Université de Nantes – U.F.R. des Sciences

Master professionnel II : Ingénierie mathématique

Simulation & Modélisation

Exercices

Anne Philippe

Le logiciel utilisé pendant les séances de TP est le logiciel libre `R`.
Quelques adresses :

`http://www.r-project.org/`
`http://www.math.sciences.univ-nantes.fr/~philippe/index.php?frame=R`

18 décembre 2007
Anne Philippe
Bureau 118 bâtiment 10
Laboratoire de Mathématiques Jean Leray

CHAPITRE 1

Générateurs de nombres pseudo aléatoires : Les premiers exemples

Exercice 1. Loi uniforme sur le disque

Soit (U, V) un couple de variables aléatoires distribuées suivant la loi uniforme sur le disque de centre $(0, 0)$ et de rayon 1

On note (R, θ) les coordonnées polaires de (U, V)

$$U = R\cos(\theta)$$
$$V = R\sin(\theta)$$

1) Déterminer la loi du couple (R, θ)
2) Les variables aléatoires R et θ sont elles indépendantes ?
3) Donner les lois marginales de R et de θ.
4) Proposer un algorithme pour simuler des nombres pseudo aléatoires suivant la loi de R, puis celle de θ.
5) En déduire un algorithme pour simuler des nombres pseudo aléatoires suivant la loi uniforme sur le disque unité

Exercice 2. Loi β

Soit $(a, b) \in \mathbb{R}_+ \times \mathbb{R}_+$.
Soit X et Y deux variables aléatoires *indépendantes*. On suppose que
– X est distribuée suivant la loi $\Gamma(a, 1)$
– Y est distribuée suivant la loi $\Gamma(b, 1)$
On rappelle que la loi $\Gamma(a, 1)$ admet pour densité

$$f(x) = \frac{1}{\Gamma(a)} e^{-x} x^{a-1} \mathbb{I}_{\mathbb{R}_+}(x).$$

La fonction Γ est définie sur $\mathbb{R}_+$ par

$$\Gamma(a) = \int_0^\infty e^{-t} t^{a-1} \, \mathrm{d}t = (a-1)\Gamma(a-1).$$

On a en particulier pour tout $a \in \mathbb{N}$, $\Gamma(a+1) = a!$

1) Écrire la densité de la loi du couple (X, Y)
2) Donner la loi du couple $(V, W) = (X + Y, \dfrac{X}{X+Y})$.
3) Les variables aléatoires V et W sont elle indépendantes ? Préciser les lois marginales de V et W.
4) En déduire un algorithme pour simuler des nombres pseudo aléatoires suivant la loi β de paramètre (a, b) utilisant un générateur de nombres pseudo aléatoires suivant des lois *Gamma*
5) *(facultatif)* En déduire une expression de $B(a, b)$.

Exercice 3. Générateur de nombres pseudo aléatoires suivant la loi uniforme

On considère le générateur

$$X_n = aX_{n-1} + c \qquad \text{modulo } M$$

On fixe $c = 1$ $a = 65$ $M = 2048$ et $X_0 = 0$.
`Indication : n%%M` retourne la valeur de `n` modulo `M`

R-1) Simuler un échantillon de taille 2048.
R-2) Représenter la distribution de cet échantillon à l'aide d'un histogramme. Commenter
R-3) Tester l'uniformité de l'échantillon à l'aide du test Kolmogorov- Smirnov (fonction `ks.test` en `R`)
R-4) Tracer le nuage de point $\{(X_i, X_{i+1}), i = 1, \ldots, 2047\}$. Commenter.
R-5) À partir de l'échantillon simulé à la question R-1), construire un échantillon suivant la loi gaussienne standard en utilisant la fonction `qnorm`. On note cet échantillon $N_1, \ldots, N_n$
R-6) Tester la normalité de l'échantillon $N_1, \ldots, N_n$ à l'aide du test Kolmogorov- Smirnov (fonction `ks.test` en `R`)
R-7) Tracer les corrélations empiriques entre $(N_i)_i$ et $(N_{i+h})_i$ en fonction du décalage (lag) h. (utiliser la fonction `acf` et son arg `lag.max`)

R-8) Tester l'indépendance des $N_1, \ldots, N_n$ à l'aide du test de Box. (utiliser la fonction `Box.test` et son argument `lag`)

R-9) Déterminer le décalage h_0 qui produit la plus forte corrélation.

R-10) Tracer le nuage de point $\{(X_i, X_{i+h_0})\}$ en utilisant la fonction `plot` avec l'argument `col` $=$ `1 : h0`. Commenter.

R-11) Comparer avec le générateur `runif` disponible dans R .

