Probabilités Appliquées
Janvier 2016

Durée 3h00 – *Une feuille de format A4 manuscrite autorisée. Pas de calculatrices. Les deux exercices sont indépendants. Le nombre de points noirs donne une indication (subjective) de la difficulté de chaque question. Veuillez inscrire votre numéro de groupe sur la copie.*

On rappelle que l'espérance de la loi exponentielle de paramètre λ est égale à $1/\lambda$ et que la variance est égale à $1/\lambda^2$.

Exercice I. Attention aux mélanges – On considère une variable aléatoire de loi de densité

$$\forall y \in \mathbb{R}, \qquad f(y) = y\mathbb{1}_{[0,1)}(y) + \frac{1}{2}e^{1-y}\mathbb{1}_{[1,\infty)}(y)$$

1) ○○ Soit $F(t)$ la fonction de répartition de cette loi. Montrer, pour tout $t \in [0,1]$, que nous avons

$$F(t) = \frac{1}{2}t^2\,,$$

et que, pour tout $t \geq 1$, nous avons

$$F(t) = 1 - \frac{1}{2}e^{1-t}\,.$$

2) ○○ Soit X une variable aléatoire de loi exponentielle de paramètre 1. Montrer que la fonction de répartition de la variable aléatoire $Y_1 = \exp(-X/2)$ vérifie

$$F_1(t) = t^2\,,$$

pour tout $t \in [0,1]$.

3) ○○ Montrer qu'il existe $p \in (0,1)$ tel que

$$\forall t \in \mathbb{R}, \qquad F(t) = pF_1(t) + (1-p)F_2(t)\,,$$

où $F_2(t)$ est la fonction de répartition de la variable aléatoire $Y_2 = 1 + X$.

4) •○ Soit p la valeur trouvée précédemment. On considère la variable aléatoire Y définie par

$$Y = V\sqrt{U} + (1-V)(1+X)$$

où U est une variable aléatoire de loi uniforme sur (0,1), V est une variable aléatoire de Bernoulli de paramètre p et U, V, X sont mutuellement indépendantes. Calculer l'espérance des variables aléatoires Y et Y^2.

5) ○○ Calculer la fonction de répartition de la variable aléatoire Y.

6) •○ On dispose d'un générateur aléatoire retournant **uniquement** des variables aléatoires indépendantes de loi exponentielle de paramètre 1. Ecrire un algorithme de simulation de la loi de densité $f(y)$ (on notera `REXP` le générateur aléatoire de loi exponentielle).

Exercice II. Dur comme le ROC – *Les fonctions de répartition des lois considérées dans cet exercice sont continues.*

On dit qu'une famille de variables aléatoires $(Z_\lambda)_{\lambda>0}$ est *renormalisable* s'il existe une variable aléatoire Z_0 de variance égale à 1 et, pour tout $\lambda > 0$, un scalaire $s(\lambda) > 0$ tels que les variables aléatoires $Z_\lambda/s(\lambda)$ et Z_0 ont même loi.

1) *Propriétés et exemples de familles renormalisables.*

   a) ∘∘ Soit (Z_λ) une famille de variables aléatoires renormalisable. Montrer que $\mathrm{Var}(Z_\lambda) = s(\lambda)^2$ pour tout $\lambda > 0$.

   b) •∘ Montrer que la famille de loi exponentielle de paramètre λ est renormalisable.

   c) •∘ Soit (Z_λ) une famille de variables aléatoires de loi normale de moyenne nulle et de variance λ^2. Montrer que cette famille est renormalisable.

   d) ∘∘ Soit (Z_λ) une famille de variables aléatoires renormalisable. Montrer, sans calcul, que la famille de variables aléatoires (Z_λ^2) est renormalisable.

On suppose désormais que (Z_λ) est une famille renormalisable de variables aléatoires **positives** telle que $s(\lambda) > 1$. On convient que Z_λ est la variable d'intérêt, et que Z_0 est une variable de référence, positive elle aussi.

2) •∘ *Dominance stochastique.* Soit $t \geq 0$. Montrer que

$$\mathrm{P}(Z_\lambda > t) \geq \mathrm{P}(Z_0 > t).$$

   En déduire que $\mathrm{E}[Z_\lambda] \geq \mathrm{E}[Z_0]$.

