Modèles de durée / Examen du 13 mai 2005

Corrigé

Durée 2h – tous les documents sont autorisés

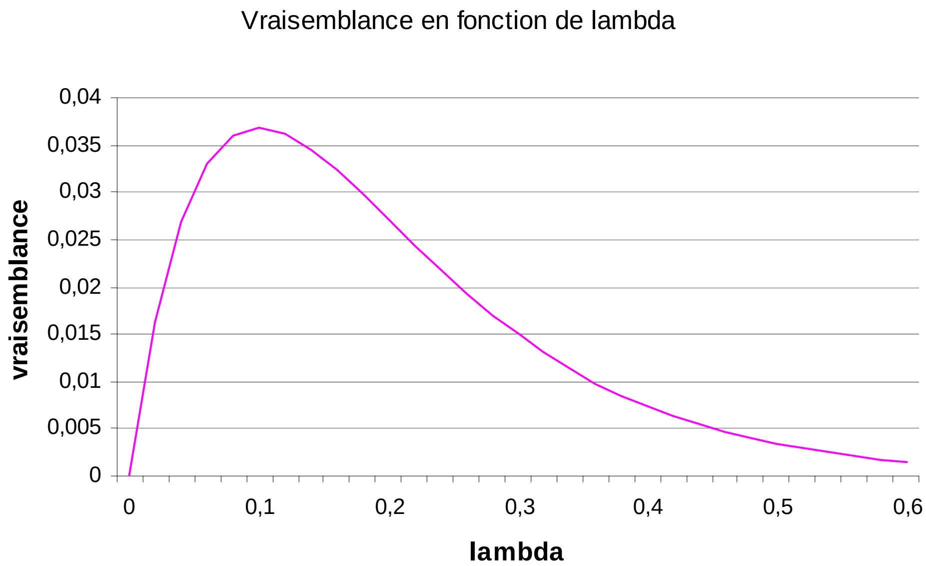
Exercice n°1 (loi exponentielle)

Sous l'hypothèse d'une arrivée aléatoire des patients dans un laboratoire d'analyses médicales, le temps passé à attendre avant d'être pris en charge est distribué selon une loi exponentielle de paramètre λ, ie telle que la fonction de hasard soit constante égale à λ.

1- *Un patient a attendu 10 minutes. Donner la vraisemblance de cette observation, et la représenter graphiquement en fonction du paramètre.*

On note T la v.a. qui mesure le temps d'attente, et t la réalisation de cette variable. Ici l'échantillon est réduit à une seule observation. La densité de probabilité de la loi exponentielle est $f(t) = \lambda e^{-\lambda t}$, donc la probabilité associée à l'échantillon est $P(T_1 \in [t_1, t_1 + dt_1[) = \lambda e^{-\lambda t_1} dt_1$. La fonction de vraisemblance correspond à cette expression, lorsqu'on la considère comme une fonction de λ: $V(\lambda) = \lambda e^{-\lambda t_1}$.

Si l'on trace cette fonction en fonction de λ, on trouve la forme caractéristique de la fonction de vraisemblance : une courbe avec un unique maximum :



L'estimation de λ au maximum de vraisemblance s'obtient en déterminant la valeur qui annule la dérivée de la vraisemblance, donc $\lambda = 1/t_1$. L'estimateur de λ au maximum de vraisemblance est donc la variable $1/T$.

2- *A la fin de la journée, on tire au sort un échantillon de 10 patients parmi ceux de la journée. Les temps d'attente observés sont (en minutes) : 1; 13; 4; 4; 3; 6; 23; 1; 16; 3. Donner la vraisemblance de cet échantillon. Quelle est l'estimation naturelle du temps moyen passé dans ce service ?*

La densité des observations est la fonction , que l'on peut calculer, sous l'hypothèse d'indépendance entre les 10 patients, comme le produit des 10 densités de probabilités individuelles.

Il est ici plus simple de passer à la log vraisemblance pour calculer l'estimation de λ. On transforme ainsi un produit en une somme de terme. Finalement on a $\lambda = n/(t_1+\cdots+ t_n)$, et l'estimateur du maximum de vraisemblance est $1/\overline{T}$ ou $\overline{T}$ est la v.a. qui donne la moyenne empirique des observations. Le temps moyen passé dans le service est de 7,4 minutes.

3- *Dans un groupe de 10 patients tirés au hasard, le plus petit temps d'attente est de 10 minutes. Que peut-on dire de plus ? (vraisemblance, temps moyen d'attente ?)*

On se trouve en situation de censure au premier décès, donc la vraisemblance s'écrit :

$$L(\lambda) = \frac{n!}{(n-1)!}\lambda\,\mathbf{exp}(-\lambda T) = n\lambda\,\mathbf{exp}(-\lambda n T_{(1)})$$

Le passage au logarithme conduit à $\mathbf{ln}\,L(\lambda) = \mathbf{ln}(n) + \mathbf{ln}(\lambda) - \lambda n T_{(1)}$, d'où $\hat{\lambda} = \frac{1}{nT_{(1)}}$; le temps moyen d'attente estimé est donc de 100 minutes !

