Sujets de révisions
Problèmes du TOP3

HEC/ESSEC I 2023 - loi normale centrée réduite, loi de Bernoulli, loi binomiale, convergence uniforme, estimation par intervalle de confiance

On s'intéresse dans ce sujet à la méthode de Stein, introduite par Charles Stein (1920/2016) en 1972, dont les développements et applications sont nombreux.

Les parties 1 et 2 concernent la justification de la méthode, elles sont indépendantes.

Dans la partie 3, on s'intéresse à l'estimation en un point d'une densité d'une loi de probabilité. Cette partie peut être traitée indépendamment des deux premières parties.

Dans la partie 4, on met en œuvre la méthode de Stein, vue dans les parties 1 et 2, pour établir des convergences « uniformes » en loi et on démontre le résultat admis dans la partie 3. Cette partie est indépendante de la partie 3 à l'exception de sa dernière question.

Dans tout le problème :

- les variables aléatoires considérées sont définies sur le même espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$.
- si X est une variable aléatoire, $\mathbb{E}(X)$ et $\mathbb{V}(X)$ désignent respectivement, lorsqu'elles existent, l'espérance et la variance de X.
- W désigne l'ensemble des fonctions h de classe $\mathcal{C}^1$ sur $\mathbb{R}$ telles que :

$$\forall x \in \mathbb{R},\ \left|h'(x)\right| \leqslant 1$$

- N est une variable aléatoire qui suit la loi normale $(0, 1)$.
- on admet que si X est une variable aléatoire possédant une espérance et $h \in W$, $\mathbb{E}(h(X))$ existe. On note en particulier c_h l'espérance de $h(N)$.
- On note Φ la fonction de répartition de la loi normale $(0, 1)$ définie par, pour tout $x \in \mathbb{R}$, $\Phi(x) = \displaystyle\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}}\, dt$. On rappelle que c'est la primitive sur $\mathbb{R}$, qui vaut $\frac{1}{2}$ en 0, de la fonction $\varphi : t \mapsto \dfrac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}}$.

Partie 1 - Transformation de Stein

Soit $h \in W$. On définit sur $\mathbb{R}$, la fonction $\theta : x \mapsto \dfrac{\Phi(x)}{\varphi(x)}$ et la fonction f_h par,

$$f_h : x \mapsto \theta(-x) \int_{-\infty}^{x} h'(t)\Phi(t)\ dt + \theta(x) \int_{x}^{+\infty} h'(t)(1 - \Phi(t))\ dt$$

lorsque ces intégrales convergent.
L'objectif principal de cette partie est d'obtenir, pour X une variable aléatoire admettant une espérance, une expression de $\mathbb{E}(h(X)) - \mathbb{E}(h(N))$ qui ne fait pas intervenir N directement.

1. ***a)*** Montrer que pour tout $x \geqslant 0$ et $t \in [x, +\infty[$, $0 \leqslant x\varphi(t) \leqslant t\varphi(t)$. En déduire que :

$$\forall x \geqslant 0,\ 0 \leqslant x(1-\Phi(x)) \leqslant \varphi(x)$$

(on remarquera que pour tout $t \in \mathbb{R}$, $\varphi'(t) = -t\varphi(t)$)

b) Procéder de façon analogue pour montrer que : $\forall x \leqslant 0$, $-\varphi(x) \leqslant x\Phi(x) \leqslant 0$.

c) En déduire à l'aide d'une intégration par parties, pour tout x réel, la convergence des intégrales qui suivent et montrer que :

$$\int_{-\infty}^{x} \Phi(t)\ dt = x\Phi(x) + \varphi(x) \quad \text{et} \quad \int_{x}^{+\infty} (1-\Phi(t))\ dt = -x(1-\Phi(x)) + \varphi(x) \qquad (R_1)$$

2. ***a)*** Montrer que pour tous réels x et y,

$$|h(x) - h(y)| \leqslant |x-y|\,, \quad \text{puis que } |h(x)| \leqslant |x| + |h(0)|$$

b) Pour tout x réel, justifier la convergence de $\displaystyle\int_{-\infty}^{x} h'(t)\Phi(t)\ dt$ et montrer que :

$$\int_{-\infty}^{x} h'(t)\Phi(t)\ dt = h(x)\Phi(x) - \int_{-\infty}^{x} h(t)\varphi(t)\ dt$$

On admet de même que, $\displaystyle\int_{x}^{+\infty} h'(t)(1-\Phi(t))\ dt$ *converge et que,*

$$\int_{x}^{+\infty} h'(t)(1-\Phi(t))\ dt = -h(x)(1-\Phi(x)) + \int_{x}^{+\infty} h(t)\varphi(t)\ dt$$

c) En déduire que, pour tout x réel :

$$-\int_{-\infty}^{x} h'(t)\Phi(t)\ dt + \int_{x}^{+\infty} h'(t)(1-\Phi(t))\ dt = c_h - h(x)$$

3. ***a)*** Établir que pour tout $x \in \mathbb{R}$,

$$\begin{aligned} \theta'(x) &= 1 + x\theta(x) \\ \theta''(x) &= x + (1+x^2)\theta(x) \\ \theta(-x)\Phi(x) &= \theta(x)(1-\Phi(x)) \end{aligned}$$

b) En déduire que f_h est une fonction de classe $\mathcal{C}^1$ sur $\mathbb{R}$ qui vérifie, pour tout x réel :

$$f_h'(x) - xf_h(x) = c_h - h(x)$$

Pourquoi peut-on alors affirmer que f_h est de classe $\mathcal{C}^2$ sur $\mathbb{R}$?

c) En conclure que, si X est une variable aléatoire admettant une espérance,

$$|\mathbb{E}(h(X)) - \mathbb{E}(h(N))| = \left|\mathbb{E}(f_h'(X) - Xf_h(X))\right|$$

4. *Majoration de* $|f_h|$.

a) Montrer, en utilisant les égalités (R_1), que pour tout x réel :

$$\theta(-x)\int_{-\infty}^{x} \Phi(t)\ dt + \theta(x)\int_{x}^{+\infty} (1-\Phi(t))\ dt = 1$$

b) En déduire que pour tout x réel : $|f_h(x)| \leqslant 1$.

5. *Majoration de* $|f_h''|$.

a) Montrer que pour tout x réel :

$$\theta''(-x)\int_{-\infty}^{x} \Phi(t)\ dt + \theta''(x)\int_{x}^{+\infty} (1-\Phi(t))\ dt = 1$$

b) Établir pour tout x réel l'égalité :

$$f_h''(x) = -h'(x) + \theta''(-x)\int_{-\infty}^{x} h'(t)\Phi(t)\ dt + \theta''(x)\int_{x}^{+\infty} h'(t)(1-\Phi(t))\ dt$$

c) Étudier les variations sur $\mathbb{R}$ de la fonction $x \mapsto \Phi(x) + \dfrac{x}{1+x^2}\varphi(x)$. En déduire son signe et le signe de θ''.
En conclure que, pour tout x réel : $|f_h''(x)| \leqslant 2$.

Partie 2 - Majoration uniforme de la distance de Kolmogorov

Dans la suite du problème, si X est une variable aléatoire de fonction de répartition F_X, on définit, pour tout x réel, $d_X(x)$ la distance de Kolmogorov au point x entre la loi de X et la loi normale centrée réduite par :

$$d_X(x) = |F_X(x) - \Phi(x)|$$

On définit, pour tout x réel, la fonction h_x sur $\mathbb{R}$ par $h_x(t) = \begin{cases} 1 & \text{si } t \leqslant x \\ 0 & \text{si } t > x \end{cases}$.

On définit aussi la fonction γ sur $\mathbb{R}$ par :

$$\gamma(t) = \begin{cases} 1 & \text{si } t < 0 \\ 0 & \text{si } t > 1 \\ 1 - 3t^2 + 2t^3 & \text{si } t \in [0,1] \end{cases}$$

Soit X une variable aléatoire.

6. Pour tout x réel, quelle est la loi de la variable aléatoire $h_x(X)$? En déduire que $\mathbb{E}(h_x(X))$ existe et vaut $F_X(x)$.

7. ***a)*** Écrire une fonction **Python** `gamma(t)` qui calcule et renvoie la valeur de $\gamma(t)$, t étant donné.

b) Utiliser la fonction précédente pour écrire un script qui affiche le graphe de γ sur le segment $[-1, 2]$ dans un repère.

8. ***a)*** Montrer que γ est continue sur $\mathbb{R}$, de classe $\mathcal{C}^1$ sur $\mathbb{R}$ privé de 0 et 1.

b) Étudier les variations de γ sur $[0,1]$ et montrer que pour tout $t \in \mathbb{R}$, $\gamma(t) \in [0,1]$.

c) Établir que γ est dérivable en 1 et que $\gamma'(1) = 0$.
On montrerait de même que γ est dérivable en 0 et que $\gamma'(0) = 0$. On l'admet.

d) Justifier que γ est de classe $\mathcal{C}^1$ sur $\mathbb{R}$ et que pour tout t réel $|\gamma'(t)| \leqslant \dfrac{3}{2}$.

On suppose dans la suite de cette partie que X admet une espérance et on considère un réel M_X qui vérifie, pour tout $h \in W$, $|\mathbb{E}(h(X)) - \mathbb{E}(h(N))| \leqslant M_X$.

9. Soit $t > 0$ et x un réel. Pour tout $y \in \mathbb{R}$, on pose $k_x(y) = \gamma\left(\dfrac{y-x}{t}\right)$.

a) Montrer que pour tout y réel, $h_x(y) \leqslant k_x(y)$.

b) *On admet l'existence de* $\mathbb{E}(k_x(X))$ *et de* $\mathbb{E}(k_x(N))$. Justifier l'inégalité suivante :

$$\mathbb{E}(h_x(X)) - \mathbb{E}(h_x(N)) \leqslant \mathbb{E}(k_x(X)) - \mathbb{E}(k_x(N)) + \mathbb{E}(k_x(N)) - \mathbb{E}(h_x(N))$$

c) Montrer que $\mathbb{E}(k_x(N)) - \mathbb{E}(h_x(N)) = \displaystyle\int_x^{x+t} k_x(u)\varphi(u)du$.

d) Établir que la fonction g, définie sur $\mathbb{R}$ par $g : u \mapsto \dfrac{2t}{3}k_x(u)$, appartient à W. En déduire que :

$$\mathbb{E}(h_x(X)) - \mathbb{E}(h_x(N)) \leqslant \frac{3}{2t}M_X + \frac{t}{\sqrt{2\pi}} \leqslant \frac{3}{2t}M_X + \frac{t}{2}$$

On admet de même, qu'en utilisant la fonction k_{x-t}*, on a :*

$$\mathbb{E}(h_x(N)) - \mathbb{E}(h_x(X)) \leqslant \frac{3}{2t}M_X + \frac{t}{2}$$

10. En étudiant la fonction $t \mapsto \dfrac{3}{2t}M_X + \dfrac{t}{2}$ sur $]0, +\infty[$, en déduire que, pour tout x réel,

$$|\mathbb{E}(h_x(X)) - \mathbb{E}(h_x(N))| \leqslant \sqrt{3M_X}, \text{ puis que } d_X(x) \leqslant \sqrt{3M_X} \qquad (R_2)$$

Partie 3 - Estimation d'une densité

On considère X une variable aléatoire à densité de fonction de répartition F et de densité de probabilité f qui dépendent d'un paramètre inconnu θ, où $\theta \in \Theta$, Θ un intervalle de $\mathbb{R}$.

Soit a un point de continuité de f, fixé. On souhaite estimer $f(a)$.

Par exemple, si X suit la loi exponentielle de paramètre θ et $a > 0$, on souhaite estimer $\theta \mathrm{e}^{-\theta a}$.

On dispose pour tout $\theta \in \Theta$, d'une suite de variables aléatoires $(X_i)_{i\geqslant 1}$ indépendantes de même loi que X.

On choisit une suite $(h_n)_{n\geqslant 1}$ de réels strictement positifs tels que :

$$\lim_{n\to+\infty} h_n = 0 \text{ et } \lim_{n\to+\infty} nh_n = +\infty$$

Pour tout $n \in \mathbb{N}^*$, et $\omega \in \Omega$, on définit :

$$C_n(\omega) \text{ comme le nombre d'indices } i \in [\![1, n]\!] \text{ tels que } X_i(\omega) \in]a - h_n, a + h_n]$$

et $f_n(\omega) = \dfrac{1}{2nh_n}C_n(\omega)$.

11. On suppose que l'on dispose d'un fichier `stats.csv` qui comporte une colonne nommée `salaire`. On considère que les valeurs de cette colonne constituent la réalisation d'un échantillon de la loi de X dont la taille dépasse 10000.

a) Après avoir exécuté `import pandas as pd`, quelle(s) instruction(s) permet(tent) de lire dans le fichier `stats.csv` les valeurs de la colonne `salaire` et d'affecter cette série `pandas` obtenue à une variable `échantillon` ?

On supposera que le fichier `stats.csv` se trouve dans le répertoire de travail.

b) On souhaite calculer et afficher $f_n(\omega)$ pour a donné, lorsque la réalisation d'un échantillon $(X_1(\omega), \ldots, X_n(\omega))$ de la loi de X est représentée en **Python** par `échantillon` et, pour tout $n \in \mathbb{N}^*$, $h_n = \dfrac{1}{\sqrt{n}}$.

