TP 5

Simulation de variables aléatoires à densité

Compétences attendues.

- ✓ Savoir simuler une loi continue usuelle à l'aide de la fonction `grand`, ou uniquement à partir de la fonction `rand()`.
- ✓ Vérifier graphiquement la pertinence d'une simulation d'une loi.

Mathieu Mansuy - Professeur de Mathématiques en ECS2 au Lycée Louis Pergaud (Besançon)
`mansuy.mathieu@hotmail.fr`

1 Simulation des lois usuelles

1.1 Fonctions Scilab

`Scilab` dispose de fonctions simulant les lois continues usuelles : les fonctions `rand` et `grand`.

Définition.

- `rand()` renvoie une réalisation d'une variable aléatoire suivant la loi $\mathscr{U}([0,1])$.

  `rand(N,M)` renvoie une matrice de taille $N \times M$ contenant des réalisations de variables aléatoires indépendantes suivant la loi $\mathscr{U}([0,1])$.

- `grand(N,M,"loi",paramètres)` renvoie une matrice de taille $N \times M$ contenant des réalisations de variables aléatoires indépendantes suivant cette loi.

  `grand(N,M,"unf",a,b)` simule la loi $\mathscr{U}([a,b])$.

  `grand(N,M,"exp",1/lambda)` simule la loi $\mathscr{E}(\lambda)$.

  `grand(N,M,"gam",nu,1)` simule la loi $\gamma(\nu)$.

  `grand(N,M,"nor",m,sigma)` simule la loi $\mathscr{N}(m,\sigma^2)$.

Dans la suite de cette partie, on explique comment simuler les lois usuelles en utilisant uniquement la fonction `rand`.

1.2 Loi uniforme

Exercice 1

Soit a et b deux réels avec $a < b$. On rappelle que si $U \hookrightarrow \mathscr{U}([0,1])$, alors $(b-a)U + a \hookrightarrow \mathscr{U}([a,b])$.

Écrire une fonction `uniforme(a,b)` simulant la loi $\mathscr{U}([a,b])$ uniquement à l'aide de la fonction `rand()`.

1.3 Loi normale

Propriété 1

- On suppose que $U_1, \ldots, U_{12}$ sont des variables aléatoires mutuellement indépendantes, suivant toutes la loi $\mathscr{U}([0,1])$. On pose $X = \sum_{i=1}^{12} U_i - 6$.

  D'après le *Théorème de la limite centrée*, on peut considérer que X suit approximativement la loi normale centrée réduite.

- Soit $m \in \mathbb{R}$ et $\sigma \in]0,+\infty[$. Si $X \hookrightarrow \mathscr{N}(0,1)$, alors $\sigma X + m \hookrightarrow \mathscr{N}(m,\sigma^2)$.

Exercice 2

1. Écrire une fonction `norcentreereduite()` simulant la loi $\mathscr{N}(0,1)$ à l'aide de la fonction `rand`.
2. Écrire une fonction `normale(m,var)` simulant la loi $\mathscr{N}(m,var)$ à partir de la fonction `norcentreereduite()`.
3. Écrire une fonction `Normale(N,m,var)` donnant un vecteur contenant N réalisations de la loi $\mathscr{N}(m,var)$.

2 Méthode d'inversion

2.1 Principe

Théorème 2

On suppose que X est une variable aléatoire à densité dont la fonction de répartition F est strictement croissante de $]a,b[$ $(-\infty \leq a < b \leq +\infty)$ sur $]0,1[$. Alors :

- F réalise une bijection de $]a,b[$ sur $]0,1[$;
- si $U \hookrightarrow \mathscr{U}(]0,1[)$, alors $F^{-1}(U)$ suit la même loi que X.

Exercice 3

1. Rappeler l'expression de la fonction de répartition de $U \hookrightarrow \mathscr{U}([0,1])$.
2. Démontrer le théorème précédent.

Méthode.

Soit X une variable aléatoire à densité dont on connait une expression de F^{-1}. Pour simuler une variable de même loi que X, on procèdera comme suit :

- *on choisit un paramètre t de manière aléatoire dans $[0,1]$ à l'aide de la fonction* `rand()` *;*
- *on retourne $F^{-1}(t)$.*

2.2 Simulation de la loi exponentielle

Exercice 4

1. Rappeler l'expression de la fonction de répartition d'une loi exponentielle, et montrer qu'elle réalise une bijection de $\mathbb{R}_+^*$ sur $]0,1[$.

