TD 1: APPROCHE PARAMÉTRIQUE DE LA SURVIE.

Modèle de duréee M1 DUAS– Semestre 2
P.-O. Goffard

---

1. L'objectif est de déterminer l'estimateur du maximum de vraisemblance de la loi exponentielle $\mathsf{Exp}(\beta)$ de densité

$$f(t) = \frac{\mathrm{e}^{-t/\beta}}{\beta}\mathbb{I}_{(0,\infty)}(t).$$

Soit $\mathcal{D} = \{t_1, \dots, t_n\}$ un échantillon i.i.d. de réalisations d'une v.a. exponentielle.

(a) Donner l'estimateur du maximum de vraisemblance $\widehat{\beta}$.

Solution: La log vraisemblance s'écrit

$$l(\mathcal{D};\beta) = -n\log(\beta) - \frac{1}{\beta}\sum t_i.$$

On a

$$\frac{\partial l}{\partial \beta}(\mathcal{D};\beta) = -n\frac{n}{\beta} + \frac{1}{\beta^2}\sum t_i.$$

On en déduit que

$$\frac{\partial l}{\partial \beta}(\mathcal{D};\beta) = 0 \Leftrightarrow \widehat{\beta} = \frac{\sum t_i}{n},$$

après avoir vérifié que

$$\frac{\partial l}{\partial \beta}(\mathcal{D};\widehat{\beta}) < 0.$$

(b) Supposons que nous soyons en présence de données censurées, avec une censure à droite, non informative, de niveaux $c_1, \dots, c_n$. Donner l'estimateur du maximum de vraisemblance de β.

Solution: Les observations sont

$$(x_i, \delta_i) = (t_i \wedge c_i, \mathbb{I}_{t_i \leq c_i}),$$

et la log vraisemblance s'écrit

$$-\sum \delta_i \log(\beta) - \frac{\sum x_i}{\beta} + \text{Cste}.$$

On en déduit que

$$\frac{\partial l}{\partial \beta}(\mathcal{D};\beta) = 0 \Leftrightarrow \widehat{\beta} = \frac{\sum x_i}{\sum \delta_i}.$$

On peut vérifier la cohérence de l'estimation. En prenant $C \to \infty$, on retombe sur l'estimateur obtenu à la question 1.

(c) Supposons que nous soyons en présence de données tronquées, avec une troncature à gauche, non informative, de niveaux c pour $i = 1, \dots, n$. Donner l'estimateur du maximum de

vraisemblance de β.

Solution: La densité de probabilité pour des observation tronquées à gauche au niveau c est donnée par

$$f_{[c,\infty)]}(t) = \frac{\mathrm{e}^{-\beta t}}{\beta \mathrm{e}^{-\beta c}} \mathbb{I}_{[c,\infty)]}(t).$$

La log-vraisemblance s'écrit

$$l(\mathcal{D};\beta) = -n\ln(\beta) - \left[\sum_i t_i - nc\right]/\beta.$$

On en déduit que

$$\frac{\partial l}{\partial \beta}(\mathcal{D};\beta) = 0 \Leftrightarrow \widehat{\beta} = \frac{\sum x_i}{n} - c.$$

On peut vérifier la cohérence de l'estimation. En prenant $c = 0$, on retombe sur l'estimateur obtenu à la question 1.

2. Soit T la durée d'un évènement, la durée de vie résiduelle X d'un évènement ayant commencé il y a θ unités de temps est définie par

$$X_\theta \sim T - \theta | T > \theta.$$

Il s'agit par exemple de la durée de vie résiduelle d'un individu d'âge θ. On notera $S_\theta(x) = \mathbb{P}(X_\theta > x)$

(a) Exprimer l'espérance de vie $\mathbb{E}(T)$ en fonction d'une intégrale de $S(t) = \mathbb{P}(T > t)$

Solution: On ré-écrit la v.a. T comme

$$T = \int_0^\infty \mathbb{I}_{[0,T]}(t)\mathrm{d}t$$

puis par définition de l'espérance et le théorème de Fubini-Tonnelli, il vient

$$\begin{aligned}\mathbb{E}(T) &= \int_\Omega \int_0^\infty \mathbb{I}_{[0,T]}(t)\mathrm{d}t\mathrm{d}\mathbb{P} \\ &= \int_0^\infty \int_\Omega \mathbb{I}_{[0,T]}(t)\mathrm{d}\mathbb{P}\mathrm{d}t \\ &= \int_0^\infty S(t)\mathrm{d}t.\end{aligned}$$

(b) Exprimer la fonction de durée de vie résiduelle $\mathbb{E}(X_\theta) := \mathbb{E}(X - \theta | X > \theta)$ en fonction d'une intégrale de $S(t)$.

