Introduction à la simulation en probabilité

Module MATH504

Année universitaire 2009–2010

L'énoncé du TP au format pdf se trouve sur le portail dans votre groupe.

1. Mise en route – Principe de simulation.

Connectez-vous puis lancez `Matlab` à partir du menu `Démarrer`.

La fonction `rand` génère des réels distribuées suivant la loi uniforme sur $[0,1]$: `rand(m,n)` fournit une matrice de taille $m \times n$ de réels distribués suivant la loi uniforme sur $[0,1]$.

Tapez `n=1000; u=rand(1,n);` pour obtenir un vecteur ligne de longueur `n`. La commande `mean(u)` donne la moyenne arithmétique du vecteur `u`. Le résultat doit être voisin de 1/2 la moyenne de la loi uniforme sur $[0,1]$. Expérimentez plusieurs valeurs de `n`.

La commande `n=1000; u=rand(1,n);` fournit un vecteur (ligne) de longueur n dont les composantes sont des réels distribués suivant la loi uniforme sur $[0,1]$. Les composantes du vecteur `u` s'interprètent comme une réalisation particulière – c'est à dire pour un $\omega \in \Omega$ donné – de n v.a. $U_1, \ldots, U_n$ indépendantes et identiquement distribuées suivant la loi uniforme sur $[0,1]$. En clair, on considère que

$$\texttt{u=(u\_1,...,u\_n)} = (U_1(\omega), \ldots, U_n(\omega)).$$

Rappelons que si U suit la loi uniforme sur $[0,1]$ et $t \in [0,1]$, $\mathbb{P}(U \leq t) = t$. Nous allons le vérifier. Pour cela, on approche la probabilité $\mathbb{P}(U \leq t)$ par la fréquence empirique de l'événement $\{U \leq t\}$:

$$\mathbb{P}(U \leq t) \simeq \frac{\text{Nombre de fois où } U_i \leq t}{\text{Nombre total d'expériences}}$$

Tapez `n=1000; u=rand(1,n);` pour obtenir un vecteur de longueur n dont les composantes sont censées être des réels distribués suivant la loi uniforme sur $[0,1]$. Pour compter le nombre de composantes de ce vecteur qui sont inférieures à t, on peut utiliser la commande `l=(u<=t)` qui renvoie un vecteur formé de 1 et de 0 suivant que la condition `u<=t` est vérifiée ou pas. Il reste à faire la somme des composantes de l, `sum(l)` pour obtenir le nombre de fois où `u_i<=t`. En résumé, la ligne de commande

```
n=10000; t=0.3; u=rand(1,n); l=(u<=t); sum(l)/n
```

devrait fournir un résultat proche de t. Faites varier n et t.

2. Simulation de variables aléatoires.

Pour simuler des variables aléatoires distribuées suivant une loi quelconque, l'idée est de se ramener à la loi uniforme sur $[0,1]$.

2.1. Variable de Bernoulli.

Une variable aléatoire de Bernoulli de paramètre $p \in]0,1[$ est une variable aléatoire X prenant deux valeurs 1 et 0 respectivement avec probabilité p et $1-p$. Vérifiez que si U suit la loi uniforme sur $[0,1]$, $X = \mathbf{1}_{U \leq p}$ suit la loi $\mathcal{B}(p)$.

Utilisez cette observation pour générer un vecteur ligne de longueur n distribué suivant la loi $\mathcal{B}(p)$.

2.2. Génération d'une binômiale.

Si $X_1, \ldots, X_n$ sont n variables aléatoires indépendantes et identiquement distribuées suivant la loi de Bernoulli $\mathcal{B}(p)$, la variable aléatoire

$$Y = X_1 + X_2 + \ldots + X_n$$

suit la loi binomiale $\mathcal{B}(n, p)$.

1. Pour générer un vecteur de m binômiales $\mathcal{B}(n, p)$, on commencera par générer un tableau $n \times m$ contenant des variables de pile ou face X_i avec $\mathbb{P}(X_i = 1) = p$. Il suffit de faire alors la somme des lignes de ce tableau.

   ```
   u=rand(n,m); l=(u<=p); y=sum(l,1);
   ```

   À l'aide de l'éditeur de `Matlab`, créer un répertoire `MATH504` puis à l'intérieur de ce répertoire un fichier `bin.m` contenant

   ```
   function y=bin(n,p,m)
       u=rand(n,m); l=(u<=p); y=sum(l,1);
   end
   ```

   Vous pourrez désormais vous servir directement de la fonction `bin`.

