

DS 10- sujet A

THÈME : ESTIMATIONS, OPTIMISATION SOUS CONTRAINTE

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

Estimation et loi log-normale

Toutes les variables aléatoires intervenant dans ce problème sont supposées être définies sur le même espace probabilisé $(\Omega, \mathcal{A}, \mathbf{P})$.

- Soit Y une variable aléatoire suivant la loi $\mathcal{N}(m, \sigma^2)$.
- On dit que la variable aléatoire $Z = \exp(Y)$ suit la loi log-normale de paramètres m et σ^2 , notée $\mathcal{LN}(m, \sigma^2)$. Ainsi, Z suit la loi $\mathcal{LN}(m, \sigma^2)$ si Z est à valeurs dans $\mathbb{R}_+^*$ et si la variable aléatoire $\ln Z$ suit la loi $\mathcal{N}(m, \sigma^2)$.

Dans la suite, on note X une variable aléatoire qui suit la loi $\mathcal{N}(0, 1)$ et Z une variable aléatoire qui suit la loi $\mathcal{LN}(m, \sigma^2)$.

• **Espérance et variance de Z**

1. a) Vérifier pour tout couple (s, t) de réels, la relation suivante : $\exp\left(-\frac{t^2}{2} + st\right) = \exp\left(\frac{s^2}{2}\right) \times \exp\left(-\frac{1}{2}(t-s)^2\right)$.
- b) Établir pour tout réel s , l'existence de $E(\exp(sX))$ et l'égalité :

$$\forall s \in \mathbb{R}, \quad E(\exp(sX)) = \exp\left(\frac{s^2}{2}\right).$$

2. En déduire les valeurs respectives de $E(Z)$ et $V(Z)$.

• On note Φ la fonction de répartition de X et F_Z la fonction de répartition de Z .

3. Montrer que la variable aléatoire Z est à densité et qu'une densité f_Z de Z est donnée sur $\mathbb{R}$ par :

$$f_Z(x) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}} \times \frac{1}{x} \times \exp\left(-\frac{(\ln x - m)^2}{2\sigma^2}\right) & \text{si } x > 0 \\ 0 & \text{sinon} \end{cases}$$

4. a) On pose : $R = \frac{1}{Z}$. Montrer que la variable aléatoire R suit la loi $\mathcal{LN}(-m, \sigma^2)$.
- b) Soit U une variable aléatoire suivant la loi $\mathcal{LN}(0, 1)$. Pour tout entier $n \geq 1$, on pose : $W_n = U^n$. Montrer que la variable aléatoire W_n suit la loi $\mathcal{LN}(0, n^2)$.

• **Estimateurs et intervalle de confiance**

Soient $\alpha \in]0; 1[$ et $(Z_k)_{k \in \mathbb{N}^*}$ une suite de variables aléatoires indépendantes suivant la même loi $\mathcal{LN}(m, \sigma^2)$, avec $0 < \sigma \leq 2$. Pour tout entier $n \geq 1$, on pose :

$$\overline{T}_n = \frac{1}{n} \sum_{k=1}^n \ln(Z_k), \quad I_n = \overline{T}_n - \frac{2}{\sqrt{n\alpha}} \quad \text{et} \quad J_n = \overline{T}_n + \frac{2}{\sqrt{n\alpha}}.$$

5. Justifier que $\overline{T}_n$ est un estimateur convergent de m .
6. En utilisant l'inégalité de Bienaymé-Tchebychev, vérifier que l'intervalle $[I_n, J_n]$ est un intervalle de confiance de m de niveau de risque α .
7. a) Reconnaître la loi de $\overline{T}_n$.
- b) Justifier qu'il existe t_α positif tel que $\Phi(t_\alpha) = 1 - \frac{\alpha}{2}$.
- c) En déduire que l'intervalle ci-dessous est un intervalle de confiance de m de niveau de risque α .

$$\left[\overline{T}_n - t_\alpha \frac{2}{\sqrt{n}}, \overline{T}_n + t_\alpha \frac{2}{\sqrt{n}} \right].$$

Problème B
Estimation d'un paramètre

Soit X une variable aléatoire de densité f définie sur $\mathbb{R}$ par

$$f(x) = \begin{cases} \frac{2}{\theta} \left(1 - \frac{x}{\theta}\right) & \text{si } 0 \leq x \leq \theta \\ 0 & \text{sinon} \end{cases}$$

où θ est un paramètre inconnu strictement positif.

On cherche à estimer θ à partir d'un échantillon $(X_1, X_2, \dots, X_n)$ de variables aléatoires indépendantes de même loi que X . On pose

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{et} \quad M_n = \max(X_1, X_2, \dots, X_n).$$

8. Déterminer la fonction de répartition F de X , et préciser l'espérance et la variance de X .
9. a) Déterminer la constante a pour que $T_n = a\overline{X}_n$ soit un estimateur sans biais du paramètre θ .
b) Préciser la variance de T_n .

• *Comparaison d'estimateurs*

10. Exprimer la fonction de répartition de la loi de M_n . Vérifier que M_n est à densité et une densité est donnée sur $[0; \theta]$ par

$$\forall t \in [0; \theta], \quad f_n(t) = \frac{2n}{\theta} \left(1 - \frac{t}{\theta}\right) \left(1 - \left(1 - \frac{t}{\theta}\right)^2\right)^{n-1}.$$

On note $I_n = \int_0^1 (1-u^2)^n du$ et on admet que $I_n \sim_{+\infty} \sqrt{\frac{\pi}{4n}}$.

