Agrégation externe de mathématiques
Épreuve de modélisation, option A : Probabilités et Statistiques

---

(*Texte public*)

Résumé : Nous abordons certaines questions relatives à l'inférence statistique de données issues de modèles de survie, lorsque ces données sont censurées, c'est à-dire partiellement observées.

Mots clefs : Loi exponentielle, loi du Chi-deux, simulation de variables aléatoires, intervalle de confiance, test.

► *Il est rappelé que le jury n'exige pas une compréhension exhaustive du texte. Vous êtes laissé(e) libre d'organiser votre discussion comme vous l'entendez. Des suggestions de développement, largement indépendantes les unes des autres, vous sont proposées en fin de texte. Vous n'êtes pas tenu(e) de les suivre. Il vous est conseillé de mettre en lumière vos connaissances à partir du fil conducteur constitué par le texte. Le jury appréciera que la discussion soit accompagnée d'exemples traités sur ordinateur.*

1. Introduction

On souhaite mesurer l'influence d'un traitement médical sur des malades. Pour cela, on observe la réalisation de n variables aléatoires qui mesurent l'intervalle de temps entre la prise du traitement et la rechute de la maladie (ou pire) de chaque patient d'un groupe de n personnes malades. L'efficacité du traitement peut se "lire" sur la fonction de risque instantané de ce temps de "survie", selon la terminologie, que l'on définit ci-dessous. On va s'intéresser plus particulièrement au cas où, pour des raisons expérimentales, certaines données sont "censurées", dans un sens que l'on précisera ci-dessous.

1.1. *Fonction de risque instantané*

Si T est une variable aléatoire positive, absolument continue, de densité continue sur $]0,+\infty[$, on définit sa fonction de risque instantané $\lambda(t)$ à l'instant t par

$$\lambda(t) = \lim_{h\to 0, h>0} \frac{1}{h}\mathbb{P}\{t \le T < t+h \mid T \ge t\}, \tag{1}$$

lorsque cela a un sens. On parle indifféremment de la fonction de risque instantané de T ou de la loi de T.

Pour alléger la terminologie et les notations, nous ne distinguerons pas une variable aléatoire de sa réalisation. Nous nous autorisons l'abus de langage (et de notation) consistant à dire que l'on a "observé une variable aléatoire T".

Lemme 1. *La fonction de risque instantané caractérise la loi de T. La densité f de T s'obtient à partir de* (1) *par la formule*

$$f(t) = \lambda(t) \exp\big(-\int_0^t \lambda(u)du\big).$$

2. Inférence statistique en l'absence de censure

2.1. *Estimation de la fonction de risque instantané constante*

On observe un n-échantillon $(T_1, \ldots, T_n)$ de temps de rechute de malades après la prise d'un traitement médical. Les variables aléatoires T_i sont indépendantes et ont la même fonction de risque instantané, supposée constante égale à λ, quantité inconnue.

Pour estimer λ à partir de l'observation des T_i, le *principe du maximum de vraisemblance* consiste à définir la *fonction de vraisemblance*

$$\lambda \mapsto L_n(\lambda; T_1, \ldots, T_n) = f_\lambda(T_1) \cdots f_\lambda(T_n) \tag{2}$$

où $f_\lambda(x) = \lambda \exp(-\lambda x)$ désigne la densité commune des T_i. Puis, la valeur $\hat{\lambda}(T_1, \ldots, T_n)$ qui maximise $L_n(\lambda; T_1, \ldots, T_n)$, lorsqu'elle est bien définie, fournit un estimateur de λ, appelé *estimateur du maximum de vraisemblance* de λ. Ainsi, l'estimateur du maximum de vraisemblance est la valeur de λ qui maximise la densité conjointe

$$f_\lambda(t_1) \cdots f_\lambda(t_n)$$

du n échantillon au point $(t_1, \ldots, t_n) = (T_1, \ldots, T_n)$. Dans notre cas,

$$L_n(\lambda; T_1, \ldots, T_n) = \lambda^n \exp\left[-\lambda V_n(T_1, \ldots, T_n)\right],$$

avec $V_n(t_1, \ldots, t_n) = \sum_{i=1}^n t_i$.