2. Estimer empiriquement les quantités $\mathbb{P}(Y = j)$ (analyser et utiliser la commande `sum(y==j)`). Les comparer à leur valeur théorique $\mathbb{P}(Y = j) = \mathrm{C}_n^j p^j (1-p)^{n-j}$. On calculera les C_n^j par la formule

   $$\mathrm{C}_n^j = \frac{\Gamma(n+1)}{\Gamma(j+1)\Gamma(n-j+1)}$$

   (la fonction gamma est disponible sous `Matlab`). On pourra prendre $m = 10000$, $n = 7$ et $p = 0.3$.

2.3. Lois amnésiques.

La densité de la variable exponentielle de paramètre c est

$$p(x) = ce^{-cx} 1_{x \geq 0}.$$

On a vu en TD qu'on peut la simuler à partir d'une variable U uniforme sur $[0, 1]$ par

$$X = \frac{-\ln(U)}{c}.$$

On a vu également que $Y = 1 + [X]$ suit une loi géométrique de paramètre $p = 1 - e^{-c}$:

$$\mathbb{P}(Y = k) = q_k = p(1-p)^{k-1}$$

1. Faire une fonction `x=expo(c,n)` qui génère une suite de variables aléatoires exponentielles de longueur n.

2. Choisir un c, simuler des échantillons de X assez long et estimer $\mathbb{P}(1 < X \leq 2)$. Le résultat doit être voisin de $e^{-c} - e^{-2c}$.

3. La loi exponentielle est une loi sans mémoire : si X est une v.a. de loi $\mathcal{E}(c)$,

   $$\mathbb{P}(X > t + h \mid X > t) = \mathbb{P}(X > h) = e^{-ch}.$$

   À l'intérieur du répertoire `MATH504`, créer un fichier `amexpo.m` Dans ce fichier, commencer à écrire les instructions répondant aux problème :

```
n=1000; c=0.5; t=2; h=1;
x=expo(c,n);
pt=exp(-c*h); p=sum(x>h)/n; disp([pt p]);
```

On peut aussi utiliser `p=mean(x>h)`.

N'oubliez pas de sauver le fichier régulièrement.

Pour exécuter ces instructions, placer vous dans `Matlab` dans le répertoire `MATH504` puis tapez simplement `amexpo`.

Modifier votre fichier afin qu'il affiche la valeur de $\mathbb{P}(X > h) = e^{-ch}$, une estimation de cette probabilité ainsi qu'une estimation de $\mathbb{P}(X > t+h \mid X > t) = \mathbb{P}(X > t+h)/\mathbb{P}(X > t)$.

Faites varier les paramètres !

4. Faire une fonction `x=geom(p,n)` qui génère une suite de variables aléatoires géométriques de longueur n. Cette fonction appellera la fonction `expo`.
5. Simuler une telle suite $(Y_i)_{1\leq i\leq n}$, et pour chaque valeur de k, estimer $\mathbb{P}(Y = k)$. On se contentera de $1 \leq k \leq 10$, et l'on prendra p aux alentours de 0,2. On appelle $\hat{q}_k$ cette estimée. Tracer le graphe $k \longmapsto \ln(\hat{q}_k)$.
6. Créer un fichier `amgeom.m` qui illustre que la loi géométrique est comme la loi exponentielle sans mémoire.

3. Histogrammes et densités.

Si X possède une densité régulière p, pour $a < b$ avec $b - a$ petit

$$\mathbb{P}(a < X < b) = \int_a^b p(x)\,dx \simeq (b-a)p\left((a+b)/2\right).$$

Si `x` est un vecteur et m un entier, l'instruction `[n,xout]=hist(x,m)` fournit un histogramme du vecteur `x` : l'intervalle $[x_{\min},x_{\max}]$ est subdivisé en m intervalles réguliers $I_1, \ldots, I_m$; le vecteur `xout` contient les centres de ces intervalles ; le vecteur `n` donne le nombre de valeurs de `x` appartenant à ces intervalles.

Écrire une fonction `[d,xout]=densite(x,m)` faisant appel à `[n,xout]=hist(x,m)` qui estime la densité de X aux centres des intervalles.

Testez votre fonction avec la loi exponentielle en faisant tracer sur la même figure la densité estimée ainsi que la vraie densité. On pourra utiliser pour le graphique la commande

`bar(xout,d); hold; plot(xout,c*exp(-c*xout); 'red');`.

Recommencez l'expérience en prenant la loi normale $\mathcal{N}(0,1)$ au lieu de la loi exponentielle. L'instruction `x=randn(1,n)` fournit un vecteur de taille `n` distribué suivant la loi $\mathcal{N}(0,1)$.

