TP2 : CONVERGENCE DE VARIABLES ALÉATOIRES

Loi des grands nombres, Théorème « central limit », Méthodes de Monte-Carlo

Résumé. Ce tp présente les fondements d'une illustration numérique de la convergence d'une suite de variables aléatoires. Les représentations par histogramme et par comparaison des fonctions de répartition, théorique et empirique, sont étudiées et comparées dans un premier temps. On s'intéresse ensuite à la quantification de la précision des résultats obtenus empiriquement, avant d'appliquer les méthodes obtenues à des cas plus concrets : ceci forme les méthodes de Monte-Carlo.

1. Représentation de la convergence de variables aléatoires

Voyez en appendice les quelques rappels sur la gestion des histogrammes.

1.1. Convergence de variables discrètes.

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. Un histogramme, une densité en bâtons, et le tour est joué, *ici*[1]...

1.2. Convergence de variables à valeurs continues.

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.
Cependant, dans un deuxième temps, demandez-vous s'il est bien raisonnable de prouver cette proximité par des histogrammes... (Voyez à cet effet l'appendice, et notamment sa section 3.2.) Quelle(s) autre(s) méthode(s), reposant sur quel(s) théorème(s) mathématique(s), proposez-vous alors ?

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[2] la normalité asymptotique de S^2 :

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

1. eh oui, ça va se gâter bientôt !
2. et c'est un très bel exercice, que je vous recommande de faire s'il vous reste du temps à la fin du tp !

où S^2 désigne indifféremment [3] la variance empirique ou $(n-1)/n$ fois la variance empirique.
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 consistents du paramètre λ ; mais, en 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.)

2. Méthodes de Monte-Carlo

2.1. **Principe de la méthode.** Soit Y une variable aléatoire. Supposons que l'on veuille calculer $\mathbb{E}[g(Y)]$, pour une certaine fonction (mesurable) g telle que $g(Y)$ soit intégrable. Cette formulation générale recouvre de nombreux problèmes particuliers.

Exemple 1 [Obtention de résultats non asymptotiques]. Supposons avoir accès à un n-échantillon $X_1^n = (X_1, \ldots, X_n)$ de loi de paramètre θ, pour lequel nous avons un intervalle de confiance $I(X_1^n)$, asymptotiquement de niveau α. 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 α ? En tout cas, il est donné par $\mathbb{E}[Y]$, où

$$Y = \mathbb{I}_{\theta \in I(X_1^n)} \ .$$

Exemple 2 [Calcul d'intégrales]. Soit g une fonction $[0,1] \to \mathbb{R}$. Calculer l'intégrale de g revient à calculer $\mathbb{E}[g(Y)]$, où Y est une variable aléatoire de loi uniforme sur [0,1].

Pour ce faire, on suppose que l'on a accès à un N-échantillon $Y_1, \ldots, Y_N$ de loi celle de Y. La loi des grands nombres assure que si $g(Y)$ est intégrable, alors

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

Le problème est que la loi des grands nombres n'indique pas de vitesse de convergence.

Exercice 4. Considérons un cas simple : g est l'indentité et Y suit une loi binômiale de paramètre p (bref, on s'intéresse [4] à l'espérance d'une variable binômiale). En statistiques, pour comparer des vitesses de convergence, on écrit des intervalles de confiance (de niveau $1-\alpha$) et on en compare les longueurs. Quels sont les intervalles que l'on déduit de l'application respective des inégalités de Markov, Tchebychev, Hoeffding 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 (surtout lorsque p est petit) ?
Ne pourrait-on pas avoir le beurre et l'argent du beurre ? Si vous n'êtes pas trop pressé par le temps, pouvez-vous essayer d'écrire une version de l'inégalité de Hoeffding telle que pour N suffisamment grand ($Np(1-p) \geq \ln(2/\alpha)$), on obtienne un intervalle de confiance du même ordre que celui du TCL (*id est*, tenant compte de la variance de Y) ? (Indication : utiliser une majoration de Chernoff et l'inégalité $\exp x \leq 1 + x + x^2$, valable pour $x \leq 1$.)

3. cf. lemme de Slutzki. Question subsidiaire : à laquelle des deux fait référence la commande Matlab `var` ?
4. oui, oui, on a déjà vu problème plus palpitant...

Exemple 3. 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 quatre inégalités considérées dans le précédent exercice donnent-elles un intervalle de confiance ? Et lorsque $g(x) = 1/\sqrt{x}$?

