Niveau : M1 IS
Cours : Introduction au logiciel R
Enseignant : Aude Illig

Université Versailles Saint-Quentin
Année 2014/2015 - 1er semestre

TP 4 : Simulation avec R

**** Manipulation des objets* `vector` *et* `data.frame` *****

Table des matières

1 Simulation de variables aléatoires

Dans certaines études statistiques, il est nécessaire de disposer d'échantillons de taille prescrite pour une loi de probabilité connue. Lorsque l'expérimentation physique ne le permet pas, on fabrique ces échantillons à l'aide d'un algorithme adapté à une implémentation informatique. La création d'un échantillon de loi donnée repose alors sur la génération de réalisations de loi uniforme $\mathcal{U}[0,1]$ obtenues à l'aide de générateurs de nombres pseudo-aléatoires.

1.1 Nombres pseudo-aléatoires

Un générateur de nombres pseudo-aléatoires est un algorithme qui génère une suite de nombres ayant en apparence certaines propriétés du hasard. Les méthodes les plus employées reposent sur une congruence linéaire (qui fournit malheureusement des suites périodiques). Ainsi, Derrick Henry Lehmer propose en 1948 la formule de récurrence suivante :

$$x_{n+1} = a\,x_n \bmod m$$

avec x_0 la "graine" (généralement un nombre premier) utilisée pour initialiser l'algorithme. La période d du générateur étant au maximum égale à m, en pratique on choisit m le plus grand possible.

Exemple. *L'algorithme Minimal Standard utilise* $m = 2^{31} - 1$ *et* $a = 7^5 = 16807$. *Sa période est* $d = m - 1$.

Enfin, les nombres renormalisés

$$u_i = \frac{x_i}{m} \quad i = 1 \dots d$$

sont des nombres compris entre 0 et 1 que l'on considère comme aléatoires. Cependant la qualité des nombres générés varie selon le générateur utilisé et peut être étudiée au moyen de tests statistiques.

Exemple. *L'algorithme Mersenne Twister (par défaut en R) inventé par Makoto Matsumoto et Takuji Nishimura en 1997 est particulièrement réputé pour sa qualité :*

1. *sa période est grande ;* $d = 2^{19937} - 1$,
2. *il est très rapide*
3. *il passe les tests diehard de Georges Marsaglia.*

1.2 Simulation par inversion

Sachant maintenant obtenir des réalisations d'une loi uniforme $\mathcal{U}[0,1]$, plusieurs méthodes peuvent être utilisées pour obtenir les réalisations d'une loi de probabilité donnée. La méthode que nous décrivons ici repose sur la connaissance de l'inverse ou de l'inverse généralisée de la fonction de répartition de la variable à simuler. Bien qu'elle s'applique à de nombreuses lois, elle ne permet pas de simuler les lois normales par exemple.

Remarque. Les lois normales et plus généralement certaines lois à densité sont simulables par la méthode d'acceptation-rejet qui consiste à majorer (à une constante près) la densité par une autre densité selon laquelle on sait déjà simuler.

1.2.1 Inverse de la fonction de répartition

Remarquons que si X est une variable aléatoire de fonction de répartition $F_X(t) = \mathbb{P}(X \leq t)$ inversible, alors la variable $U = F_X(X)$ suit une loi uniforme $\mathcal{U}[0,1]$. En effet, pour tout $y \in [0,1]$:

$$\begin{aligned} \mathbb{P}(U \leq y) &= \mathbb{P}(F_X^{-1}(U) \leq F^{-1}(y)), \\ &= \mathbb{P}(X \leq F^{-1}(y)), \\ &= F_X(F_X^{-1}(y)) = y. \end{aligned}$$

Plus généralement, on a le résultat suivant :

Proposition 1. *Soit X une variable aléatoire réelle de fonction de répartition $F_X(t) = \mathbb{P}(X \leq t)$. On définit l'inverse généralisée F_X^{-1} de F_X sur $]0,1[$ par*

$$F_X^{-1} = \inf\{t \in \mathbb{R} \,|\, F_X(t) \geq x\}.$$

Alors, si U est une variable aléatoire uniforme sur $]0,1[$, $F_X^{-1}(U)$ a même loi que X.

Ainsi, si $u_1, \ldots, u_n$ sont n réalisations de variables aléatoires indépendantes de loi $\mathcal{U}[0,1]$, $F_X^{-1}(u_1), \ldots, F_X^{-1}(u_n)$ constitue un échantillon de réalisations de la loi de X.

1.2.2 Cas d'une variable aléatoire continue

Dans le cas d'une variable aléatoire continue et plus particulièrement d'une variable à densité, il faut d'abord calculer l'inverse au sens classique F_X^{-1} et simuler X à partir d'une loi uniforme $\mathcal{U}[0,1]$.

Exemple. *Cette méthode permet de simuler un échantillon de loi exponentielle $\mathcal{E}(\lambda)$ car*

$$F^{-1}(y) = -\frac{Log(1-y)}{\lambda} \quad \forall y \in]0,1[.$$

Ainsi, si u est une réalisation d'une loi uniforme $\mathcal{U}[0,1]$, alors $-4\,Log(u)$ est une réalisation d'une loi exponentielle $\mathcal{E}(\frac{1}{4})$.