Lemme 2. *L'estimateur du maximum de vraisemblance est bien défini et vaut*

$$\hat{\lambda}_n(T_1, \ldots, T_n) = \frac{n}{V_n(T_1, \ldots, T_n)}.$$

2.2. *Construction d'intervalles de confiance*

La transformée de Laplace de λT_j est donnée par

$$\xi \mapsto \mathbb{E}\{e^{-\xi \lambda T_j}\} = (1+\xi)^{-1},$$

de sorte que la transformée de Laplace de $\lambda V_n(T_1, \ldots, T_n)$ est $\xi \mapsto (1+\xi)^{-n}$. Donc la variable aléatoire $\lambda V_n(T_1, \ldots, T_n)$ suit une loi Gamma de paramètres n et 1. De manière équivalente, $2\lambda V_n(T_1, \ldots, T_n)$ suit une loi du Chi-deux à $2n$ degrés de liberté. Ces remarques permettent de construire des intervalles de confiance (non-asymptotiques) pour la valeur λ.

Lemme 3. *Pour tout $\alpha \in]0,1[$, notons $\chi^2_\alpha(2n)$ un quantile d'ordre α de la loi du Chi-deux à $2n$ degrés de liberté, c'est-à-dire tout nombre vérifiant*

$$\mathbb{P}\{Z \leq \chi^2_\alpha(2n)\} = \alpha,$$

où Z est une variable aléatoire qui suit la loi du Chi-deux à $2n$ degrés de liberté. Alors, les intervalles

$$[0, \frac{1}{2n}\hat{\lambda}_n(T_1,\dots,T_n)\chi^2_{1-\alpha}(2n)],$$

$$[\frac{1}{2n}\hat{\lambda}_n(T_1,\dots,T_n)\chi^2_{\alpha}(2n), +\infty[,$$

et

$$[\frac{1}{2n}\hat{\lambda}_n(T_1,\dots,T_n)\chi^2_{\alpha/2}(2n), \frac{1}{2n}\hat{\lambda}_n(T_1,\dots,T_n)\chi^2_{1-\alpha/2}(2n)]$$

sont des intervalles de confiance pour λ au niveau de confiance $1-\alpha$.

3. Inférence statistique en présence de censure

On suppose désormais que les n patients sont soumis au traitement, mais que l'on arrête l'expérience lorsque les d premiers patients rechutent. En effet, si n est grand et que le traitement est efficace, le protocole consistant à attendre que les n patients aient tous rechuté peut se révéler trop long dans la pratique.

On réordonne par ordre croissant les d premiers instants de rechute parmi les observations $(T_1,\dots,T_n)$, que l'on écrit

$$T_{(1,n)} \le T_{(2,n)} \le \dots \le T_{(d,n)}.$$

Les inégalités sont en fait strictes presque-sûrement et les variables $(T_{(1,n)}, T_{(2,n)},\dots,T_{(d,n)})$ constituent l'observation de l'expérience censurée.

Il s'agit désormais de mesurer la perte d'information de l'expérience censurée, où l'on observe $(T_{(1,n)}, T_{(2,n)},\dots,T_{(d,n)})$ contre l'expérience non censurée, où l'on observe $(T_1,\dots,T_n)$, mais plus difficile à réaliser dans la pratique.

Lemme 4. *La fonction de vraisemblance censurée $L^c_{n,d}(\lambda; T_{(1,n)},\dots,T_{(d,n)})$ est donnée par*

$$L^c_{n,d}(\lambda; T_{(1,n)},\dots,T_{(d,n)}) = \frac{\lambda^d n!}{(n-d)!}\exp\left[-\lambda V^c_{n,d}(T_{(1,n)},\dots,T_{(d,n)})\right],$$

où

$$V^c_{n,d}(t_1,\dots,t_d) = \sum_{i=1}^{d} t_i + (n-d)\,t_d.$$

De, plus, l'estimateur du maximum de vraisemblance est bien défini et vaut

$$\hat{\lambda}^c_{n,d}(T_{(1,n)},\dots,T_{(d,n)}) = \frac{d}{V^c_{n,d}(T_{(1,n)},\dots,T_{(d,n)})}.$$

La preuve du lemme s'obtient en calculant la densité conjointe des d premières valeurs réordonnées de $(T_1,\dots,T_n)$, que nous notons $g(t_1,\dots,t_d)$, définie sur $0 \le t_1 \le t_2 < \dots \le t_d$. Pour cela, nous proposons le raisonnement infinitésimal (et informel) suivant : on partitionne l'axe des temps en les intervalles $[0,t_1[$, $[t_1, t_1+dt_1[$, $[t_1+dt_1, t_2[$, …, $[t_d, t_d+dt_d[$, $[t_d+dt_d, +\infty[$. Puis, on "lance" au hasard n points dans ces intervalles, suivant le schéma

multinômial suivant : on affecte respectivement les probabilités $0, 1/n, 0, \ldots, 1/n, (n-d)/n$, pour chaque intervalle, écrit dans cet ordre. Ceci donne lieu à la probabilité infinitésimale

$$\frac{n!}{(n-d)!}\exp\left[-(n-d)\lambda t_d\right]\prod_{i=1}^{d}\lambda\exp\left(-\lambda t_i\right)dt_i,$$

d'où l'on déduit la densité

$$g(t_1,\ldots,t_d)=\frac{\lambda^d n!}{(n-d)!}\exp\left[-\lambda V^c_{n,d}(t_1,\ldots,t_d)\right]. \tag{3}$$

Le lemme découle alors facilement de la formule (3).