4. Méthode de Monte-Carlo.

L'objectif est de calculer $\int_0^1 f(x)\,dx$ pour f fonction de $[0,1]$ dans $\mathbf{R}$. Le principe de la méthode de Monte-Carlo repose sur la loi forte des grands nombres. En effet, si $(U_i)_{i\geq 1}$ est une suite de v.a. i.i.d. suivant la loi uniforme sur $[0,1]$, lorsque $n \to +\infty$,

$$\frac{1}{n}\sum_{i=1}^{n} f(U_i(\omega)) \longrightarrow \mathbb{E}[f(U_1)] = \int_0^1 f(x)\,dx,$$

la convergence ayant lieu pour presque tout $\omega \in \Omega$ dès que $\int_0^1 |f(x)|\,dx < +\infty$.

Revenons à `Matlab`. La commande `n=1000; u=rand(1,n);` fournit un vecteur (ligne) de longueur n dont les composantes sont des réels distribués suivant la loi uniforme sur $[0, 1]$. Les composantes du vecteur `u` s'interprètent comme une réalisation particulière – c'est à dire pour un $\omega \in \Omega$ donné – de n v.a. $U_1, \ldots, U_n$ indépendantes et identiquement distribuées suivant la loi uniforme sur $[0, 1]$. En clair, on considère que

$$\texttt{u=(u\_1,...,u\_n)} = (U_1(\omega), \ldots, U_n(\omega)).$$

C'est parti. Tapez `n=1000; u=rand(1,n);` puis `f=u.*u;` : cette dernière commande élève tous les termes du vecteur `u` au carré. Finalement, la commande `mean(f)` retourne la moyenne arithmétique du vecteur `f`. Cette valeur est proche de 1/3 ? C'est normal puisque $\int_0^1 x^2\,dx = 1/3$. Essayez d'autres valeurs de `n`. Qu'observez-vous quand `n` grandit ?

Essayez à présent cette procédure avec la fonction $x \longmapsto \sin(\pi x)$. Pour cela, il suffit d'utiliser la commande `f=sin(pi*u); mean(f)`. Faites varier `n`.

5. Le premier 6.

Dans cet exercice, on essaie de trouver le temps moyen nécessaire à l'obtention d'un 6 lorsqu'on joue au dé. Mathématiquement, on cherche l'espérance du temps T où

$$T = \inf\{n \geq 1 : X_n = 6\},$$

avec $(X_n)_{n\geq 1}$ i.i.d. suivant la loi uniforme sur $\{1, \ldots, 6\}$.

On peut facilement simuler une v.a. X de loi uniforme sur $\{1, \ldots, 6\}$ à partir d'une v.a. U uniforme sur $[0, 1]$ via la formule $X = 1 + [6U]$.

En `Matlab`, la fonction partie entière s'appelle `floor` : les instructions

```
n=1000; u=rand(1,n); x=1+floor(6*u);
```

fournissent une vecteur `x` de longueur `n` de loi uniforme sur $\{1, \ldots, n\}$. Pour vous en convaincre, tapez la suite d'instructions

```
n=1000; u=rand(1,n); x=1+floor(6*u); mean(x)
```

en vous rappelant que la valeur moyenne d'un dé vaut 3,5. Si vous tapez les commandes

```
n=1000; u=rand(1,n); x=1+floor(6*u); mean(x==6)
```

qu'obtenez-vous ? Pourquoi ?

Nous aurons besoin dans cet exercice des boucles `for` et `while` dont la structure est la suivante :

```
for i=1:mc
    instructions
end
```

```
while condition
    instructions
end
```

Ainsi pour obtenir une réalisation de T, on peut écrire dans le fichier `premiersix.m` :

```
x=0; T=0;
while x<6
    u=rand(1,1); x=1+floor(6*u);
    T=T+1;
end
disp(T);
```

Essayez plusieurs fois en tapant `premiersix`.

Nous allons à présent estimer l'espérance de T en calculant la moyenne arithmétique de `mc` réalisations indépendantes de T. Modifiez `premiersix.m`

```
mc=10000; mT=0;
for i=1:mc
    x=0; T=0;
    while x<6
        u=rand(1,1);
        x=floor(6*u)+1;
        T=T+1;
    end
    % disp(T);
    mT=mT+T;
end
mT=mT/mc; disp(['Temps moyen du premier six : ' num2str(mT)]);
```

Modifiez ce programme pour qu'il détermine également la valeur moyenne de la somme des dés jusqu'à l'obtention du premier 6 ainsi que $\mathbb{P}(T > \mathbb{E}[T] + r)$ où r est un réel de votre choix ; normalement, vous avez constaté que $\mathbb{E}[T] = 6$.

6. Ruine du joueur.

James Bond arrive dans un Casino (Royale) avec a euros en poche. Il joue à la roulette en pariant à chaque fois un euro sur « pair ». Sa fortune S évolue au cours du temps de la manière suivante :

$$S_0 = a, \qquad S_n = a + \sum_{k=1}^{n} X_k, \quad n \geq 1,$$

où les $(X_n)_{n\geq 1}$ sont i.i.d. suivant la loi $\mathbb{P}(X_1 = 1) = 1 - \mathbb{P}(X_1 = -1) = p \in]0, 1[$. Le jeu lui est favorable si $p > 1/2$, défavorable si $p < 1/2$; si $p = 1/2$ le jeu est équitable.

