

Deuxième Année Licence M.I.A.S.H.S. 2016 – 2017

TP de Méthodes Numériques n° 7 :  
Simulation de réalisations de variables aléatoires et théorèmes  
limites

Dans ce TP, on commence par proposer des méthodes permettant de simuler le hasard à partir d'une machine déterministe qu'est un ordinateur: cela correspond à la génération de nombres pseudo-aléatoires suivant une loi uniforme sur  $[0, 1]$ . Ensuite, on utilise ces nombres pour générer des nombres pseudo-aléatoires distribués suivant une loi quelconque. Enfin on utilise ces nombres, comme réalisations de variables aléatoires indépendantes et de même loi, pour illustrer ce que sont concrètement la loi des grands nombres et le théorème de la limite centrale.

Génération de nombres pseudo-aléatoires

Le problème que nous nous posons maintenant est: comment générer des nombres pris "au hasard" entre 0 et 1 ? En termes plus mathématiques, la question se pose sous la forme suivante: si  $X : \Omega \rightarrow [0, 1]$  est une variable aléatoire de distribution uniforme sur  $[0, 1]$ , comment produire une réalisation de  $X$ , c'est-à-dire  $X(\omega)$  avec  $\omega \in \Omega$ ? Et si  $(X_1, \dots, X_n)$  sont  $n$  variables indépendantes de même loi que  $X$ , comment générer une réalisation de  $(X_1, \dots, X_n)$ , soit  $(X_1(\omega), \dots, X_n(\omega))$ ? Ces questions se posent car un processeur fonctionne de façon intégralement déterministe, donc a priori sans possibilité de produire du "hasard". Pourtant certaines fonctions déterministes peuvent approcher le "hasard" et les réalisations de  $(X_1(\omega), \dots, X_n(\omega))$  peuvent être extrêmement bien imitées par ce que l'on appelle des nombres pseudo-aléatoires...

---

Première génération de nombres pseudo-aléatoires

---

27%10 On calcule 27 modulo 10, soit le reste de la division euclidienne de 27 par 10: on obtient 7 car  $27 = 2 * 10 + 7$ .  
 19%%7 Autre exemple. Faire suffisamment d'exemples jusqu'à avoir compris %%.  
 Ouvrir un fichier *pseudo1.m* et y créer une fonction *pseudo1* telle que  $y = pseudo1(x) = 3 * x \% 10$ . Générer la suite  $(x_n)$  avec  $n = 0, 1, \dots, 20$  telle que  $x_{n+1} = pseudo1(x_n)$  et  $x_0 = 0$ . Par ce biais, on a généré des nombres qui semblent se comporter comme des chiffres entre 0 et 9 pris au "hasard" de façon équiprobable et indépendante. Cependant, en générant  $(x_n)$  pour  $n = 1, \dots, 100$  un motif semble se répéter de façon périodique. Lequel? Quelle période? Pourquoi? Ceci met en défaut la prétention de générer parfaitement ainsi des nombres "aléatoires". Représenter maintenant les points de coordonnées  $(x_n, x_{n+1})$  pour  $n = 0, 1, \dots, 99$ . Quelles sont les couples possibles pour  $(x_n, x_{n+1})$ ? Faire la même chose avec les couples  $(x_n, x_{n+2})$  et  $(x_n, x_{n+7})$ . Conclusions? Ceci met en défaut la prétention à générer des variables indépendantes.  
 Noter qu'en divisant les  $x_n$  par 10, on générerait ainsi des nombres de  $[0, 1]$ .

---

Seconde génération de nombres pseudo-aléatoires

---

Même s'il n'est pas parfait, le procédé précédent donne une idée d'un générateur de nombres pseudo-aléatoires plus performant. Définir le script *pseudo2* suivant

```
Pseudo2 = fonction(N,x0)
{ a=7^7; b=2^31-1
x=c(1:N); X=c(1:N)
x[1]=x0; X[1]=(x0%%b)/b
for (j in c(1:(N-1)))
{ x[j+1]=(a*x[j]+1)%%b
X[j+1]=x[j+1]/b }
return(X) }
X=Pseudo2(1000,123456); hist(X)
```

Implémenter ce programme. Faire des essais avec d'autres  $x_0$ . Expliquer la forme de l'histogramme. Déconnectez alors votre session R. Taper  $Y = \text{runif}(1000, 0, 1)$ , puis tracer l'histogramme. Qu'en pensez-vous? Noter les 3 premières coordonnées de  $Y$ . Fermer votre session R puis relancer la et taper à nouveau  $Y = \text{runif}(1000, 0, 1)$ . A-t-on obtenu le même résultat que celui noté? Si vous faites la même démarche avec  $X = \text{Pseudo2}(1000, 123456)$ , que remarque-t-on? Pour pallier à ce défaut, trouver un moyen de définir  $x_0$  en fonction du temps  $\text{cpu}$  de votre session R.

