DS 10 - sujet A

THÈME : ESTIMATION

La présentation, la lisibilité, l'orthographe, la qualité de la rédaction, la clarté et la précision des raisonnements entreront pour une part importante dans l'appréciation des copies. Les candidats sont invités à encadrer dans la mesure du possible les résultats de leurs calculs. Ils ne doivent faire usage d'aucun document : l'utilisation de toute calculatrice et de tout matériel électronique est interdite.

Problème A. Loi de Pareto : e^X où $X \hookrightarrow \mathscr{E}(\lambda)$

- *Loi de Pareto*

Soient λ un réel strictement supérieur positif et X une variable aléatoire suivant la loi exponentielle de paramètre λ.

1. Montrer que si $\lambda > 2$, $\mathbf{E}\left(e^X\right)$ et $\mathbf{E}\left(e^{2X}\right)$ existent et valent respectivement $\frac{\lambda}{\lambda-1}$ et $\frac{\lambda}{\lambda-2}$.

On dit que Y *suit une loi de Pareto de paramètres* λ *et 1, notée* $\mathscr{P}(\lambda,1)$ *si* Y *admet pour densité*

$$f_\lambda : t \in \mathbb{R} \mapsto \begin{cases} \lambda t^{-\lambda-1} & \text{si } t \geqslant 1 \\ 0 & \text{sinon.} \end{cases}$$

On admet que si $X \hookrightarrow \mathscr{E}(\lambda)$, *alors* $e^X \hookrightarrow \mathscr{P}(\lambda,1)$.

2. Justifier que si $Y \hookrightarrow \mathscr{P}(\lambda,1)$ alors Y admet une variance que l'on déterminera.

- *Construction d'un estimateur*

On considère une suite de variables aléatoires $(Y_i)_{i\in\mathbb{N}^*}$, indépendantes et suivant toutes la même loi de Pareto $\mathscr{P}(\lambda,1)$ que Y. On souhaite construire un estimateur du paramètre λ par une méthode dite du maximum de vraisemblance.

Soient n un entier naturel non nul et $y_1, y_2, \ldots, y_n$, n réels supérieurs strictement à 1. On pose

$$s_n = \sum_{i=1}^{n} \ln(y_i)$$

et on définit alors les fonctions H et h sur $\mathbb{R}_*^+$, par

$$H(\lambda) = \prod_{i=1}^{n} f_\lambda(y_i), \qquad h(\lambda) = \ln\big(H(\lambda)\big).$$

3. Justifier que h admet un maximum atteint en λ^* que l'on précisera.

Soit $\lambda \in \mathbb{R}_*^+$. Le calcul précédent suggère de poser

$$T_n = \frac{n}{\sum_{i=1}^{n} \ln(Y_i)}.$$

4. Reconnaitre la loi de $\lambda n/T_n$, en déduire que

$$\mathbf{E}(T_n) = \frac{n}{n-1}\lambda.$$

On admet dans la suite que T_n *admet une variance avec*

$$\mathbf{V}(T_n) = \frac{n^2}{(n-1)^2(n-2)}\lambda^2.$$

5. Déterminer un réel α_n tel que $T'_n = \alpha_n T_n$ soit un estimateur sans biais de λ.

6. Justifier que T'_n est un estimateur convergent de λ.

- *Intervalle de confiance asymptotique*

On définit, pour tout n de $\mathbb{N}^*$: $$Z_n = \sqrt{n}\left(\frac{\lambda}{T_n} - 1\right)$$

7. Justifier que la suite de variables aléatoires $(Z_n)_{n\in\mathbb{N}^*}$ converge en loi vers une variable aléatoire suivant la loi normale centrée réduite.

8. En déduire que l'intervalle

$$\left[\frac{\sqrt{n}-2}{\sqrt{n}}\cdot T_n ; \frac{\sqrt{n}+2}{\sqrt{n}}\cdot T_n\right]$$

est un intervalle de confiance asymptotique pour λ au niveau de confiance 95%.
On admettra que $\Phi(2) \geqslant 0.975$, où Φ désigne la fonction de répartition d'une variable aléatoire suivant la loi normale centrée réduite.

