ENSAE

ECOLE NATIONALE DE LA STATISTIQUE ET DE L'ADMINISTRATION ECONOMIQUE

Cursus Intégré
2004-2005

Rappels de statistique mathématique
Enoncé des travaux dirigés n°5

Guillaume Lacôte
Bureau **E03**
Guillaume.Lacote@ensae.fr
http://ensae.no-ip.com/SE222/

Enoncé de l'exercice 1

On souhaite évaluer et analyser le phénomène du chômage. Pour cela, on dispose de n obs vations sur les durées $y_i, 1 \leq i \leq n$, pendant lesquelles des individus sont restés sans emploi On suppose dans la suite que les variables aléatoires correspondantes $(Y_i)_{i\in[\![1,n]\!]}$ sont i.i.d. suivent la loi de Weibull de paramètres a et b. On rappelle que cette loi est continue sur $\mathbb{R}^+$ admet la fonction de répartition pour $y > 0$

$$F(y; a, b) = 1 - \exp\left(-ay^b\right)$$

On définit la fonction de survie par

$$S(y) = 1 - F(y)$$

et la fonction de hasard par $h(y) = \frac{f(y)}{S(y)}$.

Partie 1 **Généralités**

- Q1 Donner l'expression de la fonction de hasard du modèle.
- Q2 Quelle est en terme de chômage l'interprétation de la fonction de hasard ?
  Expliquer alors pourquoi il est important de considérer le cas particulier où cette fonction constante.
  Pour quelles valeurs des paramètres, la fonction de hasard est-elle constante ?
  Quelles sont alors les lois des durées de chômage ?
- Q3 Etudier l'évolution de la fonction de hasard en fonction de a, puis en fonction de b.

Partie 2 **Estimation contrainte**

On suppose dans cette partie $b = 1$. Le modèle est alors uniquement paramétré par a.

- Q1 Le modèle est-il exponentiel ? Si oui, expliciter une statistique exhaustive.
- Q2 Déterminer le vecteur du score et vérifier directement qu'il est centré.
- Q3 Quel est l'estimateur du maximum de vraisemblance $\hat{a}_0$ de a ?
  Est-il sans biais, y a-t-il surestimation ou sous-estimation systématique ?
- Q4 Déterminer la variance asymptotique de cet estimateur $\hat{a}_0$

Partie 3 **Estimation non contrainte**

On considère maintenant le cas où a et b peuvent a priori prendre toutes valeurs positives.

- Q1 Le modèle est-il exponentiel avec une statistique exhaustive dont la taille est indépendante nombre n d'observations ?
  Si oui, expliciter une telle statistique.
- Q2 Ecrire les équations de vraisemblance. Sont-elles résolubles sous forme analytique ?

☞ Q3 Donner la forme de la variance asymptotique de l'estimateur du maximum vraisemblance $\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix}$ du paramètre $\begin{pmatrix} a \\ b \end{pmatrix}$.

☞ Q4 En déduire la variance asymptotique de $\hat{a}$ lorsque $b = 1$.
Comparer alors les estimateurs $\hat{a}$ et $\hat{a}_0$ lorsque $b = 1$.
Quelle conclusion en tirer ?

☞ Q5 Quelle démarche pourrait-on proposer pour étudier la distribution de l'estimateur $\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix}$ lorsque l'échantillon est de petite taille, par exemple $n = 10$?

■ Partie 4 **Cas indépendant - non équidistribué**

On considère maintenant le cas de T observations $Y_1, \ldots, Y_T$ indépendantes, de lois respectives :

$$F(y; e^{\alpha t}, 1) \ , \ t \in [\![1, T]\!] \ , \ \alpha \in \mathbb{R}$$

☞ Q1 Déterminer la vraisemblance du modèle $\mathcal{L}$, et vérifier qu'elle est concave en α à $(y_1, \ldots, y_T)$ fixé.
En déduire l'équation caractérisant l'estimateur du maximum de vraisemblance $\widehat{\alpha_T}$ de α.

☞ Q2 On note $u_t = y_t - e^{-\hat{\alpha_T} t}$.
Donner l'interprétation de u_t.

☞ Q3 Montrer que l'équation de la vraisemblance correspond à la condition d'orthogonalité de $(u_1, \ldots, u_T)$ et de $1, \ldots, T$ pour un certain produit scalaire que l'on précisera.

Enoncé de l'exercice 2

On étudie entre les dates 0 et T un groupe de n individus sans emploi à la date 0, et on cherche à modéliser les durées de chômage $(T_i)_{i \in [\![1,n]\!]}$.
En pratique, on observe les durées de chômage en mois. Plus précisément, on ne dispose pas de la variable continue T_i, mais seulement de la variable discrète T_i^* donnée par

$$T_i^* = \lceil T_i \rceil$$

Autrement dit, la variable T_i^* vaut $t + 1$ si l'individu i a retrouvé du travail entre le t-ème et $(t + 1)$-ième mois.
En outre entre t et $t + 1$, on suppose que :
- l'individu i reçoit N_t^i offres d'emploi, où N_t^i est une suite de variables i.i.d. de loi de Poisson $\mathcal{P}(\lambda)$;

- si l'individu i est toujours au chômage à la date t, et si parmi les N_t^i offres qu'il reçoit, l'u au moins offre un salaire supérieur à une constante ξ_i, propre à l'individu (appelée salaire réserve), alors l'individu n'est plus au chômage : $(T_i^* = t + 1)$;
- les salaires des offres d'emploi sont tirés indépendamment des dates d'arrivée des offres et leur nombre dans une loi de fonction de répartition F.

