Devoir

La calculatrice, le polycopié de cours et les tables de la loi normale sont autorisés.
Tous les résultats doivent être justifiés et correctement rédigés.
Le barème est donné à titre indicatif.

Exercice 1 : records (10 points)

On considère des résultats numériques successifs d'une certaine expérience. On dit qu'un résultat est un record parmi N résultats s'il en est le maximum. Le but de ce problème est d'étudier quelques propriétés probabilistes des records.

Notons $(X_k)_{k\in\mathbf{N}^*}$ une suite de variables continues indépendantes de même loi. Notons f leur densité et F leur fonction de répartition.

1. **Loi du record**

   Soit $N \in \mathbf{N}^*$. Notons $R_N = \max_{k\leqslant N} X_k$. La variable R_N représente le record des N premiers tirages.

   (a) Montrer que la fonction de répartition de R_N est donnée par $F_R = F^N$.

   Soit $x \in \mathbf{R}$. Alors $F_R(x) = \mathbf{P}(R \leqslant x) = \mathbf{P}(max X_k \leqslant x) = \mathbf{P}(X_1 \leqslant x \cap \cdots \cap X_N \leqslant x)$. Comme les variables X_k sont indépendantes, on obtient

   $$F_R(x) = \mathbf{P}(X_1 \leqslant x) \times \cdots \times \mathbf{P}(X_N \leqslant x) = F^N(x).$$

   (b) Supposons jusqu'à la fin de cette partie 1 que les variables X_k suivent la loi uniforme sur $[0,1]$.
   Déterminer alors la densité de R_N et la représenter.

   La densité de la loi uniforme est $f(x) = 1$ définie sur $[0,1]$. Sa fonction de répartition est $F(x) = x$ définie sur $[0,1]$ (c'est la primitive de f s'annulant croissant de 0 vers 1).
   Donc, $F_R(x) = x^N$ définie sur $[0,1]$. Et la densité de R_N est définie par $f_R(x) = F'_R(x) = Nx^{N-1}$.

   (c) Calculer la valeur moyenne de R_N.

   Ainsi $\mathbf{E}(R_N) = \int_0^1 x f_R(x)dx = \int_0^1 Nx^N dx = \frac{N}{N+1}$.

   (d) Faisons maintenant varier N. Remarquer que la suite des valeurs prises par les variables R_N est croissante et déduire des questions précédentes qu'elle converge presque sûrement vers 1.

   On regarde la suite des résultats. Comme R_N représente le max des N premiers résultats, il est clair que si $N_1 < N_2$, $R_{N_2} \geqslant R_{N_2}$. Quelque soit la suite des résultats X_k, la suite (R_N) correspondante sera croissante. Comme elle est ici majorée par 1,

on en déduit qu'elle converge vers une limite $l \leqslant 1$.
Montrons qu'avec une probabilité 1, $l = 1$. L'idée est que la valeur moyenne des R_N converge vers 1 d'après la question précédente. Si la limite de la suite R_N n'était pas presque toujours 1, cela signifierait que les valeurs de R_N restent inférieure à une certaine valeur $a < 1$ avec une probabilité non nulle et on obtiendrait une contradiction.
Plus rigoureusement, supposons qu'il existe $a < 1$ et $\varepsilon > 1$ tels que $\mathbf{P}(l \leqslant a) = \varepsilon$. Comme les termes de la suite R_N sont toujours inférieurs à la limite l, cela implique que pour tout N $\mathbf{P}(R_N \leqslant a) \geqslant \varepsilon$. D'après l'inégalité de Markov (R_N est bien à valeurs positives ici), $\mathbf{P}(R_N > a) \leqslant \frac{\mathbf{E}(R_N)}{a}$. Donc $\mathbf{P}(R_N \leqslant a) = 1 - \mathbf{P}(R_N < a) \geqslant \frac{\mathbf{E}(R_N)}{a}$. Or $\mathbf{E}(R_N)$ tend vers 1 et $a < 1$. On obtient pour N assez grand $\mathbf{P}(R_N \leqslant a) > 1$ ce qui est absurde.
Conclusion : quitte à attendre suffisamment longtemps, il y aura toujours de nouveaux records, et ils se rapprocheront indéfiniment de 1.

2. **Nombre de tentatives pour obtenir un nouveau record**

On connaît le résultat du premier tirage X_1. On le note a. On souhaite déterminer le nombre moyen de tirages qu'il faut effectuer pour réussir à dépasser la valeur a.