Exercice n°2 (modèle à risques concurrents)

Dans cet exercice on utilise la paramétrisation suivante des lois exponentielles et de Weibull :

- ✓ loi exponentielle : $h(t) = \frac{1}{\lambda_e}$,
- ✓ loi de Weibull : $h(t) = \frac{\beta}{\lambda_w}\left(\frac{t}{\lambda_w}\right)^{\beta-1}$.

On considère pour modéliser le fonctionnement d'un appareil jusqu'à sa défaillance le modèle de durée $T = \mathbf{min}(E, W)$ où les variables aléatoires E et W sont indépendantes, E suivant une loi exponentielle et W une loi de Weibull définies comme ci-dessus.

1- *Déterminez la fonction de hasard, la densité, et la fonction de survie de T.*

La fonction de hasard s'écrit $h(t) = \frac{1}{\lambda_e} + \frac{\beta}{\lambda_w}\left(\frac{t}{\lambda_w}\right)^{\beta-1}$ ce qui est la conséquence directe du fait que la fonction de survie de T est le produit des fonctions de survie de E et W :

$$\mathbf{Pr}(T > t) = \mathbf{Pr}(\mathbf{min}(E, W) > t) = \mathbf{Pr}(E > t, W > t) = \mathbf{Pr}(E > t)\mathbf{Pr}(W > t)$$

On en déduit donc que $S_T(t) = \mathbf{exp}\left(-\frac{t}{\lambda_e} - \left(\frac{t}{\lambda_w}\right)^{\beta} \right)$; la densité s'obtient alors facilement en dérivant la fonction de survie (au signe près).

2- *On suppose $\beta = 2$ et on veut déterminer la probabilité que la défaillance provienne de E ; montrez que $\mathbf{Pr}(T = E) = \frac{\lambda_w}{\lambda_e}\frac{\sqrt{\pi}}{2} erfc\left(\frac{\lambda_w}{2\lambda_e}\right)$ avec $erfc(x) = e^{x^2}\frac{2}{\sqrt{\pi}}\int_x^{+\infty} e^{-u^2} du$ la fonction d'erreur complémentaire. On pourra conditionner par la durée exponentielle, puis faire le changement de variable $y = \frac{x}{\lambda_w}$.*

On écrit $\mathbf{Pr}(T = E) = E[\mathbf{Pr}(T = E \mid E)]$; mais $\{T = E\} = \{W > E\}$ et donc on peut écrire :

$$\mathbf{Pr}(T = E) = \int_0^{+\infty} \mathbf{exp}\left(-\left(\frac{x}{\lambda_w}\right)^2 \right) \frac{1}{\lambda_e} \mathbf{exp}\left(-\frac{x}{\lambda_e} \right) dx$$

En posant $y = \frac{x}{\lambda_w}$ on obtient $\mathbf{Pr}(T = E) = \frac{\lambda_w}{\lambda_e}\int_0^{+\infty} \mathbf{exp}\left(-y^2 - \frac{\lambda_w}{\lambda_e} y \right) dy$; mais $y^2 + \frac{\lambda_w}{\lambda_e} y = \left(y + \frac{\lambda_w}{2\lambda_e} \right)^2 - \left(\frac{\lambda_w}{2\lambda_e} \right)^2$; en posant alors $u = y + \frac{\lambda_w}{2\lambda_e}$, on trouve que :

$$\mathbf{Pr}(T = E) = \frac{\lambda_w}{\lambda_e} \mathbf{exp}\left(\left(\frac{\lambda_w}{2\lambda_e} \right)^2 \right) \int_{\frac{\lambda_w}{2\lambda_e}}^{+\infty} \mathbf{exp}\left(-u^2\right) du$$

ce qui termine la démonstration.

3- *Ecrivez la vraisemblance des données observées dans le cas d'un modèle censuré à droite ; vous rappelerez l'ensemble des notations avec précision. Quelle méthode proposez-vous pour résoudre les équations de vraisemblance ?*

On utilise l'expression générale $L(\theta) = \prod_{i=1}^{n} f_\theta(T_i)^{D_i} S_\theta(C_i)^{1-D_i}$ et les expressions obtenues à la question 1 de la densité et de la fonction de survie du modèle. Les équations de vraisemblance sont non linéaires et n'ont pas de solution analytique simple. On peut alors se tourner vers des algorithmes numériques comme par exemple l'algorithme EM, ou une résolution numérique directe des équations de vraisemblance.