Problème B. loi de Gumbel : $\ln(X)$ où $X \hookrightarrow \mathscr{E}(\lambda)$

Généralité sur la loi de Gumbel

Soient m un réel et a un réel strictement positif. On introduit la fonction $f_{m,a}$ définie sur $\mathbb{R}$ par

$$\forall x \in \mathbb{R}, \quad f_{m,a}(x) = \frac{1}{a}\exp\left(\frac{m-x}{a} - e^{\frac{m-x}{a}}\right).$$

9. Montrer que $f_{m,a}$ peut être considérée comme une densité de probabilité.
On dira qu'une variable aléatoire Z ayant pour densité $f_{m,a}$ suit la loi de Gumbel de paramètre m et a et on écrira $Z \hookrightarrow \mathscr{G}(m,a)$.

10. Déterminer l'expression de la fonction de répartition $F_{m,a}$ d'une variable aléatoire $Z \hookrightarrow \mathscr{G}(m,a)$.

11. Montrer que

$$Z \hookrightarrow \mathscr{G}(0,1) \iff aZ+m \hookrightarrow \mathscr{G}(m,a).$$

- *Simulation*
On admet que si $X \hookrightarrow \mathscr{E}(1)$, *alors* $-\ln(X) \hookrightarrow \mathscr{G}(0,1)$.
De plus, on peut simuler une loi exponentielle d'espérance α *par* `rd.exponentiel(a)`.

12. Donner une fonction Python d'en-tête `def Gumbel(m, a)` qui renvoie une simulation d'une loi de Gumbel de paramètres m et a.

13. On note Ψ la fonction de répartition de la loi $\mathscr{G}(0,1)$.
Soit $\alpha \in]0;1[$. Montrer qu'il existe deux réels c_α et d_α (que l'on déterminera) tels que

$$\Psi(c_\alpha) = 1-\alpha, \quad \text{et} \quad \Psi(d_\alpha) = \alpha.$$

Partie 2 : n-échantillon de lois exponentielles et loi de Gumbel

Soit $n \in \mathbb{N}^*$. On considère un n-échantillon $(X_1, X_2, \ldots, X_n)$ de la loi exponentielle $\mathscr{E}(\lambda)$ de paramètre $\lambda > 0$. On pose

$$L_n = \min(X_1, \ldots, X_n) \quad \text{et} \quad M_n = \max(X_1, \ldots, X_n).$$

14. Justifier que la variable $Y_n = n\lambda L_n$ suit une loi exponentielle de paramètre 1.

15. On maintenant alors $Z_n = \lambda M_n - \ln(n)$.
Justifier que Z_n converge en loi vers une variable aléatoire Z suivant une loi de Gumbel dont on précisera les paramètres.

Partie 3 : Construction de deux intervalles de confiance

La durée de vie d'un composant électrique est modélisée par une variable aléatoire $X \hookrightarrow \mathscr{E}(\lambda)$ où $\lambda > 0$ est inconnu. On cherche à estimer la durée de vie moyenne $\mu = \mathbf{E}(X) = 1/\lambda$ et on dispose d'un échantillon de n composants (dont les durées de vie sont supposées indépendantes).
Dans cette question, on suppose que la seule information dont on dispose est la durée de vie du composant qui a grillé le plus tôt.

16. À l'aide de la question 14, proposer un estimateur $\widetilde{L}_n$ de μ, construit à partir de L_n, qui soit sans biais.