Compléter le script suivant pour qu'il réalise cette tâche.

```
a = float(input('a='))
n = échantillon.count()
h = 1 / np.sqrt(n)
C = 0
for i in range(n):
    if ... and ...:
        ... += 1
print(C / ...)
```

12. Montrer que C_n suit une loi binomiale de paramètres (n, p_n) en précisant l'expression de p_n en fonction de a et h_n.

En déduire que $\mathbb{E}(f_n)$ existe et vaut : $\dfrac{F(a+h_n)-F(a-h_n)}{2h_n}$.

13. ***a)*** En utilisant la dérivabilité de F en a, montrer que $\lim\limits_{n\to+\infty} \mathbb{E}(f_n) = f(a)$.

b) Montrer que $\mathbb{V}(f_n)$ existe et que $\lim\limits_{n\to+\infty} \mathbb{V}(f_n) = 0$.

On suppose désormais, que $f(a) > 0$, que pour tout $n \in \mathbb{N}^*$, $p_n \in]0,1[$, que F est de classe $\mathcal{C}^2$ au voisinage de a et que $\lim\limits_{n\to+\infty} nh_n^3 = 0$.

On note pour tout $n \geqslant 1$, $\sigma_n = \sqrt{np_n(1-p_n)}$ et $\theta_n = \sqrt{2h_n f(a)}$.

On définit les variables aléatoires : $D_n = \dfrac{C_n - np_n}{\sigma_n}$ et $\hat{f}_n = \dfrac{\theta_n\sqrt{n}}{f(a)}(f_n - f(a))$.

14. ***a)*** En utilisant le développement limité de F à l'ordre 2 au point a, montrer que :

$$p_n \underset{n\to+\infty}{=} 2h_n f(a) + o(h_n^2) \underset{n\to+\infty}{=} \theta_n^2 + o(h_n^2)$$

b) En déduire que : $p_n \underset{n\to+\infty}{\sim} 2h_n f(a)$, puis que $\lim\limits_{n\to+\infty} np_n = +\infty$.

c) Montrer que $\hat{f}_n = \dfrac{\sigma_n}{\theta_n\sqrt{n}} D_n + \sqrt{n}\left(\dfrac{p_n}{\theta_n} - \theta_n\right)$ et que l'on a :

$$\lim_{n\to+\infty} \frac{\sigma_n}{\theta_n\sqrt{n}} = 1 \quad ; \quad \lim_{n\to+\infty} \sqrt{n}\left(\frac{p_n}{\theta_n} - \theta_n\right) = 0$$

On admet, dans la suite de cette partie, que $(\hat{f}_n)_{n\geqslant 1}$ converge en loi vers N ce qui implique que pour tout $(x,y) \in \mathbb{R}^2$, avec $x \leqslant y$, on a :

$$\lim_{n\to+\infty} \mathbb{P}(x \leqslant \hat{f}_n \leqslant y) = \Phi(y) - \Phi(x)$$

15. Soit $\alpha \in]0,1[$. On pose $\eta_\alpha = t_\alpha^2$ où t_α est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale $(0,1)$.

a) Montrer que $\lim\limits_{n\to+\infty} \mathbb{P}\left((f(a))^2 - \left(2f_n + \dfrac{\eta_\alpha}{2nh_n}\right) f(a) + f_n^2 \leqslant 0\right) = 1 - \alpha$.

b) On note, pour $n \geqslant 1$, $\Delta_n = \sqrt{\left(f_n + \dfrac{\eta_\alpha}{4nh_n}\right)^2 - f_n^2}$.

Montrer que :

$$\lim_{n\to+\infty} \mathbb{P}\left(f(a) \in \left[f_n + \frac{\eta_\alpha}{4nh_n} - \Delta_n, f_n + \frac{\eta_\alpha}{4nh_n} + \Delta_n\right]\right) = 1 - \alpha$$

Partie 4 - Convergence « uniforme » en loi vers la loi normale

Soit $n \in \mathbb{N}^*$. On considère n variables aléatoires $X_1, \dots, X_n$ indépendantes centrées qui possèdent un moment d'ordre 3. On admet alors que ces variables aléatoires possèdent une variance.

On pose, pour tout $k \in [\![1, n]\!]$, $\mathbb{E}(X_k^2) = v_k$, $S_n = \sum_{k=1}^{n} X_k$, $Y_k = S_n - X_k$ et on suppose que $\sum_{k=1}^{n} v_k = 1$.

Soit f une fonction de classe $\mathcal{C}^2$ sur $\mathbb{R}$ telle que pour tout x réel, $|f(x)| \leqslant 1$ et $|f''(x)| \leqslant 2$.

On admet que si X et Y sont des variables aléatoires possédant une espérance alors $\mathbb{E}(Yf(X))$ et $\mathbb{E}(f'(X))$ existent.

16. ***a)*** Montrer que $\sum_{k=1}^{n} v_k \mathbb{E}(f'(S_n) - f'(Y_k)) + \sum_{k=1}^{n} \mathbb{E}(X_k^2 f'(Y_k)) = \mathbb{E}(f'(S_n))$.

b) Montrer que $\sum_{k=1}^{n} \mathbb{E}(X_k [f(S_n) - f(Y_k)]) = \sum_{k=1}^{n} \mathbb{E}(X_k f(S_n)) = \mathbb{E}(S_n f(S_n))$.

c) En déduire que :

$$\mathbb{E}(f'(S_n) - S_n f(S_n)) = \sum_{k=1}^{n} v_k \mathbb{E}(f'(S_n) - f'(Y_k)) + \sum_{k=1}^{n} \mathbb{E}\left(X_k \left[X_k f'(Y_k) - (f(S_n) - f(Y_k))\right]\right)$$

17. Soit a et b deux réels.

a) Montrer que :

$$bf'(a) - (f(a+b) - f(a)) = \int_0^1 b(f'(a) - f'(a+tb))\, dt$$

b) En déduire que :

$$\left|bf'(a) - (f(a+b) - f(a))\right| \leqslant b^2$$

c) En conclure que :

$$\left|\mathbb{E}(f'(S_n) - S_n f(S_n))\right| \leqslant 2\sum_{k=1}^{n} v_k \mathbb{E}(|X_k|) + \sum_{k=1}^{n} \mathbb{E}(|X_k|^3)$$

puis, grâce à l'inégalité (R_2), que, pour tout x réel :

$$d_{S_n}(x) \leqslant \sqrt{3\left(2\sum_{k=1}^{n} v_k \mathbb{E}(|X_k|) + \sum_{k=1}^{n} \mathbb{E}(|X_k|^3)\right)} \qquad (R_3)$$

Une définition - *Dans la suite du sujet, si $(X_n)_{n\geqslant 1}$ est une suite de variables aléatoires réelles et $(\delta_n)_{n\geqslant 1}$ une suite réelle de limite nulle qui vérifient,*

$$\forall n \in \mathbb{N}^*, \forall x \in \mathbb{R}, d_{X_n}(x) \leqslant \delta_n$$

on dira alors que $(X_n)_{n\geqslant 1}$ converge uniformément en loi vers N.
On remarque, et on l'admet pour la suite, que si $(X_n)_{n\geqslant 1}$ converge uniformément en loi vers N alors $(X_n)_{n\geqslant 1}$ converge en loi vers N.

18. *Une première application.* On suppose dans cette question que $(Z_k)_{k\geqslant 1}$ est une suite de variables aléatoires indépendantes, suivant la même loi et admettant des moments d'ordre 1 à 3.

On note pour $i \in [\![1, 3]\!]$, $s_i = \mathbb{E}(|Z_k - \mathbb{E}(Z_k)|^i)$, $\sigma = \sqrt{s_2}$ et , $X_k = \dfrac{Z_k - \mathbb{E}(Z_k)}{\sigma\sqrt{n}}$ pour tout $k \in \mathbb{N}^*$.

On suppose que $\sigma \neq 0$.

On utilise les notations de la question précédente.

a) Montrer que l'on peut appliquer l'inégalité (R_3) qui donne ici :

$$d_{S_n}(x) \leqslant \sqrt{3\frac{2\sigma^2 s_1 + s_3}{\sigma^3\sqrt{n}}}$$

b) En déduire que $(S_n)_{n\geqslant 1}$ converge uniformément en loi vers N, donc converge en loi vers N. Quel résultat du cours nous aurait permis d'obtenir cette dernière convergence directement ?

19. *Une deuxième application.* On suppose dans cette question que $Z_1, \ldots, Z_n$ sont des variables aléatoires indépendantes suivant la même loi de Bernoulli de paramètre $p_n \in]0,1[$.

On pose $\sigma_n = \sqrt{np_n(1-p_n)}$ et $X_k = \dfrac{Z_k - p_n}{\sigma_n}$ pour tout $k \in [\![1, n]\!]$.

a) Montrer que $\mathbb{E}(|X_k|) = \dfrac{2\sigma_n}{n}$ et $\mathbb{E}(|X_k|^3) \leqslant \dfrac{2}{n\sigma_n}$.

b) En déduire que, pour tout x réel :

$$d_{S_n}(x) \leqslant 2\sqrt{3\left(\frac{\sigma_n}{n} + \frac{1}{2\sigma_n}\right)}$$

c) Justifier le résultat suivant :

si pour tout $n \in \mathbb{N}^*$, T_n est une variable aléatoire qui suit la loi binomiale (n, p_n) avec $\lim\limits_{n\to+\infty} p_n = 0$ et $\lim\limits_{n\to+\infty} np_n = +\infty$ alors, $\left(\dfrac{T_n - np_n}{\sqrt{np_n(1-p_n)}}\right)_{n\geqslant 1}$ converge uniformément en loi vers N.

20. *Un petit lemme.* Soit $(V_n)_{n\geqslant 1}$ une suite de variables aléatoires réelles qui converge uniformément en loi vers N. Soit $(a_n)_{n\geqslant 1}$ une suite de réels strictement positifs qui converge vers a, tel que $a > 0$, et $(b_n)_{n\geqslant 1}$ une suite de réels qui converge vers b.

a) Soit X une variable aléatoire et (α, β) un couple de réels avec $\alpha > 0$. On note $F_{\alpha X+\beta}$ et $F_{\alpha N+\beta}$ les fonctions de répartition respectives de $\alpha X + \beta$ et $\alpha N + \beta$.

Montrer que, pour tout x réel,

$$|F_{\alpha X+\beta}(x) - F_{\alpha N+\beta}(x)| = d_X\left(\frac{x-\beta}{\alpha}\right)$$

b) Montrer que pour tout x réel,

$$\lim_{n\to+\infty} \left(\mathbb{P}(a_nV_n + b_n \leqslant x) - \mathbb{P}(a_nN + b_n \leqslant x)\right) = 0$$

c) Établir que $\lim\limits_{n\to+\infty} \mathbb{P}(a_nN + b_n \leqslant x) = \mathbb{P}(aN + b \leqslant x)$ puis en déduire que $(a_nV_n + b_n)_{n\geqslant 1}$ converge en loi vers $aN + b$. Quelle est la loi de la variable aléatoire $aN + b$?

21. On reprend les notations de la partie 3.

a) Justifier que $(D_n)_{n\geqslant 1}$ converge uniformément en loi vers N.

b) En utilisant les résultats des questions ***14.*** et ***20.***, en déduire que la suite $(\hat{f}_n)_{n\geqslant 1}$ converge en loi vers N.

HEC 2010 - loi exponentielle, loi géométrique, loi de la somme, de la différence, de la valeur absolue, du max, du min, covariance, coefficient de corrélation linéaire, estimation, loi de Gumbel

- Toutes les variables aléatoires qui interviennent dans ce problème sont supposées définies sur un espace probabilisé $(\Omega, \mathscr{A}, \mathbb{P})$.
- Sous réserve d'existence, on note $\mathbb{E}(X)$ et $\mathbb{V}(X)$ respectivement l'espérance et la variance d'une variable aléatoire X, et $\mathrm{Cov}(X,Y)$ la covariance de deux v.a.r. X et Y.
- Dans les parties I et III, la fonction de répartition et une densité d'une variable aléatoire X à densité sont notées respectivement F_X et f_X.
- **On admet** que les formules donnant l'espérance et la variance d'une somme de variables aléatoires discrètes, ainsi que la définition et les propriétés de la covariance et du coefficient de corrélation linéaire de deux variables aléatoires discrètes, s'appliquent au cas de variables aléatoires à densité.
- Pour n entier supérieur ou égal à 2, on dit que les variables aléatoires à densité $X_1, X_2, \ldots, X_n$ sont indépendantes si pour tout n-uplet $(x_1, x_2, \ldots, x_n)$ de réels, les événements $[X_1 \leqslant x_1]$, $[X_2 \leqslant x_2]$, $\ldots$, $[X_n \leqslant x_n]$ sont indépendants.
- L'objet du problème est double. D'une part, montrer certaines analogies entre les lois géométriques et exponentielles, d'autre part mettre en évidence quelques propriétés asymptotiques de variables aléatoires issues de la loi exponentielle.
  La partie II est indépendante de la partie I.
  La partie III est indépendante de la partie II et largement indépendante de la partie I.

