Université Claude Bernard - Lyon 1 Année 2021-2022
ISFA - M2SAF
Frédéric Planchet et Maxime Ben-Brik

**Travaux dirigés : modèles de durée**
**Séance n°2 - Corrigé**

## Exercise 1 Censure et troncature.

Pour chacun des exemples suivants, décrire en détail le type de censure et / ou le type de troncature considérés.

1. On considère une expérience effectuée sur une population de rats pour tester la dangerosité d'une substance. La variable d'intérêt est leur durée de survie après ingestion de la substance. Cette expérience est effectuée sur 60 jours. Considérer les cas suivants :
   — Rat 1 : un rat survivant au terme des 60 jours ;
   — Rat 2 : un rat mort le $32^{\text{ème}}$ jour, mais d'un cause sans lien avec la substance ;
   — Rat 3 : un rat mort le $10^{\text{ème}}$ jour du fait de la substance.
2. On s'intéresse à la mortalité d'un portefeuille d'assurance pour une garantie en cas de décès. La population est observée entre le 01/01/2010 et le 01/01/2016. Considérer les cas suivants :
   — Assuré 1 : un assuré ayant souscrit son contrat le 01/01/2009 et décédé le 01/01/2012 ;
   — Assuré 2 : un assuré ayant souscrit son contrat le 01/01/2011 et décédé le 30/06/2016 ;
   — Assuré 3 : un assuré ayant souscrit son contrat le 01/01/2008 et décédé le 01/01/2009 ;
   — Assuré 4 : un assuré ayant souscrit son contrat le 01/01/2009 et l'ayant racheté le 01/01/2013 pour raison de santé ;
   — Assuré 5 : un assuré ayant souscrit son contrat le 01/01/2009, s'étant présenté le 01/01/2013 pour renouveler son contrat et pour lequel son épouse demande ensuite le versement de prestations le 01/01/2015.
3. On s'intéresse pour un contrat d'assurance arrêt de travail à la durée de survie en invalidité. Le contrat comprend une période de franchise de 3 mois $^*$. La population est observée entre le 01/01/2010 et le 01/01/2016. Considérer les cas suivants :
   — Assuré 1 : un assuré entré en invalidité le 01/01/2011 et décédé le 01/01/2012 ;
   — Assuré 2 : un assuré entré en invalidité le 01/01/2011 et décédé dans les trois mois ;
   — Assuré 3 : un assuré entré en invalidité le 01/12/2015.

*Réponse de l'exercice 1.*

1. — Rat 1 : Censure à droite de type I - exclu vivant ;
   — Rat 2 : Censure à droite aléatoire ;
   — Rat 3 : Observation complète.
2. — Assuré 1 : Observation à partir de l'âge atteint au 01/01/2010 (troncature gauche). ;

$*$. La date décès n'est pas communiquée si le décès survient pendant cette période.

— Assuré 2 : Observation à partir de l'âge atteint au 01/01/2011 (troncature gauche) et censure à droite de type I ;
— Assuré 3 : Observation non incluse ;
— Assuré 4 : Observation à partir de l'âge atteint au 01/01/2010 (troncature gauche). Censure à droite aléatoire potentiellement non indépendante ;
— Assuré 5 : Observation à partir de l'âge atteint au 01/01/2010 (troncature gauche). Possible censure par intervalle, le décès étant vraisemblablement survenu entre le 01/01/2013 et 01/01/2015 (on peut toutefois imaginer dans un dispositif d'assurance que la date de décès précise finira par être connue).

3. — Assuré 1 : Observation complète ;
— Assuré 2 : Censure à gauche aléatoire ;
— Assuré 3 : Troncature à droite (seul les décès de plus de 3 mois à la date de fin de l'étude peuvent être observés).

## Exercise 2 Vraisemblance et données manquantes.

Dans la plupart des analyse de survie, les observations comprennent des données manquantes. Soit un échantillon de $n$ individus de durée de vie respective $T_1, \dots, T_n$ i.i.d. Ces observations suivent une loi de densité $f$ définie à partir du paramètre $\boldsymbol{\theta}$.

