TP3 : CONVERGENCE DE VARIABLES ALÉATOIRES

Résumé. Ce tp présente une illustration numérique de la convergence d'une suite de variables aléatoires. On s'intéressera tout d'abord à la représentation graphique des théorèmes de convergence classiques. Puis les méthodes de Monte-Carlo seront formalisées et appliquées à l'étude des intervalles de confiance et au calcul d'intégrales.

1. Histogrammes

Soit $(X_1, X_2, \ldots, X_n)$ des variables aléatoires indépendantes de même loi à valeurs dans un ensemble $\mathcal{X}$. Un histogramme permet de regrouper les données par classes. On partitionne l'ensemble $\mathcal{X}$ en classes $I_1, \ldots, I_m$ et on compte le nombre de données dans chaque classe. Pour $k \in \{1, \ldots, m\}$, on note $N_k = \sum_{i=1}^n \mathbb{1}_{X_i \in I_k}$ le nombre de données dans la classe I_k. Remarquer que $\sum_{k=1}^m N_k = n$ (nombre total de données). Ainsi N_k/n représente la fréquence de la classe I_k et est un bon estimateur de $\mathbb{P}(X_1 \in I_k)$.

Matlab fournit ce comptage par la fonction `hist` : si x est le vecteur des données et m le nombre de classes, `[N,C]=hist(x,m)` renvoie la suite `N` des N_k et la suite `C` des centres des classes. On peut aussi entrer en argument le centre des classes (voir `help hist`), ou le bord des classes, avec la fonction `histc`. La représentation en barres de $I_k \mapsto N_k$ est obtenue à l'aide de plusieurs fonctions : `bar`, `stem` ou `hist`. Comme pour `plot`, il existe des options de dessin (voir l'aide). La commande `hist` représente la fonction $h_1 = \sum_{k=1}^m N_k \mathbb{1}_{I_k}$, ce qui n'est en général pas intéressant pour une comparaison avec la loi théorique car on préfère renormaliser les N_k par n.

Dans le cas de variables discrètes à valeurs dans $\mathcal{X} = \{c_1, \ldots, c_p\}$, on choisit $m = p$, $I_k = \{c_k\}$ et on représente la fonction $\sum_{k=1}^m (N_k/n)\mathbb{1}_{I_k}$. On peut ainsi visualiser la convergence de chaque N_k/n vers $\mathbb{P}(X_1 = c_k)$ (loi des grands nombres).

Dans le cas de variables continues à densité f, pour représenter la distribution des variables aléatoires, on utilise plutôt les histogrammes renormalisés :

$$h_2 = \sum_{k=1}^m \frac{N_k}{n|I_k|} \mathbb{1}_{I_k}.$$

En effet, on vérifie que $\int_{\mathcal{X}} h_2(u)du = 1$ (contrairement à $\int h_1$) et

$$h_2(u) \underset{n\to\infty}{\longrightarrow} \sum_{k=1}^m \frac{\int f \mathbb{1}_{I_k}}{|I_k|} \mathbb{1}_{I_k}(u)$$

qui est la meilleure approximation de $f(u)$ (au sens L^2) par des fonctions en escalier sur les I_k. Ainsi h_2 est un estimateur de la densité f des X_i. Avec Matlab, on peut l'obtenir à partir du vecteur `N` : si les intervalles I_k sont tous de même taille, on obtient leur longueur en écrivant `(max(x)-min(x))/m`. Une façon plus simple de procéder et d'utiliser la fonction `histo` de la *Stixbox*. La commande `histo(x,m,odd,scale,'s')` représente les points du vecteur x sous forme d'un histogramme dont les classes ont toute la même largeur. Les paramètres `m`, `odd`, `scale` et `'s'` sont optionnels :

- `x` est un vecteur de réels contenant les données,
- `m` est un entier qui permet de préciser le nombre approximatif de classes (par défaut, `m` est pris égal à $4\,n^{1/4}$ où n est la taille du vecteur `x`),
- `odd` est un nombre qui vaut 0 ou 1 et qui permet de spécifier la position des intervalles (décalage des classes d'une demi largeur vers la droite si `odd` vaut 1 par rapport à `odd` égal à 0),

- `scale` égal à 1 si on veut un histogramme tel que la somme des aires des rectangles soit 1 (ce qui est bon lorsque l'on représente sur un même graphique une densité de probabilité et un histogramme des valeurs observées),
- `'s'` est une chaîne de caractères qui spécifie le type de tracé.

Si on ne veut pas spécifier certains paramètres intermédiaires dans la liste des paramètres d'entrée, on remplace ces paramètres par `[]` (vecteur vide). En particulier, `histo(x,[],[],1)` dessine un histogramme d'aire égale à 1 pour les valeurs de `x`, c'est à dire la fonction h_2.

