

---

**Travaux dirigés : modèles de durée**  
**Séance n°4 - Corrigé**

---

**Exercice 1      Manipulation de l'estimateur de Kaplan-Meier.**

Soit un échantillon de  $n$  individus i.i.d. de durée de vie  $T_1, \dots, T_n$ . Chaque observation  $i$  est soumise à censure à droite  $C_i$ , supposée indépendante et non-informative. On observe  $(Y_i, D_i)$  pour chaque  $i$ , où  $Y_i = T_i \wedge C_i$  et  $D_i = \mathbb{1}_{\{T_i \leq C_i\}}$ . On note  $S$  la fonction de survie de  $T$  et  $G$  celle de  $C$ .

1. En notant  $H_1(t) = \mathbb{P}(Y > t, D = 1)$  et  $H(t) = \mathbb{P}(Y > t)$ , montrer que

$$\frac{dS(t)}{S(t-)} = \frac{dH_1(t)}{H(t-)}.$$

2. Retrouver à partir de cette expression, l'estimateur non-paramétrique de Nelson-Aalen pour la fonction de hasard cumulée  $\Lambda(t)$ .
3. Rappeler l'expression de la fonction de hasard  $h(t)$ , de la fonction de hasard cumulée  $\Lambda(t)$  et de la fonction de survie d'une loi discrète. En déduire l'expression de l'estimateur de Kaplan-Meier de la fonction de survie.
4. En suivant un raisonnement similaire avec  $H_1(t) = \mathbb{P}(Y > t, D = 1 | Y > L)$  et  $H(t) = \mathbb{P}(L \leq t < Y | Y > L)$ , donner l'expression de l'estimateur de Kaplan-Meier avec troncature à gauche  $L$  indépendante ( $T$  indépendant de  $(L, C)$  et  $C$  indépendant de  $L$ ).
5. En utilisant directement l'expression de la première question, exprimer l'estimateur de Kaplan-Meier sous la forme d'une somme faisant intervenir l'estimateur de  $G$ .

Réponse de l'exercice 1.

1. On a

$$\begin{aligned} H(t) &= \mathbb{P}(Y > t) = \mathbb{P}(T > t, C > t) \\ &= S(t)G(t), \end{aligned}$$

et comme  $C \perp\!\!\!\perp T$

$$\begin{aligned} H_1(t) &= \mathbb{E}[\mathbb{1}_{\{T>t\}}\mathbb{1}_{\{C\geq T\}}] \\ &= \mathbb{E}[\mathbb{1}_{\{T>t\}}\mathbb{E}[\mathbb{1}_{\{C\geq T\}}|T]] \\ &= \mathbb{E}[\mathbb{1}_{\{T>t\}}G(T-)] \\ &= \int_t^\infty G(u-)dF(u). \end{aligned}$$

On écrit alors

$$dH_1(t) = -G(t-)dF(t) = G(t-)dS(t).$$

D'où le résultat.

2. Par définition,

$$\Lambda(t) = - \int_0^t \frac{dS(t)}{S(t-)} = - \int_0^t \frac{dH_1(t)}{H(t-)}.$$

Puisque les quantités  $Y_i$  et  $D_i$  sont observées, les estimateurs empiriques de  $H_1(t)$  et  $H(t)$  sont donnés par

$$\widehat{H}_1(t) = \frac{1}{n} \sum_{i=1}^n D_i \mathbb{1}_{\{Y_i > t\}} \text{ et } \widehat{H}(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{Y_i > t\}}.$$

Naturellement, l'estimateur de Nelson-Aalen apparaît en remplaçant  $H_1(t)$  et  $H(t)$  par leur estimateur, i.e.

$$\begin{aligned} \widehat{\Lambda}(t) &= - \int_0^t \frac{d\widehat{H}_1(t)}{\widehat{H}(t-)} \\ &= \frac{1}{n} \sum_{i=1}^n \frac{D_i \mathbb{1}_{\{Y_i \leq t\}}}{\widehat{H}(Y_i-)} \\ &= \sum_{i=1}^n \frac{D_i \mathbb{1}_{\{Y_i \leq t\}}}{\sum_{j=1}^n \mathbb{1}_{\{Y_j \geq Y_i\}}} \\ &= \sum_{Y_i \leq t} \frac{D_i}{\sum_{j=1}^n \mathbb{1}_{\{Y_j \geq Y_i\}}} \end{aligned}$$

3. En temps discret, la fonction de hasard est définie par

$$h(t) = \frac{\mathbb{P}(T=t)}{S(t)},$$

et la fonction de hasard cumulée est

$$\Lambda(t) = \sum_{s \leq t} h(s).$$

La fonction de survie s'écrit

$$S(t) = \prod_{s \leq t} (1 - h(s)) = \prod_{s \leq t} (1 - \Delta \Lambda(s)).$$