11. Montrer à l'aide d'une intégration par parties que $E(M_n) = \theta(1 - I_n)$. Est-ce que M_n est un estimateur de θ sans biais? asymptotiquement sans biais¹?
12. Donner un estimateur sans biais de θ de la forme $M'_n = a_n M_n$, a_n étant un réel dépendant de n .
13. On admet que $V(M_n) = \theta^2 \left(\frac{1}{n+1} - I_n^2\right)$.
a) Déterminer la limite de $V(T_n)/V(M'_n)$ lorsque $n \rightarrow +\infty$.
b) Entre M'_n et T_n , quel estimateur choisiriez-vous, et pourquoi?

Problème C
Optimisation sous contrainte

Soit $n \in \mathbb{N}$ avec $n \geq 3$. On pose $\mathcal{D} = (\mathbb{R}_+^*)^n$ et la fonction f sur $\mathcal{D}$ par $f(x_1, \dots, x_n) = \sum_{k=1}^n x_k \ln(x_k)$.

• **Étude de f**

14. a) Étudier les variations de $g: t \in]0; +\infty[\mapsto t \ln(t)$.
b) Préciser si la fonction f est minorée, majorée ou bornée sur $\mathcal{D}$.
15. Justifier que la fonction f est de classe $\mathcal{C}^2$ sur $\mathcal{D}$.

• **Étude de f sous une contrainte linéaire**

Soit a , un réel strictement positif. On pose

$$\mathcal{C}_a = \left\{ (x_1, \dots, x_n) \in \mathbb{R}^n \mid \sum_{k=1}^n x_k = na \right\}.$$

16. Montrer que f admet un unique point critique sous la contrainte $\mathcal{C}_a$. On note A , ce point.
 17. Donner le développement limité à l'ordre 2 de f au point A . En déduire que f admet un unique minimum local sous la contrainte $\mathcal{C}_a$ que l'on explicitera.
 18. À l'aide de la formule de Taylor avec reste intégral, justifier que ce minimum local est aussi un minimum global sous la contrainte.
- Dans les deux prochaines questions, on se place dans le cas où $n = 3$.
On pose la fonction $h: (\mathbb{R}_+^*)^2 \rightarrow \mathbb{R}$ définie par :

$$h(x, y) = x \ln(x) + y \ln(y) + (3a - x - y) \ln(3a - x - y).$$

On admet que h est de classe $\mathcal{C}^2$.

1. C'est-à-dire que le biais tend vers 0 lorsque $n \rightarrow +\infty$.

19. Justifier que h a un unique point critique A' et que la matrice hessienne est donnée par $\nabla^2 h(A') = \begin{bmatrix} 2/a & 1/a \\ 1/a & 2/a \end{bmatrix}$.

20. Retrouver le résultat de la question 15.

- Dans la suite, on traite le cas où $n = 2$ et $a = 1$.

21. À l'aide du code suivant, tester la cohérence du résultat de la question 15.

Editeur

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

def g(t):
    return t*np.log(t)

def f(x,y):
    return g(x)+g(y)

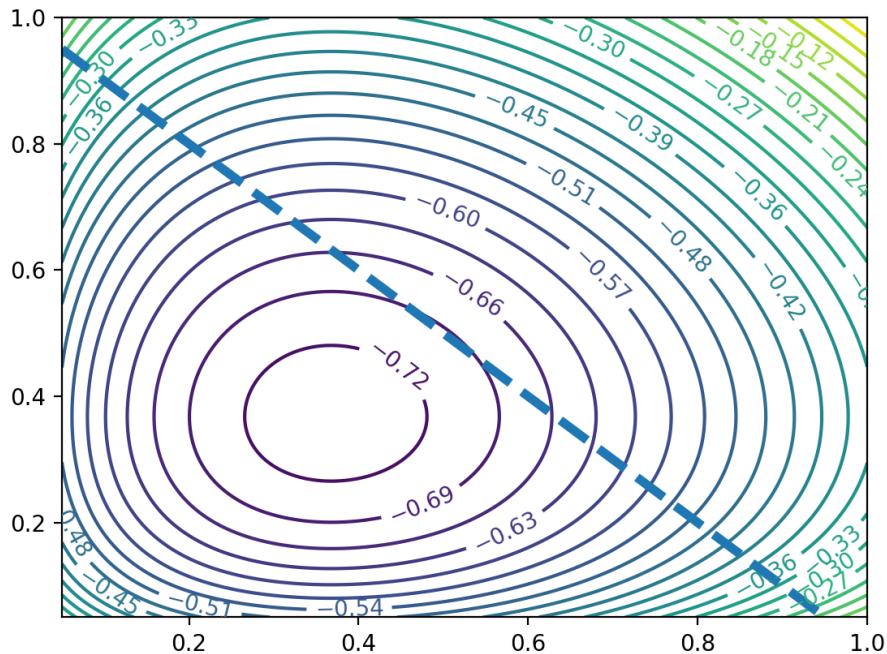
x=np.linspace(0.05,1,200)
y=np.linspace(0.05,1,200)
```

Editeur