1. Dans chacun des cas suivant, écrire l'expression générale de la vraisemblance, puis introduire progressivement (a) l'hypothèse d'indépendance du processus lié aux données manquantes et (b) le fait que ce processus ne dépende pas de $\boldsymbol{\theta}$ (non-informatif).
   — censure $C$ à droite de type I ;
   — troncature individuelle à gauche $L_i$ ;
   — censure individuelle à gauche $L_i$ et à droite $R_i$.
2. On suppose à présent que les observations sont uniquement soumises à censure **indépendante** à droite $C_i$. On fait l'hypothèse que les $T_i$ et les $C_i$ ont pour fonction de hasard respective $h_T(t) = \alpha t^{\alpha-1}$ et $h_C(t) = \beta t^{\beta-1}$ (Weibull). Écrire la log-vraisemblance du modèle et en déduire l'équation vérifie par l'estimateur du maximum de vraisemblance de $\boldsymbol{\theta} = (\alpha, \beta)$.

*Réponse de l'exercice 2.*

1. **Censure type I**
   On observe $(Y_i, D_i)$ pour chaque $i$, où $Y_i = T_i \wedge C$ et $D_i = \mathbb{1}_{\{Y_i = T_i\}}$. La contribution à la vraisemblance individuelle s'écrit

$$\begin{aligned}\mathcal{L}_i(\boldsymbol{\theta}) &= \mathbb{P}(Y_i = y_i, D_i = d_i; \boldsymbol{\theta}) \\ &= \mathbb{P}(T_i = y_i, D_i = 1; \boldsymbol{\theta})^{d_i} \mathbb{P}(C = y_i, D_i = 0; \boldsymbol{\theta})^{1-d_i} \\ &= \mathbb{P}(T_i = y_i, T_i \leq C; \boldsymbol{\theta})^{d_i} \mathbb{P}(C = y_i, C \leq T_i; \boldsymbol{\theta})^{1-d_i}.\end{aligned}$$

   Si la censure est indépendante, on a

$$\mathcal{L}_i(\boldsymbol{\theta}) = (f_T(y_i; \boldsymbol{\theta}) S_C(y_i; \boldsymbol{\theta}))^{d_i} (f_C(y_i; \boldsymbol{\theta}) S_T(y_i; \boldsymbol{\theta}))^{1-d_i}.$$

Si la censure est non-informative, on se limite à

$$\mathcal{L}_i(\boldsymbol{\theta}) \propto (f_T(y_i;\boldsymbol{\theta}))^{d_i} (S_T(y_i;\boldsymbol{\theta}))^{1-d_i}.$$

**Troncature à gauche**

Pour chaque individu i, on observe $(l_i, t_i)$ si $l_i \leq t_i$. La contribution à la vraisemblance individuelle s'écrit

$$\begin{aligned}\mathcal{L}_i(\boldsymbol{\theta}) &= \mathbb{P}(T_i = t_i L_i = l_i | l_i \leq T_i; \boldsymbol{\theta}) \\ &= \frac{1}{S_T(l_i;\boldsymbol{\theta})} f_{(T,L)}(t_i, l_i;\boldsymbol{\theta}) \mathbb{1}_{l_i \leq t_i}.\end{aligned}$$

Si la troncature est indépendante et comme on n'observe (par construction) que des durées au-delà de la date de troncature, on a

$$\mathcal{L}_i(\boldsymbol{\theta}) = \frac{f_T(t_i;\boldsymbol{\theta}) f_L(l_i;\boldsymbol{\theta})}{S_T(l_i;\boldsymbol{\theta})}.$$

Si la troncature est non-informative, on se limite à

$$\mathcal{L}_i(\boldsymbol{\theta}) = \frac{f_T(t_i;\boldsymbol{\theta})}{S_T(l_i;\boldsymbol{\theta})}.$$

*Remarque : on voit clairement qu'en présence de censure à droite et de troncature à gauche indépendantes et non-informatives, on aurait*

$$\mathcal{L}_i(\boldsymbol{\theta}) \propto \frac{1}{S_T(l_i;\boldsymbol{\theta})} (f_T(t_i;\boldsymbol{\theta}))^{d_i} (S_T(t_i;\boldsymbol{\theta}))^{1-d_i}.$$