## Simulation d'autres lois que la loi uniforme

Il est bien clair qu'en effectuant une opération simple sur un nombre pseudo-aléatoire défini comme précédemment on peut générer une "pseudo"-réalisation d'une variable suivant une loi uniforme sur  $[a, b]$  (quelle opération?).

On peut aussi générer une "pseudo"-réalisation d'une variable discrète  $Z$  telle que  $Z$  prenne ses valeurs dans  $\{z_1 < z_2 < \dots < z_n\}$  et  $\mathbb{P}(Z = z_i) = p_i$  avec  $p_1 + p_2 + \dots + p_n = 1$  (par exemple une variable de Bernoulli de paramètre  $p$ ). On suppose que  $X$  est une variable aléatoire suivant une loi uniforme sur  $[0, 1]$  et l'on pose  $W = \sum_{i=1}^n z_i \mathbb{I}_{p_0 + p_1 + \dots + p_{i-1} < X \leq p_0 + p_1 + \dots + p_i}$ , en posant par convention  $p_0 = 0$ . En considérant leurs fonctions de répartition, montrer que  $W$  a la même loi que  $Z$ . En déduire une suite de commande simple à partir de  $\text{runif}(n, 0, 1)$  pour:

1. Simuler  $n$  "pseudo"-réalisations d'une variable de Bernoulli de paramètre  $1/2$ , puis  $1/4$ . La commande  $\text{rbinom}(n, 1, 1/4)$  de R permet le même résultat.
2. Simuler  $n$  "pseudo"-réalisations des chiffres indiqués par un dé équilibré (on vérifiera également que  $1 + \text{floor}(\text{runif}(n, 0, 6))$  permet d'obtenir le même résultat).

Supposons maintenant que  $Z$  est une variable dite "continue" (on dira plutôt en L3 "absolument continue par rapport à la mesure de Lebesgue") dont on suppose la fonction de répartition  $F_Z(z) = \mathbb{P}(Z \leq x)$  strictement croissante sur l'intervalle ouvert constitué par les valeurs possibles de  $Z$ , intervalle que l'on notera  $I$  ( $I$  sera par exemple  $\mathbf{R}$  pour variable gaussienne,  $]0, \infty[$  pour une variable exponentielle). Montrer alors que  $F_Z^{-1}$  existe comme fonction de  $]0, 1[$  dans  $I$ . Montrer alors que  $W = F_Z^{-1}(X)$ , où  $X$  est une variable aléatoire suivant une loi uniforme sur  $[0, 1]$ , est une variable qui suit la même loi que  $Z$ . En déduire une suite de commande simple à partir de  $\text{runif}(n, 0, 1)$  pour:

1. Simuler  $n$  "pseudo"-réalisations d'une variable de loi exponentielle de paramètre 2. La commande  $\text{rexp}(n, 2)$  de R permet le même résultat.
2. Simuler  $n$  "pseudo"-réalisations d'une variable dite de Cauchy, dont la densité est  $f(x) = \frac{1}{\pi} (1 + x^2)^{-1}$  pour  $x \in \mathbf{R}$ . La commande  $\text{rexp}(n)$  de R permet le même résultat.

Expliquer pourquoi une telle méthode peut être utilisée pour simuler une variable gaussienne, mais avec de sérieuses difficultés...

Justement, nous allons voir un moyen simple pour générer 2 variables gaussiennes centrées réduites indépendantes. C'est ce que l'on appelle la méthode de Box-Muller.

```
U=runif(2,0,1)
V1=sqrt(-2*log(U[1]))
Z1=V1*sin(2*pi*U[2])
Z2=V1*cos(2*pi*U[2])
```

Essayer ce programme. On peut montrer que le couple  $(Z1, Z2)$  ainsi construit est un couple de pseudo-réalisations de 2 variables gaussiennes centrées réduites indépendantes. L'idée de cette simulation est le fait que dans le plan cartésien, si on considère un point  $M$  de coordonnées  $(Z1, Z2)$ , l'angle  $\widehat{Ox, OM}$  suit une loi uniforme sur  $[0, 2\pi]$  et la distance  $OM$  suit une loi exponentielle de paramètre 0.5.

Transformer le programme pour pouvoir simuler 1000 pseudo-réalisations de variables indépendantes de loi  $\mathcal{N}(-2, 4^2)$ . Afficher l'histogramme de cet échantillon. Qu'obtient-on?