Solution: On peut commencer par donner la fonction de survie de X_θ en fonction de T, il vient

$$S_\theta(x) = \mathbb{P}(X - \theta > x | X > \theta) = \mathbb{P}(X > x + \theta | X > \theta) = \frac{S(x+\theta)}{S(\theta)}.$$

On utilise ensuite la question précédente pour écrire

$$\mathbb{E}(X_\theta) = \int_0^\infty \frac{S(x+\theta)}{S(\theta)}\mathrm{d}x.$$

(c) A partir de maintenant, nous supposons que $X_\theta \sim \mathsf{Par}(\alpha, \theta)$ (loi de Pareto), avec $\alpha > 1$ tel que

$$S_\theta(x) = \theta^\alpha (x+\theta)^{-\alpha},\ x > 0.$$

Calculer $\mathbb{E}(X_\theta)$. Pensez-vous que ce modèle est adapté à la durée de vie humaine?

Solution:

$$\mathbb{E}(X_\theta) = \int_0^\infty S_\theta(x)\mathrm{dx} = \frac{\theta}{\alpha - 1}.$$

Ce modèle n'est pas adapté car la durée de vie résiduelle augmente avec l'âge !

(d) Soit un évènement ayant duré θ unités de temps, exprimer la fonction de survie $S_{Y_x}(t)$ de

$$Y_x \sim X_\theta - x | X_\theta > x$$

il s'agit de la durée résiduelle dans x unités de temps d'un évènement ayant déjà durée θ unités de temps. Quelle remarque peut-on faire?

Solution: Soit $0 < x < y$, on a

$$S_{Y_x}(y) = \frac{S_\theta(y+x)}{S_\theta(x)} = \left(\frac{\theta + x}{\theta + y}\right)^\alpha = \left(\frac{\theta + x}{\theta + x + y}\right)^\alpha.$$

On observe que la durée de vie résiduelle pour une loi de Pareto est encore distribué suivant une loi de Pareto puisque

$$Y_x \sim \mathsf{Par}(\alpha, \theta + x).$$

(e) Soit un groupe de n évènements ayant déja durés $\theta_1, \ldots, \theta_n$ unités de temps respectivement. Nous observons les durées de vie résiduelles suivantes $\mathcal{D} = \{x_1, \ldots, x_n\}$, donner l'estimateur du maximum de vraisemblance de α.

Solution: La fonction de vraisemblance s'écrit

$$L(\mathcal{D}; \alpha) = \prod_{i=1}^n f_{\theta_i}(x_i),$$

où

$$f_\theta(x) = -S'_\theta(x) = \alpha \theta^\alpha (x+\theta)^{-\alpha-1}.$$

La log-vraisemblance s'écrit donc

$$l(\mathcal{D}; \alpha) = \alpha \sum_{i=1}^n \log(\theta_i) + n \log(\alpha) - (\alpha + 1) \sum_{i=1}^n \log(\theta_i + x_i)$$

L'estimateur du maximum de vraisemblance est donnée par

$$\widehat{\alpha} = \frac{n}{\sum_{i=1}^n \log\left(1 + \frac{\theta_i}{x_i}\right)}.$$

(f) Supposons que notre groupe d'évènement n'est observé que pendant c années. Donner l'estimateur du maximum de vraisemblance de α pour les données censurées à droite avec

$$z_i = x_i \wedge c,\ i = 1, \ldots, n.$$

Solution: On note $\delta_i = \mathbb{I}_{x_i<c}$ et $\mathcal{D} = \{(z_i, \delta_i)\}_{i=1,\dots,n}$. La fonction de vraisemblance s'écrit

$$L(\mathcal{D};\alpha) = \prod_{i=1}^{n} h_{\theta_i}(z_i)^{\delta_i} S_{\theta_i}(z_i),$$

où

$$h_\theta(x) = \frac{f_\theta(x)}{S_\theta(x)} = \frac{\alpha}{x+\theta}.$$

On en déduit la log-vraisemblance avec

$$l(\mathcal{D};\alpha) = \log(\alpha)\sum_{i=1}^{n}\delta_i - \sum_{i=1}^{n}\delta_i \log(z_i+\theta_i) + \alpha\sum_{i=1}^{n}\log(\theta_i) - \alpha\sum_{i=1}^{n}\log(\theta_i+z_i).$$