*Remarque : il est aussi possible de considérer l'écriture suivante pour la vraisemblance du modèle*

$$\begin{aligned}\mathcal{L}_i(\boldsymbol{\theta}) &= \mathbb{P}(T_i = t_i L_i = l_i | L_i \leq T_i; \boldsymbol{\theta}) \\ &= \frac{\mathbb{P}(T_i = t_i L_i = l_i; \boldsymbol{\theta})}{\mathbb{P}(L_i \leq T_i; \boldsymbol{\theta})}\end{aligned}$$

*Par contre, l'évaluation du dénominateur requiert de connaître la loi de L.*

**Censure individuelle à gauche et à droite**

On introduit $(\Delta_{1,i}, \Delta_{2,i}, \Delta_{3,i})$ pour chaque $i$, où $\Delta_{1,i} = \mathbb{1}_{\{T_i \leq L_i\}}$, $\Delta_{2,i} = \mathbb{1}_{\{L_i < T_i \leq R_i\}}$ et $\Delta_{3,i} = \mathbb{1}_{\{R_i < T_i\}}$. On note $Y_i = (T_i \wedge R_i) \vee L_i$. La contribution à la vraisemblance individuelle s'écrit

$$\begin{aligned}\mathcal{L}_i(\boldsymbol{\theta}) &= \mathbb{P}(L_i = y_i, \Delta_{1,i} = 1; \boldsymbol{\theta})^{\delta_{1,i}} \mathbb{P}(T_i = y_i, \Delta_{2,i} = 1; \boldsymbol{\theta})^{\delta_{2,i}} \mathbb{P}(R_i = y_i, \Delta_{3,i} = 1; \boldsymbol{\theta})^{\delta_{3,i}} \\ &= \mathbb{P}(L_i = y_i, T_i \leq L_i; \boldsymbol{\theta})^{\delta_{1,i}} \mathbb{P}(T_i = y_i, L_i < T_i \leq R_i; \boldsymbol{\theta})^{\delta_{2,i}} \mathbb{P}(R_i = y_i, R_i < T_i; \boldsymbol{\theta})^{\delta_{3,i}}.\end{aligned}$$

Si la censure est indépendante, i.e. $T \perp\!\!\!\perp (L, R)$, on a

$$\mathcal{L}_i(\boldsymbol{\theta}) = \left(f_L(y_i;\boldsymbol{\theta})\, F_T(y_i;\boldsymbol{\theta})\right)^{\delta_{1,i}} \left(f_T(y_i;\boldsymbol{\theta})\, \mathbb{P}(L_i < y_i \leq R_i;\boldsymbol{\theta})\right)^{\delta_{2,i}} \left(f_R(y_i;\boldsymbol{\theta})\, S_T(y_i;\boldsymbol{\theta})\right)^{\delta_{3,i}}.$$

Si la censure est non-informative, on se limite à

$$\mathcal{L}_i(\boldsymbol{\theta}) \propto F_T(y_i;\boldsymbol{\theta})^{\delta_{1,i}}\, f_T(y_i;\boldsymbol{\theta})^{\delta_{2,i}}\, S_T(y_i;\boldsymbol{\theta})^{\delta_{3,i}}.$$

2. Si la censure est indépendante, la contribution individuelle à la vraisemblance s'écrit

$$\begin{aligned}\mathcal{L}_i(\boldsymbol{\theta}) &= \left(f_T(y_i;\alpha)\, S_C(y_i;\beta)\right)^{d_i} \left(f_C(y_i;\beta)\, S_T(y_i;\alpha)\right)^{1-d_i} \\ &= \left(\alpha y_i^{\alpha-1}\right)^{d_i} \left(\beta y_i^{\beta-1}\right)^{1-d_i} \exp\left(-y_i^{\alpha}\right) \exp\left(-y_i^{\beta}\right).\end{aligned}$$

La log-vraisemblance de l'échantillon est donc

$$\ln \mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} d_i \left(\ln(\alpha) + (\alpha-1)\ln(y_i)\right) + (1-d_i)\left(\ln(\beta) + (\beta-1)\ln(y_i)\right) - y_i^{\alpha} - y_i^{\beta}.$$