Dans la suite de ce cours, on utilise le générateur `runif`

Exercice 4. Loi uniforme sur le disque

On souhaite vérifier par la simulation l'affirmation suivante :

> Étant donné quatre points indépendants et distribués suivant la loi uniforme sur le disque unité, la probabilité que l'enveloppe convexe soit un triangle vaut $\frac{35}{12\pi^2}$

R-1) Écrire une fonction qui retourne un échantillon de taille n suivante la loi uniforme sur le disque unité

R-2) Soit $U_1, \ldots U_4$ quatre points i.i.d. suivant la loi uniforme sur le disque unité. On note N la variable aléatoire égale au nombre de points constituant l'enveloppe convexe de $U_1, \ldots U_4$.
Écrire une fonction qui retourne des nombres pseudo aléatoires de même loi que N.
`Indication :` la fonction `chull` calcule l'enveloppe convexe d'un nuage de points dans $\mathbb{R}^2$

```
> X <- matrix(runif(20), ncol = 2)
#X contient les coordonnées de 20 points du plan
#premiere colonne les abscisses
#seconde  colonne  les ordonnées
> X
           [,1]        [,2]
 [1,] 0.5586059 0.893774346
 [2,] 0.6389309 0.247279142
 [3,] 0.1762004 0.768140694
 [4,] 0.5311781 0.424761495
 [5,] 0.1194499 0.004410222
 [6,] 0.9765925 0.565299092
 [7,] 0.5238551 0.163369580
 [8,] 0.5505306 0.969284956
 [9,] 0.6214152 0.109622856
[10,] 0.8485409 0.422773730
> plot(X, cex = 1.5, col=3, pch=22)
#les indices des points qui forment l'enveloppe
```

```
> hpts <- chull(X)
> hpts
[1] 9 5 3 8 6

#on trace l'enveloppe
> hpts <- c(hpts, hpts[1])
> lines(X[hpts, ],lwd=2,col="chocolate3")
```

R-3) Simuler un échantillon de taille n de même loi que N ($n = 1000$).

R-4) Proposer une estimation de l'évènement $[N = 3]$ à partir de l'échantillon simulé. Commenter.

Exercice 5. Loi Γ

La loi gamma, notée $\Gamma_{a,\lambda}$ où $\lambda > 0$ et $a > 0$, admet pour densité

$$f_{a,\lambda}(x) = \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x} \mathbb{I}_{]0,+\infty[}(x).$$

1) Calculer l'espérance et la variance de cette loi.

2) Soient X_1 et X_2 deux variables aléatoires indépendantes de lois respectives $\Gamma_{a_1,\lambda}$ et $\Gamma_{a_2,\lambda}$. Montrer que $X_1 + X_2$ a une loi $\Gamma_{a_1+a_2,\lambda}$.

3) Montrer que le carré d'une variable gaussienne standard suit une loi $\Gamma_{\frac{1}{2},\frac{1}{2}}$.

4) En déduire que si $(X_1, \ldots, X_n)$ sont i.i.d. suivant la loi gaussienne $\mathcal{N}(0,1)$, alors $\sum_{j=1}^n (X_j)^2$ suit une loi $\Gamma_{\frac{n}{2},\frac{1}{2}}$. Cette loi est appelée la loi de χ^2 à n degrés de liberté et notée $\chi^2(n)$: ainsi, le carré d'une gaussienne standard suit une loi $\chi^2(1)$.

5) Soit (X, Y) un couple de variables aléatoires indépendantes telles que X suit une loi gaussienne $\mathcal{N}(0,1)$ et Y une loi de $\chi^2(n)$. Quelle est la densité de la loi de la variable

aléatoire T_n définie par

$$T_n = \frac{X}{\sqrt{Y/n}}$$

Cette loi est connue sous le nom de loi de Student de paramètre n (ou encore à n degrés de liberté).

Exercice 6. Les mélanges de lois gaussiennes

R-1) Écrire une fonction qui retourne des nombres aléatoires suivant une loi de mélanges à 2 composantes gaussiennes

$$p\mathcal{N}(m_1, \sigma_1^2) + (1-p)\mathcal{N}(m_2, \sigma_2^2)$$

R-2) Pour valider votre algorithme, simuler des échantillons et comparer la densité théorique et la densité estimée par la méthode de l'histogramme. Regarder en particulier les situations suivantes :

$m_1 \neq m_2$ et $\sigma_1 = \sigma_2$ puis $m_1 = m_2$ et $\sigma_1 \neq \sigma_2$

Exercice 7. Algorithme d'acceptation rejet (AR)

Soit f une densité de probabilité définie sur $[0, 1]$. On suppose que f est continue et strictement décroissante sur $[0, 1]$.

