Travail 2 : Partie simulations.
À déposer sur le portail de cours le 30 octobre avant 19h00

Problème 1 : Méthode de rejet pour la loi normal

Nous avons établi dans le cours que la méthode de rejet permet de simuler des données de la loi normale $\mathcal{N}(0,1)$ selon l'algorithme suivant :

Algorithme

a) Générer U_1, U_2 et U_3 uniformément sur $[0,1]$
b) $Y = -\ln(U_1)$
c) Si $U_2 \leq \exp\left[-(Y-1)^2/2\right]$ aller à $d)$ et $e)$, sinon aller à $a)$.
d) $|Z| = Y$
e) $Z = Y$ si $U_3 \leq 1/2$ sinon $X = -Y$.

Remarquons que nous pouvons simuler des données de $X \sim \mathcal{N}(\mu, \sigma^2)$ à partir de $Z \sim \mathcal{N}(0,1)$ en utilisant la relation :

$$X = \sigma Z + \mu.$$

Questions :

1. Simulez à partir de cet algorithme 10000 observations. Représentez ces données par un histogramme. Comparer ce dernier avec la densité théorique.
2. On veut utiliser cette algorithme pour estimer les moments d'une variable aléatoire de loi normale $X \sim \mathcal{N}(\mu = 24, \sigma^2 = 4)$. On sait que les deux premiers moments sont :

$$E(X) = \mu = 24 \quad \text{et} \quad E(X^2) = \mu^2 + \sigma^2 = 580.$$

   Utiliser le fait que

   (a) Utilisez l'algorithme précédent pour approcher $E(X)$ et $E(X^2)$ par simulation (avec $k = 100$ répétitions). Comparez avec les valeurs exactes. Indication : rappelons que

$$E(X^k) \approx \frac{1}{N} \sum_{i=1}^{N} X_i^k \quad k = 1, 2, \ldots$$

   où $X_1, \ldots, X_N$ des valeurs simulées à partir de la normale $\mathcal{N}\mathcal{N}(\mu = 24, \sigma^2 = 4)$.

   (b) Utilisez l'algorithme précédent pour approcher $E(X^5)$ par simulation (avec $k = 100$ répétitions).

Problème 2 : Méthode de rejet pour la loi Beta

Rappelons que la densité de la loi uniforme sur $[0,1]$ est :

$$g(x) = \begin{cases} 1 & \text{si} \quad 0 \leq x \leq 1, \\ 0 & \text{sinon.} \end{cases}$$

On veut utiliser simuler des données d'une variable aléatoire de loi beta dont la densité est :

$$f(x) = \begin{cases} 1260x^4(1-x)^5 & \text{si} \quad 0 \leq x \leq 1, \\ \\ 0 & \text{sinon.} \end{cases}$$

Ceci peut faire en utilisant la méthode de rejet basée sur la densité g pour établir un algorithme permettant de générer des données issues de la densité f.

1. Écrire en détail un algorithme permettant de simuler des données à partir de la variable aléatoire X de densité $f(x)$.
2. Simulez à partir de cet algorithme 10000 observations. Représentez ces données par un histogramme. Comparer ce dernier avec la densité théorique (Représentez ces deux quantités dans le même graphe).
3. En utilisant la simulation précédente, estimez $\mathbb{P}(0 \leq X \leq 0.70)$ avec une taille d'échantillon $N = 1000$ et $k = 100$ répétitions. Indication utiliser le fait que :

$$\mathbb{P}(0 \leq X \leq 0.70) = \frac{\text{card}\big\{i \in 1, \dots, N\big\} : 0 \leq X_i \leq 0.70\big\}}{N}$$

   où $X_1, \dots, X_N$ un échantillon simulé à partir de la loi de X.
4. En écrivant $\mathbb{P}(0 \leq X \leq 0.45)$ sous forme d'intégrale à savoir :

$$\mathbb{P}(0 \leq X \leq 0.70) = \int_0^{0.70} f(x)dx.$$

   Estimez $\mathbb{P}(0 \leq X \leq 0.70)$ en utilisant la méthode de Monte-Carlo vue dans le cours $N = 1000$ et $k = 100$ répétitions.
5. Estimez cette probabilité en utilisant la méthode du Trapèze rappelée dans le cours.
6. Finalement, calculez exactement cette probabilité. Comparez les résultats des méthodes précédentes à la valeur exacte de cette probabilité. Quelle est votre conclusion ?