L'estimateur du maximum de vraisemblance est donné par

$$\widehat{\alpha} = \frac{\sum_i \delta_i}{\log\left(1+\frac{\theta_i}{z_i}\right)}$$

3. Nous allons construire un test d'adéquation à la loi exponentielle $\mathsf{Exp}(\beta)$ de densité

$$f(t) = \frac{\mathrm{e}^{-t/\beta}}{\beta}\mathbb{I}_{(0,\infty)}(t),$$

basé sur la transformée de Laplace. La transformée de Laplace d'une v.a. T est donnée par

$$\psi(\theta) = \mathbb{E}(\mathrm{e}^{-\theta T}).$$

(a) Soit $t_1, \dots, t_n$ un échantillon i.i.d. de réalisation de $T \sim \mathsf{Exp}(\beta)$. L'estimateur du maximum de vraisemblance de β est donnée par

$$\widehat{\beta}_n = \frac{1}{n}\sum_{i=1}^{n} t_i.$$

Quelle est la distribution de $T/\widehat{\beta}_n$ lorsque $n \to \infty$?

Solution: $T/\widehat{\beta}_n \xrightarrow{D} \mathsf{Exp}(1)$, il s'agit d'une application du théorème de Slutsky `https://en.wikipedia.org/wiki/Slutsky%27s_theorem`.

(b) Donner l'expression de la transformée de Laplace $\psi(\theta) = \mathbb{E}(e^{-\theta Y})$ de $Y \sim \mathsf{Exp}(\beta = 1)$ et montrer qu'elle vérifie l'équation différentielle suivante

$$\psi'(\theta)(1+\theta) + \psi(\theta) = 0$$

Solution: On a

$$\psi(\theta) = \frac{1}{1+\theta}$$

et

$$\psi'(\theta) = -\frac{1}{(1+\theta)^2} = -\frac{1}{1+\theta}\psi(\theta)$$

(c) On définit $y_i = t_i/\widehat{\beta}_n$ et

$$\psi_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} e^{-\theta y_i}$$

A quoi correspond cette quantité?

Solution: Il s'agit d'un estimateur asymptotiquement sans biais de la transformée de Laplace d'une loi $\mathsf{Exp}(1)$.

(d) On définit la statistique de test suivante

$$S_n = n\int_0^\infty [\psi_n'(\theta)(1+\theta) + \psi_n(\theta)]^2 \mathrm{e}^{-\theta}\mathrm{d}\theta,$$

Quel est votre interprétation de cette statistique de test?

Solution: On retrouve dans le carré sous l'intégrale, l'équation différentielle satisfaite par ψ dans le cas où les données suivent une loi exponentielle. On intègre pour prendre en compte l'ensemble du domaine de définition de la transformée de Laplace. On ajoute la fonction exponentielle pour rendre la statistique de test intégrable. L'avantage est d'obtenir une statistique de test explicite, voir la question d'après.

(e) Montrer que l'on peut estimer S_n par

$$S_n = \frac{1}{n}\sum_{j,k=1}^{n}\left[\frac{(1-y_j)(1-y_k)}{y_j+y_k+1} - \frac{y_j+y_k}{(y_j+y_k+1)^2} + \frac{2y_jy_k}{(y_j+y_k+1)^2} + \frac{2y_jy_k}{(y_j+y_k+1)^3}\right].$$

Indication: Il faut développer le carré sous l'intégrale et calculer chaque terme séparément. C'est un peu fastidieux mais on y arrive.

Solution: Il s'agit d'un petit calcul intégral, je vous fait confiance.

(f) Illustrer ce test avec R (comparer son efficacité à celui de Kolmogorov-Smirnov).
On fixe le niveau du test à 0.05. Calculer la probabilité de rejeter l'hypothèse

$$(H_0) : T \sim \mathsf{Exp}(\beta)$$

lorsque les données sont des réalisations i.i.d. de loi $\mathsf{Gamma}(\alpha, 1)$. Il s'agit de la puissance du test. Evaluer la puissance du test pour $\alpha = 1/4, 1/2, 3/4, 1, 5/4, 3/2, 7/4.$ et $n = 50$. On fera le graphique de la puissance en fonction de α avec une courbe pour le test présenté dans l'exercice (baptisé LT test) et le test de Kolmogorov.