Partie I. Loi exponentielle

1. a) Rappeler la valeur de $\displaystyle\int_0^{+\infty} \mathrm{e}^{-t}\,dt$.

Établir pour tout n de $\mathbb{N}^*$ la convergence de l'intégrale $\displaystyle\int_0^{+\infty} t^n\mathrm{e}^{-t}\,dt$.

On pose alors $I_0 = \displaystyle\int_0^{+\infty} \mathrm{e}^{-t}\,dt$ et pour tout n de $\mathbb{N}^*$ $I_n = \displaystyle\int_0^{+\infty} t^n\mathrm{e}^{-t}\,dt$.

b) Soit n un entier de $\mathbb{N}^*$. À l'aide d'une intégration par parties, établir une relation de récurrence entre I_n et I_{n-1}. En déduire la valeur de I_n en fonction de n.

Soit λ un réel strictement positif. Soient X_1 et X_2 deux variables indépendantes de même loi exponentielle de paramètre λ (d'espérance $\frac{1}{\lambda}$).

On pose : $Y = X_1 - X_2$, $T = \max(X_1, X_2)$ et $Z = \min(X_1, X_2)$.

2. Justifier les relations $T + Z = X_1 + X_2$ et $T - Z = |X_1 - X_2| = |Y|$.

3. a) Rappeler sans démonstration les valeurs de $\mathbb{V}(X_1)$ et de $\mathbb{P}\big([X_1 \leqslant x]\big)$, pour tout réel x.

b) Calculer $\mathbb{E}(X_1 + X_2)$, $\mathbb{V}(X_1 + X_2)$, $\mathbb{E}(Y)$, $\mathbb{V}(Y)$.

4. Déterminer pour tout réel z, $F_Z(z)$ et $f_Z(z)$. Reconnaître la loi de Z et en déduire $\mathbb{E}(Z)$ et $\mathbb{V}(Z)$.

5. a) Montrer que pour tout réel t, on a : $F_T(t) = \begin{cases} \big(1 - \mathrm{e}^{-\lambda t}\big)^2 & \text{si } t \geqslant 0 \\ 0 & \text{si } t < 0 \end{cases}$.

Exprimer pour tout réel t, $f_T(t)$.

b) Justifier l'existence de $\mathbb{E}(T)$ et $\mathbb{V}(T)$. Montrer que $\mathbb{E}(T) = \dfrac{3}{2\lambda}$ et $\mathbb{V}(T) = \dfrac{5}{4\lambda^2}$.
(on pourra utiliser des changements de variables affine)

6. On note r le coefficient de corrélation linéaire de Z et T. Montrer que $r = \dfrac{1}{\sqrt{5}}$.

7. ***a)*** Préciser $Y(\Omega)$ et $|Y|(\Omega)$.

b) Déterminer une densité de la variable aléatoire $-X_2$.

c) Montrer que pour tout réel y, l'intégrale $\displaystyle\int_{-\infty}^{+\infty} f_{X_1}(t)\, f_{-X_2}(y-t)\, dt$ est convergente et qu'elle vaut $\dfrac{\lambda}{2}\, \mathrm{e}^{-\lambda\,|y|}$ (on distinguera les deux cas : $y \geqslant 0$ et $y < 0$).

d) Établir que la fonction $y \mapsto \dfrac{\lambda}{2}\mathrm{e}^{-\lambda\,|y|}$ est une densité de probabilité sur $\mathbb{R}$; on admet que c'est une densité de la variable aléatoire Y.

e) Déterminer pour tout y réel, $f_{|Y|}(y)$. Reconnaître la loi de $|Y| = T - Z$.

Partie II. Loi géométrique

Soit p un réel de $]0,1[$ et $q = 1-p$. Soit X_1 et X_2 deux variables indépendantes de même loi géométrique de paramètre p (d'espérance $\frac{1}{p}$).
On pose : $Y = X_1 - X_2$, $T = \max(X_1, X_2)$ et $Z = \min(X_1, X_2)$.
On rappelle que $T + Z = X_1 + X_2$ et $T - Z = |X_1 - X_2| = |Y|$.

8. ***a)*** Rappeler sans démonstration les valeurs respectives de $\mathbb{V}(X_1)$ et de $\mathbb{P}\big([X_1 \leqslant k]\big)$, pour tout k de $X_1(\Omega)$.

b) Calculer $\mathbb{E}(X_1 + X_2)$, $\mathbb{V}(X_1 + X_2)$, $\mathbb{E}(X_1 - X_2)$, $\mathbb{V}(X_1 - X_2)$.

c) Établir la relation : $\mathbb{P}\big([X_1 = X_2]\big) = \dfrac{p}{1+q}$.

9. ***a)*** Montrer que Z suit la loi géométrique de paramètre $1 - q^2$. En déduire $\mathbb{E}(Z)$, $\mathbb{V}(Z)$ et $\mathbb{E}(T)$.

b) Soit k un entier de $\mathbb{N}^*$. Justifier l'égalité : $[Z = k] \cup [T = k] = [X_1 = k] \cup [X_2 = k]$.
En déduire la relation suivante : $\mathbb{P}\big([T = k]\big) = 2\,\mathbb{P}\big([X_1 = k]\big) - \mathbb{P}\big([Z = k]\big)$.

c) Établir la formule : $\mathbb{V}(T) = \dfrac{q\,\left(2q^2 + q + 2\right)}{(1-q^2)^2}$.

10. ***a)*** Préciser $(T - Z)(\Omega)$.
Exprimer pour tout j de $\mathbb{N}^*$, l'événement $[Z = j] \cap [Z = T]$ en fonction des événements $[X_1 = j]$ et $[X_2 = j]$. En déduire pour tout j de $\mathbb{N}^*$, l'expression de $\mathbb{P}\big([Z = j] \cap [Z = T]\big)$.

b) Montrer que pour tout couple (j, l) de $(\mathbb{N}^*)^2$, on a : $\mathbb{P}\big([Z = j] \cap [T - Z = l]\big) = 2\, p^2\, q^{2j+l-2}$.

c) Montrer que pour tout k de $\mathbb{Z}$, $\mathbb{P}\big([X_1 - X_2 = k]\big) = \dfrac{pq^{|k|}}{1+q}$.
(on distinguera trois cas : $k = 0$, $k > 0$ et $k < 0$)

d) En déduire la loi de la variable aléatoire $|X_1 - X_2|$.

e) Établir à l'aide des questions précédentes que les variables Z et $T - Z$ sont indépendantes.

11. ***a)*** À l'aide du résultat de la question ***3.e***, calculer $\mathrm{Cov}(Z, T)$.
Les variables Z et T sont-elles indépendantes ?

b) Calculer en fonction de q, le coefficient de corrélation linéaire ρ de Z et T.

c) Déterminer la loi de probabilité du couple (Z, T).

d) Déterminer pour tout j de $\mathbb{N}^*$, la loi de probabilité conditionnelle de T sachant l'événement $[Z = j]$.

e) Soit j un élément de $\mathbb{N}^*$. On suppose qu'il existe une variable aléatoire D_j à valeur dans $\mathbb{N}^*$, dont la loi de probabilité est la loi conditionnelle de T sachant l'événement $[Z = j]$.
Calculer $\mathbb{E}(D_j)$.

Partie III. Convergences

Dans les questions 12 à 15, λ désigne un paramètre réel strictement positif, inconnu.
Pour n élément de $\mathbb{N}^*$, on considère un n-échantillon $(X_1, X_2, \ldots, X_n)$ de variables aléatoires à valeurs strictement positives, indépendantes, de même loi exponentielle de paramètre λ.
On pose pour tout n de $\mathbb{N}^*$: $S_n = \sum_{k=1}^{n} X_k$ et $J_n = \lambda S_n$.

12. Calculer pour tout n de $\mathbb{N}^*$, $\mathbb{E}(S_n)$, $\mathbb{V}(S_n)$, $\mathbb{E}(J_n)$ et $\mathbb{V}(J_n)$.

13. On admet qu'une densité f_{J_n} de J_n est donnée par $f_{J_n}(x) = \begin{cases} \dfrac{\mathrm{e}^{-x}x^{n-1}}{(n-1)!} & \text{si } x > 0 \\ 0 & \text{si } x \leqslant 0 \end{cases}$.

a) À l'aide du théorème de transfert, établir pour tout n supérieur ou égal à 3, l'existence de $\mathbb{E}\left(\dfrac{1}{J_n}\right)$ et de $\mathbb{E}\left(\dfrac{1}{J_n^2}\right)$, et donner leur valeurs respectives.

b) On pose pour tout n supérieur ou égal à 3 : $\widehat{\lambda_n} = \dfrac{n}{S_n}$. Justifier que $\widehat{\lambda_n}$ est un estimateur de λ. Est-il sans biais ? Calculer la limite, lorsque n tend vers $+\infty$, du risque quadratique associé à $\widehat{\lambda_n}$ en λ.

14. Dans cette question, on veut déterminer un intervalle de confiance du paramètre λ au risque α.
On note Φ la fonction de répartition de la loi normale centrée réduite, et u_α le réel strictement positif tel que $\Phi(u_\alpha) = 1 - \dfrac{\alpha}{2}$.

a) Énoncer le théorème de la limite centrée. En déduire que la variable aléatoire N_n définie par $N_n = \lambda \dfrac{S_n}{\sqrt{n}} - \sqrt{n}$ converge en loi vers la loi normale centrée réduite.

b) En déduire que pour n assez grand, on a approximativement : $\mathbb{P}\big([-u_\alpha \leqslant N_n \leqslant u_\alpha]\big) = 1 - \alpha$.

c) Montrer que pour n assez grand, l'intervalle $\left[\left(1 - \dfrac{u_\alpha}{\sqrt{n}}\right)\widehat{\lambda_n}, \left(1 + \dfrac{u_\alpha}{\sqrt{n}}\right)\widehat{\lambda_n}\right]$ est un intervalle de confiance de λ au risque α. On note λ_0 la réalisation de $\widehat{\lambda_n}$ sur le n-échantillon.

15. Avec le n-échantillon $(X_1, X_2, \ldots, X_n)$, on construit un nouvel intervalle de confiance de λ au risque β ($\beta \neq \alpha$), tel que la longueur de cet intervalle soit k ($k > 1$) fois plus petite que celle obtenue avec le risque α.

a) Justifier l'existence de la fonction réciproque Φ^{-1} de Φ.
Quel est le domaine de définition de Φ^{-1} ?

b) Établir l'égalité $\beta = 2\Phi\left(\dfrac{1}{k}\Phi^{-1}\left(\dfrac{\alpha}{2}\right)\right)$.
En déduire que $\beta > \alpha$. Ce dernier résultat était-il prévisible ?

Dans les questions 16 à 18, on suppose que $\lambda = 1$.

16. On pose pour tout n de $\mathbb{N}^*$: $T_n = \max(X_1, X_2, \ldots, X_n)$.
Pour tout n de $\mathbb{N}^*$, pour tout réel x positif ou nul, on pose :

$$g_n(x) = \int_0^x F_{T_n}(t)\, dt \quad \text{et} \quad h_n(x) = \int_0^x t\, f_{T_n}(t)\, dt$$

a) Exprimer $h_n(x)$ en fonction de $F_{T_n}(x)$ et $g_n(x)$.

b) Déterminer pour tout réel t, l'expression de $F_{T_n}(t)$ en fonction de t.
Établir pour tout n supérieur ou égal à 2, la relation : $g_{n-1}(x) - g_n(x) = \dfrac{1}{n} F_{T_n}(x)$.

c) En déduire que pour tout n de $\mathbb{N}^*$, pour tout réel x positif ou nul, l'expression de $g_n(x)$ en fonction de x, $F_{T_1}(x)$, $F_{T_2}(x)$, $\ldots$, $F_{T_n}(x)$.

d) Montrer que $F_{T_n}(x) - 1$ est équivalent à $-n\mathrm{e}^{-x}$, lorsque x tend vers $+\infty$.

e) Déduire des questions ***c)*** et ***d)*** l'existence de $\mathbb{E}(T_n)$ et montrer que $\mathbb{E}(T_n) = \sum_{k=1}^{n} \dfrac{1}{k}$.

17. On veut étudier dans cette question la convergence en loi de la suite de variables aléatoires $(G_n)_{n\geqslant 1}$ définie par : pour tout n de $\mathbb{N}^*$, $G_n = T_n - \mathbb{E}(T_n)$.
On pose pour tout n de $\mathbb{N}^*$: $\gamma_n = -\ln(n) + \mathbb{E}(T_n)$ et on admet sans démonstration que la suite $(\gamma_n)_{n\geqslant 1}$ est convergente ; on note γ sa limite.

a) Montrer que pour tout x réel et n assez grand, on a : $F_{G_n}(x) = \left(1 - \dfrac{1}{n}\, \mathrm{e}^{-(x+\gamma_n)}\right)^n$.

b) En déduire que pour tout x réel, on a : $\lim\limits_{n\to+\infty} F_{G_n}(x) = \mathrm{e}^{-\mathrm{e}^{-(x+\gamma)}}$.

c) Montrer que la fonction $F_G : \mathbb{R} \to \mathbb{R}$ définie par $F_G : x \mapsto \mathrm{e}^{-\mathrm{e}^{-(x+\gamma)}}$ est la fonction de répartition d'une variable aléatoire G à densité. Conclure.