Problème 3 : Calcul d'intégrale impropre

La méthode de Monte Carlo peut être utilisée pour approximer une intégrale impropre (c-à-d, une intégrale dont au moins une de ces bornes est infinie). Pour ce faire, considérons l'intégrale impropre suivant

$$I = \int_0^{\infty} f(x)dx < \infty.$$

L'idée principale consiste à écrire l'intégrale I sous forme d'une espérance mathématique d'une variable aléatoire X à valeur sur l'intervalle $[0, \infty[$. Ainsi, si g désigne la densité de X, alors on peut déduire que

$$I = \int_0^{\infty} \frac{f(x)}{g(x)} g(x)dx = E\left[\frac{f(X)}{g(X)}\right].$$

Par conséquent :

$$I \approx \frac{1}{N}\sum_{i=1}^{N}\frac{f(X_i)}{g(X_i)}$$

où $X_1, \ldots, X_N$ un échantillon aléatoire généré à partir de la densité g. Pour illustrer cette méthode, considérons la fameuse intégrale de Gauss :

$$I = \int_0^{\infty} e^{-x^2} dx.$$

Gauss a montré que cette l'intégrale $I = \sqrt{\pi}/2 = 0.8862269$. Notre objectif est d'évaluer cette intégrale par la méthode de Monte Carlo. Ceci peut être réalisé comme suit :

1. Montrer
$$I = E\left[e^{-X(X-1)}\right]$$
où X une varaible aléatoire de loi exponentielle de paramètre $\lambda = 1$. Rappelons que la densité d'une loi exponentielle de paramètre $\lambda = 1$ est :
$$g(x) = \begin{cases} e^{-x} & \text{si} \quad x \geq 0, \\ 0 & \text{sinon.} \end{cases}$$
2. En déduire que
$$I \approx \frac{1}{N}\sum_{i=1}^{N} e^{-\ln(U_i)(\ln(U_i)+1)}$$
où $U_1, \ldots, U_N$ un échantillon généré de la loi uniforme $[0, 1]$.
3. En prenant $N = 1000$ avec un nombre de répitition $k = 100$, calculer à l'aide de la formule précédente une approximation de I. Comparez avec la valeur exacte de I.
4. Donner une approximation de I en utilisant la formule de trapèze vu au cours en prenant le nombre de subdivision $n = 1000$ et $a = 0$ et $b = 2000$. Comparer avec la valeur exacte de I.
5. Comparez les approximations obtenues dans les questions 3) et 4).
6. En applicant la même démarche utilisée aux points 1), 2) et 3), calculer par la méthode de Monte Carlo les intégrales suivantes :

   6.1) : $J = \int_0^{\infty} e^{-x^3} dx$.
   6.2) : $K = \int_0^{\pi} \sin(x^2) dx$.

Problème 4 : Calcul d'intégrale double

La méthode de Monte Carlo peut être utilisée pour approximer une intégrale double :

$$D = \int_a^b \int_c^d f(x, y) dx dy.$$

On peut montrer que (vous pouvez ajouter la preuve comme question facultative)

$$D = (b-a)(d-c)E\big[f\big(a+(b-a)U, c+(d-c)V\big)\big].$$

avec $U \sim \mathcal{U}_{[0,1]}$, $V \sim \mathcal{U}_{[0,1]}$ et U, V indépendantes. Donc d'après la loi forte des grands nombres on peut approcher D par :

$$D \approx \frac{(b-a)(d-c)}{N} \sum_{i=1}^{N} f\big(a+(b-a)U_i, c+(d-c)V_i\big)$$

où les échantillons $U_1, \ldots, U_N$ et $V_1, \ldots, V_N$ sont indépendamment générés de $\mathcal{U}_{[0,1]}$.

Question 1 : Pour expérimenter cette méthode, on vous propose d'approcher l'intégrale suivant en utilisant les approximations ci-dessus (choisir $N = 1000$ et $k = 100$) :

$$D = \int_0^1 \int_0^1 (x^2y + xy^2)dxdy.$$

Comparez ces approximations avec la valeur exacte de D. Conclure.

Question 2 :

En s'inspirant de ce qui précède. Donner une méthode d'approximation basée sur les simulations Monte-Carlo pour approcher une intégrale double avec borne dépendantes :

$$D = \int_0^1 \int_0^x f(x,y)dxdy.$$

Expérimentez cette méthode sur l'intégrale :

$$D = \int_0^1 \int_0^y (x^2y + xy^2)dxdy.$$

Comparer avec la valeur exacte de cette intégrale.

