TP d'initiation à MATLAB n^0 6 : Loi des Grands Nombres et Théorème Limite Central

Sachant maintenant simuler des variables aléatoires, on commence par mettre en évidence la Loi des Grands Nombres et l'importance de ses hypothèses. Par la suite, on visualise des applications du Théorème de Limite Central. Enfin, on revient sur l'utilisation d'un histogramme pour visualiser la loi de probabilité d'une variable aléatoire.

Loi des Grands Nombres

On rappelle l'énoncé de ce théorème (appelé "Loi" car on a longtemps cru que c'était un résultat de physique et non de mathématiques). On considère des variables aléatoires $(X_n)_{n\in\mathbb{N}^*}$, indépendantes et de même loi qu'une variable X. Si on suppose que $\mathbb{E}|X| < \infty$, alors

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \underset{n\to+\infty}{\xrightarrow{\mathcal{P}}} m \quad \text{avec } m = \mathbb{E}X.$$

Intuitivement, cela signifie que la moyenne empirique converge vers la moyenne théorique quand on considère un nombre croissant de variables aléatoires.

	On commence par reproduire la situation d'un jeu de pile ou face. Pour cela, on considère qu'une pièce a une probabilité $p \in]0,1[$ de tomber sur pile. On jette un grand nombre de fois la pièce et on calcule la fréquence (proportion) des Piles parmi ces lancers. On modélise cette expérience par des variables indépendantes (car les lancers sont indépendants les uns des autres) de Bernoulli de paramètre p, X_i pour le ième lancer ($X_i = 1 \iff$ le ième lancer est Pile). Alors la somme des X_i pour $i = 1$ à n représente le nombre total de Piles et la moyenne empirique $\bar{X}_n$
$clear$	représente la fréquence des Piles. Efface toutes les variables de la session.
$X = binornd(1, 0.5, 1, 10)$	On simule 10 réalisations de v.a. indépendantes de Bernoulli de paramètre 0.5. Quelle est la loi de la somme de ces variables ? Comment simuler directement cette somme ? (noter que l'on a utiliser pour cette simulation *rbinornd* : il existe dans Matlab des générateurs de réalisation de v.a. diverses déjà programmés. On peut les retrouver dans *Helpwin*, puis dans *toolbox/stats*) . Faire des essais avec d'autres générateurs.
$mean(X)$	La moyenne empirique est-elle proche de la "moyenne théorique"? Recommencer en augmentant la taille de X: 100, 1000,... Conclusion? Recommencer en changeant le paramètre 0.5 en 0.2 et en 0.9.
$Y = rand(1, 10)$	Cette fois-ci on génère des v.a. de loi uniforme sur $[0, 1]$.
$m = mean(Y)$	La moyenne empirique est-elle proche de la "moyenne théorique"? Recommencer en augmentant la taille de Y: 100, 1000,... Conclusion?

Exercices

1. Faire un programme calculant la moyenne d'un nombre n (arbitraire) de réalisations de variables uniformes sur $[-2, 3]$, et qui trace en fonction de n cette moyenne. Que constate-t-on? Faîtes la même chose avec des réalisation indépendantes d'une v.a. distribuée d'abord selon une loi normale $\mathcal{N}(2, 3^2)$, puis suivant une loi de Cauchy (que l'on génère à partir de la tangente d'une variable uniforme sur $]-\pi/2, \pi/2[$). Conclusions?.

2. L'exercice précédent montrait l'importance de l'hypothèse $\mathbb{E}|X| < +\infty$. Pour montrer l'importance de la notion d'indépendance, on suppose que pour tout i, on a $X_i = X_1$. La Loi des Grands Nombres est-elle encore vérifiée ? Maintenant, on considère que les X_i sont bien indépendantes, mais on veut appliquer la Loi des Grands Nombres aux variables $Y_i = \dfrac{X_1 + .. + X_i}{i}$ (donc les Y_i ne sont pas, en général, indépendantes les unes des autres). La Loi des Grands Nombre est-elle vérifiée pour les lois considérées précédemment ?

Application de la Loi des Grands Nombres : distribution d'une variable aléatoire

Une application intéressante de la Loi des Grands Nombres est le fait que lorsque l'on a de nombreuses réalisations indépendantes d'une variable aléatoire, alors l'histogramme, s'il possède suffisamment de classes, se rapproche de la loi (densité) de probabilité. Cela s'explique par le fait que la proportion empirique de variables dans $[a, b]$ tend vers la probabilité théorique que la variable appartienne à $[a, b]$ (cette probabilité est en fait l'espérance de la fonction qui vaut 1 si la variable est dans $[a, b]$ et 0 sinon).

