Sup Galilée
MACS1

Février 2022

Initiation aux projets numériques – TP Proba. 5
Théorème Central Limite

Le contenu de ce TP n'est pas au programme des évaluations. Il sera revu en TD de Statistiques.

Avant de commencer ce TP, créer un sous-répertoire `TP_Proba_5`, et créer un script (vide) `TP.m` dans ce sous-répertoire. Copier le fichier `hist_densite.m` dans ce répertoire.

L'objectif de ce TP est d'étudier l'erreur que l'on fait dans l'approximation par la méthode de Monte-Carlo. Pour une estimation à partir de N simulations, cette erreur est de l'ordre de $1/\sqrt{N}$, et est aléatoire, avec une loi approximativement gaussienne. C'est le contenu du **théorème central limite**.

On va illustrer ce résultat, et l'utiliser pour obtenir des **intervalles de confiance** sur les résultats obtenus par la méthode de Monte-Carlo.

1 Théorème central limite et intervalles de confiance

La loi des grands nombres exprime que, si $X_1, X_2, \dots$ sont des variables aléatoires indépendantes, et de même loi qu'une variable aléatoire X intégrable, alors

$$\frac{X_1 + \cdots + X_N}{N} \underset{N\to\infty}{\longrightarrow} E[X] \qquad \text{presque sûrement.}$$

L'inégalité de Tchébychev fournit un premier contrôle de l'erreur : si X est de carré intégrable, pour tout $\alpha > 0$,

$$P\left(\left|\frac{X_1 + \cdots + X_N}{N} - E[X]\right| \geq \alpha\frac{\sigma(X)}{\sqrt{N}}\right) \leq \frac{1}{\alpha^2}.$$

En effet, $\mathrm{Var}(\frac{1}{n}(X_1 + \cdots + X_n)) = \frac{1}{n^2} n\mathrm{Var}(X) = \frac{1}{n}\mathrm{Var}(X)$ grâce à l'indépendance entre les X_i. Ainsi, avec probabilité $> 1 - \frac{1}{\alpha^2}$, le réel $E[X]$ appartient à l'intervalle aléatoire

$$\left[\overline{X}_N - \alpha\frac{\sigma(X)}{\sqrt{N}}, \overline{X}_N + \alpha\frac{\sigma(X)}{\sqrt{N}}\right],$$

où $\overline{X}_N = \frac{1}{N}(X_1 + \cdots + X_N)$. On parle d'intervalle de confiance de niveau $1 - \frac{1}{\alpha^2}$. Par exemple, pour un niveau de confiance 95%, on doit prendre $\alpha = 1/\sqrt{0.05} \simeq 4.5$. On retient en particulier que l'erreur est (au plus) d'ordre $\frac{1}{\sqrt{N}}$.

Ce résultat est seulement une majoration, mais l'erreur est en fait toujours asymptotiquement d'ordre $\frac{1}{\sqrt{N}}$, ce qui explique les convergences très lentes dans la méthode de Monte-Carlo : c'est ce qu'exprime le théorème central limite, qui donne la loi limite de l'erreur normalisée.

D'après le théorème central limite, si $X_1, X_2, \dots$ sont indépendantes, et de même loi qu'une variable aléatoire X de carré intégrable, alors

$$\frac{\sqrt{N}}{\sigma(X)}\left(\frac{X_1 + \cdots + X_N}{N} - E[X]\right) \underset{N\to\infty}{\overset{\text{(loi)}}{\longrightarrow}} \mathcal{N}(0,1),$$

où $\sigma(X) = \sqrt{\mathrm{Var}(X)}$ est l'écart-type de X. Autrement dit, pour N grand, la variable aléatoire ci-dessus suit approximativement la loi $\mathcal{N}(0,1)$. La convergence en loi peut s'écrire sous la forme de convergences classiques :

$$\text{pour tous réels } a < b, \qquad P\left(a < \frac{\sqrt{N}}{\sigma(X)}\left(\frac{X_1 + \cdots + X_N}{N} - E[X]\right) < b\right) \underset{N\to\infty}{\longrightarrow} P(a < Z < b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt,$$

où Z suit la loi $\mathcal{N}(0,1)$. Ainsi (pour $b = -a = \alpha$), avec probabilité $\simeq P(|Z| \leq \alpha)$, le réel $E[X]$ appartient à l'intervalle aléatoire

$$\left[\overline{X}_N - \alpha\frac{\sigma(X)}{\sqrt{N}}, \overline{X}_N + \alpha\frac{\sigma(X)}{\sqrt{N}}\right].$$

Par exemple, pour un niveau de confiance 95%, on doit prendre $\alpha \simeq 1.96$. C'est donc un intervalle deux fois plus étroit que celui fourni par l'inégalité de Tchebychev, mais qui est seulement valable asymptotiquement. En pratique, on l'utilise néanmoins, dès que N est assez grand (on dit souvent $N > 30$, mais cela vaut seulement si X "ne prend pas de grandes valeurs trop souvent").

