Exercice 1 : deux questions indépendantes

1. On considère les événements A : «l'étudiant **connaît** la réponse à la question posée» et J : «l'étudiant donne la **bonne réponse** à la question posée».

   L'énoncé demande de calculer la probabilité conditionnelle : $P_J(A)$.

   On dispose de $P(A) = q$ d'où $P(\overline{A}) = 1 - P(A) = 1 - q$.

   Il est clair que $P_A(J) = 1$ car si l'étudiant connaît la réponse, il donnera la bonne réponse. Mais quand l'étudiant ne connaît pas la réponse, il choisit au hasard l'une des n réponses proposées et $P_{\overline{A}}(J) = \dfrac{1}{n}$.

   Par définition, $P_J(A) = \dfrac{P(A \cap J)}{P(J)}$

   Or d'après la formule des probabilités composées, $P(A \cap J) = P(A)\,P_A(J)$ et d'après la formule des probabilités totales,

   $P(J) = P(A \cap J) + P(\overline{A} \cap J) = P(A)\,P_A(J) + P(\overline{A})\,P_{\overline{A}}(J)$

   Donc $P_J(A) = \dfrac{P(A)\,P_A(J)}{P(A)\,P_A(J) + P(\overline{A})\,P_{\overline{A}}(J)} = \dfrac{q \times 1}{q \times 1 + (1-q)\frac{1}{n}}$

   Finalement $\boxed{P_J(A) = \dfrac{n\,q}{1 + (n-1)q}}$

2. Un QCM d'examen comporte 10 questions indépendantes. 1600 étudiants passent cet examen. Chaque étudiant a 1 chance sur 3 de répondre correctement à chacune des 10 questions posées et répond à toutes les questions.

   Soit X la variable aléatoire égale au nombre de bonnes réponses obtenues pour un étudiant à ce QCM.

   (a) X suit la loi binomiale de paramètres $n_1 = 10$ et $p = 1/3$.

   (b) Posons $n = 1600$. La moyenne des notes obtenues par l'ensemble des 1600 étudiants est

   $\overline{X} = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k$ où X_k est la variable aléatoire qui représente le nombre de bonnes réponses obtenues par le $k-$ème étudiant.

   Les variables aléatoires suivent la même loi de probabilité, on considère qu'elles sont mutuellement indépendantes. Elles admettent pour espérance $\mu = E(X_1) = 10/3$ et pour variance $\sigma^2 = n_1\,p\,(1-p) = 20/9$.

   Donc, d'après le théorème central limite, $\overline{X}$ suit approximativement la loi normale $\mathcal{N}\left(\mu\,,\,\dfrac{\sigma}{\sqrt{n}}\right)$.

   Soit U une variable aléatoire de loi normale centrée réduite $\mathcal{N}(0\,,\,1)$.

   $5\% = P(\overline{X} \leqslant a) = P\left(\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leqslant \dfrac{a-\mu}{\sigma/\sqrt{n}}\right) \approx P\left(U \leqslant \dfrac{a-\mu}{\sigma/\sqrt{n}}\right) = \Phi\left(\dfrac{a-\mu}{\sigma/\sqrt{n}}\right)$

   On est donc amené à résoudre l'équation d'inconnue a :

   $\Phi\left(\dfrac{a-\mu}{\sigma/\sqrt{n}}\right) = 0,05 \iff \dfrac{a-\mu}{\sigma/\sqrt{n}} = \Phi^{-1}(0,05)$

   $\iff a = \mu + \dfrac{\sigma}{\sqrt{n}}\Phi^{-1}(0,05) \iff a = \dfrac{10}{3} + \dfrac{\sqrt{5}}{60}\Phi^{-1}(0,05)$

   Or $\Phi^{-1}(0,05) \approx -1,6449$. Par conséquent $a \approx 3,272$

Exercice 2 : indépendance de variables aléatoires

1. Chaque variable aléatoire X_k suit la loi uniforme sur l'intervalle $[0\,,\,1]$. Sa fonction de répartition F est définie par :

   $$F(x) = \begin{cases} 0 & \text{si} \quad x < 0 \\ x & \text{si} \quad 0 \leqslant x \leqslant 1 \\ 1 & \text{si} \quad x > 1 \end{cases}$$

2. (a) Soit x un nombre réel.

