Modèles de durée / Examen du 20 janvier 2012

Durée 2h – aucun document n'est autorisé

Corrigé

Problème : mesure de l'incertitude sur des taux de mortalité

On considère une variable aléatoire X décrivant une durée de vie et on note S et h respectivement la fonction de survie et la fonction de hasard associées. On utilise les notations usuelles, *ie* $T_i = X_i \wedge C_i$ et $D_i = 1_{\{X_i \leq C_i\}}$ avec C_i la variable de censure. On rappelle que l'estimateur de Greenwood de la variance de l'estimateur Kaplan-Meier de la fonction de survie S est :

$$\hat{V}\left(\hat{S}(t)\right) = \hat{S}(t)^2 \gamma(t)^2$$

où $\gamma(t) = \sqrt{\sum_{T_i \leq t} \frac{d_i}{r_i(r_i - d_i)}}$ avec d_i le nombre de sorties non censurées observées à l'instant T_i et r_i l'exposition au risque. On rappelle également l'approximation de la variance de $f(X)$ obtenue par la méthode Delta :

$$V(f(X)) \approx \left(\frac{df}{dx}(E(X))\right)^2 V(X).$$

Question n°1 (2 points) : Donner une justification de l'approximation ci-dessus et préciser dans quel cadre elle est valable.

> On met X sous la forme $X = \mu + \sigma Z$ avec Z centrée réduite, on remarque que pour une fonction $x \to f(x)$ suffisamment régulière, en effectuant le développement limité $f(\mu + h) \approx f(\mu) + h\frac{df}{dx}(\mu)$, on trouve que :
>
> $$V(f(X)) \approx V\left(f(\mu) + \sigma Z \frac{df}{dx}(\mu)\right) = \sigma^2 \frac{df}{dx}(\mu)^2.$$
>
> Cette approximation est donc valide lorsque la variance σ est petite.

Question n°2 (3 points) : En partant de l'expression suivante de l'estimateur de Kaplan-Meier, $\hat{S}(t) = \prod_{T_{(i)} \leq t}\left(1 - \frac{d_i}{r_i}\right)$, donner une justification de la formule de Greenwood (on pourra considérer d'abord $\mathbf{ln}\left(\hat{S}(t)\right)$)

> *cf.* le support de cours sur les modèles non paramétriques, section 3.2.4.

Question n°3 (2 points) : Rappeler le lien entre la probabilité conditionnelle de sortie q_x et la fonction de survie $S(x)$. En déduire un estimateur Kaplan-Meier de la probabilité conditionnelle de sortie, $\hat{q}(x)$. Quel est le lien entre q_x et la fonction de survie de la variable $X_x = (X-x) | X > x$?

> $q_x = 1 - \frac{S(x+1)}{S(x)}$ et donc $\hat{q}_x = 1 - \frac{\hat{S}(x+1)}{\hat{S}(x)}$. On a par ailleurs $1 - q_x = S_x(1)$ en notant $S_x(t) = \frac{S(x+t)}{S(x)}$ la fonction de survie de X_x.

Question n°4 (3 points) : Déduire des questions précédentes un estimateur de $\hat{V}(\hat{q}(x))$

> $\hat{q}_x = 1 - \frac{\hat{S}(x+1)}{\hat{S}(x)}$; on déduit alors de l'égalité $\hat{S}(x) = \prod_{T_{(i)} \leq x} \left(1 - \frac{d_i}{r_i}\right)$ que $1 - \hat{q}(x) = \prod_{x < T_i \leq x+1} \left(1 - \frac{d_i}{r_i}\right)$ et donc à l'aide de la formule de Greenwood :
>
> $$\hat{V}(\hat{q}(x)) = (1 - \hat{q}(x))^2 \sum_{x < T_i \leq x+1} \frac{d_i}{r_i(r_i - d_i)}$$

Question n°5 (3 points) : Comment déduire de la question précédente un intervalle de confiance (asymptotique) au niveau α pour q_x ? Donner l'expression explicite d'un tel intervalle.

> En utilisant la normalité asymptotique de l'estimateur de Kaplan-Meier, on trouve les bornes suivantes de l'intervalle de confiance au niveau α :
>
> $$\hat{q}_{\pm}(x) = 1 - (1 - \hat{q}(x)) \times \left(1 \pm u_{1-\frac{\alpha}{2}} \times \sqrt{\sum_{x < T_i \leq x+1} \frac{d_i}{r_i(r_i - d_i)}}\right)$$
>
> avec u_α le quantile d'ordre α de la loi normale centrée réduite.

On considère maintenant le nombre de sorties par décès sur $[x, x+1[$, que l'on note D_x et l'exposition au risque E_x.

Question n°6 (2 points) : Rappelez la définition (mathématique) de E_x et le lien qui existe entre E_x et D_x.

> *cf.* le support de cours sur les modèles non paramétriques, section 5.2.

Question n°7 (3 points) : Proposer dans ce cadre un estimateur non paramétrique de q_x, dont on calculera l'espérance et la variance. Quelle est la loi asymptotique de cet estimateur lorsque E_x est grand ?

On pose $\tilde{q}_x = \dfrac{D_x}{E_x}$ (estimateur de Hoem) qui est sans biais puisqu'on a vu à la question précédente que $E(D_x) = E_x \times q_x$. Comme D_x suit une loi binomiale de paramètres (E_x, q_x), on en déduit que La variance est $V(\tilde{q}_x) = \dfrac{q_x(1-q_x)}{E_x}$. Par ailleurs on peut approcher la loi binomiale de D_x par une loi normale lorsque E_x tend vers l'infini et on en déduit que $\sqrt{\dfrac{E_x}{q_x(1-q_x)}}(\tilde{q}_x - q_x) \to N(0,1)$ la convergence étant en loi.

Question n°8 (2 points) : Déduire de la question précédente un intervalle de confiance (asymptotique) au niveau α pour q_x.

D'après la question précédente on a :

$$\tilde{q}_\pm(x) = \tilde{q}(x) \pm u_{1-\frac{\alpha}{2}} \times \sqrt{\frac{q_x(1-q_x)}{E_x}}$$

et on peut montrer avec quelques calculs[1] qu'en remplaçant dans la variance du terme de droite q_x qui est inconnu par son estimateur $\tilde{q}(x)$ que l'intervalle obtenu est asymptotiquement du même niveau.

[1] Voir par exemple Saporta G. [1990] Probabilités, Analyse des données et Statistique, Paris : Technip.