17. Montrer que, pour tout $\varepsilon \in \mathbb{R}_*^+$

$$\mathbf{P}\left(\left|\widetilde{L}_n - \mu\right| > \varepsilon\right) \geqslant e^{-1-\lambda\varepsilon}.$$

18. L'estimateur $\widetilde{L}_n$ est-il un estimateur convergent de $1/\lambda$?

- *Construction d'un intervalle de confiance*

19. Soit $\alpha \in]0;1[$. Montrer que si $Y \hookrightarrow \mathscr{E}(1)$, alors

$$\mathbf{P}\left(Y < -\ln\left(1-\frac{\alpha}{2}\right)\right) = \mathbf{P}\left(Y > -\ln\left(\frac{\alpha}{2}\right)\right) = \frac{\alpha}{2}.$$

20. Montrer alors que l'intervalle $I_{\alpha,n}$ ci-dessous est un intervalle de confiance au niveau de confiance $1-\alpha$ pour μ :

$$I_{\alpha,n} = \left[\frac{\widetilde{L}_n}{-\ln(\alpha/2)}, \frac{\widetilde{L}_n}{-\ln(1-\alpha/2)}\right].$$

- *Construction d'un intervalle de confiance asymptotique*

Dans les deux prochaines questions, on suppose que l'on connait la durée de vie du dernier composant. Soit $\alpha \in]0;1[$.

21. À l'aide des questions 13 et 15, montrer que $J_{\alpha,n}$ est un intervalle de confiance asymptotique au niveau de confiance $1-\alpha$ pour μ :

$$J_{\alpha,n} = \left[\frac{M_n}{\ln(n) + c_{\alpha/2}} ; \frac{M_n}{\ln(n) + d_{\alpha/2}} \right].$$

On pourra vérifier que si $Z \hookrightarrow \mathcal{G}(0,1)$ alors

$$\mathbf{P}\left(d_{\alpha/2} \leqslant Z \leqslant c_{\alpha/2}\right) = 1-\alpha.$$

22. À l'aide des commandes suivantes, est-il préférable de connaître L_n ou M_n pour estimer $1/\lambda$?

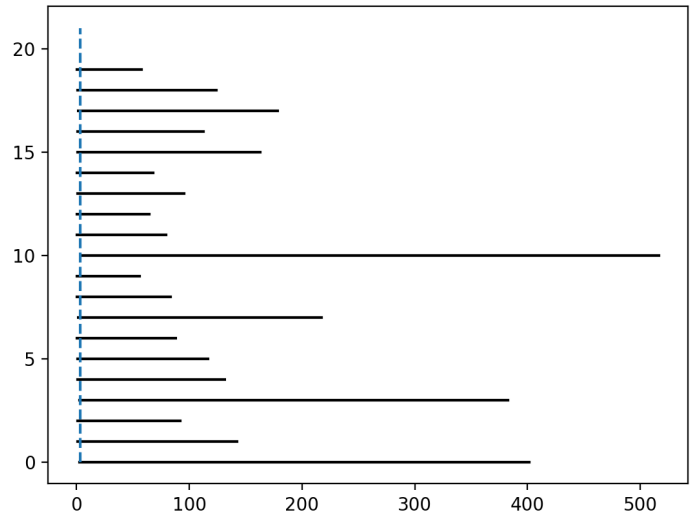
```
import numpy.random as rd
import matplotlib.pyplot as plt
import numpy as np

lbda= 1/3 # choix d'un paramètre "inconnu"

n=10
alpha=0.05 #niveau de confiance de 95%

NbreTest=20
for i in range(NbreTest):
    # Echantillon de données
    Ech=rd.exponential(1/lbda,n)
    # calcul intervalle q 20
    BorneInf=n*np.min(Ech)/(-np.log(alpha/2))
    BorneSup=n*np.min(Ech)/(-np.log(1-alpha/2))
    plt.plot([BorneInf,BorneSup],[i,i],'k')
    # tracé en noir
plt.plot([1/lbda,1/lbda],[0,21],'--')

plt.show()
```



