur contribution pour l'intervalle $[t_{i-1}, t_i[$ est donc $\frac{c_i}{2}$. Ainsi, le nb d'individus à risque pour $[t_{i-1}, t_i[$

vaut
$$r_i = n_{i-1} - \frac{c_i}{2}$$

• En notant $p_i = P(T \leq t_i | T > t_{i-1})$ la proba. de mourir dans $[t_{i-1}, t_i[$ sachant que l'on était vivant en t_{i-1} , on l'estime par $\hat{p}_i = \frac{d_i}{r_i}$.

• Ainsi,
$$\hat{S}_3(t) = \prod_{i: x_i \leq t} \left(1 - \frac{d_i}{r_i}\right),$$

et la formule de Greenwood donne:

• IC: comme avec KM.

$$\widehat{\text{Var}}(\hat{S}_3(t)) = \hat{S}_3(t)^2 \sum_{i: x_i \leq t} \frac{d_i}{r_i(r_i - d_i)}$$