```
X , Y = np.meshgrid(x,y)
Z = f(X,Y)
graphe = plt.contour(X,Y,Z,26)
plt.clabel(graphe,inline=1,fontsize=10)
# permet d'obtenir 26 lignes de niveau

plt.ylim([0.05,1])
# restriction de l'axe des ordonnées

plt.plot(x,1-x,'--')
# tracé de la droite
plt.show()
```



Problème D

Intervalle de confiance pour le paramètre d'une loi de Bernoulli

Partie A - Inégalité de Hoeffding

Soit $(\Omega, \mathcal{A}, \mathbf{P})$ un espace probabilisé. Soit X une variable aléatoire sur $(\Omega, \mathcal{A}, \mathbf{P})$, à valeurs dans $[-1, 1]$ et d'espérance nulle. On considère dans ce problème une suite $(X_i)_{i \in \mathbb{N}^*}$ de variables aléatoires discrètes sur $(\Omega, \mathcal{A}, \mathbf{P})$, mutuellement indépendantes et de même loi que X . Pour tout $n \in \mathbb{N}^*$, on note $S_n = (X_1 + \dots + X_n)/n$.

22. Montrer que pour tout $t \in \mathbb{R}_*^+$, pour tout $\varepsilon \in \mathbb{R}_*^+$ et pour tout $n \in \mathbb{N}^*$, on a :

$$\mathbf{E}(e^{tnS_n}) = \left(\mathbf{E}(e^{tX}) \right)^n, \quad \text{puis} \quad \mathbf{P}(S_n \geq \varepsilon) = \mathbf{P}(e^{tnS_n} \geq e^{tn\varepsilon}) \leq \frac{\left(\mathbf{E}(e^{tX}) \right)^n}{e^{tn\varepsilon}}.$$

- **Majoration de $\mathbf{E}(e^{tX})$**

23. Soit $a > 1$. On considère la fonction g_a définie par : $\forall x \in \mathbb{R}, \quad g_a(x) = \frac{1-x}{2}a^{-1} + \frac{1+x}{2}a - a^x.$

- a) Montrer que la fonction g_a est dérivable sur $\mathbb{R}$ et vérifier que g'_a est décroissante sur $\mathbb{R}$.
 b) Préciser $g_a(-1)$ et $g_a(1)$. Puis vérifier que pour tout $x \in [-1, 1]$, $g_a(x) \geq 0$.
24. En déduire que pour tout $t \in \mathbb{R}_+^*$ et pour tout $x \in [-1; 1]$, on a

$$e^{tx} \leq \frac{1-x}{2}e^{-t} + \frac{1+x}{2}e^t \quad \text{puis} \quad \mathbb{E}(e^{tX}) \leq \frac{e^t + e^{-t}}{2}.$$

25. En remarquant que pour tout entier $k \in \mathbb{N}$ et tout $x \in \mathbb{R}$, $\frac{x^{2k}}{(2k)!} \leq \frac{1}{k!} \left(\frac{x^2}{2}\right)^k$, justifier que pour tout $t > 0$, $\mathbb{E}(e^{tX}) \leq e^{\frac{t^2}{2}}$.
- **Majoration de $\mathbf{P}(|S_n| \geq \varepsilon)$**
 Dans ce paragraphe, on considère un entier $n \in \mathbb{N}^*$ et un réel $\varepsilon > 0$.
26. Montrer que la fonction $t \in \mathbb{R} \mapsto e^{-nt\varepsilon + n\frac{t^2}{2}}$ atteint un minimum en un point que l'on précisera.
27. En déduire que $\mathbf{P}(S_n \geq \varepsilon) \leq e^{-n\frac{\varepsilon^2}{2}}$, puis que : $\mathbf{P}(|S_n| \geq \varepsilon) \leq 2e^{-n\frac{\varepsilon^2}{2}}$.

Partie B - Intervalles de confiance, comparaisons

Soit p un paramètre réel inconnu vérifiant $0 < p < 1$. Pour $n \in \mathbb{N}^*$, on considère $B_1, B_2, \dots, B_n$, n variables aléatoires définies sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbf{P})$, indépendantes et de même loi de Bernoulli de paramètre p . On pose, pour tout $n \in \mathbb{N}^*$,

$$\overline{B}_n = \frac{1}{n} \sum_{k=1}^n B_k.$$

- **Rappels**

28. Justifier à l'aide de l'inégalité de Bienaymé-Tchebychev qu'un intervalle de confiance de p au niveau de risque α est donné par

$$\left[\overline{B}_n - \frac{1}{2\sqrt{n\alpha}}; \overline{B}_n + \frac{1}{2\sqrt{n\alpha}} \right] \quad (\text{I}_1)$$

29. Justifier à l'aide du théorème central limite qu'un intervalle de confiance asymptotique de p au niveau de risque α est donné par

$$\left[\overline{B}_n - \frac{t_\alpha}{2\sqrt{n}}; \overline{B}_n + \frac{t_\alpha}{2\sqrt{n}} \right] \quad (\text{I}_2)$$

- **Un troisième exemple**

30. Montrer que pour tout $\varepsilon > 0$, on a : $\mathbf{P}\left(\left|\overline{B}_n - p\right| \geq \varepsilon\right) \leq 2 \exp\left(-n\varepsilon^2/2\right)$.
31. En déduire que

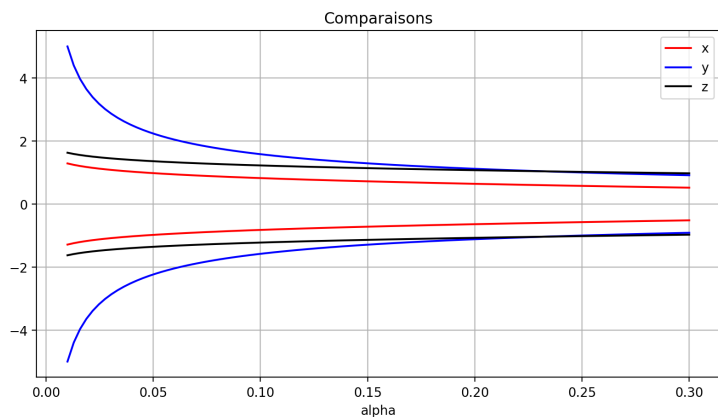
$$\left[\overline{B}_n - \sqrt{-\frac{\ln(\frac{\alpha}{2})}{2n}}; \overline{B}_n + \sqrt{-\frac{\ln(\frac{\alpha}{2})}{2n}} \right] \quad (\text{I}_3)$$

un intervalle de confiance de risque α pour le paramètre p .

32. À l'aide des courbes suivantes, quel intervalle de confiance choisiriez-vous?

Editeur

