Examen : Novembre 2015.

durée : 3h

Seul document autorisé : une feuille manuscrite recto-verso.
En particulier : calculatrice, téléphone, ordinateur portable, polycopié de cours, feuilles de TD et corrections sont interdits.

On pourra utiliser les résultats suivants sur les distributions usuelles :

1. **Loi Exponentielle** Une variable aléatoire Z suit un loi exponentielle de paramètre $\lambda > 0$, notée $\mathcal{E}(\lambda)$, si elle admet pour fonction de répartition et densité par rapport à Lebesgue, respectivement
$$F_\lambda(z) = \mathbb{1}_{z>0}(1 - e^{-\lambda z}) \quad ; \quad f_\lambda(z) = \mathbb{1}_{z>0}\lambda e^{-\lambda z}.$$
Son espérance et sa variance sont respectivement
$$\mathbb{E}_\lambda(Z) = \frac{1}{\lambda} \quad ; \quad \mathbb{V}\mathrm{ar}_\lambda(Z) = \frac{1}{\lambda^2}.$$

2. **Loi Gamma** Une variable aléatoire Y suit une loi Gamma de paramètres α et λ ($\alpha > 0$ et $\lambda > 0$), notée $\mathcal{G}amma(\alpha, \lambda)$, si elle admet une densité par rapport à la mesure de Lebesgue donnée par
$$f^{\mathcal{G}}_{(\alpha,\lambda)}(y) = \mathbb{1}_{y>0}\frac{\lambda^\alpha}{\Gamma(\alpha)}y^{\alpha-1}e^{-\lambda y}.$$
On rappelle que pour $\alpha > 0$, $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$. Si $Y \sim \mathcal{G}amma(\alpha, \lambda)$, on a
$$\mathbb{E}_{\alpha,\lambda}(Y) = \frac{\alpha}{\lambda} \quad ; \quad \mathbb{V}\mathrm{ar}_{\alpha,\beta}(Y) = \frac{\alpha}{\lambda^2}.$$

3. **Loi Inverse Gamma** Si $Y \sim \mathcal{G}amma(\alpha, \lambda)$, alors $T := \frac{1}{Y}$ suit une loi dite 'inverse gamma' $\mathcal{IG}(\alpha, \lambda)$, de densité
$$f^{\mathcal{IG}}_{\alpha,\lambda}(t) = \mathbb{1}_{t>0}\frac{\lambda^\alpha}{\Gamma(\alpha)}\frac{1}{t^{\alpha+1}}e^{-\lambda/t},$$
et l'on a, lorsque $\alpha > 1$ (*resp.* $\alpha > 2$) :
$$\mathbb{E}_{\alpha,\lambda}(T) = \frac{\lambda}{\alpha - 1} \quad ; \quad (\text{resp. } \mathbb{V}\mathrm{ar}_{\alpha,\lambda}(T) = \frac{\lambda^2}{(\alpha-1)^2(\alpha-2)}.)$$

4. **Somme d'exponentielles** Si $(Z_i)_{i=1}^n \overset{i.i.d.}{\sim} \mathcal{E}(\lambda)$, alors $Y := \sum_{i=1}^n Z_i$ suit une loi Gamma de paramètres (n, λ) :
$$\sum_{i=1}^n Z_i \sim \mathcal{G}amma(n, \lambda).$$

Pour la modélisation du trafic internet, on utilise couramment la loi de Pareto pour représenter la distribution de la taille des paquets envoyés, parmi les paquets dont la taille dépasse un certain seuil u suffisamment élevé. Par définition, une variable aléatoire X_1 suit une loi de Pareto $\mathcal{P}ar(u,\theta)$, avec $u>0$ et $\theta>0$, si la fonction de répartition de X_1 est

$$F_\theta(x)=\begin{cases}0 & \text{si } x\le u\\ 1-\left(\frac{u}{x}\right)^\theta & \text{si } x>u.\end{cases}$$

Dans toute la suite de l'exercice on fixe $u>0$, supposé connu et on considère le modèle statistique

$$\mathcal{P}=\{P_\theta,\theta\in\Theta\},\quad \text{avec } \Theta=]0,\infty[,$$

où P_θ désigne la loi de Pareto $\mathcal{P}ar(u,\theta)$. On considère alors n observations indépendantes $X=(X_1,\dots,X_n)$, où $X_i \overset{i.i.d.}{\sim} \mathrm{P}_\theta$ $(1\le i\le n)$ avec $n\ge 3$. On admet que le modèle est régulier, au sens où les hypothèses du théorème de Cramér-Rao sont satisfaites.

