Probabilités et Statistiques - Examen - 21 juin 2016 S6 Polytech

Tous les documents sont autorisés, mais aucune calculatrice, téléphone, ordinateur ou objet apparenté. Toutes les réponses doivent être justifiées, et la rédaction sera prise en compte. Les résultats de questions précédentes peuvent être utilisés même si ces questions n'ont pas été traitées.

Exercice 1. Soit X une variable aléatoire continue, dont la densité est donnée par la fonction

$$f_X(x) = \begin{cases} \frac{C}{(x+1)^4} & \text{si } x \geq 0, \\ 0 & \text{sinon.} \end{cases}$$

1. Calculer la valeur de la constante C.
2. Calculer l'espérance et la variance de X. On pourra écrire

$$\frac{x}{(x+1)^4} = \frac{1}{(x+1)^3} - \frac{1}{(x+1)^4};$$
$$\frac{x^2}{(x+1)^4} = \frac{1}{(x+1)^2} - \frac{2}{(x+1)^3} + \frac{1}{(x+1)^4}.$$

3. Soit $X_1, X_2, \ldots, X_n$ des variables aléatoires indépendantes de même loi que X, et $M_n = \frac{X_1+X_2+\cdots+X_n}{n}$. Rappeler l'énoncé du théorème central limite. Calculer la limite de

$$\mathbb{P}\left[\frac{1}{2} \leq M_n \leq \frac{1}{2}\left(1+\frac{\varepsilon}{\sqrt{n}}\right)\right]$$

lorsque n tend vers l'infini et $\varepsilon > 0$ est un paramètre fixé. On pourra donner le résultat sous la forme d'une intégrale $\int_0^\varepsilon g(y)\,dy$ d'une certaine fonction positive g.

Exercice 2. Si X et Y sont deux variables aléatoires continues indépendantes à valeurs dans $\mathbb{R}_+$, on rappelle que leur somme $Z = X+Y$ a sa loi donnée par la formule du produit de convolution :

$$\forall t \geq 0,\ \ f_Z(t) = \int_{s=0}^{t} f_X(s)\, f_Y(t-s)\, ds,$$

où f_X, f_Y et f_Z sont les densités des variables X, Y et Z (elles valent par ailleurs 0 sur $\mathbb{R}_-$).

1. On suppose que X et Y sont indépendantes et suivent des lois exponentielles de paramètre 1 :

$$\forall s \geq 0,\ \ f_X(s) = f_Y(s) = \mathrm{e}^{-s}.$$

En utilisant la formule du produit de convolution, calculer la densité de $Z = X+Y$.

2. Plus généralement, soient $X_1, X_2, \dots, X_n$ des variables aléatoires indépendantes, qui suivent toutes une loi exponentielle de paramètre 1 :
$$\forall i \in [\![1, n]\!]\,, \quad \forall s \geq 0, \quad f_{X_i}(s) = \mathrm{e}^{-s}.$$
On s'intéresse à la somme $S_n = X_1 + X_2 + \cdots + X_n$. Calculer $\mathbb{E}[X_1]$ et $\mathrm{var}(X_1)$, puis $\mathbb{E}[S_n]$ et $\mathrm{var}(S_n)$.

3. Montrer par récurrence sur n que la densité de S_n est $f_{S_n}(s) = 1_{s\geq 0}\, \frac{s^{n-1}}{(n-1)!}\, \mathrm{e}^{-s}$.

4. Que peut on dire de $M_n = \frac{S_n}{n}$ lorsque n tend vers l'infini ? A-t-on une limite ? une loi limite ?

Exercice 3. Dans cet exercice, $X_1, X_2, \dots, X_n$ sont des variables aléatoires indépendantes suivant une même loi de Poisson de paramètre $\lambda > 0$ inconnu, qu'on cherche à estimer. On rappelle que si $X \sim \mathcal{P}(\lambda)$, alors $\mathbb{E}[X] = \mathrm{var}(X) = \lambda$.

1. L'erreur quadratique moyenne d'un estimateur $\widehat{\lambda}$ du paramètre λ est définie par $e^2(\widehat{\lambda}) = \mathbb{E}[(\widehat{\lambda} - \lambda)^2]$. Rappeler le lien entre biais, variance et erreur quadratique moyenne pour un estimateur $\widehat{\lambda}$.