Il décide de s'arrêter de jouer dès qu'il aura gagné b euros à moins bien sûr qu'il ne soit ruiné avant. On s'intéresse donc à la quantité suivante :

$$T = \inf\{n \geq 1 : S_n = 0 \text{ ou } S_n = a + b\}.$$

Écrire dans le fichier `bond.m` un programme permettant d'estimer les quantités : $\mathbb{P}(S_T = 0)$ (James Bond est ruiné), $\mathbb{P}(S_T = a + b)$ (James Bond gagne la partie) ainsi que la durée moyenne de la partie à savoir $\mathbb{E}[T]$.

La théorie donne les valeurs suivantes : pour $p = 1/2$,

$$\mathbb{P}(S_T = 0) = \frac{b}{a+b}, \quad \mathbb{P}(S_T = a + b) = \frac{a}{a+b}, \quad \mathbb{E}[T] = ab,$$

pour $p \neq 1/2$, notant $r = (1 - p)/p$,

$$\mathbb{P}(S_T = 0) = \frac{r^{a+b} - r^a}{r^{a+b} - 1}, \quad \mathbb{P}(S_T = a + b) = \frac{r^a - 1}{r^{a+b} - 1}, \quad \mathbb{E}[T] = \frac{1}{2p-1}\left(\frac{(a+b)(1-r^a)}{1-r^{a+b}} - a\right).$$

Retrouvez-vous ces valeurs ?

À la roulette, il y a 37 numéros : le 0 est vert, 18 pairs rouges et 18 impairs noirs. J. Bond arrive avec 100 euros et veut gagner 10 euros. Quelle est la probabilité qu'il y parvienne ? Combien de fois va-t-il devoir miser ?

7. Le collectionneur.

Chaque paquet de céréales de marque X, contient en cadeau (empoisonné) un autocollant. Il y a m autocollants différents. Le problème est de savoir combien de paquets de céréales sont nécessaires en moyenne pour obtenir tous les autocollants.

Créer un fichier `collection.m` qui permet d'obtenir le nombre `N` de paquets nécessaires à l'obtention de la collection complète. On supposera que les autocollants se trouvant dans les paquets de céréales sont pris au hasard parmi les m autocollants possibles.

Modifiez votre programme afin qu'il fournisse une approximation du nombre moyen de paquets de céréales nécessaire. Comparez votre résultat à la valeur théorique :

$$\mathbb{E}[N] = 1 + m \sum_{k=1}^{m-1} \frac{1}{k}.$$

On prendra `m=50` pour les applications numériques.

On suppose à présent qu'un enfant collectionne les images des joueurs des équipes françaises de ligue 1. Chaque paquet contient `nv=4` images, les « doublons » étant bien évidemment permis. Sachant que le prix d'un paquet de `nv=4` vignettes est de 0,20 euros, combien faut-il dépenser en moyenne pour remplir l'album qui comporte 400 images ?

8. Polynôme de Berstein.

Soient $(U_i)_{i\geq 1}$ une suite de variables aléatoires i.i.d. de loi uniforme sur $[0,1]$ et $f : [0,1] \longrightarrow \mathbf{R}$ une fonction continue. On note $B_n(f)$ le polynôme de Bernstein

$$B_n(f)(x) = \mathbb{E}\left[f\left(M_n(x)\right)\right], \qquad M_n(x) = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{U_i \leq x},$$

et on rappelle que $B_n(f)$ converge vers f uniformément sur $[0,1]$.

1. En choisissant la fonction $f(x) = \sqrt{x}$, illustrer cette convergence. On pourra tracer sur le même graphique f ainsi que $B_n(f)$ pour différentes valeurs de n.
2. Pour $f(x) = \sqrt{x}$ puis $f(x) = |x|$ tracer le graphe $n\,|f - B_n(f)|$. Quelle conclusion en tirez-vous ?

9. Théorème central limite.

En traçant un histogramme, illustrer la convergence en loi vers une variable aléatoire de loi $\mathcal{N}(0,1)$ de

$$T_n = \frac{\sqrt{n}}{\sqrt{\sigma^2}}\left(\frac{1}{n}\sum_{k=1}^{n} X_k - \mu\right),$$

où $\mu = \mathbb{E}[X_1]$, $\sigma^2 = \mathbb{V}(X_1)$ et les $(X_i)_{i\geq 1}$ sont i.i.d..

On testera différentes loi pour X_1.