18. ***a)*** Soit X une variable aléatoire à densité de fonction de répartition F_X strictement croissante. Déterminer la loi de la variable aléatoire Y définie par $Y = F_X(X)$.

b) Écrire une fonction **Python** d'en-tête `Gumbel` qui permet de simuler la variable aléatoire G.
On supposera que la constante γ est définie en langage **Python** par une constante `gamma`.
On rappelle que la fonction **Python** `rd.random()` permet de simuler la loi uniforme sur $]0, 1[$.

ESSEC II 2010 - loi exponentielle, loi de Poisson, loi d'Erlang, loi de Weibull, loi du max, du min, fiabilité, processus de Poisson, fonction génératrice des probabilités, séries

- L'objet du problème est l'étude de la durée de fonctionnement d'un système (une machine, un organisme, un service ...) démarré à la date $t = 0$ et susceptible de tomber en panne à une date aléatoire. Après une partie préliminaire sur les propriétés de la loi exponentielle, on introduira dans la deuxième partie, les notions permettant d'étudier des propriétés de la date de première panne. Enfin, dans une troisième partie on examinera le fonctionnement d'un système satisfaisant certaines propriétés particulières.
- Les trois parties sont dans une large mesure indépendantes.
- Toutes les variables aléatoires intervenant dans le problème sont définies sur un espace probabilisé $(\Omega, \mathscr{A}, \mathbb{P})$.
- Pour toute variable aléatoire Y, on notera $\mathbb{E}(Y)$ son espérance lorsqu'elle existe.
- On adoptera les conventions suivantes :
  - × on dira qu'une fonction f continue sur $\mathbb{R}_+^*$ et continue à droite en 0 est continue sur $\mathbb{R}_+$.
  - × en outre, si T est une variable aléatoire positive dont la loi admet la densité f continue sur $\mathbb{R}_+$, sa fonction de répartition $F_T(t) = \mathbb{P}([T \leqslant t]) = \displaystyle\int_0^t f(u)\, du$, est dérivable sur $\mathbb{R}_+^*$, et dérivable à droite en 0.
  - × on conviendra d'écrire $F_T'(t) = f(t)$ pour tout $t \in \mathbb{R}_+$, $F_T'(0)$ désignant donc dans ce cas la dérivée à droite en 0.

Partie I. Généralités sur la loi exponentielle

On rappelle qu'une variable aléatoire suit la loi exponentielle de paramètre μ ($\mu > 0$) si elle admet pour densité la fonction f_μ définie par :

$$f_\mu(x) = \begin{cases} \mu\, \mathrm{e}^{-\mu x} & \text{si} \quad x \geqslant 0 \\ 0 & \text{si} \quad x < 0 \end{cases}$$

1. Soit X une variable aléatoire suivant la loi exponentielle de paramètre μ.

a) Donner l'espérance $\mathbb{E}(X)$ et la variance $\mathbb{V}(X)$.

b) Justifier que pour tout entier naturel n, X^n admet une espérance et déterminer une relation de récurrence entre $\mathbb{E}(X^{n+1})$ et $\mathbb{E}(X^n)$ pour tout entier naturel n.

c) En déduire $\mathbb{E}(X^n)$ pour tout $n > 0$.

d) Retrouver la valeur de $\mathbb{V}(X)$ à l'aide de la question précédente.

2. *Propriété caractéristique*

a) Soient $\mu > 0$ et X une variable aléatoire de loi exponentielle de paramètre μ.
Justifier que pour tout réel x positif ou nul, le nombre $\mathbb{P}([X > x])$ est non nul.
Montrer que pour tous réels positifs x et y :

$$\mathbb{P}_{[X>x]}([X > x + y]) = \mathbb{P}([X > y])$$

b) Réciproquement, soit X une variable aléatoire positive admettant une densité f continue et strictement positive sur $\mathbb{R}_+$, et telle que pour tous réels positifs x et y :

$$\mathbb{P}_{[X>x]}([X > x + y]) = \mathbb{P}([X > y])$$

(i) Soit $R(x) = \mathbb{P}([X > x])$. Justifier que $R(x)$ est non nul pour tout réel positif.

(ii) On pose $\mu = f(0)$. Montrer que pour tout x réel positif, on a la relation $R'(x) + \mu R(x) = 0$.

(iii) Calculer la dérivée de $x \mapsto R(x)\ \mathrm{e}^{\mu x}$ sur $\mathbb{R}_+$.

(iv) Déduire que X suit une loi exponentielle dont on précisera le paramètre.

3. Soient deux réels strictement positifs μ_1 et μ_2. Soient X_1 et X_2 deux variables aléatoires indépendantes suivant respectivement les lois exponentielles de paramètres μ_1 et μ_2.

a) On pose $Y = \max(X_1, X_2)$.
Déterminer la fonction de répartition F_Y de Y et en déduire la densité de la variable Y.

b) On pose $Z = \min(X_1, X_2)$.
Déterminer la fonction de répartition F_Z de Z et en déduire la loi de Z.

Partie II. Fiabilité

Soit T une variable aléatoire positive qui représente la durée de vie (c'est-à-dire le temps de fonctionnement avant la survenue d'une première panne) d'un système. On suppose que T est une variable à densité f_T continue sur $\mathbb{R}_+$ et ne s'annulant pas sur $\mathbb{R}_+^*$.
On appelle fiabilité de T la fonction R_T définie sur $\mathbb{R}_+$ par :

$$R_T(t) = \mathbb{P}([T \geqslant t]) = \mathbb{P}([T > t]) = 1 - F_T(t)$$

où F_T est la fonction de répartition de T.

4. Soient t un réel positif ou nul et h un réel strictement positif.
La dégradation du système sur l'intervalle $[t, t+h]$ est mesurée par la probabilité $\mathbb{P}([t \leqslant T \leqslant t + h])$.
Exprimer cette quantité à l'aide de la fonction R_T.

5. Montrer que, pour tout réel t positif ou nul,

$$\lim_{h \to 0^+} \frac{\mathbb{P}([t \leqslant T \leqslant t + h])}{h} = f_T(t)$$

6. ***a)*** Justifier que pour tout réel t positif, $R_T(t) > 0$.

On appelle taux de défaillance la fonction définie sur $\mathbb{R}_+$ par le rapport $\lambda(t) = \dfrac{f_T(t)}{R_T(t)}$.

b) On note : $g : t \mapsto \ln\left(\dfrac{1}{R_T(t)}\right)$. Démontrer que $\lambda = g'$.

c) Déduire l'expression de R_T en fonction de λ à l'aide d'une intégrale.

7. Soit Z une variable aléatoire réelle positive de densité g continue sur $\mathbb{R}_+$, admettant une espérance.
On pose $R_Z(t) = \mathbb{P}([Z > t])$ pour $t \geqslant 0$.

a) Soit v la fonction définie sur $\mathbb{R}_+$ par $v(t) = tR_Z(t)$.
Démontrer, pour tout $t \in \mathbb{R}_+$: $tg(t) = R_Z(t) - v'(t)$ où v' désigne la dérivée de v.

b) Montrer que $\lim\limits_{t \to +\infty} v(t) = 0$.

c) En déduire que $\mathbb{E}(Z) = \displaystyle\int_0^{+\infty} R_Z(t)\ dt$.

8. On suppose désormais que T admet une espérance. Soit t un réel positif fixé, le système ayant fonctionné sans panne jusqu'à la date t, on appelle durée de survie la variable aléatoire $T_t = T - t$ représentant le temps s'écoulant entre la date t et la première panne.
On a donc, pour tout réel x positif :

$$R_{T_t}(x) = \mathbb{P}([T_t > x]) = \mathbb{P}_{[T>t]}([T > t + x])$$

a) Démontrer, pour tout réel x positif : $R_{T_t}(x) = \dfrac{R_T(t+x)}{R_T(t)}$.

b) En déduire :

$$\mathbb{E}(T_t) = \frac{1}{R_T(t)} \int_t^{+\infty} R_T(u)\, du$$

Les questions suivantes illustrent les notions introduites précédemment pour des systèmes simples.

9. ***a)*** On suppose que T suit la loi exponentielle de paramètre μ.
Déterminer la fiabilité et le taux de défaillance.

b) On suppose que le système est composé de deux organes 1 et 2 montés en série, dont les durées de vie sont supposées indépendantes, ce qui implique qu'il tombe en panne dès que l'un d'eux tombe en panne. On note T_i la durée de vie de l'organe i, f_{T_i} la densité de sa loi qu'on suppose exponentielle de paramètre μ_i.
Déterminer la fiabilité du système et son taux de défaillance.

c) On suppose que le système est composé de deux organes 1 et 2 montés en parallèle, dont les durées de vie sont supposées indépendantes, ce qui implique qu'il tombe en panne quand les deux organes sont en panne. On note T_i la durée de vie de l'organe i, f_{T_i} la densité de sa loi qu'on suppose exponentielle de paramètre μ_i.
Déterminer la fiabilité du système.

10. Soit $\varphi_{n,\beta}$ la fonction définie par :

$$\varphi_{n,\beta} : t \mapsto \begin{cases} \dfrac{\beta}{(n-1)!}\ (\beta t)^{n-1}\ \mathrm{e}^{-\beta t} & \text{si } t \geqslant 0 \\ 0 & \text{si } t < 0 \end{cases}$$

où $\beta > 0$ est une constante strictement positive et n un entier naturel non nul.

a) Démontrer que $\varphi_{n,\beta}$ est une densité de probabilité (loi d'Erlang).

b) On suppose que T a pour densité la fonction $\varphi_{n,\beta}$. Montrer que la fiabilité à la date t est :

$$R_T(t) = \mathrm{e}^{-\beta t} \sum_{k=0}^{n-1} \frac{(\beta t)^k}{k!}$$

11. Soit $\psi_{\beta,\eta}$ la fonction définie par :

$$\psi_{\beta,\eta} : t \mapsto \begin{cases} \dfrac{\beta}{\eta} \left(\dfrac{t}{\eta}\right)^{\beta-1} \mathrm{e}^{-\left(\frac{t}{\eta}\right)^{\beta}} & \text{si } t \geqslant 0 \\ 0 & \text{sinon} \end{cases}$$

où $\beta \geqslant 1, \eta > 0$.

a) Vérifier que $\psi_{\beta,\eta}$ est une densité de probabilité (loi de Weibull).

b) On suppose que T a pour densité la fonction $\psi_{\beta,\eta}$.
Déterminer la fiabilité $R_T(t)$ et le taux de défaillance $\lambda(t)$ à la date t.

c) Étudier $\lim\limits_{t \to +\infty} \lambda(t)$ en fonction de la valeur de β.

Partie III. Système Poissonien

On considère maintenant un système dont le fonctionnement est défini comme suit : pour tout réel t positif, la variable aléatoire N_t à valeurs entières représente le nombre de pannes qui se produisent dans l'intervalle $[0,t]$. On considère que le système est réparé immédiatement après chaque panne.
On notera en particulier que pour $s \leqslant t$, on a $N_s \leqslant N_t$.
On suppose qu'on a les quatre propriétés suivantes :

- $N_0 = 0$ et $0 < \mathbb{P}([N_t = 0]) < 1$ pour tout $t > 0$.
- Pour tous réels $t_0, t_1, \ldots, t_n$ tels que $0 \leqslant t_0 < t_1 < \cdots < t_n$ les variables $N_{t_0}, N_{t_1} - N_{t_0}, N_{t_2} - N_{t_1}, \ldots, N_{t_n} - N_{t_{n-1}}$ sont mutuellement indépendantes (accroissements indépendants).
- Pour tous réels s et t tels que $0 < s < t$, $N_t - N_s$ suit la même loi que N_{t-s} (accroissements stationnaires).
- $\lim\limits_{h \to 0^+} \dfrac{\mathbb{P}([N_h > 1])}{h} = 0.$

On pose, sous réserve d'existence, pour tout $u \geqslant 0$ et pour tout s dans [0,1], $G_u(s) = \mathbb{E}(s^{N_u})$, avec la convention $0^0 = 1$.

12. ***a)*** Justifier que pour tout $u \geqslant 0$, $G_u(s)$ existe pour tout s dans [0,1] et qu'on a, pour tout $s \in [0,1]$:

$$G_u(s) = \sum_{k=0}^{+\infty} \mathbb{P}([N_u = k])\ s^k$$

b) Montrer par ailleurs que, pour tous réels u et v positifs ou nuls, et pour tout réel s tel que $0 \leqslant s \leqslant 1$, on a :

$$G_{u+v}(s) = G_u(s) G_v(s)$$

13. On fixe s tel que $0 \leqslant s \leqslant 1$.

a) Montrer que $G_1(s) > 0$.
On pose $\theta(s) = -\ln G_1(s)$ et, pour $u \geqslant 0$, $\psi(u) = G_u(s)$.

b) Montrer que $\psi(k) = \mathrm{e}^{-k\theta(s)}$ pour tout $k \in \mathbb{N}$.

c) Soit q un entier naturel non nul. En considérant $G_{\frac{1}{q}}(s)$, montrer que $\psi(\frac{1}{q}) = \mathrm{e}^{-\frac{1}{q}\theta(s)}$.

d) Montrer que si p est entier naturel et q un entier naturel non nul, on a $\psi(r) = \mathrm{e}^{-r\theta(s)}$ où on a posé $r = \frac{p}{q}$.

e) Montrer que pour tout réel positif u, $G_u(s) = \mathrm{e}^{-u\theta(s)}$.

f) En déduire que pour tout $s \in [0,1]$, $\lim\limits_{h \to 0^+} \dfrac{G_h(s) - 1}{h} = -\theta(s)$.

