

## Modèles de durée / Examen / Janvier 2017

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

**Corrigé**

### **Intervalles de confiances pour l'estimation d'une probabilité conditionnelle de sortie**

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

On considère une variable de durée  $X$  dont la fonction de hasard sous-jacente est notée  $h$  et la fonction de survie  $S$ .

**Question n°1 (2 points)** : Rappelez les définitions de la fonction de survie conditionnelle  $S_x$  et de la probabilité conditionnelle de sortie  $q_x$ ? Quel est le lien entre  $S_x$  et  $q_x$ ?

La fonction de survie conditionnelle est la fonction de survie de la variable aléatoire  $(X - x) | X > x$ , ce qui conduit à  $S_x(u) = \frac{S(x+u)}{S(x)}$ . Par définition,  $q_x = P(X \leq x | X > x)$  et donc  $q_x = 1 - S_x(1)$ .

**Question n°2 (2 points)** : Rappelez comment passer de la notion de probabilité conditionnelle de sortie,  $q_x$ , à la fonction de hasard et démontrez la relation entre la fonction de survie et la fonction de hasard.

On transforme l'expression « discrète »  $q_x = P(X \leq x | X > x)$  par son équivalent en temps continu en considérant  $\lim_{u \rightarrow 0} u^{-1} P(X \leq x + u | X > x)$ , qui admet une limite finie. On trouve facilement  $S(x) = \exp\left(-\int_0^x h(u) du\right)$ .

On se restreint à un intervalle  $[x, x+1[$ . On suppose que l'observation débute à l'instant  $e \in [x, x+1[$ .

**Question n°3 (2 points)** : Comment s'appelle l'instant  $e$ ? Quelle est la loi de durée effectivement observée? Quelle est sa fonction de survie en fonction de cette de  $X$ ?

L'instant  $e$  est une troncature gauche et la loi observée est la loi conditionnelle  $X | X > e$  de fonction de survie  $\frac{S(e+u)}{S(e)}$ .

On suppose de plus la variable  $X$  censurée à droite par une variable aléatoire  $C$ .

**Question n°4 (3 points)** : Que signifie la phrase ci-dessus ? Quelles hypothèses fait-on sur la variable  $C$  dans le contexte de la construction d'une table de mortalité d'expérience ? Par quoi sont-elles justifiées ? Quelle(s) variable(s) observe-t-on à la place de  $X$  ?

La variable  $X$  est remplacée par  $T = X \wedge C$  et  $D = 1_{\{X \leq C\}}$  ; on suppose que  $X$  et  $C$  sont indépendante et que la censure est non informative (c'est-à-dire que la loi de la censure et la loi de  $X$  n'ont pas de paramètre commun). Dans le cas de la construction d'une table de mortalité d'expérience, la censure est due au fait que l'on procède à une extraction des données à une date fixe, ce qui assure la validité des deux hypothèses.

**Question n°5 (3 points)** : calculez la probabilité qu'une observation ne soit pas censurée sur  $[x, x+1[$  en fonction de  $x$ ,  $e$  et des fonctions de survie de  $X$  et  $C$ . Comment se simplifie l'expression si la fonction de hasard est supposée constante et « petite » sur  $[x, x+1[$  ?

On écrit

$$P(D=1) = P(X \leq C \wedge (x+1) | X > e) = E_C P(X \leq C \wedge (x+1) | X > e, C)$$

d'où l'on déduit

$$P(D=1) = \int_e^{x+1} \left(1 - \frac{S_X(c)}{S_X(e)}\right) F_C(dc).$$

Si la fonction de hasard est constante sur l'intervalle, on a

$$1 - \frac{S_X(c)}{S_X(e)} = 1 - e^{-\mu_x \times (c-e)} ; \text{ en la supposant de plus petite, } 1 - e^{-\mu_x \times (c-e)} \approx \mu_x \times (c-e)$$

$$\text{et donc } P(D=1) \approx \mu_x \times \int_e^{x+1} (c-e) F_C(dc).$$

**Question n°6 (3 points)** : Comment s'interprète l'expression simplifiée obtenue ci-dessus ?

Déduisez-en un estimateur du nombre de décès observés sur  $[x, x+1[$ ,  $D = \sum_{i=1}^n D_i$ , pour un ensemble d'individus  $1 \leq i \leq n$  indépendants les uns des autres.

Le terme  $\int_e^{x+1} (c-e) F_C(dc)$  est le temps moyen passé dans l'intervalle  $[x, x+1[$ , pour lequel on dispose de l'observation  $\delta_i = s_i - e_i$  (exposition au risque) et un estimateur « naturel » de  $E(D) = \sum_{i=1}^n P(D_i = 1)$  est donc  $\mu_x \times E(x)$ ,  $E(x) = \sum \delta_i$ .

**Question n°7 (2 points)** : Quelle est la variance de  $D$  ? Comment l'estimer à partir des résultats précédents ?

On sait que comme les observations sont indépendantes,  $V(D) = \sum_{i=1}^n V(D_i)$  et comme par ailleurs les variables  $D_i$  sont des variables de Bernoulli de paramètre  $p_i = P(D_i = 1)$ ,  $V(D_i) = p_i(1 - p_i)$ , un estimateur de la variance est

$$V(D) = \mu_x \times E(x) - \mu_x^2 \times \sum_{i=1}^n \delta_i^2.$$

**Question n°8 (3 points)** : à partir des résultats des questions 6 et 7, proposez un intervalle de confiance pour l'estimateur de Hoem  $\hat{q}_x^H = \frac{D_x}{E_x}$  ; en exploitant la question n°1, proposez un intervalle de confiance pour l'estimateur de la probabilité conditionnelle de sortie construit à partir de l'estimateur de Kaplan-Meier de la fonction de survie  $\hat{S}_{KM}(t) = \prod_{T_i \leq t} \left(1 - \frac{d_i}{r_i}\right)$ .

| Voir le cours.