   Déterminer sa bijection réciproque.

2. (a) Écrire une fonction `exponentielle(lambda)` simulant une loi $\mathscr{E}(\lambda)$ à partir de la fonction `rand()`.

   (b) Écrire une fonction `Exponentielle(N,lambda)` donnant un échantillon de taille N de la loi $\mathscr{E}(\lambda)$.

   (c) Créer un vecteur de taille 10000 contenant 10000 simulations d'une variable aléatoire suivant la loi $\mathscr{E}(1/2)$.

   En utilisant les commandes `mean` et `stdev`, vérifier que la moyenne et l'écart-type empiriques (c'est-à-dire de ce vecteur) sont bien conformes à ce qu'on attend.

3. (a) Écrire une fonction `gamma(n)` simulant la loi $\gamma(n)$ pour tout $n \in \mathbb{N}^*$.

   (b) Écrire une fonction `Gamma(N,n)` renvoyant un vecteur contenant N réalisations de la loi $\gamma(n)$.

2.3 Simulation de la loi de Cauchy

Exercice 5

On rappelle qu'une variable X suit une loi de Cauchy si elle admet pour fonction de répartition la fonction :

$$F : x \in \mathbb{R} \mapsto \frac{1}{\pi}\left(\arctan(x) + \frac{\pi}{2}\right).$$

Une densité de X est alors la fonction $f : t \mapsto \frac{1}{\pi}\frac{1}{1+t^2}$.

1. Montrer que F réalise une bijection de $\mathbb{R}$ sur $]0,1[$, et déterminer F^{-1}.
2. (a) Écrire une fonction `cauchy()` simulant une loi de Cauchy à partir de la fonction `rand`.
   (b) Écrire une fonction `Cauchy(N)` donnant un échantillon de taille N de la loi de Cauchy.
3. (a) Créer un vecteur de taille 10000 contenant 10000 simulations d'une variable aléatoire suivant la loi de Cauchy.
   (b) Utiliser la commande `mean` avec ce vecteur pour évaluer l'espérance d'une loi de Cauchy. Recommencer avec plusieurs échantillons. Que constatez-vous ? Une variable suivant la loi de Cauchy admet-elle une espérance ?

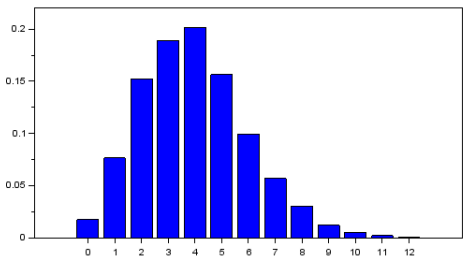
3 Représentations graphiques

Soit X une variable aléatoire à densité. Supposons qu'on dispose d'une fonction `Loi` permettant de simuler la loi de X. Pour juger de la qualité de cette simulation, on va utiliser des représentations graphiques. Pour cela, on procèdera comme suit :

- on crée un échantillon de taille N, c'est-à-dire un vecteur ligne `x` contenant N réalisations de la fonction `Loi` ;
- on compare graphiquement les fréquences empiriques obtenues grâce à l'échantillon avec les probabilités théoriques pour vérifier la pertinence de la simulation.

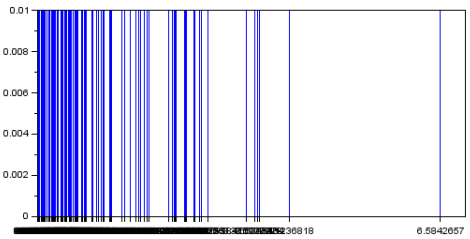
Remarque. Pour une variable discrète, afin de comparer nos résultats à la théorie, nous avions regroupé les simulations par modalités, et tracé le diagramme en bâtons des fréquences. Par exemple, si `x` est un vecteur contenant 10000 simulations d'une loi de Poisson de paramètre 4, on peut procéder ainsi :

```
x = grand(1,100,'poi',1)
y = tabul(x,'i')
bar(y(:,1),y(:,2)/10000)
```



Nous n'allons pas pouvoir procéder de cette manière dans le cas d'une variable à densité X. En effet, on a $P(X = x) = 0$ pour tout $x \in \mathbb{R}$, et chaque modalité de notre échantillon risque donc d'avoir un effectif égal à 1. Le tri par modalités n'est donc pas adapté ici, et la représentation à l'aide d'un diagramme en bâtons non plus. Voici par exemple ce qu'on obtient pour la loi $\mathscr{E}(1)$:

```
x = grand(1,100,'exp',1)
y = tabul(x,'i')
bar(y(:,1),y(:,2)/100)
```



Ce graphique est bien difficile à interpréter, et ne nous permet pas de reconnaître une loi exponentielle.

