Université d'Angers

Master Data Science – Statistique et science des données pour la biologie

Examen Final (2h)

Le barème est donné à titre indicatif.
Les exercices sont indépendants.

Il sera tenu compte de la rédaction.

Introduction à l'Analyse de Survie

Exercice 1. (4 points)
On suppose que le temps de survie X admet pour densité

$$f(t) := \begin{cases} \dfrac{\theta}{t^{\theta+1}} & \text{si } t \geqslant 1, \\ 0 & \text{si } t < 1, \end{cases} \tag{1}$$

où $\theta > 1$.

On définit par X le temps de survie aléatoire et C la censure aléatoire et on suppose que l'on travaille avec un modèle de censure à droite, c'est-à-dire, que l'on observe les instants $T := \min(X, C)$ muni de l'indicateur de censure $\delta := \mathbb{1}_{\{X \leqslant C\}}$. On note aussi par la suite $(T_i, \delta_i)_{1 \leqslant i \leqslant n}$ un échantillon aléatoire *i.i.d.* issu de (T, δ).

1. Montrer que la fonction de survie associée à X est donnée par $S(t) = t^{-\theta}$ pour $t \geqslant 1$ et 1 sinon.
2. Proposer une transformation de $\widehat{S}_n$ pour justifier l'adéquation de X à la distribution décrite par f dans (1) et expliquer comment l'appliquer (on pourra chercher à obtenir une transformation qui rend S linéaire en θ).
3. Calculer la fonction de log-vraisemblance pour la distribution (1).
4. En déduire l'estimateur du maximum de vraisemblance $\widehat{\theta}_n$.

Exercice 2. (6 points)

On définit par X le temps de survie aléatoire et C la censure aléatoire et on suppose que l'on travaille avec un modèle de censure à droite, c'est-à-dire, que l'on observe les instants $T := \min(X, C)$ muni de l'indicateur de censure $\delta := \mathbb{1}_{\{X \leqslant C\}}$. On note aussi par la suite $(T_i, \delta_i)_{1 \leqslant i \leqslant n}$ un échantillon aléatoire *i.i.d.* issu de (T, δ).

On suppose que X et C sont continus et indépendants et on rappelle que l'estimateur de Kaplan-Meier pour la fonction de survie de X est alors donné par

$$\widehat{S}(t) = \prod_{T_{(i)} \leqslant t} \left(1 - \frac{1}{n-i+1}\right)^{\delta_{(i)}}$$

où $\delta_{(i)}$ est l'indicateur de censure associé à $T_{(i)}$.

1. Expliquer en quoi l'estimateur

$$\widehat{S}_C(t) = \prod_{T_{(i)} \leqslant t} \left(1 - \frac{1}{n-i+1}\right)^{1-\delta_{(i)}}$$

   est un candidat pour estimer la fonction de survie de C.

2. Soit $k \in \{1, 2, \ldots, n-1\}$ et $t > 0$ un réel tel que $T_{(k)} \leqslant t < T_{(k+1)}$.

   Montrer que

$$\prod_{T_{(i)} \leqslant t} 1 - \frac{1}{n-i+1} = \frac{n-k}{n}$$

   et en déduire que $\forall t > 0$

$$\prod_{T_{(i)} \leqslant t} 1 - \frac{1}{n-i+1} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{T_i > t\}}.$$

3. A l'aide des questions précédentes, montrer que le produit de $\widehat{S}$ et $\widehat{S}_C$ renvoi l'estimateur empirique pour la fonction de survie de T. Pouvait-on s'y attendre en regardant l'expression théorique de $\mathbb{P}(T > t)$, $\forall t > 0$?

Exercice 3 (5 points)
On suppose que le temps de survie X admet pour densité

$$f(t) := \begin{cases} 0 & \text{si } t > \tau, \\ \beta \exp(-\beta(\tau - t)) & \text{si } t \leqslant \tau, \end{cases}$$

où β et τ sont des paramètres strictement positifs.

On définit par C la censure aléatoire et on suppose que l'on travaille avec un modèle de censure à gauche, c'est-à-dire, que l'on observe les instants $T := \max(X, C)$ muni de l'indicateur de censure $\delta := \mathbb{1}_{\{X \geqslant C\}}$. On note aussi par la suite $(T_i, \delta_i)_{1 \leqslant i \leqslant n}$ un échantillon aléatoire *i.i.d.* issu de (T, δ).

1. Calculer les fonctions de survie S et de risque instantané λ associées à X.
2. Montrer que la variable aléatoire X suit la même loi que $\tau - Y$ où Y est une variable aléatoire dont on précisera la loi.

On suppose à présent que $C \leqslant \tau$ presque sûrement.

3. Proposer modèle de censure à droite à partir de T et δ.
4. Montrer que $\widehat{\tau}_n := \max(T_i, i = 1, \ldots, n)$ est un estimateur consistant de τ.
5. Déduire des questions précédentes une procédure d'estimation pour les paramètres τ et β à partir de l'échantillon $(T_i, \delta_i)_{1 \leqslant i \leqslant n}$. On ne demande pas de montrer la consistance de chacun des estimateurs.