L'EMV de $\boldsymbol{\theta}$ s'obtient numériquement en résolvant les deux équations

$$\frac{\partial \ln \mathcal{L}(\boldsymbol{\theta})}{\partial \alpha} = \sum_{i=1}^{n} \frac{d_i}{\alpha} + d_i \ln(y_i) - y_i^{\alpha} \ln(y_i) = 0\,,$$

et

$$\frac{\partial \ln \mathcal{L}(\boldsymbol{\theta})}{\partial \beta} = \sum_{i=1}^{n} \frac{1-d_i}{\beta} + (1-d_i)\ln(y_i) - y_i^{\beta} \ln(y_i) = 0.$$

**Exercise 3 Modèle de Gompertz-Makeham.**

Soit $T$ la variable aléatoire positive représentant la durée de vie d'un individu que l'on suppose définie à partir du modèle de Gompertz-Makeham dont la fonction de hasard vaut $h(t) = a + bc^t$, avec $a$, $b$ et $c$ trois paramètres strictement positifs.

1. Rappeler pour ce modèle l'expression de sa fonction de survie, sa densité et sa fonction de survie conditionnelle à l'évènement $T > t_0$.
2. Ce modèle vous semble t-il adapté à la modélisation de la vie humaine ?
3. En présence de censure indépendante à droite non-informative, écrire la log-vraisemblance du modèle et calculer sa dérivée première pour une échantillon $(y_1, d_1, \ldots, y_n, d_n)$.

*Réponse de l'exercice 3.*

1. — fonction de survie : $S(t) = \exp\left(-\int_0^t a + bc^u du\right) = \exp\left(-at - \frac{b}{\ln(c)}\left(c^t - 1\right)\right)$ ;
   — densité : $f(t) = \left(a + bc^t\right) \exp\left(-at - \frac{b}{\ln(c)}\left(c^t - 1\right)\right)$ ;

— fonction de survie conditionnelle : $\dfrac{S(t+t_0)}{S(t_0)} = \exp\left(-at - \dfrac{bc^{t_0}}{\ln(c)}\left(c^t - 1\right)\right)$.

2. Dans la suite, on note $H(t) = \int_0^t h(u)\,du$ la fonction de hasard cumulée.
   La loi de Gompertz-Makeham est usuellement utilisée pour modéliser la durée de vie humaine. Le second terme du taux de hasard ($bc^t$) correspond à la loi de Gompertz originale (1825) et est croissant pour $c > 1$ et $b > 0$. Il permet de traduire le vieillissement progressif de l'organisme. Le paramètre $a$ ajouté par Makeham intègre les décès accidentels survenant aux âges plus jeunes. Si ce modèle est retenu pour la population générale, il ne permettra pas de prendre en compte les âges de la vie où le taux de hasard est potentiellement décroissant (mortalité infantile, bosses dues aux accidents chez les jeunes adultes...). Notons qu'il n'existe pas d'expression explicite pour la loi de Makeham-Gompertz pour l'espérance de vie et les moments d'ordre supérieurs.

3. La log-vraisemblance du modèle s'écrit

$$\ln\left(\mathcal{L}(a,b,c)\right) = \sum_i^n d_i \ln\left(a + bc^{y_i}\right) - ay_i - \frac{b}{\ln(c)}\left(c^{y_i} - 1\right).$$

   Ainsi, ses dérivées première valent

$$\frac{\partial \ln\left(\mathcal{L}(a,b,c)\right)}{\partial a} = \sum_{i=1}^n \frac{d_i}{a + bc^{y_i}} - y_i \ ,$$

$$\frac{\partial \ln\left(\mathcal{L}(a,b,c)\right)}{\partial b} = \sum_{i=1}^n \frac{d_i c^{y_i}}{a + bc^{y_i}} - \frac{c^{y_i} - 1}{\ln(c)} \ ,$$

$$\frac{\partial \ln\left(\mathcal{L}(a,b,c)\right)}{\partial c} = \sum_{i=1}^n \frac{d_i y_i b c^{y_i - 1}}{a + bc^{y_i}} - \frac{y_i b c^{y_i - 1}}{\ln(c)} + \frac{b\left(c^{y_i} - 1\right)}{c\left(\ln(c)\right)^2}.$$

   L'estimateur de maximum de vraisemblance satisfait le système où ces trois équations sont nulles. Le système se résout numériquement par le biais d'un procédure itérative (ex : algorithme de Newton-Raphson).