Les parties 1,2,3,4 et 5 sont "indépendantes", au sens où il n'est pas nécessaire d'avoir répondu aux questions des parties précédentes pour traiter une partie donnée. Cependant certaines parties font appel aux résultats (donnés dans l'énoncé) de parties précédentes.

1 Estimateur du maximum de vraisemblance

1. Montrer que la loi $\mathcal{P}ar(u,\theta)$ admet comme densité par rapport à la mesure de Lebesgue :

$$p_\theta(x)=\mathbb{1}_{\{x>u\}}\,\theta\frac{u^\theta}{x^{\theta+1}}$$

2. Exprimer la log-vraisemblance $\log p_\theta^{\otimes n}(\mathbf{x})$, pour $\mathbf{x}=(x_1,\dots,x_n)\in U\subset\mathbb{R}^n$ où U est le domaine (que l'on précisera), où cette quantité est finie.
3. Montrer que l'estimateur de maximum de vraisemblance $\widehat{\theta}_{MV}(X)$ pour le paramètre θ s'écrit

$$\widehat{\theta}_{MV}(X)=\frac{n}{\sum_{i=1}^n\log\left(\frac{X_i}{u}\right)}.$$

2 Risque quadratique

On se demande maintenant si cet estimateur est « bon », au sens du risque quadratique

1. Montrer que $Y_i=\log(X_i/u)$ suit une loi exponentielle dont on précisera le paramètre.
2. Montrer que si une variable aléatoire $V\sim\mathcal{G}amma(\alpha,\lambda)$, alors $\lambda V\sim\mathcal{G}amma(\alpha,1)$.
3. Montrer ensuite que $\widehat{\theta}_{MV}(X)\sim\mathcal{IG}(n,n\theta)$.
4. Quelle est le biais de $\widehat{\theta}_{MV}$? Quel est son risque quadratique ?
5. Montrer que l'estimateur 'corrigé'

$$\widehat{\theta}_c(X)=\frac{n-1}{\sum_{i=1}^n\log(X_i/u)}$$

est un estimateur sans biais de θ.

6. Calculer l'information de Fisher $I_n(\theta)$ pour le modèle $\mathcal{P}^{\otimes n}$ à n observations indépendantes.
7. Calculer la variance de $\widehat{\theta}_c(X)$.
8. Les estimateurs $\widehat{\theta}_{MV}$ et $\widehat{\theta}_c$ sont -ils efficaces ?
9. Si l'utilisateur est sensible au risque quadratique sur l'erreur d'estimation, quel estimateur doit-il préférer : $\widehat{\theta}_c$ ou $\widehat{\theta}_{MV}$?

3 Construction d'un intervalle de confiance

1. En s'appuyant sur le résultat des questions 2.1 et 2.2, donner une fonction pivotale $\varphi : \mathbb{R}^n \times \Theta \to \mathbb{R}$ telle que en posant $Z = \varphi(X, \theta)$, on ait pour tout θ,
$$Z \sim \mathcal{G}amma(n, 1).$$
2. Pour $\alpha \in]0, 1[$, on note $q_n(\alpha)$ le quantile d'ordre α de la loi $\mathcal{G}amma(n, 1)$. Autrement dit, $q_n(\alpha)$ est le nombre réel tel que,
$$\text{si } Z \sim \mathcal{G}amma(n, 1), \text{ alors} \quad \mathbb{P}(Z \leq q_n(\alpha)) = \alpha.$$
Justifier l'existence et l'unicité de $q_n(\alpha)$.
3. Donner un intervalle de confiance $I(X) = [L(X), R(X)]$ pour θ, de niveau $1 - \alpha$, basé sur l'observation X, et faisant intervenir deux quantiles $q_n(\alpha_1)$ et $q_n(1 - \alpha_1)$. On précisera la valeur de α_1.