   $Y_n \leqslant x \iff \max\{X_1\,,\,X_2\,,\,\ldots\,,\,X_n\} \leqslant x \iff \forall\,k \in [\![1\,,\,n]\!],\ X_k \leqslant x$.

   Donc $[Y_n \leqslant x] = \displaystyle\bigcap_{k=1}^{n}[X_k \leqslant x]$

   (b) Soit x un nombre réel.

   $G_n(x) = P([Y_n \leqslant x]) = P\left(\displaystyle\bigcap_{k=1}^{n}[X_k \leqslant x]\right) = \displaystyle\prod_{k=1}^{n} P([X_k \leqslant x])$ car les variables aléatoires X_k sont mutuellement indépendantes.

   D'où $G_n(x) = \displaystyle\prod_{k=1}^{n} F(x) = F(x)^n = \begin{cases} 0 & \text{si} \quad x < 0 \\ x^n & \text{si} \quad 0 \leqslant x \leqslant 1 \\ 1 & \text{si} \quad x > 1 \end{cases}$

3. La fonction de répartition G_n de Y_n est continue sur $\mathbb{R}$, de classe $\mathcal{C}^1$ sur une partie de $\mathbb{R}$ contenant $\mathbb{R} \setminus \{0\,,\,1\}$ (en fait G_n est de classe $\mathcal{C}^1$ sur $\mathbb{R} \setminus \{1\}$ dès que $n \geqslant 2$). Donc Y_n est une variable aléatoire à densité, dont une densité g_n vérifie :

   $$\forall\,t \in \mathbb{R} \setminus \{0\,,\,1\}\,,\quad g_n(t) = G_n'(t) = \begin{cases} 0 & \text{si} \quad t < 0 \\ n\,t^{n-1} & \text{si} \quad 0 < t < 1 \\ 0 & \text{si} \quad t > 1 \end{cases}$$

On choisit de prendre $g_n(0) = 0$ et $g_n(1) = n$.

4. Comme la densité g_n est nulle en dehors du segment $[0\,,\,1]$, Y_n admet une espérance mathématique égale à :

$$E(Y_n) = \int_0^1 t\,g_n(t)\,\mathrm{d}t = n\int_0^1 t^n\,\mathrm{d}t = n\left[\frac{t^{n+1}}{n+1}\right]_0^1 = n\frac{1^{n+1}}{n+1}$$

$$\text{Donc} \quad \boxed{E(Y_n) = \frac{n}{n+1}}$$

5. Comme la densité g_n est nulle en dehors du segment $[0\,,\,1]$, Y_n admet un moment d'ordre 2 et une variance. D'après la formule de König-Huyghens,

$$V(Y_n) = E(Y_n^2) - E(Y_n)^2 = \int_0^1 t^2\,g_n(t)\,\mathrm{d}t - E(Y_n)^2$$

$$= n\int_0^1 t^{n+1}\,\mathrm{d}t - E(Y_n)^2 = n\left[\frac{t^{n+2}}{n+2}\right]_0^1 - E(Y_n)^2$$

$$= \frac{n}{n+2} - \frac{n^2}{(n+1)^2} = \frac{n(n+1)^2 - n^2(n+2)}{(n+2)(n+1)^2} = \frac{n[(n+1)^2 - n(n+2)]}{(n+2)(n+1)^2}$$

Ainsi $V(Y_n) = \dfrac{n[n^2+2n+1-n^2-2n)]}{(n+2)(n+1)^2} = \dfrac{n}{(n+1)^2(n+2)}$.

Exercice 3 : estimation par intervalle de confiance

$(X_n)_{n\in\mathbb{N}^*}$ est une suite de variables aléatoires mutuellement indépendantes qui suivent toutes la loi exponentielle de paramètre $\lambda > 0$.