(a) Soit $k > 1$. Exprimer la probabilité $p_a = \mathbf{P}(X_k > a)$ en fonction de f ou de F.

$$p_a = \mathbf{P}(X_k > a) = 1 - F(a) = \int_a^{+\infty} f(x)dx.$$

Notons Y le nombre de tirages nécessaires pour dépasser a :
pour tout nombre entier $n \geqslant 1$, $Y = n$ si $\forall k \in \{2, 3, \ldots, n\}$, $X_k \leqslant a$ et $X_{n+1} > a$.

(b) Quelle est la loi de Y ?
Y représente le nombre de tirages avant d'obtenir un certain résultat, les différents tirages étant de même loi et indépendants. On reconnaît la loi géométrique qui est la loi du premier succès. La probabilité de succès pour un tirage est p_a. Donc $Y \sim \mathcal{G}(p_a)$.

(c) En déduire le nombre moyen de tentatives nécessaire pour dépasser a.

On connaît l'espérance de la loi géométrique : $\mathbf{E}(Y) = \frac{1}{p_a}$.

(d) Supposons qu'un nouveau record ait été établi ; notons b sa valeur. Que peut-on dire de p_b par rapport à p_a ?

Puisque b est le nouveau record, on a $b > a$. Alors comme F est croissante, $F(b) \geqslant F(a)$, donc $p_b \leqslant p_a$.

(e) Que peut-on en déduire sur les temps écoulés entre les nouveaux records successifs ?
On sait que le nombre moyen de tentatives nécessaire pour battre le record initial a est $\frac{1}{p_a}$. Une fois un nouveau record b établi, le nombre moyen de tentatives nécessaire pour le dépasser sera $\frac{1}{p_b} \geqslant \frac{1}{p_a}$. Si on obtient un nouveau record $c > b$, le nombre moyen de tentatives nécessaire pour dépasser c sera alors $\frac{1}{p_c} \geqslant \frac{1}{p_b}$.
On montre ainsi récursivement que, <u>en moyenne</u>, les nombres de tentatives séparant les différents records successifs sont de plus en plus élevés.
Conclusion : tout record sera battu, mais le temps nécessaire pour établir un nouveau record sera de plus en plus élevé.

Exercice 2 (4 points)

1. Représenter le graphe d'une fonction de densité f (discrète ou continue) associée à une variable aléatoire X telle que $\mathbf{E}(X) = 0$ et $\mathbf{V}(X) \approx 10$.

La densité doit être paire. Pour assurer la variance demandée, le plus simple est de prendre une variable discrète à valeurs dans $\pm\sqrt{10}$ avec probabilités $0,5 - 0,5$.

2. Représenter le graphe d'une fonction de densité f (discrète ou continue) associée à une variable aléatoire X telle que $\mathbf{E}(X) \approx 0$, $\mathbf{V}(X) \approx 10$ et $\mathbf{P}(X \leqslant 2) \approx 0,75$.

Un premier exemple reprenant un peu le précédent : une variable à valeurs dans $-a, 0, a$ avec probabilités $0,25 - 0,5 - 0,25$. Reste à choisir a pour que la variance soit bonne. On trouve $a = \sqrt{5}$.
Un autre exemple : une variable à valeurs dans $\{-3, 2, 4\}$ avec probabilités $\{0.5, 0.25, 0.25\}$.

Exercice 3 : estimation d'un paramètre. (8 points)

Soit $(X_k)_{k\geqslant 1}$ une suite de variables aléatoires indépendantes de même loi exponentielle $\mathcal{E}(\lambda)$, où λ est un paramètre strictement positif. Le but de cet exercice est d'estimer la valeur du paramètre λ à partir des tirages des variables X_k.

On rappelle que la loi exponentielle a pour densité $f(x) = \lambda e^{-\lambda x}$ définie pour $x \geqslant 0$. Son espérance est $\frac{1}{\lambda}$ et sa variance est $\frac{1}{\lambda^2}$.

Enfin, nous considérons un entier $n \geqslant 100$ et notons $\overline{X}_n$ la moyenne des n premiers tirages X_k :

$$\overline{X}_n = \frac{\sum_{k=1}^n X_k}{n}.$$

1. Que peut-on dire du comportement de la suite $\overline{X}_n$ lorsque n tend vers $+\infty$?

   D'après la loi forte des grands nombres, la suite converge presque sûrement vers l'espérance $1/\lambda$. Les hypothèses du théorème sont bien satisfaites : les variables X_k sont iid.