Q-1 Montrer que la loi de densité f peut être simulée par la méthode d'acceptation rejet associée à la loi uniforme sur $[0, 1]$.

Q-2 Montrer que le taux d'acceptation est égal à $\dfrac{1}{f(0)}$

Soit $p \in]0, 1[$. On considère un mélange de deux lois uniformes qui admet pour densité

$$m(x) = 2p\, \mathbb{I}_{[0,1/2]}(x) + 2(1-p)\, \mathbb{I}_{[1/2,1]}(x)$$

Q-3 Montrer que la loi de densité f peut être simulée par la méthode d'acceptation rejet associée à la loi de densité m.

Q-4 Montrer que le taux d'acceptation est égal

$$\frac{1}{\max\left(\frac{1}{2p}f(0)\ ,\ \frac{1}{2(1-p)}f(1/2)\right)}$$

Q-5 En déduire que le choix optimal de p est $p^{optimal} = \dfrac{f(0)}{f(0) + f(1/2)}$

Q-6 Comparer l'algorithme associé à la valeur de $p^{optimal}$ avec celui décrit à la question 1).

Exercice 8. Algorithme AR pour les lois gaussiennes tronquées

1) Proposer un algorithme pour simuler un échantillon suivant la loi de Cauchy.
Indication : calculer la fonction de répartition.

2) Montrer que l'on peut simuler la loi gaussienne par un algorithme d'acceptation rejet associé à une loi de Cauchy.

3) préciser le taux d'acceptation

R 4) Programmer en R cet algorithme. Valider l'algorithme (test d'adéquation, outils graphiques)

R 5) Calculer une estimation du taux d'acceptation.

6) Construire un algorithme d'acceptation rejet pour simuler suivant la loi gaussienne tronquée sur $[2, +\infty[$, c'est à dire la loi dont la densité est proportionnelle à

$$h(x) = e^{-x^2/2}\, \mathbb{I}_{[2,+\infty[}(x)$$

en utilisant comme loi instrumentale
- La loi normale standard
- la loi de Cauchy.

R 7) Programmer et tester ces deux algorithmes.

R 8) Comparer les taux d'acceptation.

9) Proposer un algorithme pour simuler un échantillon suivant la loi de Cauchy tronquée sur $[2, +\infty[$.
Indication : calculer la fonction de répartition.

10) Construire un algorithme d'acceptation rejet pour simuler suivant la loi gaussienne tronquée sur $[2, +\infty[$ en utilisant comme loi instrumentale la loi de Cauchy tronqué.

R 11) Comparer le taux d'acceptation de cet algorithme aux deux précédents.

Exercice 9. Loi Gamma et algorithme d'acceptation rejet

L'objectif de cet exercice est de construire un générateur de nombres pseudo aléatoires suivant une loi $\Gamma(a, b)$ $(a > 0,\, b > 0)$.

Pour $a = 1$, on retrouve la loi exponentielle que l'on peut facilement simuler par la méthode d'inversion.

1. Montrer que si $X \sim \Gamma(a, 1)$ alors X/b suit une loi $\Gamma(a, b)$.

   On peut donc supposer dans la suite de l'exercice que $b = 1$.

2. Montrer que si $X \sim \Gamma(a + 1, 1)$, $U \sim \mathcal{U}(0, 1)$ et si les variables aléatoires X et U sont indépendantes alors $XU^{1/a}$ suit une loi gamma $\Gamma(a, 1)$

   On peut donc supposer dans la suite de l'exercice que $a > 1$.

3. Montrer que sous ces hypothèses, la loi $\Gamma(a, 1)$ peut être simulée par la méthode d'acceptation rejet associée à une loi exponentielle $\mathcal{E}(\lambda)$.

4. Calculer la probabilité d'acceptation en fonction de (λ, a)

5. En déduire la valeur optimale de λ en fonction de a

R 6. Représenter la probabilité d'acceptation associée à la valeur optimale de λ en fonction de $a > 1$. Commenter.

Lorsque $a \in \mathbb{N}$, la loi $\Gamma(a, 1)$ peut être simulée à partir de a variables aléatoires iid suivant la loi exponentielle de paramètre 1. Voir Exercice 5.

7. Soit $a > 1$ et $a \notin \mathbb{N}$. Montrer que la loi $\Gamma(a, 1)$ peut être simulée par la méthode d'acceptation rejet associée à une loi gamma $\Gamma([a], \lambda)$ où $[a]$ est la partie entière de a

8. Calculer la valeur optimale de λ

R 9. Représenter la probabilité d'acceptation associée à la valeur optimale en fonction de $a > 1$. Commenter et comparer avec les performances de l'algorithme précédent.

