

## Modèles de durée / Examen du 13 janvier 2015

**Durée 2h – aucun document n'est autorisé**

**Corrigé**

### Quelques propriétés des estimateurs non paramétriques

La qualité de la rédaction, des justifications apportées et de la présentation de la copie seront prises en compte dans la notation.

On considère un modèle de durée homogène dont la fonction de hasard sous-jacente est notée  $h$  et la fonction de survie  $S$ . On rappelle que le processus d'événements non censurés  $N^1(t) = 1_{\{T \leq t, D=1\}}$  est compensé par

$$\Lambda^1(t) = \int_0^t R(u)h(u)du,$$

avec  $R(t) = 1_{\{T \geq t\}} \dots$

**Question n°1 (2 points)** : Que signifie la phrase ci-dessus ?

Elle signifie que  $\Lambda^1(t)$  est un processus prévisible et que  $N^1(t) - \Lambda^1(t)$  est une martingale centrée.

**Question n°2 (2 points)** : Rappelez et démontrez la relation entre la fonction de survie et la fonction de hasard.

On sait que  $S(x) = \exp\left(-\int_0^x h(u)du\right)$ . La preuve est immédiate en calculant

$$\lim_{u \rightarrow 0} u^{-1} P(X \leq x+u | X > x).$$

Pour un échantillon  $(T_i, D_i)_{1 \leq i \leq n}$ , on note  $\bar{R}(t) = \sum_{i=1}^n R_i(t)$  et  $\bar{N}^1(t) = \sum_{i=1}^n N_i^1(t)$ .

**Question n°3 (4 points)** : En utilisant le processus  $X(t) = \frac{1_{\{\bar{R}(t)>0\}}}{\bar{R}(t)}$  et le résultat de la question

n°1, construisez un estimateur  $\hat{H}(t)$  de la fonction de hasard cumulée  $H(t) = \int_0^t h(u)du$ .

Vous exprimerez cet estimateur en fonction des couples  $(d_i, r_i)$ .

De la première question, on déduit que  $M(t) = \bar{N}^1(t) - \int_0^t \bar{R}(u)h(u)du$  est une martingale centrée et

$$\int_0^t \frac{1_{\{\bar{R}(u)>0\}}}{\bar{R}(u)} dM(u) = \int_0^t \frac{1_{\{\bar{R}(u)>0\}}}{\bar{R}(u)} d\bar{N}^1(u) - \int_0^t h(u) du = \int_0^t \frac{1_{\{\bar{R}(u)>0\}}}{\bar{R}(u)} d\bar{N}^1(u) - H(t)$$

est également une martingale, ce qui suggère de proposer comme estimateur

$$\hat{H}(t) = \int_0^t \frac{1_{\{\bar{R}(u)>0\}}}{\bar{R}(u)} d\bar{N}^1(u).$$

$$\text{On en déduit } \hat{H}(t) = \sum_{T_i \leq t} \frac{d_i}{r_i}.$$

**Question n°4 (2 points)** : Montrez que l'estimateur ci-dessus présente un biais de sous-estimation de la fonction de survie.

De l'expression de l'estimateur, on déduit

$$E[\hat{H}(t)] = \int_0^t P[\bar{R}(u) > 0] h(u) du \leq H(t).$$

**Question n°5 (2 points)** : Rappelez l'expression de l'estimateur de Kaplan-Meier de la fonction de survie en fonction des couples  $(d_i, r_i)$ .

$$\hat{S}(t) = \prod_{T_i \leq t} \left(1 - \frac{d_i}{r_i}\right).$$

**Question n°6 (3 points)** : À partir de l'estimateur du hasard cumulé introduit à la question 3, proposez un estimateur de la fonction de survie et montrez qu'il présente un biais de surestimation de celle-ci (vous pourrez utiliser l'inégalité de Jensen).

On pose  $\hat{S}_{HF}(t) = \exp(-\hat{H}(t))$ .

Comme  $E[\hat{H}(t)] \leq H(t)$  et que la fonction  $g(x) = e^{-x}$  est convexe, on en déduit que

$$\begin{aligned} E(\hat{S}_{HF}(t)) &= E(g(\hat{H}(t))) \geq g(E(\hat{H}(t))) = \exp(-E(\hat{H}(t))) \\ &\geq \exp(-H(t)) = S(t) \end{aligned}$$

**Question n°7 (3 points)** : Montrez que l'estimateur de la fonction de survie défini à la question précédente est toujours supérieur à l'estimateur de Kaplan-Meier de celle-ci.

Les deux estimateurs s'écrivent respectivement, après transformation par le logarithme  $\ln \hat{S}_{KM}(t) = \sum_{T_i \leq t} \ln \left(1 - \frac{d_i}{r_i}\right)$  et  $\ln \hat{S}_{HF}(t) = - \sum_{T_i \leq t} \frac{d_i}{r_i}$  et donc

$$\ln \hat{S}_{KM}(t) - \ln \hat{S}_{HF}(t) = \sum_{T_{(i)} \leq t} \left( \ln \left( 1 - \frac{d_i}{r_i} \right) + \frac{d_i}{r_i} \right).$$

On vérifie aisément que la fonction  $f(x) = \ln(1-x) + x$  est toujours négative et donc  $\hat{S}_{KM}(t) \leq \hat{S}_{HF}(t)$ .

**Question n°8 (2 points) :** En utilisant la méthode Delta, rappelez la construction de l'estimateur de Greenwood de la variance de  $\hat{S}_{KM}(t)$ .

| Voir le cours.