$X = rand(100, 1)$	On simule 100 réalisations de v.a. indépendantes de loi uniforme sur $[0, 1]$.
hist(X)	Histogramme de ces réalisations. Déterminer théoriquement la densité d'une v.a. de loi uniforme sur $[0, 1]$. Conclusion ? Changer le nombre de classes (20, 5) et expliquer les résultats. Faire la même chose avec 1000 puis 10 réalisations. Conclusion ?
$X = randn(100, 1)$	On refait la même chose avec une v.a. normale centrée réduite. Répondre aux mêmes questions. Recommencer avec une v.a. binomiale $\mathcal{B}(5, 0.5)$, une variable exponentielle de paramètre 2 et enfin une v.a. de Cauchy. Expliquer pourquoi cela "marche" pour Cauchy alors que $E\|X\| = \infty$. On s'intéresse maintenant à la démarche inverse. A partir d'un histogramme, on peut avoir une idée d'une v.a. modélisant un problème. Par exemple, voici le diamètre de 100 écrous obtenu en tapant : $!cp\ \sim bardet/Ecrous.mat\ Ecrous.mat$. Vous avez ainsi recopié le fichier *Ecrous.mat*
load Ecrous	dans votre répertoire. On charge ce fichier, qui contient la variable *Ecrous*.
hist(Ecrous)	Retrouver la distribution de la variable D : diamètre d'un écrou.

Théorème de Limite Central

Rappelons d'abord son énoncé qui est une sorte de raffinement par rapport à la Loi des Grands Nombres : on considère des variables aléatoires $(X_n)_{n\in\mathbb{N}^*}$, indépendantes et de même loi qu'une variable X. Si on suppose que $\mathbb{E}|X| < \infty$, et que $E|X|^2 < \infty$, alors

$$\sqrt{n}\left(\frac{\bar{X}_n - m}{\sigma^2}\right) = \sqrt{n}\left(\frac{\frac{1}{n}(X_1 + X_2 + \cdots + X_n) - m}{\sigma^2}\right) \xrightarrow[N\to+\infty]{\mathcal{L}} \mathcal{N}(0,1) \quad \text{avec } m = \mathbb{E}X \text{ et } \sigma^2 = \mathbb{E}X^2.$$

Cela signifie, de façon intuitive, que quand n est grand alors la moyenne empirique converge vers la moyenne théorique suivant une répartition gaussienne $\mathcal{N}(0, \frac{\sigma^2}{n})$.

	On recommence la situation d'un jeu de pile ou face. Cela correspond aussi à la situation d'un sondage d'opinion pour choisir entre A et B. Ainsi, si la proportion d'intention de vote pour A sur l'ensemble de la population est p, on peut modéliser chaque votant par X_i, et lorsque $X_i = 1$ se traduit par un vote pour A. Alors la somme des X_i pour $i = 1$ à n représente le nombre total d'intentions de vote pour A parmi n personnes et la moyenne empirique $\bar{X}_n$ représente la fréquence des intentions de vote pour A parmi les n personnes.
$X = binornd(1, 0.5, 10, 100)$	On considère 100 fois 10 personnes sondées.
$mean(X)$	On détermine la moyenne empirique de chaque groupe de 10 personnes.
hist(X)	On visualise la distribution de la moyenne empirique à partir de l'histogramme de 100 réalisations indépendantes. A-t-on quelque chose de proche d'une distribution normale. Quels sont alors les paramétres de cette distribution ? Comparer avec la théorie. Recommencer avec des groupes de 100, puis 1000 personnes. Conclusions ? Recommencer en changeant le paramètre 0.5 en 0.2 et en 0.9.

Exercices

1. Refaire des histogrammes (comme ci-dessus) sur la moyenne empirique dans le cas d'une loi exponentielle de paramètre 2, puis une loi de Cauchy. Conclusions ?

2. On a interrogé 1000 personnes avant une élection entre A et B. Les résultats de ce sondage (1 lorsque la personne veut voter pour A et 0 lorsqu'elle veut voter pour B) sont dans le fichier *Votes.mat* à recopier comme précédemment. Déterminer un intervalle de confiance sur le pourcentage de vote pour A. Déterminer empiriquement la probabilité que A soit élu.