1.2.3 Cas d'une variable aléatoire discrète

Dans le cas d'une variable aléatoire discrète, la méthode est canonique. En particulier, si X est une variable aléatoire discrète prenant un nombre fini de valeurs $\{x_1, \ldots, x_r\}$ telle que pour tout $j \in 1 \ldots r$, $\mathbb{P}(X = x_j) = p_j$, alors l'inverse de la fonction de répartition est

$$F_X^{-1}(u) = \sum_{j=1}^{r} x_j 1_{[p_0 + \cdots + p_{j-1}, p_1 + \cdots + p_j[}(u)$$

où $p_0 = 0$.

Exemple. *D'après ce qui précède, l'inverse généralisée d'une variable aléatoire X de loi de Bernoulli $\mathcal{B}(p)$ est $F_X^{-1} = 1_{[1-p,1[}$. Ainsi, si u est une réalisation d'une loi uniforme $\mathcal{U}[0,1]$, alors $1_{[1-p,1[}(u)$ est une réalisation d'une loi de Bernoulli $\mathcal{B}(p)$.*

2 Test de Kolmogorov

De nombreux tests statistiques permettent de décider si les réalisations obtenues par les méthodes décrites précédemment sont conformes à la loi désirée. Parmi les plus connus, citons le test d'adéquation du Khi-Deux et le test de Kolmogorov. Nous détaillons dans cette section le test de Kolmogorov. Il s'agit d'un test non paramétrique permettant, sur la base d'un échantillon $X_1, \dots X_n$ de fonction de répartition F, de tester l'adéquation à une loi de fonction de répartition F_0 donnée. Il repose sur le théorème de Glivenko et Cantelli établissant la convergence uniforme de la fonction de répartition empirique $F_n(x) = \frac{1}{n}\sum_{i=1}^n 1_{\{X_i \leq x\}}$:

$$D_n = \sup_x |F_n(x) - F(x)| \xrightarrow[n\to\infty]{p.s.} 0.$$

Par ailleurs, le théorème de Kolmogorov précise la distribution asymptotique de D_n (indépendante de la loi de l'échantillon initial) :

$$\lim_{n\to\infty} \mathbb{P}(\sqrt{n}D_n \leq y) = K(y) = \sum_{k=-\infty}^{+\infty} (-1)^k \exp(-2k^2y^2).$$

Pour tester,

$$H_0 : F = F_0 \text{ contre } H_1 : F \neq F_0.$$

- au niveau asymptotique α (lorsque $n \geq 80$) : la région critique est définie par $\sqrt{n}D_n > C$ avec C le quantile d'ordre $1-\alpha$ de la loi de fonction de répartition K.
- au niveau α (lorsque $n < 80$) : la région critique est définie par $D_n > d_n$ où d_n est le quantile d'ordre $1-\alpha$ de D_n (tabulée jusqu'à $n = 100$).

Exemple. *En R, on teste l'adéquation d'un vecteur de données* `X` *à la loi normale centrée réduite, en tapant la commande*

```
> ks.test(X,''pnorm'',0,1)
```

S'affiche alors la valeur observée $(D_n)_{obs}$ *de* D_n *et la p-value* $\mathbb{P}(D_n > (D_n)_{obs})$.

Remarque. Une extension de ce résultat est le test de Kolmogorov-Smirnov qui permet de tester si deux échantillons indépendants $X_1, \dots X_n$ et $Y_1, \dots, Y_m$ sont issus d'une même loi. Notons F_n et G_m les fonctions de répartition empriques des deux échantillons et considérons la variable aléatoire :

$$D_{n\,m} = \sup_x |F_n(x) - G_m(x)|.$$

Si les deux échantillons sont issus d'une même loi,

$$\mathbb{P}(\sqrt{\frac{n\,m}{n+m}}D_{n\,m} \leq y) \xrightarrow[n,\,m\to\infty]{} K(y)$$

où

$$K(y) = \sum_{k=-\infty}^{+\infty} (-1)^k \exp(-2k^2y^2).$$

Pour tester

$$H_0 : F = G \text{ contre } H_1 : F \neq G,$$

au niveau asymptotique α, on rejette H_0 lorsque $\sqrt{\frac{n\,m}{n+m}}D_{n\,m} > C$ avec C le quantile d'ordre $1-\alpha$ de la loi de fonction de répartition K. En R, la commande `ks.test` appliquée à deux vecteurs de données permet également d'effectuer le test de Kolmogorov-Smirnov.

3 Distributions de probabilité en R

Le logiciel R permet pour un certain nombre de lois de probabilité (c.f. tableau TAB. 1) d'évaluer en un point la fonction de répartition et la fonction de densité, de calculer les quantiles et de générer des réalisations.