On pose pour tout entier $n \geqslant 3$, $S_n = \displaystyle\sum_{k=1}^{n} X_k$; $J_n = \lambda S_n$ et $T_n = \dfrac{n}{S_n}$.

1. On a $E(S_n) = \displaystyle\sum_{k=1}^{n} E(X_k) = n/\lambda$, et $V(S_n) = \displaystyle\sum_{k=1}^{n} V(X_k) = n/\lambda^2$ par indépendance.

   $E(J_n) = \lambda E(S_n) = n$ et $V(J_n) = \lambda^2 V(S_n) = n$

2. (a) T_n est une fonction du $n-$échantillon. Donc c'est un estimateur de λ.

   $$T_n = \lambda\frac{n}{\lambda S_n} = \lambda n\frac{1}{J_n} \text{ donc } E(T_n) = \lambda n\,E\left(\frac{1}{J_n}\right) = \lambda n\frac{1}{n-1} \neq \lambda$$

   Conclusion : $\boxed{T_n \text{ est biaisé.}}$

   Mais $E(T_n) = \dfrac{n}{n-1}\lambda = \dfrac{1}{1-1/n}\lambda \underset{(n\to+\infty)}{\longrightarrow} \lambda$

   Donc T_n est asymptotiquement sans biais.

   (b) Le risque quadratique associé à T_n est

   $r(T_n) = V(T_n) + b(T_n)^2 = V(T_n) + (E(T_n) - \lambda)^2$

   Or $V(T_n) = V\left(\lambda\,n\dfrac{1}{J_n}\right) = \lambda^2 n^2\,V\left(\dfrac{1}{J_n}\right)$

   avec

   $$V\left(\frac{1}{J_n}\right) = E\left(\frac{1}{J_n^2}\right) - E\left(\frac{1}{J_n}\right)^2 = \frac{1}{(n-1)(n-2)} - \frac{1}{(n-1)^2}$$
   $$= \frac{1}{(n-1)^2(n-2)}$$
   $$V(T_n) = \frac{\lambda^2 n^2}{(n-1)^2(n-2)}$$

   Donc le risque quadratique est

   $$r(T_n) = V(T_n) + \left(\frac{n}{n-1}\lambda - \lambda\right)^2$$
   $$= \frac{\lambda^2 n^2}{(n-1)^2(n-2)} + \left(\frac{\lambda}{n-1}\right)^2$$
   $$= \frac{\lambda^2(n^2+n-2)}{(n-1)^2(n-2)} = \frac{\lambda^2(n-1)(n+2)}{(n-1)^2(n-2)}$$
   $$= \frac{\lambda^2(n+2)}{(n-1)(n-2)}$$

   (c) $r(T_n) = \dfrac{n+2}{n^2-3n+2}\lambda^2 \underset{(n\to+\infty)}{\sim} \dfrac{n}{n^2}\lambda^2 = \dfrac{\lambda^2}{n}$ avec $\displaystyle\lim_{n\to+\infty}\frac{\lambda^2}{n} = 0$.

   Par conséquent $\boxed{\displaystyle\lim_{n\to+\infty} r(T_n) = 0}$ le risque quadratique de T_n tend vers 0 quand n tend vers $+\infty$.

3. (a) Selon le théorème de la limite centrée, étant donnée une suite de variables aléatoires (X_n) mutuellement indépendantes, de même loi et de variance non nulle, la somme centrée réduite des n premiers termes converge en loi vers une variable aléatoire U de loi normale $\mathcal{N}(0,1)$.

   C'est-à-dire que la fonction de répartition de la somme centrée réduite tend vers Φ.

   On a vu que $S_n = \displaystyle\sum_{k=1}^{n} X_k$ avait pour espérance $\dfrac{n}{\lambda}$ et pour variance $\dfrac{n}{\lambda^2}$

Les $(X_k)_{k\in N^*}$ sont indépendantes et ont une variance non nulle. Donc la centrée réduite $\dfrac{S_n - \frac{n}{\lambda}}{\frac{\sqrt{n}}{\lambda}} = \dfrac{\lambda S_n - n}{\sqrt{n}} = \dfrac{\lambda S_n}{\sqrt{n}} - \sqrt{n} = S_n^*$ converge en loi vers U de loi normale centrée réduite.