```
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as ss
alpha=np.linspace(0.01,0.3,100)
x=ss.norm.ppf(1-alpha/2,0,1)/2
y=1/(2*np.sqrt(alpha))
z=np.sqrt(-np.log(alpha)/2)
plt.plot(alpha,x,'r',label='x')
plt.plot(alpha,y,'b',label='y')
plt.plot(alpha,z,'k',label='z')
plt.plot(alpha,-x,'r')
plt.plot(alpha,-y,'b')
plt.plot(alpha,-z,'k')
plt.title("Comparaisons")
plt.xlabel("alpha")
plt.legend()
plt.grid()
plt.show()
```



La commande `ss.norm.ppf(t,0,1)` renvoie une approximation de $\Phi^{-1}(t)$ où Φ^{-1} désigne l'application réciproque de la fonction de répartition de la loi normale centrée réduite.

– FIN –

DS 10 - sujet *

THÈME : ESTIMATIONS

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

Intervalle de confiance pour le paramètre d'une loi de Bernoulli

Partie A - Inégalité de Hoeffding

Soit $(\Omega, \mathcal{A}, \mathbf{P})$ un espace probabilisé. Soit X une variable aléatoire sur $(\Omega, \mathcal{A}, \mathbf{P})$, à valeurs dans $[-1, 1]$ et d'espérance nulle. On considère dans ce problème une suite $(X_i)_{i \in \mathbb{N}^*}$ de variables aléatoires discrètes sur $(\Omega, \mathcal{A}, \mathbf{P})$, mutuellement indépendantes et de même loi que X . Pour tout $n \in \mathbb{N}^*$, on note :

$$S_n = \frac{X_1 + \dots + X_n}{n}.$$

1. Montrer que pour tout $t \in \mathbb{R}_+^*$, pour tout $\varepsilon \in \mathbb{R}_+^*$ et pour tout $n \in \mathbb{N}^*$, on a :

$$\mathbf{E}(e^{tnS_n}) = \left(\mathbf{E}(e^{tX}) \right)^n, \quad \text{puis} \quad \mathbf{P}(S_n \geq \varepsilon) = \mathbf{P}(e^{tnS_n} \geq e^{tn\varepsilon}) \leq \frac{\left(\mathbf{E}(e^{tX}) \right)^n}{e^{tn\varepsilon}}.$$

• **Majoration de $\mathbf{E}(e^{tX})$**

2. Soit $a > 1$. On considère la fonction g_a définie par :

$$\forall x \in \mathbb{R}, \quad g_a(x) = \frac{1-x}{2} a^{-1} + \frac{1+x}{2} a - a^x.$$

- a) Montrer que la fonction g_a est dérivable sur $\mathbb{R}$ et vérifier que g'_a est décroissante sur $\mathbb{R}$.
 b) Préciser $g_a(-1)$ et $g_a(1)$. Puis vérifier que pour tout $x \in [-1, 1]$, $g_a(x) \geq 0$.

3. En déduire que pour tout $t \in \mathbb{R}_+^*$ et pour tout $x \in [-1, 1]$, on a

$$e^{tx} \leq \frac{1-x}{2} e^{-t} + \frac{1+x}{2} e^t \quad \text{puis} \quad \mathbf{E}(e^{tX}) \leq \frac{e^t + e^{-t}}{2}.$$

4. En remarquant que pour tout entier $k \in \mathbb{N}$ et tout $x \in \mathbb{R}$, $\frac{x^{2k}}{(2k)!} \leq \frac{1}{k!} \left(\frac{x^2}{2} \right)^k$, justifier que pour tout $t > 0$

$$\mathbf{E}(e^{tX}) \leq e^{\frac{t^2}{2}}.$$

• **Majoration de $\mathbf{P}(|S_n| \geq \varepsilon)$**

Dans ce paragraphe, on considère un entier $n \in \mathbb{N}^*$ et un réel $\varepsilon > 0$.

5. Montrer que la fonction $t \in \mathbb{R} \mapsto e^{-nt\varepsilon + n\frac{t^2}{2}}$, atteint un minimum en un point que l'on précisera.
 6. En déduire que $\mathbf{P}(S_n \geq \varepsilon) \leq e^{-n\frac{\varepsilon^2}{2}}$, puis que :

$$\mathbf{P}(|S_n| \geq \varepsilon) \leq 2e^{-n\frac{\varepsilon^2}{2}}.$$

Partie B - Intervalles de confiances, comparaisons

Soit p un paramètre réel inconnu vérifiant $0 < p < 1$. Pour $n \in \mathbb{N}^*$, on considère $B_1, B_2, \dots, B_n$, n variables aléatoires définies sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbf{P})$, indépendantes et de même loi de Bernoulli de paramètre p . On pose, pour tout $n \in \mathbb{N}^*$,

$$\overline{B}_n = \frac{1}{n} \sum_{k=1}^n B_k.$$

• **Rappels**

7. Justifier à l'aide de l'inégalité de Bienaymé-Tchebychev qu'un intervalle de confiance de p au niveau de risque α est donné par

$$\left[\overline{B}_n - \frac{1}{2\sqrt{n\alpha}}; \overline{B}_n + \frac{1}{2\sqrt{n\alpha}} \right] \quad (I_1)$$

8. Justifier à l'aide du théorème central limite qu'un intervalle de confiance asymptotique de p au niveau de risque α est donné par

$$\left[\overline{B}_n - \frac{t_\alpha}{2\sqrt{n}}; \overline{B}_n + \frac{t_\alpha}{2\sqrt{n}} \right] \quad (I_2)$$

• **Un troisième exemple**

9. Montrer que pour tout $\varepsilon > 0$, on a : $\mathbf{P}\left(\left|\overline{B}_n - p\right| \geq \varepsilon\right) \leq 2 \exp(-2n\varepsilon^2)$.

10. En déduire que

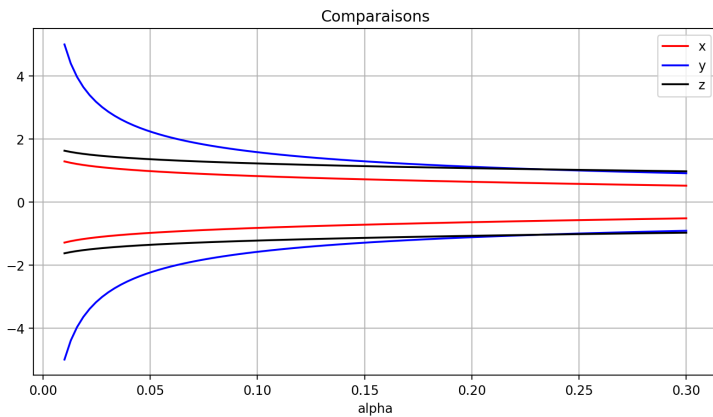
$$\left[\overline{B}_n - \sqrt{-\frac{\ln(\frac{\alpha}{2})}{2n}}; \overline{B}_n + \sqrt{-\frac{\ln(\frac{\alpha}{2})}{2n}} \right] \quad (I_3)$$

un intervalle de confiance de risque α pour le paramètre p .

11. À l'aide des courbes suivantes, quel intervalle de confiance choisirez-vous?

Editeur