On retrouve ainsi l'expression de l'estimateur de Kaplan-Meier

$$\begin{aligned} \widehat{S}(t) &= \prod_{Y_i \leq t} \left(1 - \Delta \widehat{\Lambda}(T_i)\right) \\ &= \prod_{Y_i \leq t} \left(1 - \frac{D_i}{\sum_{j=1}^n \mathbb{1}_{\{Y_j \geq Y_i\}}}\right). \end{aligned}$$

4. On a immédiatement

$$\begin{aligned} H(t) &= \mathbb{P}(L \leq t < Y | Y > L) = \frac{\mathbb{P}(T > t, C > t, L \leq t)}{\mathbb{P}(Y > L)} \\ &= S(t) G(t) \frac{\mathbb{P}(L \leq t)}{\mathbb{P}(Y > L)}. \end{aligned}$$

En conditionnant par  $T$  et par indépendance avec  $L$

$$\begin{aligned} H_1(t) &= \frac{\mathbb{E} [\mathbb{1}_{\{T>t\}} \mathbb{1}_{\{C \geq T\}} \mathbb{1}_{\{T>L\}}]}{\mathbb{P}(Y > L)} \\ &= \frac{\mathbb{E} [\mathbb{1}_{\{T>t\}} G(T-) \mathbb{P}(L \leq T)]}{\mathbb{P}(Y > L)} \end{aligned}$$

On écrit alors

$$dH_1(t) = \frac{1}{\mathbb{P}(Y > L)} G(t-) \mathbb{P}(L \leq t) dS(t).$$

Ainsi, on montre que

$$\frac{dH_1(t)}{H(t-)} = \frac{dS(t)}{S(t-)}.$$

En considérant les estimateurs empiriques de  $H_1(t)$  et  $H(t)$ , on en déduit l'estimateur de Kaplan-Meier en présence de troncature à gauche tel que

$$\widehat{S}(t) = \prod_{Y_i \leq t} \left( 1 - \frac{D_i}{\sum_{j=1}^n \mathbb{1}_{\{L_i \leq Y_i \leq Y_j\}}} \right).$$

5. On écrit

$$dS(t) = \frac{dH_1(t)}{G(t-)}.$$

Ainsi en intégrant, il est possible d'exprimer l'estimateur de Kaplan-Meier sous la forme

$$\begin{aligned} \widehat{S}(t) &= \frac{1}{n} \sum_{i=1}^n \frac{D_i}{\widehat{G}(Y_i-)} \mathbb{1}_{\{Y_i > t\}} \\ &= \sum_{i=1}^n W_{i,n} \mathbb{1}_{\{Y_i > t\}}. \end{aligned}$$

On remarque que l'estimateur de Kaplan-Meier attribue des poids  $W_{i,n}$  aux observations qui ne sont pas censurées. On peut montrer que ces poids en présence de censure s'écrivent, en ordonnant les valeurs de  $Y$ ,  $Y_{(1)} \leq \dots \leq Y_{(n)}$ , et avec  $D_{[i]}$  la valeur associée à  $Y_{(i)}$

$$W_{[i],n} = \frac{D_{[i]}}{n-i+1} \prod_{j=1}^{i-1} \frac{n-j}{n-j+1}.$$

Cette expression sous forme de somme est commode car elle permet de fournir un estimateur pour les quantités du type  $\Psi_\phi = \mathbb{E}[\phi(T)]$  avec  $\phi$  une fonction intégrable (on parle d'intégrales Kaplan-Meier)

$$\widehat{\Psi}_\phi = \sum_{i=1}^n W_{i,n} \phi(Y_i).$$

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

Soit un échantillon de  $n$  individus i.i.d. de durée de vie  $T_1, \dots, T_n$ . Chaque observation  $i$  est soumise à censure à droite  $C_i$ , supposée indépendante et non-informative. On observe  $(Y_i, D_i)$  pour chaque  $i$ , où  $Y_i = T_i \wedge C_i$  et  $D_i = \mathbb{1}_{\{T_i \leq C_i\}}$ . On note  $S$  la fonction de survie de  $T$  et  $G$  celle de  $C$ .

On rappelle que l'estimateur de Kaplan-Meier vérifie

$$\sqrt{n} (\widehat{S} - S) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathbf{U},$$

avec  $\mathbf{U}$  un processus gaussien centrée de variance-covariance

$$\rho(s, t) = -S(s) S(t) \int_0^{s \wedge t} \frac{dS(u)}{S(u)^2 G(u)}.$$

1. On note  $H_1(t) = \mathbb{P}(Y > t, D = 1)$  et  $H(t) = \mathbb{P}(Y > t)$ . Puisque

$$\frac{dS(t)}{S(t-)} = \frac{dH_1(t)}{H(t-)},$$

donner un estimateur de la variance de  $\widehat{S}(t)$ .