2. On suppose dans cette question $\lambda = 0,5$ et $n = 100$. Estimer à l'aide du théorème central limite la probabilité

   $$\mathbf{P}(1,9 \leqslant \overline{X_{100}} \leqslant 2,1).$$

   D'après le TCL, pour n assez grand (100>30), $\frac{\overline{X}_n - 1/\lambda}{\frac{1}{\lambda\sqrt{n}}}$ suit à peu près la loi normale centrée réduite. Ainsi, en remplaçant λ et n par leurs valeurs, on obtient

   $$\mathbf{P}(1,9 \leqslant \overline{X_{100}} \leqslant 2,1) = \mathbf{P}(-0,5 \leqslant \frac{\overline{X}_n - 2}{0,2} \leqslant 0,5) \approx 2F_{\mathcal{N}}(0,5) - 1 \approx 0,4.$$

3. Déterminer des nombres α_n et β_n (ne dépendant que de n) tels que

   $$\mathbf{P}(\frac{\alpha_n}{\lambda} \leqslant \overline{X}_n \leqslant \frac{\beta_n}{\lambda}) \approx 0,99.$$

   Reprenons le TCL utilisé précédemment : pour n assez grand, $\sqrt{n}(\lambda\overline{X}_n - 1) \sim \mathcal{N}(0,1)$ à peu près. On souhaite obtenir un encadrement de $\overline{X}_n$ à 99%. Partons d'un intervalle à 99% pour la loi normale. D'après les tables, $F_{\mathcal{N}}(2,57) \approx 0,995$. Alors, en utilisant la TCL, on peut écrire

$$\mathbf{P}(-2,57 \leqslant \sqrt{n}(\lambda\overline{X}_n - 1) \leqslant 2,57) \approx 0,99.$$

Donc

$$\mathbf{P}(\frac{1-2,57/\sqrt{n}}{\lambda} \leqslant \overline{X}_n \leqslant \frac{1+2,57/\sqrt{n}}{\lambda}) \approx 0,99.$$

On obtient bien un encadrement sous la forme demandée avec $\alpha_n = 1 - 2,57/\sqrt{n}$ et $\beta_n = 1 + 2,57/\sqrt{n}$.

Cet encadrement permet de donner une estimation de λ. À partir de n tirages X_k obtenus, on calcule leur moyenne $\overline{X}_n$ et on estime au risque de 1% que λ est dans l'intervalle $[\frac{\alpha_n}{\overline{X}_n}, \frac{\beta_n}{\overline{X}_n}]$ appelé intervalle de confiance.

4. On a effectué 100 tirages et on a obtenu $\overline{X}_{100} = 2,23$. Donner l'intervalle de confiance correspondant pour λ.

$\frac{\alpha_{100}}{\overline{X}_{100}} = \frac{1-0,257}{2,23} \approx 0,33$ et $\frac{\beta_{100}}{\overline{X}_{100}} = \frac{1+0,257}{2,23} \approx 0,56$. Ainsi, au risque de 1%, $\lambda \in [0,33; 0,56]$.

5. On souhaite obtenir une estimation de λ avec une précision de 0,05. On sait seulement que $\lambda < 0,8$.
D'après la question 3, à partir de quelle valeur de n peut-on raisonnablement considérer que $\overline{X}_n$ est supérieur à 1.
Proposer alors une valeur de n pour laquelle l'intervalle de confiance de λ au risque de 1% est de longueur inférieure à 0,05.

On a obtenu à la question 3 un encadrement à 99% de $\overline{X}_n$. On peut raisonnablement espérer que le tirage $\overline{X}_n$ qu'on obtiendra sera bien dans cet intervalle. Pour $n > 166$, on a $\frac{\alpha_n}{\lambda} > \frac{1-2,57/\sqrt{166}}{0,8} > 1$. Ainsi, pour $n > 166$, on peut considérer que $\overline{X}_n$ est supérieur à 1.

ON souhaite trouver n tel que l'intervalle $[\frac{\alpha_n}{\overline{X}_n}, \frac{\beta_n}{\overline{X}_n}]$ soit de longueur inférieure à 0,05. Si $n > 166$, il est de longueur $\frac{\beta_n - \alpha_n}{\overline{X}_n} = \frac{2\times 2,57/\sqrt{n}}{\overline{X}_n} < \frac{5,14}{\sqrt{n}}$. Pour $n > 10600$, il est de longueur inférieure à 0,05.