2. Représentation de la convergence de variables aléatoires

Exercice 1. Soit $X_1, \dots, X_n$ un n-échantillon de loi μ donnée par $\mu(1) = 0.2$, $\mu(2) = 0.3$, $\mu(3) = 0.1$, $\mu(4) = 0.4$. Simuler un tel n-échantillon (avec différentes valeurs de n) et représenter la convergence des fréquences empiriques vers les fréquences théoriques, telle qu'assurée par la loi des grands nombres. On peut utiliser un histogramme pour représenter les fréquences empiriques et une densité en bâtons pour la loi limite. On peut également observer, pour $k = 1, \dots, 4$, la convergence des fonctions $n \mapsto n^{-1} \sum_{i=1}^{n} \mathbb{1}_{X_i=k}$ vers $\mu(k)$.

Exercice 2. Soit $X_1, X_2, \dots, X_n$ un n-échantillon de loi uniforme sur $[-1/2, 1/2]$. Ecrire ce que donne le théorème central limit [TCL] dans ce cas. Renormalisez la statistique apparaissant dans le TCL pour avoir une convergence vers une loi gaussienne standard. Que remarque-t-on lorsque $n = 12$ pour la statistique renormalisée ? Est-ce à dire que pour $n = 12$, cette statistique est gaussienne ? Certes non, mais on dit souvent que l'on peut simuler une gaussienne par 12 uniformes, c'est-à-dire qu'avec $n = 12$, la loi de la statistique renormalisée est déjà très proche d'une loi normale. (Remarquez que le choix $n = 12$ est dicté dans un premier temps pour des raisons de simplifications !) Vérifiez cette affirmation expérimentalement, par exemple par une représentation par histogrammes.
Quelle autre méthode peut-on proposer pour comparer les deux distributions ?

Exercice 3. Soit $X_1, \dots, X_n$ un n-échantillon de loi commune ν (possédant un moment d'ordre 4). On note, pour $i = 1, \dots, 4$,

$$\mu_i = \mathbb{E}\left[(X_1 - \mathbb{E}[X_1])^i\right] .$$

On peut alors montrer la normalité asymptotique de la variance empirique S^2 :

$$\sqrt{n}\left(S^2 - \mu_2\right) \rightsquigarrow \mathcal{N}(0, \mu_4 - \mu_2^2).$$

Supposons que ν soit la loi de Poisson de paramètre λ. Ecrire les formules de normalité asymptotique pour la moyenne empirique $\bar{X}_n$ et la variance empirique S^2 (on admettra ou calculera $\mu_4 = 3\lambda^2 + \lambda$). Remarquez que pour une loi de Poisson, ces deux variables aléatoires sont des estimateurs consistants du paramètre λ ; mais, au vu des normalités asymptotiques, lequel choisir ? Prenez un paramètre λ assez grand et illustrez graphiquement que celui dont vous avez dit qu'il était préférable l'est effectivement. (Ceci est un résultat général, des deux estimateurs en présence, le meilleur vient d'une estimation par maximum de vraisemblance.)

3. Méthodes de Monte-Carlo

3.1. **Principe de la méthode.**

Soit Y une variable aléatoire. Les méthodes de Monte-Carlo permettent de calculer numériquement la quantité $I = \mathbb{E}[g(Y)]$, qui s'écrit aussi $\int gf$ lorsque Y est de densité f, pour toute fonction g telle que $g(Y)$ soit intégrable, et ce même sans connaître f. Il suffit pour cela de disposer d'un N-échantillon $Y_1, \dots, Y_N$ de loi celle de Y. La loi des grands nombres assure que si $g(Y)$ est intégrable, alors

$$\hat{I}_N = \frac{1}{N} \sum_{k=1}^{N} g(Y_k) \xrightarrow{\mathbb{P}-ps} \mathbb{E}\left[g(Y)\right] .$$

On estime donc $\mathbb{E}[g(Y)]$ par $\hat{I}_N$. Il s'agit alors de quantifier l'erreur commise $|I - \hat{I}_N|$ (pour cela la loi des grands nombres ne suffit pas car il faut la vitesse de convergence).

Cette méthode a de nombreuses applications, pour calculer des intégrales bien sûr, mais aussi pour obtenir des résultats non asymptotiques. Supposons avoir accès à un N-échantillon $X_1^N = (X_1, \dots, X_N)$ de loi de paramètre θ, pour lequel nous avons un intervalle de confiance $I(X_1^N)$, asymptotiquement de niveau $1-\alpha$. Le problème se pose de connaître le niveau exact de cet intervalle à un rang fixé, $N = 1000$ par exemple : est-il déjà proche de $1-\alpha$? En tout cas, il est donné par $\mathbb{E}[Y]$, où $Y = \mathbb{1}_{\theta \in I(X_1^N)}$. Donc on peut calculer ce niveau par la méthode de Monte-Carlo.

En conclusion, les méthodes de Monte-Carlo sont utiles dans les cas (nombreux) où l'on peut simuler facilement Y, mais où il est difficile de calculer précisément sa loi, et permettent de calculer précisément des intégrales. On peut montrer (voir l'exercice suivant) que dans les bons cas, et notamment ceux où la fonction à intégrer est en fait de carré intégrable, la précision de la méthode est (avec grande probabilité) en $O(1/\sqrt{N})$ pour une complexité de calcul linéaire, alors que les méthodes numériques donnent une précision similaire pour une complexité N^d exponentielle en la dimension d de l'espace dans lequel vit g, et sous des conditions souvent fortes de régularité.