En conclusion, les méthodes de Monte-Carlo sont utiles dans deux types de problèmes :
- ceux où l'on peut simuler facilement Y, mais où il est difficile de calculer précisément sa loi ;
- pour calculer des intégrales : les exercices précédents montrent 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.

2.2. **Variables faciles à simuler mais de loi difficile à estimer.**

Exercice 5. Nous allons comparer le niveau exact d'intervalles de confiance tout 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

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

où $\bar{X}_n$ est la moyenne empirique. On voudrait un intervalle de confiance pour λ. Ecrivez ce que donne l'application du lemme de Slutzki à la convergence précédente, lorsque vous divisez le premier membre par $1/\sqrt{\bar{X}_n}$ (resp. $1/\sqrt{S^2}$) ; quels sont les intervalles de confiance (asymptotiques, au niveau $1-\alpha$) que l'on obtient ?
Intuitivement, on sent que la méthode précédente (dite de *plug-in*) est illusoire : on voulait étudier les déviations de $\bar{X}_n$ et au dernier moment, on rajoute un terme $\bar{X}_n$ ou S^2 qu'on ne relie pas à ses déviations. Voici une manière plus propre de faire : on peut montrer que

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

Quel est l'intervalle de confiance (asymptotique au niveau $1-\alpha$) que l'on en déduit ?
Vient maintenant le temps de la comparaison du niveau de ces intervalles, à un rang fixé, par exemple aux rangs `n=[10 15 25 50 100]` (et pour $\lambda = 3$). Quelle est Y ? Combien de simulations N faut-il faire pour avoir (au moins 94% du temps[5]) des valeurs approchées *simultanément* à 10^{-2} près des valeurs exactes aux différents rangs `n` ?

2.3. **Calcul d'intégrales.**

Exercice 6 [Calcul de π]. Voilà un exercice « tarte à la crème », qu'il faut avoir fait une fois dans sa vie, mais à ne jamais ressortir lors d'une leçon d'agrégation ! 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^{-4} près de π par méthode de Monte-Carlo. Simuler avec Matlab et conclure : êtes-vous dans les 5% de malchanceux[6] ?
Alternative : reprendre l'exercice avec $g(x,y) = 4\mathbb{I}_{x^2+y^2\leq 1}$...

5. attention ! Il faut donc un rang N tel que chacun des niveaux soit 98% du temps proche à 10^{-2} près de sa vraie valeur...

6. en fait, il faudrait aussi tenir compte du fait que personne, pas même Matlab, ne fait des calculs exacts, que le générateur de nombres aléatoires n'est pas parfait, *etc.* Du coup, on est sans doute plus souvent malchanceux que 5% du temps...

Exercice 7 [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. Quand donc cette méthode est-elle une amélioration du schéma ordinaire ?

Application (idiote mais spectaculaire) : $g(x) = \exp x$. En utilisant pour une fois le TCL, puisque l'on travaille à réduire la variance, combien de simulations 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 et dites-moi si vous êtes dans les malchanceux (deux calculs à faire, donc (presque) deux fois plus de chances d'être malchanceux...).

3. Appendice : Rappels sur la gestion des histogrammes

3.1. **Histogrammes non renormalisés (cas discret, essentiellement).** Soit $x = (x_1,x_2,\ldots,x_n)$ une réalisation de la suite de variables aléatoires $X = (X_1,X_2,\ldots,X_n)$ et C l'ensemble des valeurs possibles pour les données.

(1) L'ensemble C est découpé en $|I|$ classes disjointes : $C = \bigcup_{i\in I} C_i$.

(2) Le nombre de données appartenant à chaque classe C_i est $N(i)$:

$$N(i) = \sum_{j=1}^{n} \mathbb{I}_{\{x_j \in C_i\}}.$$

Ce comptages sont obtenus directement par la fonction `hist` :
» `[N,C]=hist(x,n)`
où `n` désigne le nombre $|I|$ de classes souhaitées, `N` est la suite des $N(i)$ et `C` désigne les centres des classes.

(3) La représentation en barres de $C_i \mapsto N(i)$ est obtenue à l'aide de plusieurs fonctions :
- `bar` : » `bar(C,N)`, on a aussi `barh`,
- `stem` : » `stem(C,N)` (elle s'utilise plutôt pour une suite discrète),
- `hist` : » `hist(x,C)`.

Comme pour `plot`, il existe des options de dessin (voir l'aide).