14. Montrer par ailleurs que pour tout $s \in [0,1]$,

$$G_h(s) - 1 = \mathbb{P}([N_h = 1])\ (s-1) + \sum_{k=2}^{+\infty} \mathbb{P}([N_h = k])\ (s^k - 1)$$

15. Montrer que pour tout $s \in [0,1]$: $\lim\limits_{h \to 0^+} \dfrac{\sum\limits_{k=2}^{+\infty} \mathbb{P}([N_h = k])\ (s^k - 1)}{h} = 0.$

16. ***a)*** En déduire qu'il existe $\alpha \geqslant 0$ tel que $\alpha = \lim\limits_{h \to 0^+} \dfrac{\mathbb{P}([N_h = 1])}{h}$ et que pour pour tout $s \in [0,1]$:

$$\theta(s) = \alpha(1-s)$$

b) En considérant $G_u(0)$, montrer que $\alpha > 0$.

c) On fixe un temps $u > 0$. Montrer que pour tout $s \in [0,1]$,

$$G_u(s) = \sum_{k=0}^{+\infty} \mathbb{P}([N_u = k])\ s^k = \sum_{k=0}^{+\infty} \left[\mathrm{e}^{-\alpha u} \frac{(\alpha u)^k}{k!} \right] s^k$$

d) Déduire que pour tout $u > 0$, la variable aléatoire N_u suit la loi de Poisson de paramètre αu.
Une famille de variables aléatoires ayant les mêmes caractéristiques que la famille $(N_t)_{t \geqslant 0}$ est un **processus de Poisson** et la constante α s'appelle le **paramètre** du processus de Poisson.

17. Soit T la variable aléatoire désignant la date de la première panne. Soit $t > 0$.
Comparer les événements $[T > t]$ et $[N_t = 0]$.
En déduire que T suit la loi exponentielle de paramètre α.

18. Pour t positif fixé, on pose pour h réel positif, $\tilde{N}_h = N_{t+h} - N_t$.

a) Montrer que $\tilde{N}_h$ est la variable aléatoire qui représente le nombre de pannes survenues dans l'intervalle de temps $]t, t+h]$.

b) Montrer que la famille $(\tilde{N}_h)_{h \geqslant 0}$ est un processus de Poisson de paramètre α.

c) En déduire que la première panne survenant après la date t se produit à une date suivant la loi exponentielle de paramètre α.

d) En déduire que le processus de Poisson a la propriété que, pour chaque date t donnée, le taux de défaillance du système après t est constant.

ESSEC II 2009 - estimation ponctuelle, loi de Poisson, loi de Bernoulli, loi binomiale, loi normale, information de Fisher

Notations

- Tout au long du sujet $(\Omega, \mathcal{F}, \mathbb{P})$ désignera un espace probabilisé et les variables aléatoires utilisées seront toutes définies sur cet espace probabilisé. Sous réserve d'existence, l'espérance mathématique d'une variable aléatoire réelle X sera notée $\mathbb{E}(X)$ et sa variance sera notée $\mathbb{V}(X)$.
- Pour un événement A, on notera $\mathbb{P}_B(A)$ la probabilité conditionnelle de A sachant B où B est un événement non négligeable.

Le sujet est composé de quatre parties. Les parties I, II, III et IV.1 sont **indépendantes**.
Il s'agit de variations autour de la notion de risque quadratique en théorie de l'estimation.

Partie I. Premier problème d'estimation

Dans ce premier problème d'estimation, on dispose d'une seule observation notée X.
On suppose que X admet pour densité f_θ définie sur $\mathbb{R}$ par :

$$f_\theta : x \mapsto \begin{cases} \dfrac{k+1}{\theta^{k+1}}\, x^k & \text{si } x \in [0, \theta] \\ 0 & \text{sinon} \end{cases}$$

où k est un entier naturel non nul et θ un paramètre réel inconnu strictement positif que l'on souhaite estimer.

1. Montrer que f_θ est bien une densité de probabilité.

2. Calculer $\mathbb{E}(X)$.

3. Déterminer λ_0 un réel dépendant uniquement de k tel que $\lambda_0 X$ soit un estimateur de θ sans biais.

4. Calculer $\mathbb{V}(X)$.

5. On définit le risque quadratique de T estimateur de θ par :

$$r_\theta(T) = \mathbb{E}\left((T-\theta)^2\right)$$

Redémontrer le résultat du cours précisant que pour tout T estimateur de θ :

$$r_\theta(T) = \big(\mathbb{E}(T) - \theta\big)^2 + \mathbb{V}(T)$$

6. Donner la valeur de $r_\theta(\lambda_0 X)$.
Le but de la fin de cette partie I est de déterminer un estimateur de θ ayant un plus petit risque quadratique que celui de $\lambda_0 X$.

7. En utilisant I.5 montrer que pour tout λ réel

$$r_\theta(\lambda X) = \theta^2\, Q(\lambda)$$

où Q est un polynôme de degré 2 dont les coefficients ne dépendent que de k.

8. Montrer que la fonction $\lambda \mapsto Q(\lambda)$ atteint son minimum en un unique réel noté λ^* que l'on exprimera en fonction de k.

9. Conclure sur le but recherché.

Partie II. Second problème d'estimation

Dans ce second problème d'estimation, on dispose de n observations indépendantes $(n \geqslant 2)$ notées $X_1, \ldots, X_n$ de même loi de Poisson de paramètre θ inconnu ($\theta \in]0, +\infty[$). On souhaite estimer le paramètre $\exp(-\theta)$. On définit pour tout i élément de $[\![1, n]\!]$ la variable aléatoire Y_i par :

$$Y_i \ : \ \omega \ \mapsto \ \begin{cases} 1 & \text{si } X_i(\omega) = 0 \\ 0 & \text{sinon} \end{cases}$$

Puis on note :

$$\overline{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

10. Pour tout i élément de $[\![1, n]\!]$, donner la loi de Y_i.

11. Donner la loi de $\sum_{i=1}^{n} Y_i$, puis montrer que $\mathbb{E}\big(\overline{Y}_n\big) = \exp(-\theta)$.

On dira dans ce cas que $\overline{Y}_n$ est un estimateur sans biais de $\exp(-\theta)$.

12. Calculer $\mathbb{V}\left(\overline{Y}_n\right)$.

Pour tout k élément de $[\![1, n]\!]$ on définit $S_k = \sum_{i=1}^{k} X_i$.

13. Rappeler sans démonstration la loi de S_k pour tout k élément de $[\![1, n]\!]$.

On définit jusqu'à la fin de cette partie II pour tout j entier naturel :

$$\varphi(j) = \mathbb{P}_{[S_n = j]}\big([X_1 = 0] \big)$$

14. Montrer que pour tout j entier naturel :

$$\varphi(j) = \left(1 - \frac{1}{n}\right)^j$$

On a donc $\varphi(j)$ indépendant du paramètre θ inconnu.

D'après la question II.13, on peut définir l'estimateur :

$$\varphi(S_n) = \left(1 - \frac{1}{n}\right)^{S_n}$$

15. Montrer que $\varphi(S_n)$ admet une espérance et que $\mathbb{E}\big(\varphi(S_n)\big) = \exp(-\theta)$. On dira dans ce cas que $\varphi(S_n)$ est un estimateur sans biais de $\exp(-\theta)$.

16. Montrer que $\varphi(S_n)$ admet une variance vérifiant :

$$\mathbb{V}\big(\varphi(S_n)\big) = \exp(-2\theta) \ \left(\exp\left(\frac{\theta}{n}\right) - 1\right)$$

17. On souhaite comparer les performances de $\overline{Y}_n$ et $\varphi(S_n)$ en tant qu'estimateurs de $\exp(-\theta)$.

a) En utilisant le théorème des accroissements finis, démontrer :

$$1 \leqslant \frac{\exp(\theta) - 1}{\theta} \leqslant \exp(\theta)$$

b) Soit la fonction $h : [0, 1] \to \mathbb{R}$ définie par :

$$h(t) = t \exp(\theta) + (1 - t) - \exp(t\theta)$$

pour tout $t \in [0, 1]$. Étudier les variations de h.

c) En déduire :

$$\exp\left(\frac{\theta}{n}\right) \leqslant \frac{\exp(\theta)}{n} + \frac{n-1}{n}$$

puis l'inégalité :

$$\mathbb{V}\big(\varphi(S_n)\big) \leqslant \mathbb{V}\big(\overline{Y}_n\big)$$

d) On définit le risque quadratique de T_n estimateur de $\exp(-\theta)$ par :

$$r_\theta(T_n) \;=\; \mathbb{E}\Big(\big(T_n - \exp(-\theta)\big)^2\Big)$$

Comparer les risques quadratiques de $\overline{Y}_n$ et $\varphi(S_n)$.

On reprendra à la fin de la partie IV l'étude de $\varphi(S_n)$.

Partie III. Information de Fisher

A. Cas discret

Dans cette section III.1, on considère I un intervalle de $\mathbb{R}$, θ un paramètre inconnu appartenant à I et X une variable aléatoire à valeurs dans $\mathbb{N}$ ($X(\Omega) \subset \mathbb{N}$). On suppose qu'il existe une fonction p définie sur $I \times X(\Omega)$ telle que pour tout k élément de $X(\Omega)$:

$$\mathbb{P}\big([X=k]\big) = p(\theta, k)$$

et vérifiant pour tout k de $X(\Omega)$, $\theta \mapsto p(\theta, k)$ dérivable sur I.
On note de plus : $h : (\theta, k) \mapsto \ln\big(p(\theta, k)\big)$.
On définit enfin, sous réserve d'existence l'**information de Fisher** de X par :

$$I_X(\theta) = \sum_{k \in X(\Omega)} \big(\partial_1(\ln \circ p)(\theta, k)\big)^2 p(\theta, k)$$

18. Dans cette question 18, on considère X une variable aléatoire qui suit la loi de Bernoulli de paramètre θ (où $\theta \in]0,1[$).
On a alors $X(\Omega) = \{0,1\}$, $\mathbb{P}\big([X=1]\big) = p(\theta,1) = \theta$, $\mathbb{P}\big([X=0]\big) = p(\theta,0) = 1-\theta$ et :

$$I_X(\theta) = \big(\partial_1(h)(\theta,1)\big)^2 p(\theta,1) + \big(\partial_1(h)(\theta,0)\big)^2 p(\theta,0)$$

Montrer :

$$I_X(\theta) = \frac{1}{\theta\,(1-\theta)}$$

19. Dans cette question 19, on considère X une variable aléatoire qui suit la loi binomiale de paramètres N et θ ($N \in \mathbb{N}^*$, $\theta \in]0,1[$).

a) Montrer :

$$I_X(\theta) = \frac{1}{\big(\theta\,(1-\theta)\big)^2} \sum_{k=0}^{N} (k - N\theta)^2 \binom{N}{k} \theta^k (1-\theta)^{N-k}$$

b) En déduire :

$$I_X(\theta) = \frac{\mathbb{V}(X)}{\big(\theta\,(1-\theta)\big)^2}$$

puis donner la valeur de $I_X(\theta)$.

20. Dans cette question 20, on considère X une variable aléatoire qui suit la loi de Poisson de paramètre θ ($\theta \in]0, +\infty[$). Puisque $X(\Omega) = \mathbb{N}$, on a sous réserve de convergence :

$$I_X(\theta) = \sum_{k=0}^{+\infty} \big(\partial_1(h)(\theta, k)\big)^2 p(\theta, k)$$

a) Montrer que la série de terme général $\big(\partial_1(h)(\theta, k)\big)^2 p(\theta, k)$ converge et calculer sa somme $I_X(\theta)$.

b) Justifier :

$$I_X(\theta) = \mathbb{E}\left(\big(\partial_1(h)(\theta, X)\big)^2\right)$$

B. Cas d'une variable gaussienne

Soit X une variable aléatoire qui suit la loi normale de moyenne θ ($\theta \in \mathbb{R}$) et de variance 1 dont la densité est notée $x \mapsto f(\theta, x)$. On définit sous réserve de convergence l'**information de Fisher** de X par :

$$I_X(\theta) = \int_{-\infty}^{+\infty} \big(\partial_1(\ln \circ f)(\theta, x)\big)^2 f(\theta, x)\ dx$$

21. Montrer que sous réserve de convergence :

$$I_X(\theta) = \int_{-\infty}^{+\infty} (x - \theta)^2 f(\theta, x)\ dx$$

22. En déduire l'existence et la valeur de $I_X(\theta)$.

23. Justifier :

$$I_X(\theta) = \mathbb{E}\left(\big(\partial_1(\ln \circ f)(\theta, X)\big)^2\right)$$

Partie IV. Minoration du risque quadratique

A. Inégalité de Cramer-Rao

Dans cette section IV.1, on considère I un intervalle de $\mathbb{R}$, θ un paramètre inconnu appartenant à I et X une variable aléatoire telle que $X(\Omega) = [\![0, N]\!]$ ($N \in \mathbb{N}$). On suppose qu'il existe une fonction p définie sur $I \times X(\Omega)$ telle que pour tout $k \in [\![0, N]\!]$:

$$\mathbb{P}\big([X = k]\big) = p(\theta, k)$$

et vérifiant :

- pour tout $k \in [\![0, N]\!]$, $\theta \mapsto p(\theta, k)$ dérivable sur I,
- l'information de Fisher de X notée $I_X(\theta)$ définie dans la partie III est non nulle pour tout $\theta \in I$.