2. Soient $X_1, \dots, X_n$ des variables indépendantes et de même loi. Exprimer en fonction de $\mathrm{var}(X_1)$ les variances de $S_n = X_1 + X_2 + \cdots + X_n$ et de $M_n = \frac{X_1+X_2+\cdots+X_n}{n}$.

3. Calculer le biais de M_n (vu comme estimateur de λ), puis son erreur quadratique moyenne $\mathbb{E}[(M_n - \lambda)^2]$.

4. Soit $V_n = \frac{1}{n-1}\sum_{i=1}^n (X_i - M_n)^2$. On admet que $(V_n)_{n\in\mathbb{N}}$ est une suite consistante d'estimateurs sans biais du paramètre λ, et que sa variance est donnée par la formule
$$\mathrm{var}(V_n) = \frac{1}{n-1}\left(\frac{n-1}{n}\mathbb{E}[(X_1 - \mathbb{E}[X_1])^4] - \frac{n-3}{n}\left(\mathbb{E}[(X_1 - \mathbb{E}[X_1])^2]\right)^2\right).$$
On donne aussi $\mathbb{E}[(X_1 - \mathbb{E}[X_1])^4] = 3\lambda^2 + \lambda$ pour une variable de Poisson $X_1 \sim \mathcal{P}(\lambda)$. Montrer que l'erreur quadratique moyenne de V_n vaut
$$\frac{\lambda}{n} + \frac{2\lambda^2}{n-1}.$$

5. Quel estimateur de λ est préférable : M_n ou V_n ? Justifier votre réponse.

6. On se donne $n = 20$ tirages indépendants de la loi de Poisson $\mathcal{P}(\lambda)$:
$$0, 0, 1, 4, 1, 1, 0, 2, 2, 5, 1, 1, 3, 1, 2, 3, 2, 3, 1, 1.$$
Donner une estimation du paramètre λ.

Corrigé exercice 1. Comme f_X est la densité d'une variable aléatoire,

$$1 = \int_{\mathbb{R}} f_X(x)\,dx = C\int_0^\infty \frac{1}{(x+1)^4} = C\left[-\frac{1}{3(x+1)^3}\right]_0^\infty = \frac{C}{3},$$

donc $C = 3$. On calcule ensuite :

$$\begin{aligned}
\mathbb{E}[X] &= \int_{\mathbb{R}} x\, f_X(x)\,dx = \int_0^\infty \frac{3\,x}{(x+1)^4}\,dx = \int_0^\infty \frac{3}{(x+1)^3} - \frac{3}{(x+1)^4}\,dx \\
&= \left[-\frac{3}{2(x+1)^2} + \frac{1}{(x+1)^3}\right]_0^\infty = \frac{3}{2} - 1 = \frac{1}{2}; \\
\mathbb{E}[X^2] &= \int_{\mathbb{R}} x^2\, f_X(x)\,dx = \int_0^\infty \frac{3\,x^2}{(x+1)^4}\,dx = \int_0^\infty \frac{3}{(x+1)^2} - \frac{6}{(x+1)^3} + \frac{3}{(x+1)^4}\,dx \\
&= \left[-\frac{3}{x+1} + \frac{3}{(x+1)^2} - \frac{1}{(x+1)^3}\right]_0^\infty = 1; \\
\mathrm{var}(X) &= 1 - \frac{1}{4} = \frac{3}{4}.
\end{aligned}$$

Par le théorème central limite,

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}\left[\frac{1}{2} \leq M_n \leq \frac{1}{2}\left(1+\frac{\varepsilon}{\sqrt{n}}\right)\right] &= \lim_{n\to\infty} \mathbb{P}\left[0 \leq \sqrt{\frac{n}{\mathrm{var}(X_1)}}\,(M_n - \mathbb{E}[X_1]) \leq \frac{\varepsilon}{\sqrt{3}}\right] \\
&= \frac{1}{\sqrt{2\pi}}\int_0^{\frac{\varepsilon}{\sqrt{3}}} \mathrm{e}^{-\frac{x^2}{2}}\,dx = \frac{1}{\sqrt{6\pi}}\int_0^{\varepsilon} \mathrm{e}^{-\frac{y^2}{6}}\,dy.
\end{aligned}$$

Corrigé exercice 2. Si X et Y sont des variables exponentielles indépendantes et si $Z = X + Y$, alors

$$\begin{aligned}
f_Z(t) &= \int_{s=0}^{t} f_X(s)\, f_Y(t-s)\,ds = \int_{s=0}^{t} \mathrm{e}^{-s}\,\mathrm{e}^{-(t-s)}\,ds \\
&= \int_{s=0}^{t} \mathrm{e}^{-t}\,ds = t\,\mathrm{e}^{-t}
\end{aligned}$$

en utilisant la formule du produit de convolution.