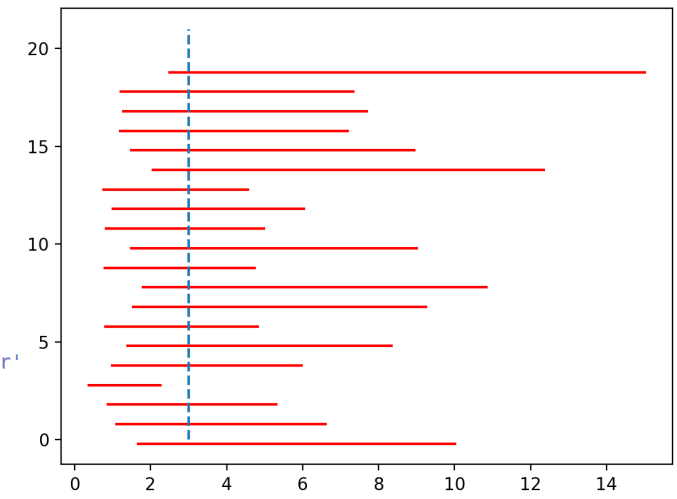

```
lbda= 1/3 # choix d'un paramètre "inconnu"

n=10
alpha=0.05 #niveau de confiance de 95%

NbreTest=20

for i in range(NbreTest):
    # Echantillon de données
    Ech=rd.exponential(1/lbda,n)
    # calcul intervalle q 21
    c=-np.log(-np.log(1-alpha/2))
    d=-np.log(-np.log(alpha/2))
    BorneInf2=np.max(Ech)/(np.log(n)+c)
    BorneSup2=np.max(Ech)/(np.log(n)+d)
    plt.plot([BorneInf2,BorneSup2],[i-0.2,i-0.2],'r'
    # tracé en rouge
plt.plot([1/lbda,1/lbda],[0,21],'--')

plt.show()
```



– FIN –

.

DS 10 - sujet *

THÈME : ESTIMATION

La présentation, la lisibilité, l'orthographe, la qualité de la rédaction, la clarté et la précision des raisonnements entreront pour une part importante dans l'appréciation des copies. Les candidats sont invités à encadrer dans la mesure du possible les résultats de leurs calculs. Ils ne doivent faire usage d'aucun document : l'utilisation de toute calculatrice et de tout matériel électronique est interdite.

Les tables de mortalité sont utilisées en démographie et en actuariat pour prévoir l'espérance de vie des individus d'une population. On s'intéresse dans ce problème à un modèle qui permet d'ajuster la durée de vie à des statistiques portant sur les décès observés au sein d'une génération. Dans tout le problème, on note :

- a et b deux réels strictement positifs ;
- $(\Omega, \mathscr{A}, \mathbf{P})$ un espace probabilisé sur lequel sont définies toutes les variables aléatoires du problème ;
- $\mathrm{G}_{a,b}$ la fonction définie sur $\mathbb{R}^+$ par : $\mathrm{G}_{a,b}(x) = \exp\left(-ax - \frac{b}{2}x^2\right)$.

Partie I. Loi exponentielle linéaire

- *Bijectivité de la fonction* $\mathrm{G}_{a,b}$

1. Montrer que la fonction $\mathrm{G}_{a,b}$ réalise une bijection de $\mathbb{R}^+$ sur l'intervalle $]0,1]$.

2. On note $\mathrm{G}_{a,b}^{-1}$ la bijection réciproque de $\mathrm{G}_{a,b}$. Quelle est, pour tout $u \in [0,1[$, l'expression de $\mathrm{G}_{a,b}^{-1}(1-u)$?

- *Convergence de l'intégrale de* $\mathrm{G}_{a,b}$

3. Justifier la convergence de l'intégrale $\displaystyle\int_0^{+\infty} \mathrm{G}_{a,b}(x)\,\mathrm{d}x$.

Soit f la fonction définie sur $\mathbb{R}$ par : $f(x) = \sqrt{\dfrac{b}{2\pi}} \times \exp\left(-\dfrac{1}{2}b\left(x+\dfrac{a}{b}\right)^2\right)$.

4. Montrer que f est une densité d'une variable aléatoire suivant une loi normale dont on précisera les paramètres (espérance et variance).