```
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as ss
alpha=np.linspace(0.01,0.3,100)
x=ss.norm.ppf(1-alpha/2,0,1)/2
y=1/(2*np.sqrt(alpha))
z=np.sqrt(-np.log(alpha/2)/2)
plt.plot(alpha,x,'r',label='x')
plt.plot(alpha,y,'b',label='y')
plt.plot(alpha,z,'k',label='z')
plt.plot(alpha,-x,'r')
plt.plot(alpha,-y,'b')
plt.plot(alpha,-z,'k')
plt.title("Comparaisons")
plt.xlabel("alpha")
plt.legend()
plt.grid()
plt.show()
```



La commande `ss.norm.ppf(t,0,1)` renvoie une approximation de $\Phi^{-1}(t)$ où Φ^{-1} désigne l'application réciproque de la fonction de répartition de la loi normale centrée réduite.

Problème B

Poissons de mars

On cherche à évaluer le nombre N de poissons dans un étang.

On prélève dans l'étang un échantillon de m poissons. On les marque et on les remet dans l'étang. On propose deux méthodes d'estimation de N .

• **Première méthode**

On prélève des poissons dans l'étang, au hasard et avec remise. On note X_n la variable aléatoire égale au nombre de poissons qu'il a été nécessaire de pêcher pour obtenir n poissons marqués, pour tout $n \in \llbracket 1, m \rrbracket$.

Pour tout $i \in \llbracket 2, n \rrbracket$, on pose :

$$D_i = X_i - X_{i-1} \quad \text{et} \quad D_1 = X_1.$$

On suppose que les D_i , pour $i \in \llbracket 2, n \rrbracket$ sont des variables aléatoires indépendantes.

12. Reconnaître la loi de D_1 .

13. a) Reconnaître la loi de D_i pour tout $i \in \llbracket 2, n \rrbracket$. Préciser son espérance et sa variance.

b) Vérifier que

$$\mathbf{E}(X_n) = \frac{nN}{m} \quad \text{et} \quad \mathbf{V}(X_n) = n \left(1 - \frac{m}{N}\right) \left(\frac{N}{m}\right)^2.$$

14. Déterminer, A_n , un estimateur sans biais de N qui soit fonction de X_n .
15. A_n est-il un estimateur convergent de N ?
16. Pour n assez grand, par quelle loi peut-on approcher la loi de A_n ?
17. a) On suppose que $\sigma(A_n) \leq 100$. Déterminer un intervalle de confiance asymptotique de N , au niveau de confiance 0,9.
On donne $\Phi(1,64) \approx 0,95$ où Φ est la fonction de répartition de la loi normale centrale réduite.
- b) *Application numérique*
On a marqué 200 poissons, puis effectué 450 prélèvements pour obtenir 50 poissons marqués. Déterminer l'intervalle de confiance de N obtenu à la question précédente pour l'échantillon étudié.
- **Deuxième méthode**
On prélève successivement et avec remise n poissons. Soit Y_n la variable aléatoire égale au nombre de poissons marqués parmi eux.
18. Déterminer, A_n , un estimateur sans biais de $\frac{1}{N}$ qui soit fonction de Y_n .
19. Pour quelle raison évidente ne peut-on prendre $1/A_n$ comme estimateur de N ?
On pose alors : $B_n = \frac{n(n+1)}{Y_n+1}$.
20. a) Calculer l'espérance de B_n .
b) Quelle qualité peut-on en déduire sur cet estimateur de N ?

Problème C Propagation d'un virus

Dans tout le problème, N désigne un entier naturel fixé supérieur ou égal à 2, et p un réel fixé de l'intervalle $]0, 1[$. On pose : $q = 1 - p$. Soit n un entier naturel quelconque.

Dans une population de N individus, on s'intéresse à la propagation d'un certain virus. Chaque jour, on distingue dans cette population trois catégories d'individus : en premier lieu, les individus sains, c'est-à-dire ceux qui ne sont pas porteurs du virus, ensuite les individus qui viennent d'être contaminés et qui sont inoffensifs pour les autres, et enfin, les individus contaminés par le virus et qui sont contagieux.

Ces trois catégories évoluent jour après jour selon le modèle suivant ;