3) *Notion de u-valeur.* Soit $F_0(t)$ la fonction de répartition de la variable Z_0.

   a) •∘ On pose $U_0 = 1 - F_0(Z_0)$. Montrer que U_0 suit la loi uniforme sur (0,1).

   b) •• On pose $U_1 = 1 - F_0(Z_\lambda)$. On note $G_1(\theta)$ la fonction de répartition de la variable aléatoire U_1. Montrer que

$$\forall \theta \in (0,1), \quad G_1(\theta) \geq \theta\,.$$

   Note : On pourra poser $t = F_0^{-1}(1-\theta)$ et utiliser le résultat de la question 2.

4) *Taux de faux positifs, puissance et courbe ROC.* Soit un entier $n \geq 2$ et $\lambda > 0$. On suppose l'ensemble $\{1, \ldots, n\}$ partitionné en deux sous-ensembles I_0 et I_1 de cardinaux respectifs n_0 et n_1 ($n_1 = n - n_0$). Les sous-ensembles I_0 et I_1 sont inconnus. On considère n variables aléatoires indépendantes, $\tilde{Z}_1, \ldots, \tilde{Z}_n$, définies de la manière suivante. Sachant $i \in I_0$, $\tilde{Z}_i$ est une variable aléatoire de même loi que Z_0. Sachant $i \in I_1$, $\tilde{Z}_i$ est une variable aléatoire de même loi que Z_λ.

On considère les u-valeurs associées aux variables $\tilde{Z}_i$ de la manière suivante

$$\forall i = 1, \ldots, n, \quad V_i = 1 - F_0(\tilde{Z}_i)\,.$$

L'objectif de cette question est détecter les variables d'intérêt ($i \in I_1$) à l'aide d'un test portant sur les u-valeurs, puis d'évaluer la performance du test par un indice d'aire.

Soit $\theta \in (0, 1)$. Un test est défini de la manière suivante. Pour tout i, on déclare que i est *positif* si ($V_i \leq \theta$) et on déclare que i est *négatif* sinon. De plus, on dit que i est un *faux positif* si i est positif et si $i \in I_0$. On dit que i est un *vrai positif* si i est positif et si $i \in I_1$.

a) •◦ On définit le taux de faux positifs, $\mathrm{TFP}(\theta)$, comme le nombre de faux positifs divisé par n_0. Montrer que $\mathrm{TFP}(\theta)$ est une variable aléatoire étagée. Montrer que son espérance est égale à

$$\mathrm{E}[\mathrm{TFP}(\theta)] = \theta\,,$$

et calculer sa variance.

b) •◦ On définit la puissance, $\mathrm{Pow}(\theta)$, comme le nombre de vrais positifs divisé par n_1. Montrer que $\mathrm{Pow}(\theta)$ est une variable aléatoire étagée. Montrer que son espérance est égale à

$$\mathrm{E}[\mathrm{Pow}(\theta)] = G_1(\theta)\,,$$

et calculer sa variance.

c) •• On définit la courbe ROC comme l'espérance de la puissance exprimée en fonction de θ. Représenter graphiquement la forme de cette courbe. Montrer que l'aire sous la courbe, notée AUC, est égale à

$$\mathrm{AUC} = \mathrm{E}[\mathrm{Pow}(U_0)],$$

où U_0 est une variable aléatoire de loi uniforme sur (0,1) indépendante des V_i.

d) •• Montrer que

$$1 - \mathrm{AUC} = \mathrm{E}[U_1].$$

En déduire l'égalité suivante

$$\mathrm{AUC} = \mathrm{P}(Z_\lambda > Z_0),$$

où Z_λ et Z_0 sont respectivement des copies indépendantes de la variable aléatoire d'intérêt et de la variable de référence.

Note : On pourra se souvenir de la loi de la variable aléatoire $F_0^{-1}(1-U_0)$ présentée lors de la description de l'algorithme d'inversion.

e) ◦◦ Montrer que AUC est comprise entre 1/2 et 1. Que peut-on dire de la détection des variables d'intérêt dans les deux cas extrêmes ?