TD et TP n° 4 : Estimateurs

Exercice 1. La loi Gamma de paramètres $\alpha > 0$ et $\beta > 0$, notée $\gamma(\alpha, \beta)$, admet pour densité la fonction

$$f : x \in \mathbb{R} \longmapsto \frac{1}{\Gamma(\alpha)} \beta^\alpha x^{\alpha-1} e^{-\beta x} \mathbb{1}_{\{x>0\}}.$$

Les deux premiers moments d'une variable aléatoire Z de loi Gamma $\gamma(\alpha, \beta)$ valent $\mathbb{E}(Z) = \alpha/\beta$ et $\mathbb{E}(Z^2) = \alpha(\alpha+1)/\beta^2$.

Questions théoriques (TD) :

1. Calculer les estimateurs $\hat{\alpha}_{MM}$ et $\hat{\beta}_{MM}$ de α et β obtenus par la méthode des moments (préciser le modèle paramétrique).
2. Donner une expression des estimateurs du maximum de vraisemblance $\hat{\alpha}_{ML}$ et $\hat{\beta}_{ML}$ de α et β.

Les lois Gamma sont utilisées pour modéliser une grande variété de phénomènes aléatoires, lorsque les variables mesurées sont positives (données de survie, données météorologiques, données financières, etc...). Dans cet exercice, les données étudiées sont des quantités d'eau de pluie (en *inches*) mesurées lors de $n = 223$ orages survenus dans l'état de l'Illinois entre 1960 et 1964 (publiées par Le Cam et Neyman en 1967). Ces données sont accessibles (fichier `Orages.txt`) sur mon site internet.

Questions pratiques (TP) :

3. Représenter les données à l'aide d'un histogramme normalisé avec 16 classes.
4. Calculer la valeur de la moyenne empirique ? de la médiane empirique ? Pourquoi ces deux valeurs différent-elles autant ?
5. Calculer les valeurs des estimateurs $\hat{\alpha}_{MM}$, $\hat{\beta}_{MM}$, $\hat{\alpha}_{ML}$ et $\hat{\beta}_{ML}$. Pour ces derniers, on pourra utiliser les fonctions `scipy.special.gamma` (fonction Gamma d'Euler) et `scipy.optimize.minimize_scalar` (minimise une fonction réelle ; la valeur de l'argmin est alors obtenu par `minimize_scalar(fun).x`, avec `fun` la fonction à minimiser).
6. Tracer sur un même graphique l'histogramme et la densité de la loi gamma de paramètres $\hat{\alpha}_{MM}$ et $\hat{\beta}_{MM}$. Faire de même avec la fonction de répartition empirique et la fonction de répartition de la loi gamma. La loi gamma ainsi ajustée est-elle un bon modèle pour la loi des donnée ?
7. Répéter la dernière question avec les estimateurs du maximum de vraisemblance.

Exercice 2. Risque d'estimateurs et borne de Cramér-Rao. Soit $X_1, \ldots, X_n$ un échantillon de loi $\mathcal{N}(\theta, \theta^2)$, où θ est un réel strictement positif. La densité correspondante est notée $p_\theta(x)$.

Questions théoriques (TD) :

1. Donner deux estimateurs « naturels » pour estimer le paramètre θ.
2. Calculer la log-vraisemblance $l_\theta(x) = \ln p_\theta(x)$ associée aux observations pour le paramètre θ et en déduire que le maximum de vraisemblance $\hat{\theta}_n$ est défini par :

$$\hat{\theta}_n = -\frac{\overline{X_n}}{2} + \sqrt{\overline{X_n^2} + \frac{1}{4}\overline{X_n}^2},$$

   où $\overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i$ et $\overline{X_n^2} = \frac{1}{n}\sum_{i=1}^n X_i^2$ désignent respectivement la moyenne empirique et le moment empirique d'ordre 2.
3. Montrer que l'estimateur $\hat{\theta}_n$ est consistant et asymptotiquement sans biais.

On rappelle que le *risque* d'un estimateur Θ_n de θ est la quantité

$$R(\hat{\theta}_n, \theta) = \mathbb{E}[(\Theta_n - \theta)^2].$$

On souhaite étudier numériquement le risque de l'estimateur $\hat{\theta}_n$ ainsi que des estimateurs de la première question.

Questions pratiques (TP) :

4. Montrer qu'on peut se ramener à $\theta = 1$. On va supposer par la suite que $\theta = 1$.
5. Tracer $M = 5$ réalisations de l'estimateur $\hat{\theta}_n$, pour $n = 1, \dots, 1000$. Pourquoi cela illustre-t-il la consistance ?
6. Pour $n = 2^3, \dots, 2^{10}$, simuler $M = 1000$ réalisations de l'estimateur $\hat{\theta}_n$, puis estimer $R(\hat{\theta}_n, \theta)$ par la moyenne empirique de $(\hat{\theta}_n - \theta)^2$; on obtient alors une valeur estimée $\hat{R}_n$. Tracer $n\hat{R}_n$ en fonction de n (on pourra mettre l'abscisse en échelle logarithmique). Faire de même pour les estimateurs de la première question et superposer les tracés.
7. En vue des résultats de la dernière question, comparer les estimateurs.

Questions avancées (TD et TP) : Notons $\dot{l_\theta}(x) = \frac{\partial}{\partial\theta} l_\theta(x)$. Cette quantité s'appelle *fonction score* pour l'estimation de θ. On note aussi $I_\theta = \mathbb{E}[(\dot{l_\theta}(X))^2]$, où X suit la loi $\mathcal{N}(\theta, \theta^2)$. Cette quantité est l'*information de Fisher* associée au modèle au point θ. La *borne de Cramér-Rao* stipule que le risque d'un estimateur non-biaisé $\Theta_n = \Theta_n(X_1, \dots, X_n)$ de θ admet la borne inférieure suivante :

$$R(\hat{\theta}_n, \theta) \geq \frac{1}{nI_\theta}.$$

8. (TD) Montrer que $I_\theta = 3/\theta^2$.
9. (TP) Déduire des calculs numériques que l'estimateur $\hat{\theta}_n$ atteint asymptotiquement la borne de Cramér-Rao, c'est-à-dire que son risque vaut asymptotiquement

$$R(\hat{\theta}_n, \theta) = \mathbb{E}[(\hat{\theta}_n - \theta)^2] \sim \frac{1}{nI_\theta}, \quad n \to \infty.$$

10. (TP) On soupçonne que $\sqrt{nI_\theta}(\hat{\theta}_n - \theta)$ converge en loi vers la loi normale standard quand $n \to \infty$. Illustrer ce fait numériquement.
11. (TD) Déduire de la dernière question un intervalle de confiance asymptotique de niveau de confiance $1 - \alpha$ pour θ.
12. (TP) Calculer l'intervalle de confiance en Python. Faire 10 essais et compter le nombre de fois que θ est inclus dans l'intervalle de confiance (avec $\alpha = 0.2$, par exemple).

Pour continuer :

— Démontrer l'asymptotique de $R(\hat{\theta}_n, \theta)$ ainsi que la limite en loi de la question 10.

— Etudier (théoriquement et/ou numériquement) les risques des estimateurs de l'Exercice 1.

— Essayer de démontrer la borne de Cramér-Rao.