Nous proposons ici deux façons d'évaluer la qualité de nos simulations dans le cas de variables aléatoires continues :

- en comparant l'*histogramme des fréquences d'un échantillon* à la densité théorique ;
- en comparant la *fonction de répartition empirique d'un échantillon* à la fonction de répartition théorique de la loi.

3.1 Comparaison histogramme des fréquences / densité

Le nombre de modalités de notre échantillon `x` étant a priori très grand, chacune avec un effectif de 1, nous allons les regrouper par classe. Afin de définir ces classes, on fixe une suite de réels strictement croissante :

$$c = (c_0 < c_1 < \cdots < c_p).$$

On va alors comparer :

- l'*histogramme des fréquences de l'échantillon*, défini par les rectangles de base les classes $[c_i, c_{i+1}]$ et d'aires les fréquences d'appartenance à ces classes pour notre échantillon `x` ;
- la courbe représentative d'une densité f de la loi de X.

Si notre simulation suit bien la loi attendue, on doit constater que :

Théorème 3 (*Théorème d'or de Bernoulli*)

Pour N « suffisamment grand », la fréquence observée pour la classe $[c_i, c_{i+1}]$ est proche de la probabilité théorique $P(c_i \leq X \leq c_{i+1}) = \displaystyle\int_{c_i}^{c_{i+1}} f(t)\,\mathrm{d}t$.

Graphiquement, on devrait donc observer que l'aire du rectangle de base $[c_i, c_{i+1}]$, qui est égale à la fréquence d'appartenance à cette classe, est proche de l'aire sous la courbe représentative de f entre c_i et c_{i+1}.

On va pour cela avoir besoin des commandes suivantes.

Définition.

- Si `x` et `c` sont des vecteurs, `c` à composantes strictement croissantes définissant les classes, `histplot(c,x)` trace l'histogramme des fréquences de `x` pour les classes définies par `c`.
- Si `x` est un vecteur et `n` un entier naturel non nul, `histplot(n,x)` trace l'histogramme des fréquences de `x` à `n` classes (qui sont alors équiréparties entre la plus petite et la plus grande valeur de `x`).

Exercice 6

1. Simuler avec la fonction `Exponentielle` $N = 10000$ valeurs de la loi $\mathscr{E}(0.5)$.
2. Tracer la courbe représentative de la densité f de la loi $\mathscr{E}(0.5)$.
3. Tracer l'histogramme des fréquences de l'échantillon obtenu (on prendra pour cela une subdivision c de l'intervalle $[0, 10]$ en $p = 100$ intervalles de même longueur).

   Comparer l'histogramme des fréquences de l'échantillon à la courbe représentative de f. Qu'en pensez vous ?
4. Procéder de même pour la loi $\gamma(3)$.

3.2 Comparaison fonction de répartition empirique / théorique

On propose dans cette section une deuxième méthode pour juger de la qualité de la simulation d'une loi de probabilité à densité. On va comparer :

- la *fonction de répartition empirique* : il s'agit de la fonction en escaliers qui à un réel x associe la fréquence d'apparition des nombres inférieurs ou égaux à x dans l'échantillon `x` ;
- la fonction de répartition théorique.

Exemple. Prenons $X \hookrightarrow \mathscr{U}([0,1])$ et un échantillon `x` de taille $N = 6$ de la loi de X à l'aide de la fonction `rand`.