- chaque jour n , chaque individu sain peut être contaminé par n'importe lequel des individus contagieux ce jour avec la même probabilité p , ces contaminations éventuelles étant indépendantes les unes des autres ;
- un individu contaminé le jour n devient contagieux le jour $n + 1$;
- chaque individu contagieux le jour n redevient sain le jour $n + 1$.

On note alors X_n le nombre aléatoire d'individus contagieux le jour n . On remarquera que si, pour un certain entier naturel i , on a $X_i = 0$, alors on a aussi $X_{i+1} = 0$. Les variables aléatoires $X_0, X_1, \dots, X_n, \dots$ sont supposées définies sur un même espace probabilisé $(\Omega, \mathcal{A}, \mathbf{P})$, et $E(X_n)$ désigne, pour tout n de $\mathbb{N}$, l'espérance de X_n .

Partie A

On suppose que pour tout entier naturel n et pour tout entier i de $[[0, N]]$, on a : $\mathbf{P}([X_n = i]) > 0$.

On suppose également que pour tout couple (i, j) de $[[0, N]]^2$, le réel $q_{i,j}$ défini par :

$$q_{i,j} = \mathbf{P}_{[X_n=i]}([X_{n+1} = j]),$$

est indépendant de n .

Soit Q la matrice de $\mathcal{M}_{N+1}(\mathbb{R})$ définie par : $Q = (q_{i,j})_{0 \leq i, j \leq N}$.

21. a) Déterminer, pour tout j de $[[0, N]]$, les probabilités $q_{0,j}$ et $q_{N,j}$. De même, déterminer pour tout i de $[[0, N]]$, la probabilité $q_{i,N}$.
b) Justifier que si l'on a $j > N - i$, alors $q_{i,j} = 0$.
c) Montrer que pour tout i de $[[1, N - 1]]$, la loi de probabilité conditionnelle de X_{n+1} sachant $[X_n = i]$, est une loi binomiale dont on déterminera les paramètres.
22. a) Montrer que 1 est valeur propre de la matrice Q .
b) Soit λ une valeur propre de Q , et

$$V = \begin{bmatrix} V(0) \\ V(1) \\ \vdots \\ V(N) \end{bmatrix}$$

un vecteur propre associé à λ .

On pose : $|V(i)| = \max_{0 \leq j \leq N} |V(j)|$. Justifier que la composante $V(i)$ n'est pas nulle, puis, en examinant la ligne i du système $QV = \lambda V$, montrer que l'on a : $|\lambda| \leq 1$.

23. On pose pour tout entier naturel n :

$$U_n = \begin{bmatrix} \mathbf{P}(X_n = 0) \\ \mathbf{P}(X_n = 1) \\ \vdots \\ \mathbf{P}(X_n = N) \end{bmatrix}.$$

Déterminer en fonction de Q une matrice M de $\mathcal{M}_{N+1}(\mathbb{R})$ indépendante de n et vérifiant, pour tout n de $\mathbb{N}$, la relation : $U_{n+1} = MU_n$.
On suppose jusqu'à la fin de cette partie que la matrice M est diagonalisable, et que $\mathcal{B} = (V_0, \dots, V_N)$ est une base de vecteurs propres de M telle que, pour tout k de $\llbracket 0, N \rrbracket$, le vecteur propre V_k est associé à une valeur propre λ_k .
De plus, on suppose que : $\lambda_0 = 1$,

$$V_0 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

et que pour tout k de $\llbracket 1, N \rrbracket$, on a : $|\lambda_k| < 1$.

24. On décompose alors le vecteur U_0 sur la base $\mathcal{B}$: $U_0 = \sum_{k=0}^N \alpha_k V_k$.

- Déterminer, pour tout n de $\mathbb{N}$, la décomposition du vecteur U_n sur la base $\mathcal{B}$.
- On note, pour tout couple (k, i) de $\llbracket 0, N \rrbracket^2$, $V_k(i)$ la $(i+1)$ -ième composante du vecteur V_k . Exprimer, pour tout n de $\mathbb{N}$ et pour tout i de $\llbracket 0, N \rrbracket$, la probabilité de l'événement $[X_n = i]$ en fonction des réels α_k , λ_k et $V_k(i)$ ($k \in \llbracket 0, N \rrbracket$).
- Montrer que, pour tout i de $\llbracket 1, N \rrbracket$, on a : $\lim_{n \rightarrow +\infty} \mathbf{P}(X_n = i) = 0$.
- En déduire que le virus finit par disparaître presque sûrement, quelle que soit la loi de la variable aléatoire initiale X_0 .

Partie B. Estimations ponctuelle et par intervalle de confiance de p