2. En notant  $Y_{(i)}$  la statistique d'ordre  $i$  des  $Y$  et  $D_{[i]}$  la valeur de  $D$  associée, réécrire cet estimateur de la variance et vérifier qu'il correspond à l'estimateur de Greenwood.
3. Proposer un intervalle de confiance de niveau  $\alpha$  pour l'estimateur de Kaplan-Meier.
4. Déterminer la distribution asymptotique de  $\widehat{p}_t = \frac{\widehat{S}(t+1)}{\widehat{S}(t)}$ , puis en déduire l'estimateur de sa variance (de type Greenwood).

### Réponse de l'exercice 2.

1. La variance asymptotique de l'estimateur de Kaplan-Meier s'écrit

$$\begin{aligned} \text{Var}(\widehat{S}(t)) &= -S(t)^2 \int_0^t \frac{dS(u)}{S(u)^2 G(u)} \\ &= -S(t)^2 \int_0^t \frac{dH_1(u)}{H(u) H(u-)}. \end{aligned}$$

En remplaçant  $H_1(t)$  et  $H(t)$  par leur estimateur empirique

$$\widehat{H}_1(t) = \frac{1}{n} \sum_{i=1}^n D_i \mathbb{1}_{\{Y_i > t\}} \text{ et } \widehat{H}(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{Y_i > t\}},$$

on obtient l'estimateur suivant

$$\begin{aligned} \widehat{\text{Var}}(\widehat{S}(t)) &= -\frac{\widehat{S}(t)^2}{n} \int_0^t \frac{d\widehat{H}_1(u)}{\widehat{H}(u) \widehat{H}(u-)} \\ &= \frac{\widehat{S}(t)^2}{n} \frac{1}{n} \sum_{i=1}^n \frac{D_i \mathbb{1}_{\{Y_i \leq t\}}}{\left( \sum_{j=1}^n \mathbb{1}_{\{Y_j > Y_i\}} \right) \left( \sum_{j=1}^n \mathbb{1}_{\{Y_j \geq Y_i\}} \right)} \\ &= \widehat{S}(t)^2 \sum_{i=1}^n \frac{D_i \mathbb{1}_{\{Y_i \leq t\}}}{\left( \sum_{j=1}^n \mathbb{1}_{\{Y_j > Y_i\}} \right) \left( \sum_{j=1}^n \mathbb{1}_{\{Y_j \geq Y_i\}} \right)} \end{aligned}$$

2. En notant  $Y_{(i)}$  la statistique d'ordre  $i$  des  $Y$  et  $D_{[i]}$  la valeur de  $D$  associée, ainsi que  $r_{[i]}$  le nombre d'individus à risque juste avant la date  $T_{(i)}$  et  $d_{[i]}$  le nombre de décès à cette date, on obtient

$$\widehat{\text{Var}}\left(\widehat{S}(t)\right) = \widehat{S}(t)^2 \sum_{i:T_{(i)} \leq t} \frac{d_{[i]}}{(r_{[i]} - d_{[i]}) r_{[i]}} = \widehat{S}(t)^2 \gamma(t)^2.$$

On reconnaît l'estimateur de Greenwood (qui est consistant).

3. On obtient un intervalle de confiance asymptotique d'ordre  $\alpha$  pour l'estimateur de Kaplan-Meier

$$\left[\widehat{S}(t) \left(1 - \phi^{-1}\left(1 - \frac{\alpha}{2}\right) \gamma(t)\right), \widehat{S}(t) \left(1 + \phi^{-1}\left(1 - \frac{\alpha}{2}\right) \gamma(t)\right)\right],$$

avec  $\phi$  la fonction de répartition d'un loi normale centrée réduite.

4. En remarquant que

$$\widehat{p}_t = \prod_{i:t < Y_i \leq t+1} \left(1 - \frac{D_i}{\sum_{j=1}^n \mathbb{1}_{\{Y_i \leq Y_j\}}}\right) = \prod_{i:t < Y_{(i)} \leq t+1} \left(1 - \frac{d_{[i]}}{r_{[i]}}\right),$$

on en déduit facilement que

$$\widehat{\text{Var}}(\widehat{p}_t) = \widehat{p}_t^2 \sum_{i:t < Y_{(i)} \leq t+1} \frac{d_{[i]}}{(r_{[i]} - d_{[i]}) r_{[i]}}.$$