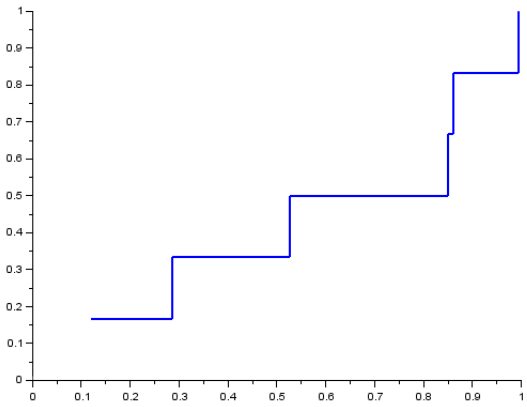
```
--> x = rand(1,6)
ans =
            column 1 to 5

    0.1205996    0.2855364    0.8607515    0.8494102    0.5257061

         column 6

    0.9931210
```

On représente ci-dessous la courbe de la fonction de répartition empirique associée à cet échantillon.



On notera qu'elle présente une discontinuité en chaque modalité, et que la hauteur des sauts est constante égale à $\frac{1}{N}$. En particulier :

- l'abscisse en une modalité est égale à sa fréquence cumulée, c'est-à-dire à la somme de toutes les fréquences des modalités qui lui sont inférieures ;
- un point d'abscisse x de la courbe a pour ordonnée la fréquence d'apparition des simulations inférieures ou égales à x.

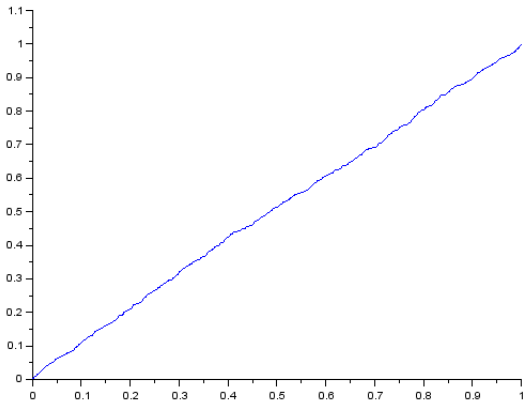
On doit observer que :

Théorème 4 (*Théorème d'or de Bernoulli*)

Pour N « suffisamment grand », la fréquence observée des modalités plus petites que x est proche de la probabilité $P(X \leq x)$.

Graphiquement, la courbe (en escalier) de la fonction de répartition empirique doit donc être proche de la courbe de la fonction de répartition théorique si notre simulation correspond à la loi attendue.

Exemple. Reprenons le cas où $X \hookrightarrow \mathscr{U}([0,1])$, avec cette fois un échantillon `x` de taille $N = 1000$ de la loi de X obtenu à l'aide de la fonction `rand`. La courbe de la fonction de répartition empirique donne alors :



Cette courbe étant proche de celle de la fonction de répartition théorique de la loi $\mathscr{U}([0,1])$, on peut donc conclure que les simulations obtenues à l'aide de la fonction `rand` suivent bien la loi $\mathscr{U}([0,1])$.

Pour tracer la fonction de répartition empirique, nous aurons pour cela besoin de la commande suivante.

Définition.

`plot2d2(x,y)` effectue un tracé en escalier d'abscisse `x` et d'ordonnée `y`.

Méthode.

Pour tracer la fonction de répartition empirique d'un échantillon `x`*, on procède ainsi :*

```
y=tabul(x,"i")
plot2d2(y(:,1),cumsum(y(:,2)/length(x)))
```

Exercice 7

1. Expliquer les lignes de commandes proposées pour tracer la fonction de répartition empirique.
2. Simuler avec la fonction `Cauchy` $N = 10000$ réalisations de la loi de Cauchy.
3. Tracer la fonction de répartition empirique de l'échantillon obtenu et la fonction de répartition théorique de cette loi. Comparer.

   On pourra utiliser la commande `atan` *qui calcule l'arctangente.*

On va maintenant tester notre simulation de la loi $\mathscr{N}(m,\sigma^2)$. Il nous faut pour cela tracer la fonction de répartition théorique de cette loi. On utilise la commande `Scilab` suivante.

Définition.

`[P,Q]=cdfnor("PQ",x,m,sigma)` donne $P = F(x)$ où F est la fonction de répartition de la loi $\mathscr{N}(m,\sigma^2)$, et $Q = 1 - P$.

Exercice 8

1. Calculer $\Phi(0), \Phi(1), \Phi(1.96)$.
2. Soit X une variable aléatoire suivant la loi $\mathscr{N}(3,4)$. Calculer $P(X > 10)$, $P(0 \leq X < 3)$.

