Statistiques TP 4
Méthode de MonteCarlo, intervalles de confiance et tests

Rappel : taper `krdc vnc://nom_du_serveur` dans un terminal.
Les sujets et les corrigés des TPs sont mis le lendemain des séances sur ma page :
http://www.math.u-bordeaux1.fr/~chabanol/stat.html

1. Intervalle de confiance

(1) *Écrire une fonction* `Intconf` *qui prend en paramètre un échantillon X, un écart-type exact ou estimé s et un réel a et fournit un intervalle de confiance (a priori asymptotique) pour l'espérance $E[X]$ de risque a (on peut utiliser la fonction* `qnorm`*)*

2. Méthode de Monte Carlo

Soit $(U_1, \dots, U_d)$ un vecteur de d variables aléatoires indépendantes toutes de loi uniforme sur $[0,1]$, et f une fonction de carré intégrable sur $[0,1]^d$. On considère la variable aléatoire $X = f(U_1, \dots, U_d)$. D'après la loi forte des grands nombres, si X_n est une suite de variables aléatoires indépendantes toutes de même loi que X, $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n X_i = E(f(U_1, \dots, U_d)) = \int_{[0,1]^d} f(u_1, \dots, u_d) du_1 \dots du_d$. La méthode de Monte Carlo consiste à utiliser ceci pour calculer une approximation de $\int_{[0,1]^d} f(u_1, \dots, u_d) du_1 \dots du_d$. Le thérorème central limite permet de plus d'avoir une idée de l'ordre de grandeur de l'erreur : la convergence est en $\frac{1}{\sqrt{n}}$, indépendamment de d, et on peut obtenir un intervalle de confiance asymptotique pour l'intégrale, (au besoin en estimant σ).
L'intérêt par rapport à d'autres méthodes déterministes, qui approchent l'intégrale par exemple par des sommes de Riemann, se fait donc surtout sentir pour des intégrales multiples. En effet l'erreur avec ce genre de méthodes est en général en (Pas de la subdivision)$^{-k}$, où k est fixé et dépend de la méthode; donc pour une précision donnée, plus la dimension d de l'espace sur lequel on intègre est grand, plus il faudra de points.

On va se servir de la méthode de Monte-Carlo pour calculer une approximation de π. On suppose donc bien sûr dans les questions qui suivent que π est inconnu.

(2) On va d'abord utiliser $\pi = \int_0^1 4\sqrt{1-x^2}dx$. *Générer un* 500 *échantillon* $[U_1, \dots, U_{500}]$ *de loi uniforme sur* $[0,1]$. *Obtenir l'échantillon* $[X_1, \dots, X_{500}]$ *où* $X_i = 4\sqrt{1-U_i^2}$, *puis donner une estimation de* $\pi = E[X]$. *Obtenir une estimation de l'écart-type de* X. *Donner un intervalle de confiance pour* π *au risque 0.05.*

(3) *Recommencer l'opération précédente 10 fois. Combien de fois votre intervalle contient-il π ? Faire une boucle qui répète la question précédente 100 fois et compte le nombre de fois où π est dans votre intervalle. Quelle est la loi (théorique) de ce nombre ? Ce que vous observez est-il conforme à cette loi ?*

(4) On va déterminer un intervalle de confiance pour π en utilisant $\pi = 4\int_{[0,1]^2} 1_{x^2+y^2\leq 1} dxdy$. Si (U_1, U_2) suit une loi uniforme sur $[0,1]^2$, quelle est la loi de $X = 4 1_{U_1^2+U_2^2\leq 1}$? Donner son espérance et son écart-type en fonction de π.

Obtenir de même que précédemment une estimation de π et de l'écart-type de X_i, et donner un intervalle de confiance pour π au risque 0.05.

(5) *(Question subsidiaire) Déterminer de même un intervalle de confiance pour π en utilisant* $\pi = 6\int_{[0,1]^3} 1_{x^2+y^2+z^2\leq 1} dxdydz$.

(6) *(Question subsidiaire) Ecrire* $J = \int_{S^2} x^2 dxdydz$ *où* S_2 *désigne la boule unité de* $\mathbb{R}^3$ *comme l'espérance d'une variable aléatoire, et l'estimer par MonteCarlo, puis obtenir un intervalle de confiance. Votre intervalle contient-il la vraie valeur ?*

3. Test d'ajustement de Kolmogorov Smirnov

Soit $(X_1, \ldots, X_n)$ un n-échantillon de loi inconnue μ. On veut tester H_0: "$\mu = \nu$", contre H_1: "$\mu \neq \nu$", où ν est une loi de probabilité fixée et connue.
On suppose que la fonction de répartition F_ν est continue.
Intuitivement, on peut essayer de voir si la fonction de répartition empirique F_n des X_i est proche de la fonction de répartition F_ν. On sait en effet (`Glivenki Cantelli`) que quand n tend vers $+\infty$, la fonction de répartition empirique converge p.s vers la fonction de répartition de μ, et que cette convergence est uniforme, c'est-à-dire que $\sup_x |F_n(x) - F_\mu(x)|$ tend vers 0. De plus, à n fini la loi ne dépend pas de F_μ, seulement de n, et enfin on connaît la loi limite de $\sqrt{n}\sup_x |F_n(x) - F_\mu(x)|$ lorsque n tend vers l'infini : elle a pour fonction de répartition $F_{KS}(t) = 1 + 2\sum_{k=1}^{\infty}(-1)^k \exp(-2k^2t^2)$. La fonction `pks` de matlab calcule cette fonction de répartition asymptotique.
Donc sous H_0, $D_n = \sqrt{n}\sup_x |F_n(x) - F_\nu(x)|$ suit une loi qui ne dépend que de n, et dont l'asymptotique est connue. Sous H_1, D_n tend vers $+\infty$. Le test d'ajustement de Kolmogorov Smirnov consiste à prendre une région critique de la forme $\{D_n > K_\alpha\}$.

(1) On note $(X_{(1)}, \ldots, X_{(n)})$ le n-échantillon X_i réordonné dans l'ordre croissant. Le calcul de D_n peut se faire simplement en remarquant que les fonctions de répartition sont croissantes, ainsi :
$\forall x \in [X_{(i)}, X_{(i+1)}[,\ \frac{i}{n} - F_\nu(X_{(i+1)}) \leq F_n(x) - F_\nu(x) \leq \frac{i}{n} - F_\nu(X_{(i)})$.
Finalement, $D_n = \sqrt{n}\max_{1\leq i\leq n}(\max(\frac{i-1}{n} - F_\nu(X_{(i)}), F_\nu(X_{(i)}) - \frac{i}{n}))$.

Écrire une fonction `FKS` *qui prend en paramètre un échantillon* X*, un paramètre* α*, et le nom d'une fonction* F*, calcule* D_n *et affiche OUI ou NON suivant le résultat du test d'ajustement de* X *à la loi de fonction de répartition* F *au risque* α*, et renvoie également le niveau critique* α_c *auquel on aurait changé de conclusion.*

Tester ainsi le générateur de loi normale de matlab, le générateur de loi uniforme, votre générateur de loi exponentielle.

(2) (Question subsidiaire) *Générer un échantillon de loi normale centrée réduite, et chercher l'ensemble des paramètres* m *pour lesquels votre test de Kolmogorov Smirnov répond oui à "la loi est-elle une loi normale d'espérance* m *et de variance 1" ? Comparer avec l'intervalle de confiance pour* m *que vous obtenez pour votre échantillon (au même risque alpha).*