5. Soit Φ la fonction de répartition de la loi normale centrée réduite. Déduire l'égalité :

$$\int_0^{+\infty} \mathrm{G}_{a,b}(x)\,\mathrm{d}x = \sqrt{\frac{2\pi}{b}} \times \exp\left(\frac{a^2}{2b}\right) \times \Phi\left(-\frac{a}{\sqrt{b}}\right).$$

- *Loi exponentielle linéaire*

6. Pour tout $a > 0$ et pour tout $b > 0$, on pose :

$$f_{a,b}(x) = \begin{cases} (a+bx)\exp\left(-ax - \frac{b}{2}x^2\right) & \text{si } x \geqslant 0 \\ 0 & \text{si } x < 0 \end{cases}$$

Justifier que $f_{a,b}$ est une densité de probabilité.

On dit qu'une variable aléatoire suit la loi exponentielle linéaire de paramètres a et b, notée $\mathscr{E}_\ell(a,b)$ si elle admet $f_{a,b}$ pour densité.

7. Soit X une variable aléatoire suivant la loi $\mathscr{E}_\ell(a,b)$. Justifier que X admet une espérance avec :

$$\mathbf{E}(\mathrm{X}) = \int_0^{+\infty} \mathrm{G}_{a,b}(x)\,\mathrm{d}x.$$

- *Simulation*

8. On note U une variable aléatoire suivant la loi uniforme sur $[0,1[$. Déterminer la loi de la variable aléatoire $G_{a,b}^{-1}(1-U)$.

9. En déduire un programme d'entête `simuE(a,b)` qui simule la loi exponentielle linéaire de paramètres a et b.

Dans la suite du problème, on note $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires indépendantes suivant chacune la loi exponentielle linéaire $\mathscr{E}_\ell(a,b)$ dont les paramètres $a>0$ et $b>0$ sont inconnus.
Soit h un entier supérieur ou égal à 2. On suit pendant une période de h années une "cohorte" de n individus de même âge au début de l'étude et on modélise leurs durées de vie respectives à partir de cette date par les variables $X_1, X_2, \ldots, X_n$.

Partie II. Premier décès et intervalle de confiance de a

Pour tout $n\in\mathbb{N}^*$, on définit les variables aléatoires M_n, H_n et U_n par :

$$M_n = \min(X_1, X_2, \ldots, X_n), \quad H_n = \min(h, X_1, X_2, \ldots, X_n) \quad \text{et} \quad U_n = nH_n$$

10. Reconnaître la loi de la variable aléatoire M_n.

11. Pour tout $n\in\mathbb{N}^*$, on note F_{U_n} la fonction de répartition de la variable aléatoire U_n. Vérifier que pour tout $x\in\mathbb{R}$, on a :

$$F_{U_n}(x) = \begin{cases} 0 & \text{si } x<0 \\ 1-\exp\left(-ax-\frac{b}{2n}x^2\right) & \text{si } 0\leqslant x<nh \\ 1 & \text{si } x\geqslant nh \end{cases}$$

12. Est-ce que la variable U_n est une variable à densité ?

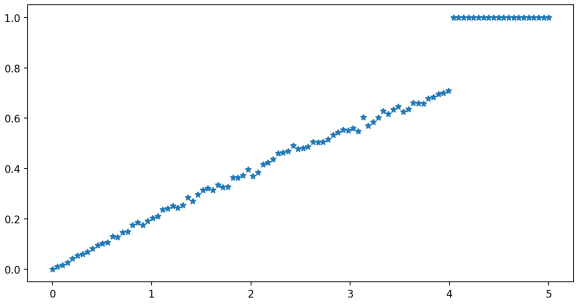
13. **a)** Écrire une fonction `simuU` qui prend en argument a, b, n et h et simule la variable U_n.

b) Expliquer le code suivant et vérifier la cohérence avec la question 12.

Editeur

```
a=0.2; b=1/10
h=2; n=2
x=np.linspace(0,5,100)
F=np.zeros(100)
for i in range(100):
    c=0
    for j in range(2000):
        c+=simuU(a,b,h,n)<x[i]
    F[i]=c/m

plt.plot(x,F,'*')
plt.show()
```

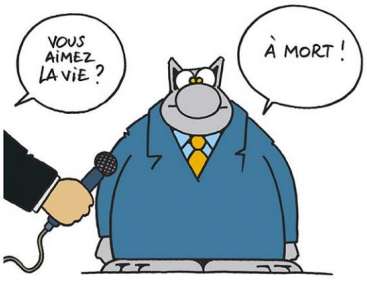


14. Montrer que la suite de variables aléatoires $(U_n)_{n\in\mathbb{N}^*}$ converge en loi vers une variable aléatoire dont on précisera la loi.