Plus généralement, si $X_1, \ldots, X_n$ sont des variables exponentielles indépendantes, montrons par récurrence sur n, $S_n = X_1 + \cdots + X_n$ suit une loi de densité $1_{s\geq 0}\,\frac{s^{n-1}}{(n-1)!}\,\mathrm{e}^{-s}$. Comme toutes les variables X_i sont positives, S_n est également positive avec probabilité 1, donc sa densité porte sur $\mathbb{R}_+$ et est nulle sur $\mathbb{R}_-$. En supposant le résultat établi pour S_{n-1}, on a $S_n = S_{n-1} + X_n$ qui est une somme de deux variables aléatoires indépendantes, donc, par la formule du produit de convolution, pour tout $s \geq 0$,

$$\begin{aligned}
f_{S_n}(t) &= \int_{s=0}^{t} f_{S_{n-1}}(s)\, f_{X_n}(t-s)\,ds = \int_{s=0}^{t} \frac{s^{n-2}}{(n-2)!}\,\mathrm{e}^{-s}\,\mathrm{e}^{-(t-s)}\,dx \\
&= \mathrm{e}^{-t}\int_{s=0}^{t} \frac{s^{n-2}}{(n-2)!}\,ds = \frac{t^{n-1}}{(n-1)!}\,\mathrm{e}^{-t}.
\end{aligned}$$

D'autre part, comme l'espérance et la variance d'une loi exponentielle de paramètre 1 sont égales à 1, $\mathbb{E}[S_n] = \mathrm{var}(S_n) = n$. Finalement, par la loi des grands nombres, $M_n = \frac{S_n}{n}$ tend vers 1, et par le théorème central limite,

$$\sqrt{n}\,(M_n - 1) = \frac{S_n - n}{\sqrt{n}} \to_{\mathrm{loi},\, n\to+\infty} \mathcal{N}(0,1).$$

Corrigé exercice 3. Si $M_n = \frac{X_1+\cdots+X_n}{n}$, alors

$$\mathbb{E}[M_n] = \frac{\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]}{n} = \frac{n\lambda}{n} = \lambda,$$

car l'espérance d'une variable de Poisson de paramètre λ est λ. Donc, M_n est un estimateur sans biais de λ. Comme l'erreur moyenne au carré est la somme de la variance et du carré du biais, on obtient :

$$\mathbb{E}[(M_n - \lambda)^2] = \mathrm{var}(M_n) + 0 = \frac{\mathrm{var}(X_1) + \cdots + \mathrm{var}(X_n)}{n^2} = \frac{n\lambda}{n^2} = \frac{\lambda}{n}.$$

De même, V_n est un estimateur sans biais de λ, donc

$$\begin{aligned}
\mathbb{E}[(V_n - \lambda)^2] &= \mathrm{var}(V_n) + 0 \\
&= \frac{1}{n-1}\left(\frac{n-1}{n}\,\mathbb{E}[(X_1 - \mathbb{E}[X_1])^4] - \frac{n-3}{n}\left(\mathbb{E}[(X_1 - \mathbb{E}[X_1])^2]\right)^2\right) \\
&= \frac{1}{n-1}\left(\frac{n-1}{n}(3\lambda^2 + \lambda) - \frac{n-3}{n}\,\lambda^2\right) = \frac{\lambda}{n} + \frac{2\lambda^2}{n-1}.
\end{aligned}$$

Comme l'erreur moyenne au carré de V_n est supérieure à celle de M_n, le meilleur estimateur de λ est M_n. Avec les données de l'énoncé, on obtient une moyenne empirique de $\frac{34}{20} = 1.7$, donc on peut donner comme estimée de λ la valeur 1.7.