On suppose dans un premier temps pour simplifier que $T = +\infty$.

☞ Q1 Calculer $\mathbb{P}\left(T_i^* = t + 1 \,|\, T_i > t\right)$ en fonction de F, ξ_i, λ .

☞ Q2 En déduire la vraisemblance de $(T_1^*, \ldots, T_n^*)$.

☞ Q3 On suppose que tous les individus ont le même salaire de réserve : $\xi_i = \xi$ et que ce salaire réserve commun ξ est connu, ainsi que la fonction de répartition F.
Trouver l'estimateur du maximum de vraisemblance de λ.
Montrer directement que cet estimateur est convergent et asymptotiquement efficace qua $n \to +\infty$.

☞ Q4 En pratique, l'enquête se termine à la fin du T-ième mois, $T < +\infty$: à cette date, certa individus sont encore au chômage ; on n'observe donc que :

$$\begin{aligned} T_i^{**} &= T_i^* \text{ si } T_i^* \leq T \\ &= T + 1 \text{ si } T_i^* > T \end{aligned}$$

Ecrire la vraisemblance des observations $(T_1^{**}, \ldots, T_n^{**})$.
Donner l'estimateur du maximum de vraisemblance de λ.
Remarque : On suppose toujours le salaire de réserve commun ξ et la fonction de répartitior connue.

☞ Q5 On suppose désormais que F s'écrit :

$$F(x) = \left(1 - e^{-\gamma(\xi - \xi_0)}\right) \mathbb{1}_{\xi \geq \xi_0}$$

où γ est un paramètre inconnu à estimer et ξ_0 est connu.
Le couple (λ, γ) est-il identifiable ?

Enoncé de l'exercice 3

■ Partie 1 **Préliminaire**

☞ Q1 On rappelle que la loi exponentielle de paramètre λ admet la densité $f(y, \lambda) = \begin{cases} \lambda e^{(-\lambda y)}, & \text{si } y \\ 0 \text{ sinon.} \end{cases}$
Soient n variables aléatoires $Y_1 \ldots Y_n$ i.i.d. de densité $f(\cdot, \lambda)$.
Montrer que $\sum_{i=1}^{n} Y_i$ suit une loi de densité $\lambda^n \dfrac{y^{n-1}}{(n-1)!} e^{-\lambda y} \mathbb{1}_{[0,+\infty]}(y)$ (on pourra commencer le montrer pour $n = 2$ puis procéder par récurrence).

Dans tout le problème, Z_i et C_i, pour $i \in [\![1,n]\!]$ désignent des variables aléatoires indépendantes suivant des lois exponentielles de paramètres respectifs $\lambda, \mu \in \mathbb{R}^{+*}$.

■ Partie 2 Observation parfaite

On dispose d'un échantillon d'observations $(z_i, c_i)_{i \in [\![1,n]\!]}$

☞ Q1 Ecrire le modèle statistique correspondant.
S'agit-il d'une famille exponentielle ? Si oui, peut-on exhiber une statistique exhaustive ?

☞ Q2 Quel est l'estimateur du maximum de vraisemblance $\begin{pmatrix} \hat{\lambda} \\ \hat{\mu} \end{pmatrix}$ du paramètre $\begin{pmatrix} \lambda \\ \mu \end{pmatrix}$?

☞ Q3 Déterminer la loi asymptotiquement de $\begin{pmatrix} \hat{\lambda} \\ \hat{\mu} \end{pmatrix}$.

☞ Q4 Déterminer la loi (à distance finie) de $\begin{pmatrix} \hat{\lambda} \\ \hat{\mu} \end{pmatrix}$.
Est-il biaisé à distance finie ?
Calculer sa matrice de variance-covariance.

☞ Q5 Proposer des estimateurs sans biais optimaux de λ et μ, si possible efficaces.

■ Partie 3 Observation imparfaite

On suppose dans cette seconde partie que les seules observations disponibles portent sur $X_i = \mathrm{Min}\,(Z_i, C_i)\,,\ i = 1, ..., n.$

☞ Q1 Calculer la fonction de répartition de la variable X_i pour $i \in [\![1,n]\!]$

☞ Q2 Ecrire le modèle statistique correspondant et déterminer les fonctions identifiables du paramètre $\begin{pmatrix} \lambda \\ \mu \end{pmatrix}$.

☞ Q3 Quels sont les estimateurs du maximum de vraisemblance de $\gamma = \lambda + \mu$ fondés :
i) sur $(X_i)_{i \in [\![1,n]\!]}$;
ii) sur $(Z_i, C_i)_{i \in [\![1,n]\!]}$?
Est-il naturel que ces estimateurs soient différents ?

☞ Q4 Comparer les propriétés asymptotiques de ces estimateurs.

■ Partie 4 Conclusion

☞ Q1 Déduire des parties **I** et **II** l'expression de l'espérance conditionnelle

$$\mathrm{E}\left(\frac{1}{\sum_{i=1}^n \min(Z_i, C_i)} \,\middle|\, \sum_{i=1}^n Z_i, \sum_{i=1}^n C_i \right)$$