15. Soit $\alpha\in]0,1[$.
Soit Y une variable aléatoire qui suit la loi exponentielle de paramètre 1. Trouver deux réels c et d strictement positifs tels que :

$$\mathbf{P}([c\leqslant Y\leqslant d]) = 1-\alpha \quad \text{et} \quad \mathbf{P}([Y\leqslant c]) = \frac{\alpha}{2}.$$

16. Montrer que $\left[\frac{c}{U_n}, \frac{d}{U_n}\right]$ est un intervalle de confiance asymptotique de a, de niveau de confiance $1-\alpha$.



Partie III. Nombre de survivant et estimateur convergent de b

Pour tout $i \in \mathbb{N}^*$, soient S_i et D_i les variables aléatoires telles que :

$$S_i = \begin{cases} 1 & \text{si } X_i \geqslant h \\ 0 & \text{sinon} \end{cases} \quad \text{et} \quad D_i = \begin{cases} 1 & \text{si } X_i \leqslant 1 \\ 0 & \text{sinon} \end{cases}$$

Pour tout $n \in \mathbb{N}^*$, on pose :

$$\overline{S_n} = \frac{1}{n}\sum_{i=1}^{n} S_i \quad \text{et} \quad \overline{D_n} = \frac{1}{n}\sum_{i=1}^{n} D_i.$$

17. Justifier que pour tout $i \in [\![1;n]\!]$, on a $\mathbf{E}(S_i) = G_{a,b}(h)$ et calculer $\mathbf{E}(S_i D_i)$.

18. Pour quels couples $(i,j) \in [\![1,n]\!]^2$, les variables S_i et D_j sont-elles indépendantes ?

19. En déduire l'expression de la covariance $\mathrm{Cov}\left(\overline{S_n}, \overline{D_n}\right)$ de $\overline{S_n}$ et $\overline{D_n}$ en fonction de n, $G_{a,b}(h)$ et $G_{a,b}(1)$.

20. Montrer que $\overline{S_n}$ est un estimateur sans biais et convergent du paramètre $G_{a,b}(h)$.
De quel paramètre, $\overline{D_n}$ est-il un estimateur sans biais et convergent ?

21. À l'aide de l'inégalité de Bienaymé-Tchebychev, donner un intervalle de confiance de $G_{a,b}(h)$.

22. On pose :

$$z(a,b) = \ln\left(G_{a,b}(1)\right) \quad \text{et} \quad r(a,b) = \ln\left(G_{a,b}(h)\right).$$

Pour tout $n \in \mathbb{N}^*$, on pose aussi

$$Z_n = \ln\left(1 - \overline{D_n} + \frac{1}{n}\right) \quad \text{et} \quad R_n = \ln\left(\overline{S_n} + \frac{1}{n}\right).$$

Justifier que Z_n et R_n sont des estimateurs convergents de $z(a,b)$ et $r(a,b)$ respectivement.

23. Soient ε, λ et μ des réels strictement positifs. Justifier l'inégalité suivante :

$$\mathbf{P}\left(\left[\left|\left(\lambda Z_n - \mu R_n\right) - \left(\lambda z(a,b) - \mu r(a,b)\right)\right| \geqslant \varepsilon\right]\right) \leqslant \mathbf{P}\left(\left[|Z_n - z(a,b)| \geqslant \frac{\varepsilon}{2\lambda}\right]\right) + \mathbf{P}\left(\left[|R_n - r(a,b)| \geqslant \frac{\varepsilon}{2\mu}\right]\right).$$

24. Pour tout $n \in \mathbb{N}^*$, on pose :

$$B_n = \frac{2}{h-1} Z_n - \frac{2}{h(h-1)} R_n.$$

Montrer que B_n est un estimateur convergent du paramètre b.

– FIN –