Pour construire un intervalle de confiance de λ à partir de $\hat{\lambda}^c_{n,d}(T_1,\ldots,T_n)$, on a besoin, comme précedemment, de la loi de $V^c_{n,d}(T_{(1,n)},\ldots,T_{(d,n)})$. Le changement de variable

$$u_i=(n-i+1)(t_i-t_{i-1}),\ \ i=1,\ldots,d \tag{4}$$

où l'on a posé $t_0=0$, a pour jacobien $(n-d)!/n!$. Il s'ensuit que la loi conjointe des variables aléatoires $U_i=(n-i+1)(T_{(i,n)}-T_{(i-1,n)})$, pour $i=1,\ldots,d$ et en convenant $T_{(0,n)}=0$, a pour densité

$$(u_1,\ldots,u_d)\mapsto\prod_{i=1}^{d}\lambda e^{-\lambda u_i}$$

pour $u_i\in\mathbb{R}_+$. Donc

$$2\lambda V^c_{n,d}(T_{(1,n)},\ldots,T_{(d,n)})=2\lambda\sum_{i=1}^{d}U_i$$

suit une loi du Chi-deux à $2d$ degrés de liberté, et la construction d'un intervalle de confiance est la même que pour le cas non-censuré. Il est remarquable que, dans ce cas précis où la fonction de risque instantané est constante, la même précision statistique est obtenue en observant d patients jusqu'à ce qu'ils aient tous rechuté, ou n patients jusqu'au temps de rechute des d premiers.

4. Vers un test d'exponentialité dans le cas de censure

La transformation (4) du paragraphe précédent permet d'aller plus loin dans l'analyse de la fonction de risque instantané. En écrivant, pour $i=1,\ldots,d$

$$T_{(i,n)}=\frac{U_1}{n}+\frac{U_2}{n-1}+\ldots+\frac{U_i}{n-i+1},$$

on montre que

$$\mathbb{E}\{T_{(i,n)}\}=\frac{1}{\lambda}\sum_{j=1}^{i}\frac{1}{n-j+1}.$$

Ceci suggère une méthode simple pour tester l'hypothèse que la fonction de risque instantanée $\lambda(t)$ est constante en présence de censure : le nuage de points des variables $T_{(i,n)}$ en fonction des $\sum_{j=1}^{i}(n-j+1)^{-1}$ pour $i=1,\ldots,d$ est grossièrement sur une droite.

Suggestions pour le développement

► *Soulignons qu'il s'agit d'un menu à la carte et que vous pouvez choisir d'étudier certains points, pas tous, pas nécessairement dans l'ordre, et de façon plus ou moins fouillée. Vous pouvez aussi vous poser d'autres questions que celles indiquées plus bas. Il est très vivement souhaité que vos investigations comportent une partie traitée sur ordinateur et, si possible, des représentations graphiques de vos résultats.*

— On pourra préciser les preuves des lemmes 1, 2 et 3.
— On pourra donner une preuve de la convergence de $\hat{\lambda}_n(T_1,\dots,T_n)$ vers λ (dans un sens que l'on précisera) et exhiber la loi limite de $\alpha_n(\hat{\lambda}_n(T_1,\dots,T_n)-\lambda)$, pour une certaine normalisation $\alpha_n \to +\infty$ lorsque $n \to \infty$, que l'on pourra préciser.
— On pourra simuler les variables $(T_1,\dots,T_n)$, et se convaincre de la pertinence de l'estimateur du maximum de vraisemblance ainsi que des intervalles déduits dans le lemme 3 (en présence ou absence de censure).
— On pourra donner une preuve précise du lemme 4 (sans nécessairement suivre la méthode suggérée).
— On pourra mettre en oeuvre le test d'exponentialité du paragraphe 4, en réfléchissant (en particulier) à la simulation d'une variable aléatoire positive dont la fonction de risque instantané $\lambda(t)$ n'est pas constante.