Exercice 4. Considérons un cas simple : g est l'identité et Y suit une loi de Bernoulli de paramètre p. On s'intéresse donc à l'estimation de p par la moyenne empirique des observations. En statistiques, pour comparer des vitesses de convergence, on écrit des intervalles de confiance et on en compare les longueurs. Quels sont les intervalles de niveau de confiance $1-\alpha$ pour l'estimation de p que l'on déduit de l'application respective des inégalités de Tchebychev, Hoeffding (voir la proposition 1 ci-dessous) et du TCL ? Lesquels sont asymptotiques, lesquels sont exacts ? Si l'on ne tient pas compte du désavantage certain que l'intervalle associé au TCL est asymptotique, quel est cependant son avantage ?

Proposition 1. Inégalité de Hoeffding
Soit (X_k) une suite i.i.d. et $a \leq b$ des constantes telles que $\mathbb{P}(a \leq X_k \leq b) = 1$, alors

$$\mathbb{P}\left(|S_n - \mathbb{E}S_n| \geq n\varepsilon\right) \leq 2\exp\left(-2n\varepsilon^2/(b-a)^2\right)$$

où S_n est la somme des n premières variables de la suite (X_k).

Exercice 5. Considérons un cas plus compliqué, le calcul d'une valeur approchée, sur $[0,1]$, de l'intégrale d'une fonction réelle g. Si $g(x) = x^{-1/4}$, lesquelles des trois inégalités considérées dans le précédent exercice donnent-elles un intervalle de confiance ? Et lorsque $g(x) = 1/\sqrt{x}$?

Exercice 6. Nous allons comparer le niveau exact d'intervalles de confiance tous asymptotiquement de niveau $1-\alpha$. Soit $X_1, \dots, X_n$ un n-échantillon de la loi de Poisson de paramètre λ. Le TCL assure que

$$\sqrt{n}\left(\overline{X}_n - \lambda\right) \rightsquigarrow \mathcal{N}(0, \lambda) \ ,$$

où $\overline{X}_n$ est la moyenne empirique. On voudrait un intervalle de confiance pour λ. Ecrivez ce que donne l'application du lemme de Slutski à la convergence précédente, lorsque vous divisez le premier membre par $\sqrt{\overline{X}_n}$ (resp. $\sqrt{S^2}$) ; quels sont les intervalles de confiance asymptotiques que l'on obtient ?
En réalité, on dispose d'un résultat de convergence qui évite les substitutions précédentes :

$$2\sqrt{n}\left(\sqrt{\overline{X}_n} - \sqrt{\lambda}\right) \rightsquigarrow \mathcal{N}(0,1) \ .$$

Quel est l'intervalle de confiance asymptotique que l'on en déduit ?
Comparer le niveau réel (c'est-à-dire non-asymptotique) de ces intervalles, à un rang fixé, par exemple aux rangs `n=[10 15 25 50 100]`, et pour $\lambda = 3$. Estimez pour cela $\mathbb{E}(\mathbb{1}_{\lambda \in \mathrm{ICA}(\alpha)})$.

Exercice 7 [Calcul de π]. Donner une fonction $g : [0,1] \to \mathbb{R}$ d'intégrale π. Calculer N tel qu'avec N variables uniformes sur $[0,1]$, on obtienne (au moins 95% du temps) une approximation bonne à 10^{-2} près de π par la méthode de Monte-Carlo. Simuler avec Matlab et comparer avec le résultat théorique.
Alternative : reprendre l'exercice avec $g(x,y) = 4\mathbb{1}_{x^2+y^2\leq 1}$.

Exercice 8 [Réduction de la variance]. Voici une méthode dite de réduction de la variance. Notez que nous avons vu plus haut que la précision d'un calcul de Monte-Carlo dépend de manière essentielle de la variance de $g(Y)$. Réduire cette variance, c'est gagner en précision ! Soit à intégrer g sur $[0,1]$ contre la mesure de Lebesgue. Si X est uniforme sur $[0,1]$, alors

$$\int_{[0,1]} g(x)\,\mathrm{d}x = \mathbb{E}[g(X)] = \mathbb{E}[g(1-X)] = \mathbb{E}\left[\frac{1}{2}\left(g(X)+g(1-X)\right)\right] .$$

Ainsi, on peut penser à utiliser la méthode de Monte-Carlo avec $h(X) = (g(X)+g(1-X))/2$. Que peut-on dire des variances de $g(X)$ et de $h(X)$? Notez que la méthode avec h, appelée réduction par variables antithétiques, requiert deux fois plus de calculs.
Application (simple mais spectaculaire) : $g(x) = \exp x$. En utilisant, puisque l'on travaille à réduire la variance, le TCL ou l'inégalité de Tchebytchev, combien de simulations faut-il pour obtenir une valeur approchée de l'intégrale (dont la valeur est $e-1$) à 10^{-3} près, dans le cas où l'on utilise la méthode standard et celle des variables antithétiques (au moins 95% du temps, asymptotiquement) ? Simulez avec Matlab, comparez à la vraie valeur.