Exercice 1 : Mise en évidence du TCL

On se propose d'observer la convergence de l'erreur vers la loi normale, dans le cas où X suit la loi uniforme sur $[0,1]$. On rappelle $E[X] = \frac{1}{2}$ et $\mathrm{Var}(X) = \frac{1}{12}$.

1. (Dans `ex1a.m`) On veut estimer la loi de $Z_N = \frac{\sqrt{N}}{\sigma(X)}\left(\frac{X_1+\cdots+X_N}{N} - E[X]\right)$ pour $N = 1000$. Pour cela, remplir un tableau `tirages` contenant $M = 10000$ variables aléatoires suivant cette loi, et représenter leur fonction de répartition empirique. *Vectoriser le calcul de Z_N pour accélérer le calcul.*

2. (Dans `ex1b.m`) Représenter un histogramme normalisé pour comparer la loi de Z_N à la loi $\mathcal{N}(0,1)$.

3. (Dans `ex1b.m`) À l'aide du vecteur `tirages`, estimer la probabilité $P(|Z_N| \leq 1.96)$.

Exercice 2 : Intervalle de confiance dans l'estimation de probabilités

Souvent, on ne connaît pas $\sigma(X)$ (vu que l'on veut estimer $E[X]$). Pour calculer un intervalle de confiance, on peut cependant estimer $\sigma(X)$ en même temps que l'on estime $E[X]$.

Quand on utilise la méthode de Monte-Carlo pour estimer une probabilité $P(A) = p$, on a $X = \mathbf{1}_A$ donc $E[X] = p$ et $\mathrm{Var}(X) = p(1-p)$. Comme on sait estimer p par $\overline{X}_N = \frac{1}{N}(X_1 + \cdots + X_N)$, on peut donc estimer $\sigma(X) = \sqrt{p(1-p)}$ par $\sqrt{\overline{X}_N(1-\overline{X}_N)}$.

1. (Dans `ex2.m`) Estimer $P(U^2 + V^2 < 1)$, où U et V suivent la loi uniforme sur $[0,1]$ et sont indépendantes, à l'aide de la méthode de Monte-Carlo à partir de $N = 10000$ tirages, et donner un intervalle de confiance de niveau 95% sur le résultat. Est-ce que l'intervalle contient bien la valeur théorique ?

2. (Dans `ex2.m`) Estimer, par la méthode de Monte-Carlo, le niveau de confiance de l'intervalle fourni par le programme précédent : répéter $M = 1000$ fois la simulation précédente et calculer la proportion de fois où $E[X]$ appartient à l'intervalle calculé.

Exercice 3 : Intervalle de confiance dans l'estimation d'espérances

Quand on utilise la méthode de Monte-Carlo pour estimer une espérance $E[X]$, on peut aussi estimer $\mathrm{Var}(X) = E[X^2] - E[X]^2$ à l'aide de la variance empirique :

$$V_N = \frac{1}{N}(X_1^2 + \cdots + X_N^2) - \left(\frac{X_1 + \cdots + X_N}{N}\right)^2.$$

Par la loi des grands nombres, V_N converge presque sûrement vers $\mathrm{Var}(X)$. On peut justifier que l'on peut remplacer $\sigma(X)$ dans l'intervalle de confiance par $\sqrt{V_N}$, autrement dit : avec probabilité $\simeq P(|Z| < \alpha)$, $E[X]$ appartient à l'intervalle aléatoire

$$\left[\overline{X}_N - \alpha\frac{\sqrt{V_N}}{\sqrt{N}}, \overline{X}_N + \alpha\frac{\sqrt{V_N}}{\sqrt{N}}\right].$$

En pratique, on estime simultanément $E[X]$ et $E[X^2]$, ce qui permet de calculer V_N et donc l'intervalle ci-dessus.

(Dans `ex3.m`) Estimer $E[(U+V)^2]$ où U et V sont indépendantes et de loi exponentielle de paramètre 1, à l'aide de la méthode de Monte-Carlo à partir de $N = 10000$ tirages, et donner un intervalle de confiance de niveau 95% sur le résultat. Est-ce que l'intervalle contient bien la valeur théorique ?