3. Tester les simulations obtenues des lois $\mathcal{N}(0,1)$ et $\mathcal{N}(2,9)$ en traçant les fonctions de répartition théoriques et empiriques.

4. Comparer ces résultats avec ceux obtenus par la fonction `grand`.

4 Exercices

Exercice 9 (★★)

Soit $a \in \mathbb{R}_+^*$ et $(X_1, \dots, X_n)$ une famille de variables aléatoires indépendantes, identiquement distribuées suivant une loi uniforme sur $[0, a]$. On pose :

$$U = \min(X_1, \dots, X_n) \quad \text{et} \quad V = \max(X_1, \dots, X_n).$$

1. Déterminer les lois de U et V.

2. On considère la fonction suivante :

```
function couple =simulation(a,n)
    nb_sim = 10000;
    R = grand(nb_sim,n,"unf",0,a);
    couple = [min(R,"c") max(R,"c")];
endfunction
```

   Expliquer la fonction `simulation` (on pourra consulter l'aide des fonctions `min` et `max` pour cela).

3. Prenons $n = 10$ et $a = 1$. Comparer graphiquement la qualité de cette simulation (fonction de répartition empirique/théorique).

Exercice 10 (★★ - Simulation de la loi normale à l'aide du TLC)

Soit $(X_k)_{k\geq 1}$ une famille de variables aléatoires mutuellement indépendantes et toutes de même loi. Notons $\overline{X_n} = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k$. D'après le *Théorème de la limite centrée*, la variable centrée réduite $\overline{X_n}^* = \dfrac{\overline{X_n} - E(\overline{X_n})}{\sqrt{V(\overline{X_n})}}$ suit approximativement une loi normale $\mathcal{N}(0,1)$ lorsque n est suffisamment grand.

1. On suppose dans cette question que les variables X_k suivent toutes la loi $\mathscr{U}(\llbracket 1, 6 \rrbracket)$.

   (a) Écrire une fonction `norcentreereduite2(N,n)` donnant N réalisations de la variable $\overline{X_n}^*$ dans ce cas.

   (b) Évaluer la qualité de cette simulation de la loi $\mathcal{N}(0,1)$ selon les valeurs de n. Pour quelles valeurs de n cette simulation semble pertinente ?

2. Mêmes questions en supposant que les variables X_k suivent toutes la loi $\mathscr{E}(1)$.

3. Comparer les différentes méthodes de simulation de la loi $\mathcal{N}(0,1)$ ainsi obtenues.

Exercice 11 (★★★ - Différentes simulations de la loi de Pareto)

Soit $k \in \mathbb{N}^*$ et $\lambda > 0$.

1. Déterminer la valeur de r pour laquelle $f_\lambda : x \to \begin{cases} 0 & \text{si } x \leq \lambda \\ \dfrac{r}{x^{k+1}} & \text{sinon} \end{cases}$ est une densité de probabilité.

   Si X admet pour densité f_λ, on dit que X suit la loi de Pareto de paramètre λ et k.

2. Déterminer la fonction de répartition d'une variable suivant la loi de Pareto de paramètres λ et k.

3. En utilisant la méthode d'inversion, simuler une variable aléatoire suivant une loi de Pareto de paramètres λ et k.

4. Soient $X_1, \cdots, X_k$ k variables aléatoires indépendantes suivant toutes la loi uniforme sur [0,1]. On pose alors $Y = \dfrac{\lambda}{\max(X_1, \cdots, X_k)}$.

   (a) Montrer que Y suit une loi de Pareto de paramètres λ et k.

   (b) En déduire une autre méthode pour simuler la loi de Pareto.

5. Comparons à présent les deux méthodes obtenues de simulation d'une loi de Pareto.

   (a) En simulant 100000 réalisations d'une loi de Pareto à l'aide de chacune de ces méthodes, déterminer laquelle des deux est la plus rapide.

   On pourra à cet effet utiliser les commandes `tic` *et* `toc`*. Pour cela, on exécute la commande* `tic()` *juste avant le début de la simulation, puis la commande* `toc()` *juste après. Le résultat retourné est alors le temps (en secondes) qui a été nécessaire à l'exécution du calcul.*

   (b) Prenons $\lambda = 1$. Comparer graphiquement la qualité de chacune des deux simulations d'une loi de Pareto (histogramme des fréquences/densité théorique).