Problème 5 : Estimation des probabilités stationnaires

Soit $\mathcal{S} = \{1, 2, 3, 4\}$ l'espace des états d'une chaîne de Markov de matrice de transition :

$$\mathcal{P} = \begin{pmatrix} 2/7 & 0 & 2/7 & 3/7 \\ 1/5 & 2/5 & 0 & 2/5 \\ 1/6 & 1/3 & 1/3 & 1/6 \\ 4/10 & 3/10 & 0 & 3/10 \end{pmatrix}.$$

1. Rappelons que la loi stationnaire $\pi = (\pi_1, \pi_2, \pi_3, \pi_4)$ est déterminée en résolvant le système d'équations :

$$\pi = \pi\mathcal{P} \quad \text{et} \quad \pi_1 + \pi_2 + \pi_3 + \pi_4 = 1.$$

2. Monter que le système précédent est équivalent à

$$\begin{aligned}
\frac{2}{7}\pi_1 + \frac{1}{5}\pi_2 + \frac{1}{6}\pi_3 + \frac{4}{10}\pi_4 &= \pi_1 \\
\frac{2}{5}\pi_2 + \frac{1}{3}\pi_3 + \frac{3}{10}\pi_4 &= \pi_2 \\
\frac{2}{7}\pi_1 + \frac{1}{3}\pi_3 &= \pi_3 \\
\frac{3}{7}\pi_1 + \frac{2}{5}\pi_2 + \frac{1}{6}\pi_3 + \frac{3}{10}\pi_4 &= \pi_4 \\
\pi_1 + \pi_2 + \pi_3 + \pi_4 &= 1.
\end{aligned}$$

3. En résolvant le système d'équations précédent, trouver les valeurs des probabilités stationnaires π_1, π_2, π_3 et π_4. Vous pouvez résoudre ces équation avec un logiciel.
4. Maintenant notre objectif est d'estimer par simulation la loi stationnaire précédente. Pour cela, on suppose que la loi initiale de la chaîne est uniformément distribuée sur $\mathcal{S} = \{1, 2, 3, 4\}$, c'est-à-dire :

$$P(X_0 = i) = 1/4, \quad i = 1, 2, 3, 4.$$

— Étape 1 : on simule N fois les 20 premiers états de la chaîne $\{X_n, n \geq 0\}$ (prenez, par exemple, $N = 10000$).

— Étape 2 : on stocke à chaque fois le vingtième état simulé.

— Étape 3 : on fait ensuite un tableau de fréquence des 10000 états obtenus à l'étape précédente.

— Étape 4 : si n_i désigne la fréquence de l'état i obtenu à l'étape 3, alors la probabilité stationnaire π_i sera estimée par :

$$\hat{\pi}_i = \frac{n_i}{N}.$$

— Étape 5 : pour raffiner cette estimation, on répète cette procédure plusieurs fois ($k = 100$ fois par exemple), puis on prend la moyenne des $\hat{\pi}_i$ obtenues comme estimation de π_i.

— Comparez les probabilités estimées $\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3$ et $\hat{\pi}_4$ aux probabilités théoriques π_1, π_2, π_3 et π_4.

6. Estimation du temps de retour

Soit $\mathcal{S} = \{1, 2, 3\}$ l'espace des états d'une chaîne de Markov de matrice de transition :

$$\mathcal{P} = \begin{pmatrix} 1 & 0 & 0 \\ 1/3 & 1/6 & 1/2 \\ 2/5 & 2/5 & 1/5 \end{pmatrix}.$$

1. Expliquez pourquoi les états 2 et 3 de cette chaîne de Markov sont transitoires.
2. Rappelons que le nombre moyen de visite d'un état i est donné par :

$$E(N_i|X_0 = i) = \sum_{k=0}^{\infty} P_{ii}^{(k)} \quad \text{avec} \quad P_{ii}^{(0)} = 1.$$

   En calculant les matrices $\mathcal{P}^k$, $k = 1, 2, 3, \ldots,$. Évaluez avec une précision de 10^{-5}
   — la valeur de $E(N_i|X_0 = i)$ pour l'état 2,
   — la valeur de $E(N_i|X_0 = i)$ pour l'état 3.
3. Soit f_i la probabilité de retour à l'état i sachant que la chaîne a débuté de i. Dans le cas où i est transitoire, cette probabilité est liée à $E(N_i|X_0 = i)$ par l'équation :

$$E(N_i|X_0 = i) = \frac{1}{1 - f_i}.$$

   Évaluez avec une précision de 10^{-5}
   — la valeur de f_i pour l'état 2,
   — la valeur de f_i pour l'état 3.