On suppose que le paramètre p , qui exprime la probabilité qu'un individu contagieux transmette le virus à un individu sain, est inconnu, et on cherche à l'estimer. On rappelle que : $q = 1 - p$.

Pour m entier supérieur ou égal à 1, on considère un m -échantillon $(Y_1, Y_2, \dots, Y_m)$ de variables aléatoires indépendantes, de même loi de Bernoulli de paramètre p , définies sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbf{P})$. On pose : $\tilde{Y}_m = \frac{1}{m} \sum_{i=1}^m Y_i$.

Dans toute la suite de cette partie, on note ε un réel strictement positif quelconque.

25. a) Montrer que $\tilde{Y}_m$ est un estimateur sans biais de p ; déterminer son risque quadratique défini par

$$r_p(\tilde{Y}_m) = b_p(\tilde{Y}_m)^2 + \mathbf{V}(\tilde{Y}_m)$$

où $b_p(\tilde{Y}_m)$ désigne le biais de l'estimateur.

- A l'aide de l'inégalité de Bienaymé-Tchebychev, montrer que l'intervalle $\left[\tilde{Y}_m - \sqrt{\frac{5}{m}}, \tilde{Y}_m + \sqrt{\frac{5}{m}} \right]$ est un intervalle de confiance de p au niveau de confiance 0.95.

26. Soit θ un réel positif.

La question suivante propose une démonstration différente de l'inégalité de Hoeffding du premier problème connue aussi sous le nom d'inégalité de Chernoff

- Établir l'égalité suivante :

$$\mathbf{P}([\tilde{Y}_m - p \geq \varepsilon]) = \mathbf{P}\left(\left[e^{m\theta\tilde{Y}_m} \geq e^{m\theta(p+\varepsilon)} \right] \right).$$

- Montrer que si T est une variable aléatoire discrète finie à valeurs positives d'espérance $\mathbf{E}(T)$, et a un réel strictement positif, on a l'inégalité : $\mathbf{P}([T \geq a]) \leq \frac{\mathbf{E}(T)}{a}$.
- Soit g la fonction définie sur $\mathbb{R}^+$ par : $g(x) = \ln(pe^x + q)$. Déduire des questions précédentes, l'inégalité suivante :

$$\mathbf{P}([\tilde{Y}_m - p \geq \varepsilon]) \leq e^{m(g(\theta) - \theta(p+\varepsilon))}.$$

- Montrer que la fonction g est de classe C^2 sur $\mathbb{R}^+$ et vérifie, pour tout x de $\mathbb{R}^+$, l'inégalité : $|g''(x)| \leq \frac{1}{4}$.
- En déduire l'inégalité suivante : $g(\theta) \leq \theta p + \frac{\theta^2}{8}$.
- Étudier les variations de la fonction h définie sur $\mathbb{R}^+$ par : $h(x) = \frac{x^2}{8} - \varepsilon x$. En déduire l'inégalité :

$$\mathbf{P}([\tilde{Y}_m - p \geq \varepsilon]) \leq e^{-2m\varepsilon^2}.$$

27. a) Déduire des questions précédentes, l'inégalité suivante :

$$\mathbf{P}([| \tilde{Y}_m - p | \geq \varepsilon]) \leq 2e^{-2m\varepsilon^2}.$$

- Sachant que $\ln(0.025) \approx -3.688$, calculer $2e^{-2m\varepsilon^2}$ pour $\varepsilon = \sqrt{\frac{1.844}{m}}$. En déduire un nouvel intervalle de confiance de p au niveau de confiance 0.95. Comparer cet intervalle de confiance à celui obtenu à la question 25.b. Conclure.

– FIN –

DS 10* - solution

Problème A

1. Le fait que $tn > 0$, la croissance de l'exponentielle, puis l'inégalité de Markov, appliquée à la variable aléatoire positive bornée (donc admettant une espérance) e^{tnS_n} , donnent

$$\begin{aligned} \mathbf{P}(S_n \geq \varepsilon) &= \mathbf{P}(tnS_n \geq tn\varepsilon) \\ &= \mathbf{P}(e^{tnS_n} \geq e^{tn\varepsilon}) \leq \frac{\mathbf{E}(e^{tnS_n})}{e^{tn\varepsilon}}. \end{aligned}$$

Or, $e^{tnS_n} = \prod_{i=1}^n e^{tX_i}$ et les variables aléatoires e^{tX_i} sont bornées, donc admettent une espérance. Alors, l'indépendance des X_i et donc des variables e^{tX_i} (lemme des coalitions) donne

$$\mathbf{E}(e^{tnS_n}) = \prod_{i=1}^n \mathbf{E}(e^{tX_i}) = [\mathbf{E}(e^{tX})]^n.$$

Ce qui conclut.

2. Par différence d'une fonction polynomiale et d'une fonction puissance, g_a est de classe $\mathcal{C}^\infty$ et pour tout réel x ,

