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.

Exercice 1. (3 points)
Montrer que pour tout variable aléatoire continue X, $\Lambda(X)$ suit une loi exponentielle standard. Montrer également que pour $T := \min(X, c)$, alors $\mathbb{E}[\Lambda(T)] = F(c)$ où $c \in \mathbb{R}_+$.

Exercice 1. (5 points)
On suppose que le temps de survie X admet pour densité

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

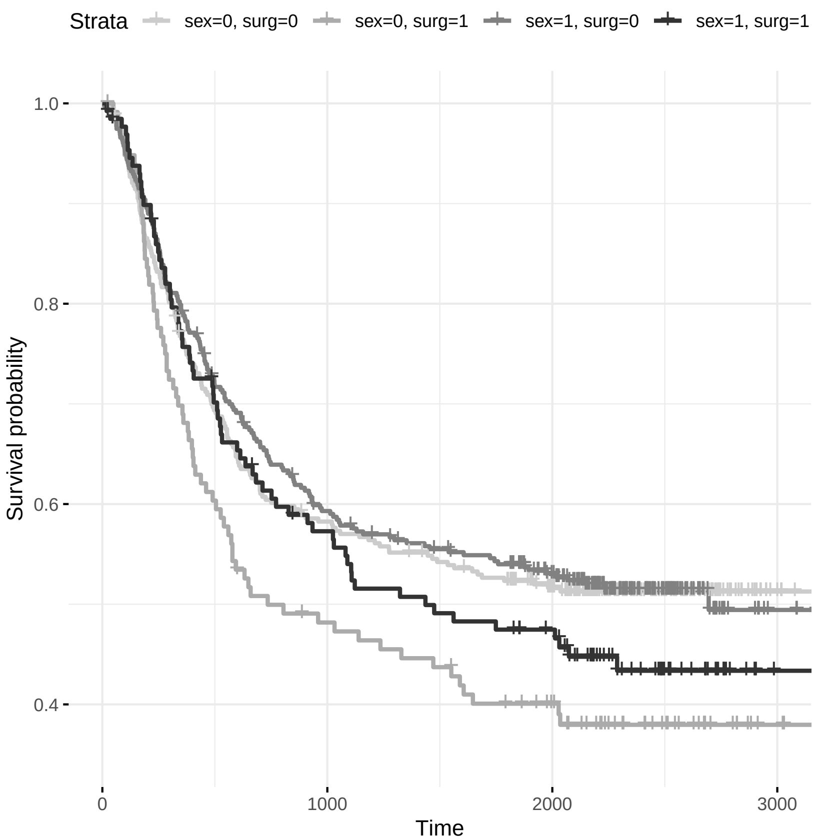
où $\theta > 1$.

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 considère un jeu de données issu d'une étude de partients atteints du cancer du colon qui suivent également un traitement de chimiothérapie. Pour simplifier l'étude, on distingue un patient selon son sexe (`sex` $= 0$ si c'est une femme et 1 sinon) et si le temps entre son opération et la prise en charge est considéré comme court ou long (`surg` $= 0$ pour court et 1 sinon).

On s'intéresse à l'instant de récidive du cancer, considéré comme censuré à droite. Une première estimation des fonctions de survie via l'estimateur de Kaplan-Meier retourne les graphes suivants :



D'après cette figure

1. Que pouvez-vous dire de l'influence du temps de prise en charge sur les probabilités de survie ?
2. De même, y a t-il une différence significative entre les femmes et les hommes ?
3. De manière globale, quelle catégorie de patients est la moins avantagée ? Justifier.

On essaye à présent de regarder la qualité d'ajustement des données pour un modèle de Cox. Le résultat est le suivant :

```
coxph(formula = Surv(time, status) ~ surg + sex, data = colon2)

  n= 929, number of events= 468

         coef exp(coef) se(coef)      z Pr(>|z|)
surg  0.25703   1.29308  0.10081  2.550   0.0108 *
sex  -0.08860   0.91521  0.09251 -0.958   0.3382
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
     exp(coef) exp(-coef) lower .95 upper .95
surg    1.2931     0.7733    1.0612     1.576
sex     0.9152     1.0926    0.7634     1.097

Concordance= 0.53  (se = 0.013 )
Likelihood ratio test= 7.09  on 2 df,   p=0.03
Wald test            = 7.31  on 2 df,   p=0.03
Score (logrank) test = 7.35  on 2 df,   p=0.03
```

4. Sous un modèle de Cox, qu'elle est la forme de la fonction de survie conditionnelle $S(t|x)$, pour $t \in \mathbb{R}_+$ et $x \in \mathbb{R}^p$?

5. D'après ces résultats, donner les valeurs estimées des paramètres du modèle.

6. Que peut-on dire des covariables `sex` et `surg` ?

7. Rappeler les hypothèses nulle et alternative liées au test du Log Rank. Qu'elle conclusion semble suggérer ces résultats ?

8. Pour conclure et d'après les questions précédentes, pensez-vous que le modèle de Cox est adéquate aux données ?

Exercice 3 (6 points)
On suppose que le temps de survie X admet pour densité

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

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.