(b) Donc, pour n assez grand (en général $n > 50$),

$$\begin{aligned} P\left(-t_\alpha \leqslant S_n^* \leqslant t_\alpha\right) &\simeq P(-t_\alpha \leqslant U \leqslant t_\alpha) \\ &= \Phi\left(t_\alpha\right) - \Phi\left(-t_\alpha\right) \\ &= \Phi\left(t_\alpha\right) - \left[1 - \Phi\left(t_\alpha\right)\right] \\ &= 2\Phi\left(t_\alpha\right) - 1 \\ &= 2\left(1 - \frac{\alpha}{2}\right) - 1 = 1 - \alpha \end{aligned}$$

(c) On résout : $\lambda \in \left[\left(1 - \dfrac{t_\alpha}{\sqrt{n}}\right) T_n, \left(1 + \dfrac{t_\alpha}{\sqrt{n}}\right) T_n\right]$

$\Longleftrightarrow \left(1 - \dfrac{t_\alpha}{\sqrt{n}}\right) \dfrac{n}{S_n} \leqslant \lambda \leqslant \left(1 + \dfrac{t_\alpha}{\sqrt{n}}\right) \dfrac{n}{S_n}$

$\Longleftrightarrow n\left(1 - \dfrac{t_\alpha}{\sqrt{n}}\right) \leqslant \lambda S_n \leqslant n\left(1 + \dfrac{t_\alpha}{\sqrt{n}}\right)$

$\Longleftrightarrow \sqrt{n}\left(1 - \dfrac{t_\alpha}{\sqrt{n}}\right) \leqslant \dfrac{\lambda S_n}{\sqrt{n}} \leqslant \sqrt{n}\left(1 + \dfrac{t_\alpha}{\sqrt{n}}\right)$

$\Longleftrightarrow \sqrt{n} - t_\alpha \leqslant \dfrac{\lambda S_n}{\sqrt{n}} \leqslant \sqrt{n} + t_\alpha$

$\Longleftrightarrow -t_\alpha \leqslant \dfrac{\lambda S_n}{\sqrt{n}} - \sqrt{n} \leqslant t_\alpha$

$\Longleftrightarrow -t_\alpha \leqslant S_n^* \leqslant t_\alpha$

Donc $P\left(\lambda \in \left[\left(1 - \dfrac{t_\alpha}{\sqrt{n}}\right) T_n\,,\, \left(1 + \dfrac{t_\alpha}{\sqrt{n}}\right) T_n\right]\right) = P\left(-t_\alpha \leqslant S_n^* \leqslant t_\alpha\right)$ $\simeq 1 - \alpha$

Ainsi $\left[\left(1 - \dfrac{t_\alpha}{\sqrt{n}}\right) T_n \;\; \left(1 + \dfrac{t_\alpha}{\sqrt{n}}\right) T_n\right]$ est un intervalle de confiance **aléatoire** pour λ au niveau de risque α quand n est grand.

4. La commande `sum(grand(1,100,"'exp"',1/lambda)` renvoie 14.646

(a) $t_\alpha = \Phi^{-1}\left(1 - \dfrac{\alpha}{2}\right) = \Phi^{-1}(0.975)$.

Avec la calculatrice ou la commande `cdfnor("X",0,1,1-0.05/2,0.05/2)` sur Scilab, on obtient : $t_\alpha \approx 1.959964 \approx 1.96$

(b) Avec l'échantillon observé où $n = 100$, une réalisation de T_n est :

$T_n(\omega) = \dfrac{n}{\sum_{k=1}^n x_k} = \dfrac{100}{14.646}$

Donc un intervalle de confiance **observé** pour λ au niveau de confiance 95% est :

$$I_{95\%,obs} = \left[\left(1 - \frac{t_\alpha}{10}\right) \frac{100}{14.646}\,,\, \left(1 + \frac{t_\alpha}{10}\right) \frac{100}{14.646}\right] = [5.489\,,\, 8.166]$$