Le but de la section IV.1 est de démontrer l'inégalité suivante dûe à Cramer et Rao.

Théorème 1. *(de Cramer-Rao)*
Soit $f(X)$ un estimateur sans biais de $g(\theta)$ à savoir tel que $\mathbb{E}\big(f(X)\big) = g(\theta)$ où g est dérivable sur I. On a alors :

$$\mathbb{V}\left(f(X)\right) \geqslant \frac{\big(g'(\theta)\big)^2}{I_X(\theta)}$$

24. Montrer que pour tout θ élément de I :

$$\sum_{k=0}^{N} \partial_1(p)(\theta, k) = 0$$

25. En déduire que pour tout θ élément de I :

$$\mathbb{E}\left(\partial_1(h)(\theta, X)\right) = 0 \qquad (E)$$

26. En dérivant partiellement par rapport à θ les deux membres de l'égalité (E), montrer que pour tout θ élément de I :

$$\mathbb{E}\left(\partial_{1,1}^2(h)(\theta, X)\right) = -\mathbb{E}\left(\left(\partial_1(h)(\theta, X)\right)^2\right)$$

27. Montrer que pour tout θ élément de I :

$$g'(\theta) = \sum_{k=0}^{N} f(k)\ \left(\partial_1(h)(\theta, k)\right)\ p(\theta, k)$$

puis :

$$g'(\theta) = \mathbb{E}\left(\big(f(X) - g(\theta)\big)\big(\partial_1(h)(\theta, X)\big)\right)$$

28. On pose pour tout t réel :

$$L(t) = \mathbb{E}\left(\left(\big(f(X) - g(\theta)\big) + t\big(\partial_1(h)(\theta, X)\big)\right)^2\right)$$

a) Développer le polynôme L suivant les puissances décroissantes de t.

b) Calculer le discriminant Δ de L et justifier : $\Delta \leqslant 0$.

c) En déduire l'inégalité de Cramer-Rao.

B. Extension du théorème de Cramer-Rao

On reprend dans cette section IV.2 les notations et hypothèses de la partie II. On admet que, dans ce contexte, le théorème de Cramer-Rao peut se généraliser comme suit :

Théorème 2. *(de Cramer-Rao)*
Soit $T_n = f(X_1, \ldots, X_n)$ un estimateur sans biais de $g(\theta)$ à savoir tel que $\mathbb{E}\big(f(X_1, \ldots, X_n)\big) = g(\theta)$ où g est dérivable sur $]0, +\infty[$. On a alors :

$$\mathbb{V}(T_n) \geqslant \frac{\big(g'(\theta)\big)^2}{n\, I_{X_1}(\theta)}$$

où $I_{X_1}(\theta)$ est l'information de Fisher d'une variable aléatoire de loi de Poisson de paramètre θ définie et calculée à la partie III.

Il s'agit dans cette section d'exploiter cette nouvelle inégalité de Cramer-Rao. On note :

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

29. Calculer $\mathbb{E}\big(\overline{X}_n\big)$ et $\mathbb{V}\big(\overline{X}_n\big)$.

30. Déduire de la généralisation de Cramer-Rao, que $\overline{X}_n$ a le plus petit risque quadratique parmi les estimateurs sans biais de θ.

31. Montrer que pour $g(\theta) = \exp(-\theta)$ où $\theta \in\,]0, +\infty[$:

$$\mathbb{V}\big(\varphi(S_n)\big) \underset{n\to+\infty}{\sim} \frac{\big(g'(\theta)\big)^2}{n\, I_{X_1}(\theta)}$$

32. Que prouve ce résultat en terme d'optimalité de $\varphi(S_n)$ dans l'estimation de $\exp(-\theta)$?

33. À la lumière de la partie II, peut-on conclure que lorsque n est grand $\varphi(S_n)$ est le meilleur estimateur de $\exp(-\theta)$ en terme de risque quadratique ?

HEC 2019 - loi de Rademacher, loi binomiale, loi uniforme, loi normale, loi de Poisson, convergence en loi, fonctions génératrice des moments, des cumulants et des probabilités

Dans ce problème, on définit et on étudie les fonctions génératrices des cumulants de variables aléatoires discrètes ou à densité.
Les cumulants d'ordre 3 et 4 permettent de définir des paramètres d'asymétrie et d'aplatissement qui viennent compléter la description usuelle d'une loi de probabilité par son espérance (paramètre de position) et sa variance (paramètre de dispersion) ; ces cumulants sont notamment utilisés pour l'évaluation des risques financiers.

Dans tout le problème :

- on note $(\Omega, \mathscr{A}, \mathbb{P})$ un espace probabilisé et toutes les variables aléatoires introduites dans l'énoncé sont des variables aléatoires réelles définies sur $(\Omega, \mathscr{A})$;
- sous réserve d'existence, l'espérance et la variance d'une variable aléatoire X sont respectivement notées $\mathbb{E}(X)$ et $\mathbb{V}(X)$;
- pour tout variable aléatoire X et pour tout réel t pour lesquels la variable aléatoire $\mathrm{e}^{t\,X}$ admet une espérance, on pose :
$$M_X(t) = \mathbb{E}\left(\mathrm{e}^{t\,X}\right) \quad \text{et} \quad K_X(t) = \ln\left(M_X(t)\right);$$
(les fonctions M_X et K_X sont respectivement appelées la *fonction génératrice des moments* et la *fonction génératrice des cumulants* de X)
- lorsque, pour un entier $p \in \mathbb{N}^*$, la fonction K_X est de classe $\mathcal{C}^p$ sur un intervalle ouvert contenant l'origine, on appelle *cumulant d'ordre p de X*, noté $Q_p(X)$, la valeur de la dérivée $p^{\text{ème}}$ de K_X en 0 :
$$Q_p(X) \;=\; K_X^{(p)}(0)$$

Partie I. Fonction génératrice des moments de variables aléatoires discrètes

Dans toute cette partie :

- on note n un entier supérieur ou égal à 2 ;
- toutes les variables aléatoires considérées sont discrètes à valeurs entières ;
- on note S une variable aléatoire à valeurs dans $\{-1, 1\}$ dont la loi est donnée par :
$$\mathbb{P}([S=-1]) \;=\; \mathbb{P}([S=+1]) \;=\; \frac{1}{2}$$

1. Soit X une variable aléatoire à valeurs dans $[\![-n, n]\!]$.

a) Pour tout $t \in \mathbb{R}$, écrire $M_X(t)$ sous la forme d'une somme et en déduire que la fonction M_X est de classe $\mathcal{C}^\infty$ sur $\mathbb{R}$.

b) Justifier pour tout $p \in \mathbb{N}^*$, l'égalité : $M_X^{(p)}(0) = \mathbb{E}(X^p)$.

c) Soit Y une variable aléatoire à valeurs dans $[\![-n, n]\!]$ dont la fonction génératrice des moments M_Y est la même que celle de X.
On note G_X et G_Y les deux polynômes définis par :
$$\forall x \in \mathbb{R},\ \begin{cases} G_X(x) \;=\; \displaystyle\sum_{k=0}^{2n} \mathbb{P}([X=k-n])\, x^k \\ G_Y(x) \;=\; \displaystyle\sum_{k=0}^{2n} \mathbb{P}([Y=k-n])\, x^k \end{cases}$$

(i) Vérifier pour tout $t \in \mathbb{R}$, l'égalité : $G_X(\mathrm{e}^t) = \mathrm{e}^{nt}\, M_X(t)$.

(ii) Justifier la relation : $\forall t \in \mathbb{R}$, $G_X(\mathrm{e}^t) = G_Y(\mathrm{e}^t)$.

(iii) En déduire que la variable aléatoire Y suit la même loi que X.

2. Dans cette question, on note X_2 une variable aléatoire qui suit la loi binomiale $\mathcal{B}\left(2, \frac{1}{2}\right)$.
On suppose que les variables aléatoires X_2 et S sont indépendantes et on pose $Y_2 = S\,X_2$.

a) (i) Préciser l'ensemble des valeurs possibles de la variable aléatoire Y_2.

(ii) Calculer les probabilités $\mathbb{P}([Y_2 = y])$ attachées aux diverses valeurs possibles y de Y_2.

b) Vérifier que la variable aléatoire $X_2 - (S+1)$ suit la même loi que Y_2.

3. Le script **Python** suivant permet d'effectuer des simulations de la variable aléatoire Y_2 définie dans la question précédente.

```
1  n = 10
2  X = rd.binomial(2,0.5,[n,2])
3  B = rd.binomial(1,0.5,[n,2])
4  S = 2*B - np.ones([n,2])
5  Z1 = [S[:,0]*X[:,0], X[:,0] - S[:,0] - np.ones(n)]
6  Z2 = [S[:,0]*X[:,0], X[:,1] - S[:,1] - np.ones(n)]
```

a) Que contiennent les variables `X` et `S` après l'exécution des quatre premières instructions ?

b) Expliquer pourquoi, après l'exécution des six instructions, chacun des coefficients des matrices `Z1` et `Z2` contient une simulation de la variable aléatoire Y_2.

c) On modifie la première ligne du script précédent en affectant à `n` une valeur beaucoup plus grande que 10 (par exemple, 100000) et en lui adjoignant les deux instructions 7 et 8 suivantes :

```
7  p1 = len(np.argwhere(Z1[0] == Z1[1])) / n
8  p2 = len(np.argwhere(Z2[0] == Z2[1])) / n
```

Quelles valeurs numériques approchées la loi faible des grands nombres permet-elle de fournir pour `p1` et `p2` après l'exécution des huit lignes du nouveau script ?
Dans le langage **Python**, la fonction `len` fournit la « longueur » d'un vecteur, d'une liste ou d'une matrice carrée et la fonction `np.argwhere` calcule les positions des coefficients d'une matrice pour lesquels une propriété est vraie, comme l'illustre le script suivant :

```
-->  A = np.array([1,2,0,4])
-->  B = np.array([2,2,4,3])
-->  len(A)
     ans = 4.
-->  np.argwhere(A < B)
     = [[0]
        [2]]
     # car 1 < 2 et 0 < 4, alors que 2 ⩾ 2 et 4 ⩾ 3
```

4. Dans cette question, on note X_n une variable aléatoire qui suit la loi binomiale $\mathcal{B}\left(n, \frac{1}{2}\right)$.
On suppose que les variables aléatoires X_n et S sont indépendantes et on pose $Y_n = S\,X_n$.

a) Justifier que la fonction M_{X_n} est définie sur $\mathbb{R}$ et calculer $M_{X_n}(t)$ pour tout $t \in \mathbb{R}$.

b) Montrer que la fonction M_{Y_n} est donnée par : $\forall t \in \mathbb{R}$, $M_{Y_n}(t) = \dfrac{1}{2^{n+1}}\left((1+\mathrm{e}^t)^n + (1+\mathrm{e}^{-t})^n\right)$.

c) En utilisant l'égalité $(1+\mathrm{e}^{-t})^n = \mathrm{e}^{-nt}\,(1+\mathrm{e}^{t})^n$, montrer que Y_n suit la même loi que la différence $X_n - H_n$, où H_n est une variable aléatoire indépendante de X_n dont on précisera la loi.

Partie II. Propriétés générales des fonctions génératrices des cumulants et quelques exemples

5. Soit X une variable aléatoire et $\mathcal{D}_X$ le domaine de définition de la fonction K_X.

a) Donner la valeur de $K_X(0)$.

b) Soit $(a,b) \in \mathbb{R}^2$ et $Y = a\,X + b$. Justifier pour tout réel t pour lequel $a\,t$ appartient à $\mathcal{D}_X$, l'égalité :

$$K_Y(t) \;=\; b\,t + K_X(a\,t)$$

c) On suppose ici que les variables aléatoires X et $-X$ suivent la même loi.
Que peut-on dire dans ce cas des cumulants d'ordre impair de la variables aléatoire X ?

6. Soit X et Y deux variables aléatoires indépendantes et $\mathcal{D}_X$ et $\mathcal{D}_Y$ les domaines de définition respectifs des fonctions K_X et K_Y.

a) Monter que pour tout réel t appartenant à la fois à $\mathcal{D}_X$ et $\mathcal{D}_Y$, on a : $K_{X+Y}(t) = K_X(t) + K_Y(t)$.

b) En déduire une relation entre les cumulants des variables aléatoires X, Y et $X+Y$.

7. Soit U une variable aléatoire suivant la loi uniforme sur l'intervalle $[0,1]$.

a) Montrer que la fonction M_U est définie sur $\mathbb{R}$ et donnée par : $\forall t \in \mathbb{R}$, $M_U(t) = \begin{cases} \dfrac{\mathrm{e}^t - 1}{t} & \text{si } t \neq 0 \\ 1 & \text{si } t = 0 \end{cases}$.

b) Calculer la dérivée de la fonction M_U en tout point $t \neq 0$.

c) Trouver la limite du quotient $\dfrac{M_U(t) - 1}{t}$ lorsque t tend vers 0.

d) Montrer que la fonction M_U est de classe $\mathcal{C}^1$ sur $\mathbb{R}$.