Distribution	Nom	Paramètre(s)	Valeurs par défaut
Beta	`beta`	`shape1, schape2`	
Binomiale	`binom`	`size, prob`	
Binomiale Négative	`nbinom`	`size, prob`	
Cauchy	`cauchy`	`location, scale`	0, 1
Khi-Deux	`chisq`	`df`	
Exponentielle	`exp`	`rate=1/mean`	1
Fischer	`f`	`df1, df2`	
Gamma	`gamma`	`shape, rate=1/scale`	-, 1
Géométrique	`geom`	`prob`	
Hypergéométrique	`hyper`	`m, n, k`	
Log-Normale	`lnorm`	`meanlog, sdlog`	0, 1
Logistique	`logis`	`location, scale`	0, 1
Normale	`norm`	`mean, sd`	0, 1
Poisson	`pois`	`lambda`	
Student	`t`	`df`	
Uniforme	`unif`	`min, max`	0, 1
Weibull	`weibull`	`shape, scale`	
Wilcoxon	`wilcox`	`m, n`	

TABLE 1 – Distributions en R

Plus précisément, pour chacune des distributions du tableau TAB. 1, quatre commandes R préfixées par les lettres `d, p, q, r` sont disponibles. Ainsi pour la distribution `xxx` :

- `dxxx` donne la fontion de densité $f(x)$ pour une loi continue et la fonction de probabilité $\mathbb{P}(X = k)$ pour une loi discrète.
- `pxxx` donne la fonction de répartition $F(x) = \mathbb{P}(X \leq x)$.
- `qxxx` renvoie le quantile q tel que $F(q) = \alpha$ pour une distribution continue et le plus petit entier u tel que $F(u) \geq \alpha$ pour une distribution discrète.
- `rxxx` génère des réalisations aléatoires indépendante de loi `xxx`.

Ci-dessous, un petit exemple d'utilisation de ces commandes :

```
> # Calcul de la densité d'une loi N(0,1) aux points x
> x=seq(-4,4,0.1)
> y=dnorm(x)
> plot(x,y, type="l", col=3)
> # Calcul de la fonction de répartition d'une loi E(3) aux points a
> a=seq(0,2,0.1)
> b=pexp(a,3)
> plot(a,b,type="l",col=4)
> # Quantile d'ordre 95\% d'une loi de Student à 10 degrés de liberté
> qt(0.95,10)
> # Création d'un vecteur X de taille 50 de réalisations d'une loi Gamma(2,3)
> X=rgamma(50,2,3)
```

4 Exercices

4.1 Exercice 1

1. Attachez les données `faithful` de type `data.frame` et affichez les noms des variables.
2. Affichez les paramètres de position et de dispersion des données `eruptions`.
3. Tracez sur le même graphique l'histogramme d'aire égale à 1 des données `eruptions`.
   Remarque. Le tracé de l'histogramme fait apparaître deux sous-populations.
4. Enregistrez dans un vecteur `longtime` les données `eruptions` strictement supérieures à 3. Créez un vecteur d'abscisses `abs=seq(min(eruptions),max(eruptions),0.1)`.
5. Représentez sur un nouveau graphique la fonction de répartition empirique de `longtime` (vous pourrez utiliser la commande `ecdf` pour calculer la fonction de répartition empirique et la commande `plot` avec les options `verticals=TRUE` et `do.points=FALSE`). Sur le même graphique, tracez à l'aide de la fonction `lines` la fonction de répartition d'une loi normale de moyenne `mean(longtime)` et d'écart-type `sd(longtime)` aux points d'abscisses `longabs=abs[abs>3]`.
6. Testez au moyen du test de Kolmogorov l'adéquation de l'échantillon à une loi normale de moyenne `mean(longtime)` et d'écart-type `sd(longtime)`.
7. Utilisez les commandes `qqnorm` et `qqline` pour tester la normalité de l'échantillon `longtime` au moyen de la droite de Henry.
8. Détachez les données `faithful`.

4.2 Exercice 2

1. Considérons un étudiant qui tente de deviner les réponses aux questions d'un test.
   (a) Pour chaque question, supposons qu'il a 20% de chances de répondre correctement. A l'aide de la procédure de simulation des lois discrètes et de la commande `runif`, simuler les réponses (correcte/incorrecte) d'un tel étudiant à un test comportant 20 questions.
   (b) Ecrire une fonction R permettant de simuler les réponses d'un étudiant qui tente de deviner les réponses aux n questions d'un test avec une probabilité de bonne réponse égale à p.
2. Supposons qu'une classe de 100 étudiants passe un test "vrai ou faux" comportant 20 questions et que tous les étudiants répondent au hasard à chaque question.
   (a) Au moyen d'une simulation, donnez une estimation de la note moyenne et une estimation de l'écart-type des notes d'une telle classe. Comparez les résultats obtenus aux valeurs théoriques.
   (b) Donnez une estimation de la proportion d'étudiants qui obtiendraient un pourcentage de bonnes réponses supérieur à 30%.
   (c) Indiquez sur le vecteur contenant les notes des 100 élèves que Marie, Jérôme et Rémi possèdent respectivement la 1 ère, la 20 ème et la 89 ème note. Rémi a-t-il un poucentage de bonnes réponses supérieur à 30% ?