Exemple 4 [Cas discret]. On observe $n = 30$ vaches au hasard et on s'intéresse au caractère race qui prend 3 modalités : « normande », « frisonne », « autre ». Soit x le vecteur tel que la i-ième composante est j si la i-ème vache est de race j ; `x` est de la forme `[1 1 2 3 2 1 1 3 2 ...]`. Il y a 16 normandes, 9 frisonnes et 5 autres : on note `vache=[16 9 5]` le vecteur des comptages.

- Si on dispose des comptages, on peut obtenir une représentation en barres par :
  - la fonction `bar` :
    » `bar(vache)` retourne trois barres dont la hauteur est le comptage pour chaque modalité. (En écrivant `bar(1:3,vache)`, on obtient le même graphe.)
    » `bar(vache/30)` retourne trois barres dont la hauteur est la fréquence pour chaque modalité.
    On peut aussi construire k histogrammes côte à côte. Dans ce cas l'argument pour `bar` est une matrice à k colonnes, chaque colonne donnant lieu à un histogramme.
  - la fonction `stem` :
    » `stem(vache)`
- Si on ne dispose pas des comptages (mais simplement du vecteur x des observations) :
  » `vache=histc(x,1:3)` effectue le comptage pour chaque modalité et on peut utiliser le paragraphe précédent ;
  à défaut, on peut toujours taper » `[N,C]=hist(x,1:3)` puis
  » `bar(C,N)`.

Conclusion : **Représentation d'une loi empirique à l'aide d'un histogramme renormalisé – Cas d'une loi discrète sur $C = \{c_i, i \in I\}$.**

La loi des grands nombres assure la convergence de la suite des fréquences empiriques :

$$\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}_{[X_j=c_i]} \overset{\mathbb{P}-ps}{\to} \mathbb{P}[X=c_i].$$

Donc, pour représenter empiriquement la loi, et avec les notations précédentes, on :

- pose $C_i = \{c_i\}$,
- représente $C_i \mapsto \frac{N(i)}{n}$.

3.2. **Histogrammes renormalisés (cas continu, essentiellement).** `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.

`[edge,h]=histo(x,m,odd,scale,'s')` ne trace pas l'histogramme mais retourne le résultat du tri des valeurs de x :

- `edge` vecteur contenant les valeurs des extrémités des classes,
- `h` est un vecteur tel que sa i-ème composante $h(i)$ est le nombre de valeurs de x compris dans l'intervalle $[\texttt{edge}(i), \texttt{edge}(i+1)[$.

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

C'est cela qu'il faut faire pour représenter un estimateur empirique de la densité.

Conclusion : **Représentation d'une loi empirique à l'aide d'un histogramme renormalisé – Cas d'une loi continue sur l'intervalle $C = [a,b]$ de $\mathbb{R}$.**

On va établir un résultat similaire à celui se trouvant à la fin du paragraphe 3.2.
Si la densité f est continue et si $h_n \to 0$ de sorte que $nh_n \to \infty$, alors

$$\hat{f}_n(x) = \frac{1}{2h_n}(F_n(x+h_n) - F_n(x-h_n)) \overset{\mathbb{P}\text{–proba}}{\longrightarrow} f(x)$$

où $F_n(x) = \frac{1}{n}\sum_{j=1}^{n}\mathbb{I}_{\{X_j\le x\}}$ est la fonction de répartition empirique.

Remarquez que $\hat{f}_n(x)$ n'est pas la densité empirique, celle-ci n'existant même pas ! C'est simplement une statistique qui est raisonnablement proche de la notion intuitive qu'on pourrait avoir de

« densité empirique ». En particulier, ne dites jamais, jamais, que l'on vient de prouver une quelconque convergence des densités empiriques ; on a juste remarqué que ce que fait Matlab[7], appelé avec `histo(x,[],[],1)`, est mathématiquement fondé, à savoir :

- considérer $h_n = 4n^{1/4}$,
- décomposer $C = [\min(x), \max(x)]$ en $4n^{1/4}$ intervalles de même longueur notée $2h_n$,
- représenter $C_i \mapsto (\frac{N(i)/n}{2h_n})$.

Moralité : `histo` permet de gérer efficacement et à votre place la discrétisation qui est un préalable à l'application des méthodes du paragraphe 3.1.

(Bon, en fait Matlab considère des histogrammes là où nous avions une fenêtre glissante ; notre $\hat{f}_n$ n'était pas constante par morceaux. Mais asymptotiquement, c'est la même chose.)

Note : on trace une fonction de répartition empirique avec `stairs`.

7. Pensez bien à mettre 1 ! On veut une aire égale à 1, parce que l'on représente une densité de probabilité.