## Loi des grands nombres et théorème de la limite centrale

On rappelle d'abord un résultat essentiel qui légitime le recours possible des probabilités pour traiter des données.

**Loi faible des Grands Nombres:** Soit  $(X_n)_{n \in \mathbf{N}}$  une suite de variables aléatoires indépendantes et identiquement distribuées sur le même espace de probabilité  $(\Omega, \mathcal{A}, \mathbb{P})$ . On suppose que  $\mathbb{E}(|X_0|) < \infty$ . Alors  $\bar{X}_n = \frac{1}{n} (X_1 + X_2 + \dots + X_n) \xrightarrow[n \rightarrow +\infty]{\mathcal{P}} m = \mathbb{E}(X_0)$ .

On rappelle que la convergence en probabilité signifie heuristiquement que quand  $n$  devient grand alors  $\bar{X}_n$  se rapproche de  $m$ , avec une probabilité tendant vers 0 de s'en écarter d'une distance plus grande que  $\varepsilon$ , et ceci pour tout  $\varepsilon > 0$  (ou encore  $\mathbb{P}(|\bar{X}_n - m| \geq \varepsilon) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} 0$ ).

Vérifier cette propriété en simulant d'abord  $N = 10^4$  pseudo-réalisations de toutes les lois précédemment définies, puis en calculant la moyenne empirique  $\bar{X}_n$  (à l'aide de la commande `mean(.)`) pour tout  $n \in \{1, 2, \dots, N\}$ . On tracera alors les points  $(n, \bar{X}_n)$  pour  $n \in \{1, 2, \dots, N\}$  et on comparera aux valeurs de  $m = \mathbb{E}(X_0)$  calculées théoriquement. Est-ce que l'on observe une convergence pour toutes les lois?

La Loi des Grands Nombres permet de relier (asymptotiquement) une quantité calculée à partir de données (une moyenne empirique) à une valeur purement théorique (l'espérance). C'est le début de ce que l'on appelle la statistique inférentielle, qui permet par exemple de dire qu'à une élection présidentielle entre 2 candidats, si l'on interroge 1000 personnes choisies au hasard, alors le pourcentage de voix obtenu pour le candidat  $A$  ne sera pas très éloigné de celui qu'il aura sur l'ensemble des votants.

Si l'on veut aller un peu plus loin et par exemple obtenir des intervalles de confiance ou mettre en place des tests, il peut être intéressant de mesurer l'écart entre la moyenne empirique et la moyenne théorique qu'est l'espérance. C'est ce que précise le Théorème de la Limite Centrale:

**Théorème de la Limite Centrale:** Soit  $(X_n)_{n \in \mathbb{N}}$  une suite de variables aléatoires indépendantes et identiquement distribuées sur le même espace de probabilité  $(\Omega, \mathcal{A}, \mathbb{P})$ . On suppose que  $\sigma^2 = \text{var}(|X_0|) < \infty$ . Alors  $\sqrt{n}(\bar{X}_n - m) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2)$ .

Heuristiquement, ce théorème signifie que quand  $n$  est grand, alors la loi de la variable aléatoire  $\sqrt{n}(\bar{X}_n - m)$  est à peu près une loi gaussienne centrée de variance  $\sigma^2$ . Pour illustrer ce résultat, un seul échantillon  $(X_1, \dots, X_n)$  ne signifie pas. Simuler ainsi  $M$  fois des pseudo-échantillons  $(X_1, \dots, X_n)$  qui suivent une des lois précédentes (par exemple la loi exponentielle de paramètre 2). On supposera  $M$  et  $n$  grands (par exemple  $M = n = 1000$ ). Tracer alors l'histogramme des  $M$  valeurs  $\sqrt{n}(\bar{X}_n - m)$  calculées à partir des échantillons. On retrouve alors un histogramme de type courbe en cloche pour toutes les lois simulées sauf une (laquelle?).

## Exercices

1. Simuler une pseudo-réalisation de loi binomiale de paramètre  $(100, 1/3)$  à partir de pseudo-réalisations de variables uniformes sur  $[0, 1]$ .
2. On suppose que la durée de vie d'une ampoule en mois suit une loi exponentielle de paramètre  $\theta = 1/15$ . On voudrait estimer ce paramètre  $\theta$  en ayant observé 30 ampoules, qui sont toutes décédées. Pour cela, simuler 30 pseudo-réalisations indépendantes de loi exponentielle de paramètre  $\theta$ . En utilisant les théorèmes précédents, estimer  $\theta$  à partir de ces 30 observations et vérifier que  $\theta = 1/15$  est bien dans l'intervalle de confiance à 95% obtenu à partir du Théorème de la Limite Centrale.