8. Soit α et β deux réels tels que $\alpha < \beta$.
Dans cette question, on note X une variable aléatoire qui suit la loi uniforme sur l'intervalle $[\alpha, \beta]$.

a) Exprimer K_X en fonction de M_U, où la variable aléatoire U a été définie dans la question ***7.***

b) Justifier que la fonction K_X est de classe $\mathcal{C}^1$ sur $\mathbb{R}$ et établir l'égalité : $Q_1(X) = \mathbb{E}(X)$.

9. Soit un réel $\lambda > 0$ et soit T une variable aléatoire qui suit la loi de Poisson de paramètre λ.

a) Déterminer les fonctions M_T et K_T.

b) En déduire les cumulants de T.

10. Soit Z une variable aléatoire qui suit la loi normale centrée réduite.

a) Justifier pour tout $t \in \mathbb{R}$, la convergence de l'intégrale $\displaystyle\int_{-\infty}^{+\infty} \exp\left(t\,x - \frac{x^2}{2}\right)\,dx$.

b) Montrer que la fonction M_Z est définie sur $\mathbb{R}$ et donnée par : $\forall t \in \mathbb{R}$, $M_Z(t) = \exp\left(\dfrac{t^2}{2}\right)$.

c) En déduire la valeur de tous les cumulants d'une variable aléatoire qui suit une loi normale d'espérance $\mu \in \mathbb{R}$ et d'écart-type $\sigma \in \mathbb{R}_+^*$.

11. Soit $(T_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires telles que, pour tout $n \in \mathbb{N}^*$, la variable aléatoire T_n suit la loi de Poisson de paramètre n. Pour tout $n \in \mathbb{N}^*$, on pose : $W_n = \dfrac{T_n - n}{\sqrt{n}}$.

a) Justifier la convergence en loi de la suite de variables aléatoires $(W_n)_{n\in\mathbb{N}^*}$ vers une variable aléatoire W.

b) Déterminer la fonction K_{W_n}.

c) Montrer que pour tout $t \in \mathbb{R}$, on a : $\lim\limits_{n\to+\infty} K_{W_n}(t) = K_W(t)$.

Partie III. Cumulant d'ordre 4

Dans cette partie, on considère une variable aléatoire X telle que M_X est de classe $\mathcal{C}^4$ sur un intervalle ouvert I contenant l'origine.
On admet alors que X possède des moments jusqu'à l'ordre 4 qui coïncident avec les dérivées successives de la fonction M_X en 0. Autrement dit, pour tout $k \in [\![1, 4]\!]$, on a : $M_X^{(k)}(0) = \mathbb{E}(X^k)$.

De plus, on pose : $\mu_4(X) = \mathbb{E}\left(\big(X - \mathbb{E}(X)\big)^4\right)$.

12. Justifier les égalités : $Q_1(X) = \mathbb{E}(X)$ et $Q_2(X) = \mathbb{V}(X)$.

13. Soit X_1 et X_2 deux variables aléatoires indépendantes et de même loi que X. On pose : $S = X_1 - X_2$.

a) Montrer que la variable aléatoire S possède un moment d'ordre 4 et établir l'égalité :

$$\mathbb{E}(S^4) \;=\; 2\,\mu_4(X) + 6\,\big(\mathbb{V}(X)\big)^2$$

b) Montrer que les fonctions M_S et K_S sont de classe $\mathcal{C}^4$ sur I et que pour tout $t \in I$, on a :

$$M_S^{(4)}(t) \;=\; K_S^{(4)}(t)\,M_S(t) + 3\,K_S^{(3)}(t)\,M_S'(t) + 3\,K_S''(t)\,M_S''(t) + K_S'(t)\,M_S^{(3)}(t)$$

c) En déduire l'égalité : $\mathbb{E}(S^4) \;=\; Q_4(S) + 3\,\big(\mathbb{V}(S)\big)^2$.

14. Justifier que le cumulant d'ordre 4 de X est donné par la relation : $Q_4(X) \;=\; \mu_4(X) - 3\,\big(\mathbb{V}(X)\big)^2$.

HEC 2020 - couple de variables aléatoires à densité, loi exponentielle, loi de Bernoulli, loi binomiale, inégalité de Boole

On s'intéresse dans ce sujet au problème de la *double dépense* de *bitcoins* par un groupe d'individus mal intentionnés.
On rappelle que le bitcoin est une monnaie virtuelle dont l'utilisation pour des transactions est associée à une structure unique appelée *blockchain*, partagée sur le réseau des usagers de cette monnaie et ayant pour but de sécuriser ces transactions.
La modélisation étudiée ne nécessite pas de connaissances particulières sur le *bitcoin* et la *blockchain*.

Partie I - Deux résultats généraux

On démontre dans cette partie deux résultats préliminaires, aux questions ***5.*** et ***6.***. Ces résultats seront utilisés dans la suite du sujet et pourront être admis.

Calcul d'une probabilité

Soient X et Y deux variables aléatoires sur un espace probabilisé, à densité et indépendantes.
On note F_X et F_Y les fonctions de répartition de X et Y.
On suppose que Y est à valeurs positives et possède une densité f_Y dont la restriction à $[0, +\infty[$ est continue sur cet intervalle.
Pour tout $x \in \mathbb{R}_+$, on pose : $H(x) = \mathbb{P}(\,[X \leqslant Y] \cap [Y \leqslant x]\,)$.

1. ***a)*** Montrer que H est une fonction croissante sur $\mathbb{R}_+$ qui admet une limite finie en $+\infty$.

b) En utilisant la suite $\big(H(n)\big)_{n\in\mathbb{N}}$, démontrer : $\lim\limits_{x\to+\infty} H(x) = \mathbb{P}(\,[X \leqslant Y]\,)$.
Que vaut $H(0)$?

2. Soit (u, v) un couple de réels positifs tels que : $u < v$.

a) Montrer : $H(v) - H(u) = \mathbb{P}(\,[X \leqslant Y] \cap [u < Y \leqslant v]\,)$. Puis :

$$F_X(u)\,\frac{F_Y(v) - F_Y(u)}{v-u} \;\leqslant\; \frac{H(v) - H(u)}{v-u} \;\leqslant\; F_X(v)\,\frac{F_Y(v) - F_Y(u)}{v-u}$$

b) En déduire que pour tout $x \in \mathbb{R}_+$, H est dérivable en x et : $H'(x) = F_X(x)\,f_Y(x)$.

c) En conclure que pour tout x réel positif : $H(x) = \displaystyle\int_0^x F_X(t)\,f_Y(t)\,dt$.

3. Démontrer : $\mathbb{P}(\,[X \leqslant Y]\,) = \displaystyle\int_0^{+\infty} F_X(t)\,f_Y(t)\,dt$.

4. En utilisant la fonction $K : x \mapsto \mathbb{P}(\,[X < Y] \cap [Y \leqslant x]\,)$, on montrerait de même et nous l'admettrons :

$$\mathbb{P}(\,[X < Y]\,) \;=\; \int_0^{+\infty} F_X(t)\,f_Y(t)\,dt \;=\; \mathbb{P}(\,[X \leqslant Y]\,)$$

Que peut-on en déduire pour $\mathbb{P}(\,[X = Y]\,)$?

5. *Application aux lois exponentielles*
On suppose que U et V sont deux variables aléatoires indépendantes suivant des lois exponentielles de paramètres respectifs λ et μ, réels strictement positifs.
Soit θ un réel positif ou nul.

a) Déterminer la fonction de répartition de la variable aléatoire $X = U - \theta$.

b) En déduire que pour tout $\theta \geqslant 0$:

$$\mathbb{P}(\,[U - \theta \leqslant V]\,) \;=\; 1 - \frac{\mu}{\lambda + \mu}\,\mathrm{e}^{-\lambda\,\theta}$$

Inégalité de Boole

6. On considère $(B_k)_{k\in\mathbb{N}^*}$ une famille d'événements d'un espace probabilisé.

a) Montrer par récurrence sur $n \in \mathbb{N}^*$: $\mathbb{P}\left(\bigcup_{k=1}^{n} B_k\right) \leqslant \sum_{k=1}^{n} \mathbb{P}(B_k)$.

b) On suppose que la série $\sum_{k\geqslant 1} \mathbb{P}(B_k)$ converge. Démontrer :

$$\mathbb{P}\left(\bigcup_{k=1}^{+\infty} B_k\right) \;\leqslant\; \sum_{k=1}^{+\infty} \mathbb{P}(B_k)$$

Partie II - Une compétition entre deux groupes

Dans toute la suite du sujet, on désigne par p un réel de l'intervalle $]0,1[$ et on pose $q = 1 - p$.
On modélise une compétition entre deux groupes d'individus A et B avec les règles suivantes.

- Le groupe A doit résoudre une suite de problèmes $(P_k)_{k\geqslant 1}$ dans l'ordre des indices. Au temps $t = 0$, le groupe commence la résolution du problème P_1, ce qui lui prend un temps représenté par la variable aléatoire X_1. Une fois P_1 résolu, le groupe aborde immédiatement le problème P_2, et on note X_2 le temps consacré à la résolution de P_2 par le groupe A, et ainsi de suite.
  Pour tout $k \in \mathbb{N}^*$, on note X_k la variable aléatoire donnant le temps consacré à la résolution du problème P_k par le groupe A.
- De même, le groupe B doit résoudre dans l'ordre une suite de problèmes $(Q_k)_{k\geqslant 1}$; la résolution du premier problème Q_1 commence au temps $t = 0$ et on note, pour tout $k \in \mathbb{N}^*$, Y_k la variable aléatoire donnant le temps consacré par le groupe B à la résolution du problème Q_k.
- À ce jeu est associé un espace probabilisé $(\Omega, \mathscr{A}, \mathbb{P})$ sur lequel sont définies les suites de variables aléatoires $(X_k)_{k\geqslant 1}$ et $(Y_k)_{k\geqslant 1}$, et on fait les hypothèses suivantes :
  - × pour tout $k \in \mathbb{N}^*$, X_k suit la loi exponentielle de paramètre p, notée $\mathcal{E}(p)$, et Y_k suit la loi exponentielle $\mathcal{E}(q)$;
  - × pour tout $k \in \mathbb{N}^*$, les variables aléatoires $X_1, \ldots, X_k, Y_1, \ldots, Y_k$ sont indépendantes.
- On établit alors la liste de tous les problèmes résolus *dans l'ordre où ils le sont par les deux groupes.* En cas de simultanéité temporelle de la résolution par les deux groupes d'un de leurs problèmes, on placera d'abord le problème résolu par A dans la liste puis celui résolu par B.
  Pour tout $n \in \mathbb{N}^*$, on note U_n la variable aléatoire de Bernoulli associée à l'événement « le $n^{\text{ème}}$ problème placé dans la liste est un problème résolu par le groupe A ».
  Par exemple, si la liste des cinq premiers problèmes résolus est $(P_1, P_2, Q_1, P_3, Q_2)$, alors $U_1 = 1$, $U_2 = 1$, $U_3 = 0$, $U_4 = 1$ et $U_5 = 0$.
- Pour tout $n \geqslant 0$, on note aussi S_n la variable aléatoire donnant le nombre de problèmes qui ont été résolus par A présents dans la liste des n premiers problèmes résolus. En particulier, S_0 vaut toujours 0.

7. ***a)*** Que représente la variable aléatoire $\sum_{k=1}^{n} X_k$?

b) On suppose que $X_1 = 5$, $X_2 = 2$, $X_3 = 3$, $X_4 = 2$, $Y_1 = 2$, $Y_2 = 2$, $Y_3 = 4$ et $Y_4 = 2$.
Déterminer $U_1, \ldots, U_7$.
Peut-on aussi en déduire la valeur de U_8 ?

c) Compléter le script **Python** suivant pour qu'il simule le jeu et, pour n, p donnés, affiche la liste des valeurs $U_1, U_2, \ldots, U_n$:

```
p = float(input('p = '))
n = int(input('n = '))
q = 1 - p
U = np.zeros(n)
sommeX = rd.exponential(1/p)
sommeY = rd.exponential(1/q)
mini = min(sommeX, sommeY)
for k in range(n):
    if sommeX == ...:
        U[k] = ...
        sommeX = sommeX + rd.exponential(1/p)
    else:
        sommeY = ...
    mini = min(sommeX, sommeY)
...
```

d) Quelle(s) instruction(s) faut-il ajouter pour afficher la valeur de S_n ?

8. *Loi de U_n*
Dans cette question, on démontre par récurrence sur $n \geqslant 1$: $\mathbb{P}([U_n = 1]) = p$.

a) Démontrer : $\mathbb{P}([U_1 = 1]) = \mathbb{P}([X_1 \leqslant Y_1]) = p$.

b) (i) Démontrer, pour tout réel $x < 0$: $\mathbb{P}_{[U_1=1]}([Y_1 - X_1 \leqslant x]) = 0$.