4 Tests d'hypothèses

Le gestionnaire du réseau s'intéresse à la probabilité que la réseau sature (pour une observation X_1). En l'état actuel, son réseau sature lorsque $X_1 > s$, où $s > u$ est connu. Soit $g(\theta) = \mathbb{P}_\theta(X_1 > s)$ (où θ est inconnu). La réglementation autorise le réseau à saturer "rarement", c'est-à-dire, elle impose que $g(\theta) \leq \rho_0$ où $0 < \rho_0 < 1$ est petit. La question est de savoir si le gestionnaire doit redimensionner son installation ou non. L'hypothèse nulle est que tout va bien :
$$H_0 : \{g(\theta) \leq \rho_0\}.$$
(l'hypothèse alternative est donc $H_1 : \{g(\theta) > \rho_0\}$).

1. Donner l'expression de $g(\theta)$ en fonction de θ, s et u.
2. Montrer que H_0 est vérifiée si et seulement si
$$\theta \in \Theta_0 = [\theta_0, +\infty[\quad \text{où } \theta_0 = \frac{\log(\rho_0)}{\log(u/s)}$$
3. Considérons pour commencer le test d'hypothèses simples $\tilde{H}_0 : \{\theta = \theta_0\}$ contre $\tilde{H}_1 : \{\theta = \theta_1\}$, où $\theta_1 < \theta_0$. Écrire la statistique du rapport de vraisemblance et montrer que le test de Neyman Pearson revient à comparer la variable aléatoire
$$W = \sum_{i=1}^{n} \log(X_i/u)$$
à un seuil c (qu'on ne calculera pas pour l'instant).
4. Soit $\alpha \in]0, 1[$. Déterminer le seuil c tel que le test
$$\delta(X) = \begin{cases} 1 & \text{si } W \geq c \\ 0 & \text{si } W < c \end{cases}$$
soit un test uniformément plus puissant (U.P.P) au niveau α (c'est-à-dire, de risque de première espèce égal à α) pour l'hypothèse $\tilde{H}_0$ contre $\tilde{H}_1$. On exprimera c en fonction des quantiles de la loi $\mathcal{G}amma(n, 1)$.
indication : *c.f.* question 3.1.
5. Soit $t > \theta_0$ considérons le test de l'hypothèse $H_0(t) : \{\theta = t\}$ contre $\tilde{H}_1$. Montrer que le risque de première espèce du test δ construit à la question 4 est strictement inférieur à α.
6. En déduire que δ est U.P.P. de niveau α pour tester $H_0 : \{\theta \geq \theta_0\}$ contre $\tilde{H}_1$.
7. En déduire que δ est U.P.P. de niveau α pour H_0 contre H_1.

5 Approche Bayésienne

On dispose d'une information a priori π sur le paramètre θ. On note π la loi a priori et $\boldsymbol{\theta} \sim \pi$ la variable aléatoire à valeurs dans $\Theta =]0, \infty[$ associée. On choisit pour π une loi Gamma, $\pi = \mathcal{G}amma(a, \lambda)$ avec $a > 0$, $\lambda > 0$ des quantités fixées par l'utilisateur en fonction de sa connaissance a priori sur $\boldsymbol{\theta}$. Soit $\pi(\,\cdot\,|x)$ la loi a posteriori sachant l'observation $X = x$. Dans la suite, on notera $\pi(\theta)$ la densité de la loi a priori évaluée en $\theta \in \Theta$ et $\pi(\theta|x)$ la densité de la loi a posteriori, sachant l'observation $X = x \in \mathbb{R}^n$.

1. Donnez un choix de (a, λ) tel que $\mathbb{E}_\pi(\boldsymbol{\theta}) = 2$ et $\mathbb{V}\mathrm{ar}_\pi(\boldsymbol{\theta}) = 100$.

   Dans la suite on n'utilisera pas le résultat numérique ci-dessus, on donnera tous les résultats en fonction de a et λ sans plus de calcul.

2. Montrer que pour $x = (x_1, \ldots, x_n)$ tel que $x_i > u$ pour $1 \leq i \leq n$, la loi a posteriori $\pi(\,\cdot\,|x)$ est une loi Gamma dont on précisera les paramètres en fonction de x, a, λ.

3. Pour x comme ci-dessus, quelle est l'espérance a posteriori $\mathbb{E}(\boldsymbol{\theta}|x)$?

4. Soit $p \in]0, 1[$. Donner une borne inférieure $m(x)$ telle que

$$\pi\Big([m(x), \infty[\ \Big|X = x\Big) = \mathbb{P}_\pi\Big(\boldsymbol{\theta} \geq m(x) \mid X = x\Big) = 1 - p,$$

   en fonction des quantiles d'une loi $\mathcal{G}amma(\beta, 1)$ (préciser β).