$$g_a''(x) = -(\ln a)^2 a^x < 0,$$

ce qui montre que g_a' est strictement décroissante.

Comme

$$g_a(-1) = a^{-1} - a^{-1} = 0 = a - a = g_a(1),$$

le théorème de Rolle entraîne que g_a' s'annule au moins une fois sur $] -1, 1[$. Comme g_a' est strictement décroissante, g_a' s'annule une seule fois et est d'abord positive, puis négative. Ainsi, il existe $x_0 \in] -1, 1[$ tel que g_a est strictement croissante sur $[-1, x_0]$ et strictement décroissante sur $[x_0, 1]$. En particulier, g_a est positive sur $[-1, 1]$.

3. Soit $t > 0$. En prenant $a = e^t > 1$, l'inégalité $g_a(x) \geq 0$ donne

$$\frac{1-x}{2}e^{-t} + \frac{1+x}{2}e^t - e^{tx} \geq 0$$

d'où l'inégalité demandée.

X est à valeurs dans $[-1; 1]$, on a donc

$$e^{tX} \leq \frac{1-X}{2}e^{-t} + \frac{1+X}{2}e^t$$

Par croissance et linéarité de l'espérance

$$\mathbf{E}(e^{tX}) \leq \frac{1-\mathbf{E}(X)}{2}e^{-t} + \frac{1+\mathbf{E}(X)}{2}e^t.$$

Puis, en utilisant le fait que $\mathbf{E}(X) = 0$, il vient

$$\mathbf{E}(e^{tX}) \leq \frac{e^{-t} + e^t}{2}.$$

4. Si $k \in \mathbb{N}^*$,

$$(2k)! = k! \prod_{j=1}^k (k+j) \geq k! \prod_{j=1}^k 2 = 2^k k!,$$

et l'inégalité est aussi vraie pour $k = 0$. En en prenant l'inverse, puis en multipliant par la quantité positive t^{2k} , il vient

$$\frac{t^{2k}}{(2k)!} \leq \frac{1}{k!} \left(\frac{t^2}{2} \right)^k.$$

En utilisant alors la question précédente et le développement en série de la fonction exponentielle

$$\mathbf{E}(e^{tX}) \leq \sum_{k=0}^{\infty} \frac{t^{2k}}{(2k)!} \leq \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{t^2}{2} \right)^k = e^{t^2/2}.$$

5. Posons $\varphi(t) = e^{-nt\varepsilon + n\frac{t^2}{2}}$. Alors, φ est de classe $\mathcal{C}^\infty$ et $\varphi'(t) = n(t - \varepsilon)\varphi(t)$ est du signe de $t - \varepsilon$. Il s'ensuit que φ admet un minimum en ε et que ce minimum vaut $\varphi(\varepsilon) = e^{-n\frac{\varepsilon^2}{2}}$.

6. Les questions précédentes donnent une majoration de $\mathbf{P}(S_n \geq \varepsilon)$ valable pour tout $t > 0$:

$$\mathbf{P}(S_n \geq \varepsilon) \stackrel{q.1}{\leq} e^{-nt\varepsilon} \mathbf{E}(e^{tX})^n \stackrel{q.4}{\leq} e^{-nt\varepsilon} \times e^{\frac{nt^2}{2}}$$

En particulier, pour $t = \varepsilon$, il vient

$$\mathbf{P}(S_n \geq \varepsilon) \leq e^{-n\frac{\varepsilon^2}{2}}.$$

Les variables aléatoires $-X_i$ vérifient les mêmes hypothèses que les X_i (elles sont à valeurs dans $[-1, 1]$ et indépendantes). Il s'ensuit que l'on peut appliquer la majoration ci-dessus à la variable $-S_n$, ce qui donne

$$\mathbf{P}(-S_n \geq \varepsilon) \leq e^{-n\frac{\varepsilon^2}{2}}.$$

Ainsi

$$\mathbf{P}(|S_n| \geq \varepsilon) = \mathbf{P}(S_n \leq -\varepsilon) + \mathbf{P}(S_n \geq \varepsilon) \leq 2e^{-n\frac{\varepsilon^2}{2}}.$$

- 7.8. Cours.

9. Voir question 6.

10. Avec le choix

$$\varepsilon = \sqrt{-\ln(\alpha/2)/2n}$$

$$\text{on a } 2e^{-n\varepsilon^2} = 2e^{+\ln(\alpha/2)} = \alpha.$$

Ainsi, par passage au complémentaire

$$\begin{aligned} & \mathbf{P}(\bar{B}_n - \varepsilon \leq p \leq \bar{B}_n + \varepsilon) \\ &= \mathbf{P}(-\varepsilon \leq p - \bar{B}_n \leq \varepsilon) \\ &= \mathbf{P}(|p - \bar{B}_n| \leq \varepsilon). \\ &\geq \mathbf{P}(|p - \bar{B}_n| < \varepsilon) \\ &\geq 1 - \mathbf{P}(|p - \bar{B}_n| \geq \varepsilon) \\ &\geq 1 - 2 \exp(-n2\varepsilon^2) = 1 - \alpha. \end{aligned}$$

D'où le résultat.

11. À risque α fixé, il faut choisir, l'intervalle avec la plus petite longueur, Pour n suffisamment grand, I_2 semble donc le meilleur choix.