(ii) Soit x un réel positif ou nul.
Établir : $\mathbb{P}_{[U_1=1]}([Y_1 - X_1 \leqslant x]) = \dfrac{1}{p}\,\mathbb{P}([X_1 \leqslant Y_1 \leqslant X_1 + x])$,
puis calculer $\mathbb{P}_{[U_1=1]}([Y_1 - X_1 \leqslant x])$.

c) On peut interpréter ce résultat en disant que la *loi conditionnelle de* $Y_1 - X_1$ *sachant* $[U_1 = 1]$ est une loi exponentielle. Quelle est son paramètre ?
Par analogie, quelle est la loi conditionnelle de $X_1 - Y_1$ sachant $[U_1 = 0]$? (on n'attend pas une démonstration précise mais un argument de bon sens pour justifier le résultat proposé).

d) On suppose que $n \in \mathbb{N}^*$ et $\mathbb{P}([U_1 = 1]) = p$.
Déduire de cette hypothèse et de la question précédente :
$\mathbb{P}_{[U_1=1]}([U_{n+1} = 1]) = p$ et $\mathbb{P}_{[U_1=0]}([U_{n+1} = 1]) = p$.

e) Conclure.

9. On montrerait aussi par récurrence, et nous l'admettrons, que pour tout $n \in \mathbb{N}^*$, les variables aléatoires $U_1, \ldots, U_n$ sont mutuellement indépendantes.
En déduire la loi de S_n.

Soit $r \in \mathbb{N}$, on s'intéresse, dans les questions qui suivent, à la probabilité a_r de l'événement :

$$A_r \ : \ \begin{array}{l} \text{« il existe un } n \geqslant r \text{ tel que, lorsque } n \text{ problèmes} \\ \text{en tout ont été résolus, le groupe } A \text{ en a résolu} \\ r \text{ de plus que le groupe } B \text{ »} \end{array}$$

10. ***a)*** Justifier : $a_0 = 1$.

b) Démontrer, pour tout $r \geqslant 1$:

$$\mathbb{P}_{[U_1=1]}(A_r) = \mathbb{P}(A_{r-1}) \quad \text{et} \quad \mathbb{P}_{[U_1=0]}(A_r) = \mathbb{P}(A_{r+1})$$

c) En déduire, pour tout $r \geqslant 1$: $a_{r+1} = \dfrac{1}{q}\, a_r - \dfrac{p}{q}\, a_{r-1}$.

d) En remarquant que $1 - 4pq = (1-2p)^2$, donner une expression de a_r en fonction de p, q, r et de deux constantes que l'on introduira.

11. Le cas $p \geqslant \dfrac{1}{2}$.

Montrer que, dans les cas $p = \dfrac{1}{2}$ et $p > \dfrac{1}{2}$, la suite $(a_r)_{r\in\mathbb{N}}$ est constante et égale à 1.

12. Le cas $p < \dfrac{1}{2}$.

a) Soit k un entier naturel.

(i) Établir : $A_{2k} = \bigcup_{i \geqslant k} [S_{2i} = i + k]$.

(ii) Montrer que pour tout $i \geqslant k$, on a : $\mathbb{P}(\,[S_{2i} = i + k]\,) = \binom{2i}{i+k} p^{i+k}\, q^{i-k}$.

(iii) Après avoir donné la valeur de la somme $\sum_{j=0}^{2i} \binom{2i}{j}$, démontrer :

$$\forall i \geqslant k, \ \binom{2i}{i+k} \leqslant 4^i$$

(iv) En déduire l'inégalité :

$$\sum_{i=k}^{+\infty} \mathbb{P}(\,[S_{2i} = k + i]\,) \ \leqslant \ \left(\frac{p}{q}\right)^k \ \frac{(4pq)^k}{1-4pq}$$

b) Montrer en utilisant l'inégalité de Boole (voir question ***6.***) que si $p < \dfrac{1}{2}$, alors : $\lim\limits_{k\to+\infty} a_{2k} = 0$.

c) Conclure en utilisant la question ***10.d)***, que si $p < \dfrac{1}{2}$, alors :

$$\forall r \in \mathbb{N}, \ a_r = \left(\frac{p}{q}\right)^r$$

On a ainsi établi dans les questions ***11.*** et ***12.*** :

$$\forall r \in \mathbb{N}, \quad a_r = \begin{cases} \left(\dfrac{p}{q}\right)^r & \text{si } p < \dfrac{1}{2} \\ 1 & \text{si } p \geqslant \dfrac{1}{2} \end{cases}$$

Ce résultat pourra être admis et utilisé dans la suite du sujet.

Partie III - La *blockchain* et la stratégie de la *double dépense*

On utilise, dans cette partie, les notations et résultats de la partie **II**.
Soit n un entier supérieur ou égal à 1.
La *blockchain* est formée d'une suite de blocs, chacun associé à plusieurs transactions. Elle contient l'historique de toutes les transactions effectuées depuis la création du *bitcoin*.
Avant d'être placé dans la *blockchain*, un nouveau bloc doit être validé. Cette validation nécessite la mise en oeuvre d'une grande puissance de calcul pour résoudre un problème dépendant fortement du contenu du bloc et des blocs qui le précèdent.
Les individus qui valident les blocs sont appelés mineurs.
Il est possible qu'à un instant donné, coexistent sur le réseau deux *blockchains*, valides et différentes. Dans ce cas, le réseau choisira celle qui comporte le plus de blocs et l'autre sera abandonnée.
Par prudence, lorsqu'un bloc est validé, il est recommandé d'attendre que $n-1$ blocs le suivant soient aussi validés pour considérer que les transactions incluses dans le bloc soient honnêtes.
Un groupe de mineurs mal intentionnés, noté A, peut essayer de dépenser deux fois les mêmes *bitcoins* en procédant ainsi :

- le groupe A demande la validation de l'achat d'un bien d'un montant de s *bitcoins* qu'il a en sa possession.
- lorsque le bloc K incluant cette transaction est proposé à la validation sur le réseau, A modifie ce bloc en K', qu'il ne diffuse pas, en remplaçant l'achat par une vente des s *bitcoins* en euros à son profit par exemple. Il se met alors à la validation de ce nouveau bloc et crée ainsi une deuxième instance de la *blockchain* qu'il continue à développer sans la diffuser.
- lorsque le groupe B, représentant l'ensemble des autres mineurs du réseau, a validé K ainsi que les $n-1$ blocs suivants, le vendeur du bien considère que la transaction est valide et fournit le bien.
- le groupe A attend alors d'avoir une *blockchain* plus longue que celle de B, qui est publique, pour la diffuser donc invalider la *blockchain* publique et l'achat du bien. Le crédit en *bitcoins* du vendeur du bien est alors annulé.

On reprend et on complète la modélisation de la partie précédente pour déterminer la probabilité que la stratégie de la *double dépense* réussisse et le choix de n pour que cette probabilité soit faible.
Une première phase du jeu, décrit dans la partie **II**, s'achève à l'instant aléatoire t où le problème Q_n est ajouté à la liste des problèmes résolus.
Le groupe de mineurs A est ensuite déclaré vainqueur s'il se trouve un instant $t' \geqslant t$ où le nombre de problèmes résolus par A dans la liste des problèmes résolus depuis le début du jeu, est strictement supérieur au nombre de ceux résolus par B dans cette même liste. On note G_n cet événement.
On détermine, dans cette partie, la probabilité de G_n en fonction de n et de p.

13. On s'intéresse tout d'abord à la loi de la variable aléatoire T_n égale au nombre de problèmes résolus par le groupe A lorsque l'on place Q_n dans la liste des problèmes résolus.

a) Démontrer, pour tout $k \in \mathbb{N}$: $[T_n = k] = [S_{n+k-1} = k] \cap [U_{n+k} = 0]$.

b) En déduire : $\mathbb{P}(\,[T_n = k]\,) = \binom{n+k-1}{k} p^k\, q^n$.

14. a) En utilisant la formule des probabilités totales, établir :

$$\mathbb{P}(G_n) \;=\; \mathbb{P}(\,[T_n \geqslant n+1]\,) + \sum_{k=0}^{n} \mathbb{P}(\,[T_n = k]\,)\, a_{n+1-k}$$

b) Dans le cas où $p \geqslant \dfrac{1}{2}$, en déduire : $\mathbb{P}(G_n) = 1$.

c) De même lorsque $p < \dfrac{1}{2}$, démontrer :

$$\mathbb{P}(G_n) \;=\; 1 - \sum_{k=0}^{n} \binom{n+k-1}{k} \left(p^k\, q^n - p^{n+1}\, q^{k-1}\right)$$

15. *Une meilleure expression de $\mathbb{P}(G_n)$ lorsque $p < \dfrac{1}{2}$*

Pour tout $x \in [0,1]$ et $n \in \mathbb{N}^*$, on pose :

$$u_n(x) \;=\; (1-x)^n \sum_{k=0}^{n} \binom{n+k-1}{k} x^k$$

a) Vérifier que pour tout $n \in \mathbb{N}^*$: $\mathbb{P}(G_n) = 1 - u_n(p) + \dfrac{p}{q}\, u_n(q)$.

b) Pour tout $x \in [0,1]$ et $n \in \mathbb{N}^*$, établir la relation :

$$u_{n+1}(x) \;=\; u_n(x) + (1-x)^n\, x^{n+1} \left(\binom{2n}{n+1} - \binom{2n+1}{n+1} x \right)$$

c) En déduire, pour tout $n \in \mathbb{N}^*$:

$$\mathbb{P}(G_{n+1}) \;=\; \mathbb{P}(G_n) - \left(1 - \frac{p}{q}\right)\, (pq)^{n+1} \binom{2n+1}{n+1}$$

d) Montrer par récurrence, pour tout $n \in \mathbb{N}^*$:

$$\mathbb{P}(G_n) \;=\; \frac{p}{q} - \left(1 - \frac{p}{q}\right) \sum_{k=1}^{n} \binom{2k-1}{k} (pq)^k$$

16. *Application à la sécurisation des transactions*

Connaissant $p < \dfrac{1}{2}$, on cherche à limiter le risque que la stratégie mise en place par le groupe de mineurs A réussisse.

a) Après avoir établi la formule $\binom{n}{k} = \dfrac{n}{k} \binom{n-1}{k-1}$ lorsque $k \in [\![1, n]\!]$, écrire une fonction **Python** qui calcule les coefficients binomiaux.

b) Écrire un script **Python** qui détermine n_p, le plus petit entier n tel que $\mathbb{P}(G_n) \leqslant \varepsilon$ pour $p < \dfrac{1}{2}$ et $\varepsilon > 0$ saisis au clavier par l'utilisateur.

NB : Pour $\varepsilon = 10^{-4} = 0,1\%$ et p variant entre 10% et 32%, on obtient pour la représentation de n_p en fonction de p :

HEC 2008 (Exercice) - fonction de deux variables, méthode des moindres carrés, droite de régression linéaire

Étant donné un entier n supérieur ou égal à 2, on considère un nuage de n points du plan, c'est-à-dire un n-uplet $\big((x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\big)$ d'éléments de $\mathbb{R}^2$. On suppose que les réels x_1, x_2, …, x_n (resp. y_1, y_2, …, y_n) ne sont pas tous égaux.
On appelle moyenne arithmétique $\overline{x}$ et écart-type σ_x du n-uplet $x=(x_1,\ldots,x_n)$, les réels suivants :

$$\overline{x} \;=\; \frac{1}{n}\sum_{k=1}^{n} x_k \qquad \text{et} \qquad \sigma_x \;=\; \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(x_k-\overline{x}\right)^2}$$

On définit de même la moyenne arithmétique $\overline{y}$ et l'écart-type σ_y du n-uplet $y=(y_1,\ldots,y_n)$.
La covariance $\mathrm{cov}(x,y)$ et le coefficient de corrélation linéaire $r(x,y)$ du couple (x,y) sont donnés par :

$$\mathrm{cov}(x,y) \;=\; \frac{1}{n}\sum_{k=1}^{n}\left(x_k-\overline{x}\right)\left(y_k-\overline{y}\right) \qquad \text{et} \qquad r(x,y)=\frac{\mathrm{cov}(x,y)}{\sigma_x\times\sigma_y}$$

Soit f la fonction définie sur $\mathbb{R}^2$ à valeurs réelles qui, à tout couple (a,b) de $\mathbb{R}^2$, associe le réel $f(a,b)$ tel que :

$$f(a,b) \;=\; \sum_{k=1}^{n}\left(a\,x_k+b-y_k\right)^2$$

1. Justifier que f est de classe $\mathcal{C}^2$ sur $\mathbb{R}^2$.

2. ***a)*** Écrire le système d'équations (S) permettant de déterminer les points critiques de f.

b) Résoudre le système (S). En déduire que f admet un unique point critique $(\hat{a},\hat{b})$ que l'on exprimera en fonction de $\overline{x}$, $\overline{y}$, σ_x^2 et $\mathrm{cov}(x,y)$.

c) Montrer que ce point critique correspond à un minimum local de f.

d) Établir la formule suivante : $f(\hat{a},\hat{b})=n\,\sigma_y^2\left(1-r^2(x,y)\right)$.

3. ***a)*** Montrer que l'on a : $|r(x,y)|\leqslant 1$.

b) Que peut-on dire du nuage de points lorsque $|r(x,y)|=1$?