Indications: Voici les étapes de l'algorithme

1. Simuler $t_i \sim \mathsf{Gamma}(\alpha, 1)$ pour $i = 1, \ldots, n$
2. Estimer β par $\widehat{\beta}$
3. Définir $y_i = t_i/\widehat{\beta}$ et calculer S_n
4. Simuler $\tilde{t}_i \sim \mathsf{Exp}(\widehat{\beta})$
5. Estimer β par $\tilde{\beta}$
6. Définir $\tilde{y}_i = \tilde{t}_i/\tilde{\beta}$ et calculer $\tilde{S}_n$ sur la base des $\tilde{y}_i$

Répéter ces étapes $J = 1000$ fois. On a une suite de valeur de test statistiques S_n^j et $\tilde{S}_n^j$ pour $j = 1, \ldots, J$. La valeur critique du test est donnée par

$$S_{0.95} = \text{Quantile}(\tilde{S}_n^j,\ j = 1, \ldots, J; 0, 95),$$

qui correspond au quantile empirique d'ordre 95% des $\tilde{S}_n^j$. La puissance du test est alors

$$\frac{1}{J} \sum_{j=1}^{J} \mathbb{I}_{S_n^j > S_{0.95}}.$$

Vous devriez obtenir un résultat proche de celui de la Figure 1.

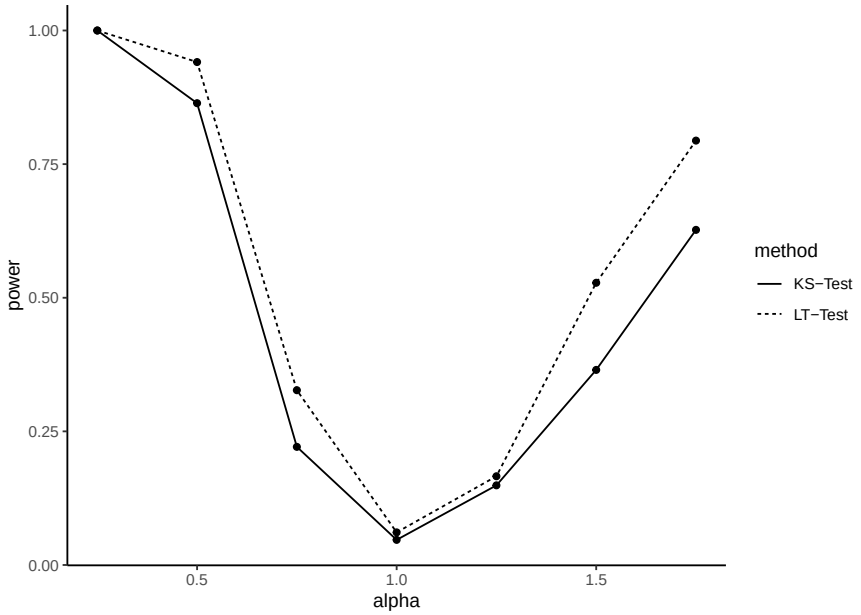

Figure 1: Puissance du LT test et du test de K-S.

(g) Commenter le résulat de la Figure 1.

- La forme de la courbe est elle celle attendue?
- Quel est le meilleur test?

Pour plus d'information sur ce test, on pourra se référer au travail de Henze et Meintanis [1]

Solution:

- La forme de la courbe est celle attendue, pour $\alpha = 1$ les données provienne du modèle exponentielle, la puissance des test doit donc atteindre le niveau fixé à 0.05 et augmenté de part et d'autre de $\alpha = 1$
- Le LT test semble supérieur. Il est tout à fait possible qu'un résultat inverse soit obtenu si les données proviennent d'un modèle de Weibull par exemple. On peut conclure que le LT test semble meilleur pour discriminer des données qui proviennent de la loi gamma.

References

[1] Norbert Henze and Simos G. Meintanis. Tests of fit for exponentiality based on the empirical laplace transform. *Statistics*, 36(2):147–161, jan 2002.