Exercice 10. Algorithmes SIR : choix de m en fonction de n

On cherche à simuler par la méthode SIR un échantillon de taille n suivant la loi $\Gamma(3.5, 1)$ à partir d'un échantillon de taille m suivant le loi $\Gamma(3, 1)$.

R 1. Programmer cet algorithme.
Indication : utiliser la fonction `sample` pour l'étape de reéchantillonnage.

On cherche à calibrer m en fonction de n.

R 2. Pour évaluer les performances de cet algorithme, on teste l'ajustement à la loi $\Gamma(3.5, 1)$, par le test de Komogorov de niveau 5% pour $N = 500$ échantillons simulés de façon indépendante.
Comparer la proportion d'échantillons rejetés avec l'erreur de première espèce 5% dans les situations suivantes

m	500	5000	10 000

et

n/m	0.01	0.05	0.10	0.15	0.20	0.50

Exercice 11. Algorithmes SIR : application à la loi slash

Soient N et U deux variables aléatoires indépendantes, $N \sim \mathcal{N}(0, 1)$ et $U \sim \mathcal{U}[0, 1]$.
La loi du rapport $S = N/U$ est appelée la loi slash.

1. Montrer que la densité de la loi slash est égale à

$$s(x) = \begin{cases} \frac{1-e^{-x^2/2}}{x^2\sqrt{2\pi}} & \text{si } x \neq 0 \\ \frac{1}{2\sqrt{2\pi}} & \text{sinon} \end{cases}$$

2. Vérifier que le moment d'ordre 1 de la loi slash est infini.

R 3. Programmer et tester un algorithme SIR pour simuler un échantillon suivant la loi slash à partir de la loi gaussienne standard.

R 4. Programmer et tester un algorithme SIR pour simuler un échantillon suivant la loi gaussienne à partir de la loi slash.

CHAPITRE 2

Valider un test par la simulation

Exercice 1. courbes niveau/puissance

Soit $X_1, \ldots, X_n$ n variables aléatoires iid. Soit $T_n = T_n(X_1, \ldots, X_n)$ une statistique, on considère la procédure de test associée à la région critique[1] $\{T_n > c(\alpha)\}$, où $c(\alpha)$ est le quantile supérieur de la loi limite F_0 de T_n sous H_0. On suppose que F_0 est inversible et donc $c(\alpha) = F_0^{-1}(1-\alpha)$.

Ayant observé l'échantillon $(X_1, \ldots, X_n)$, la p-value p_n est la variable aléatoire définie par $p_n = 1 - F_0(T_n(X_1, \ldots, X_n))$.

1) Soit $\alpha \in]0,1[$ Montrer que si $p_n(x_1, \ldots, x_n) > \alpha$ alors on accepte l'hypothèse nulle H_0 au niveau α.

2) Soit T une variable aléatoire de loi F_0. Montrer que $1 - F_0(T)$ est distribuée suivant la loi uniforme sur $]0,1[$.

3) Pour évaluer les performances d'un test sur des petits échantillons, on propose la démarche suivante
   - on simule N échantillons de taille n vérifiant H_0, on calcule la p-value pour chacun des échantillons (notée $p^{(1)}, \ldots, p^{(N)}$)
   - on trace la fonction de répartition empirique de l'échantillon $p^{(1)}, \ldots, p^{(N)}$ c'est à dire la fonction
   $$\hat{F}_N(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{]-\infty,x]}(X_i)$$
   - si la courbe est proche de la droite $y = x$ alors le niveau du test est satisfaisant pour les petits échantillons.

   Justifier cette démarche

4) Quelle doit être l'allure de la fonction $\hat{F}_N$ sous des hypothèses alternatives lorsque le test est sans biais, lorsque la puissance du test converge vers 1 quand $n \to \infty$

Exercice 2. Le test de Kolmogorov

H_0 "les variables aléatoires sont distribuées suivant une loi exponentielle"

R 5) Simuler $N = 500$ échantillons de taille $n = 50$ suivant la loi exponentielle de paramètre 1 et calculer la p-value de chacun des échantillons, on obtient $(p^{(1)}, \ldots, p^{(N)})$.

[1] région où l'on rejette l'hypothèse nulle

R 6) Tracer la fonction de répartition empirique associée à l'échantillon $(p^{(1)}, \ldots, p^{(N)})$. Le niveau du test est-il satisfaisant ? `Indication` utiliser la fonction `ecdf` de la librairie `Hmisc`.

R 7) On veut maintenant évaluer la puissance du test lorsque les variables aléatoires sont distribuées suivant un mélange de lois exponentielles

$$p\mathcal{E}(1) + (1-p)\mathcal{E}(2)$$

Simuler N échantillons de taille n suivant la loi $p\mathcal{E}(1) + (1-p)\mathcal{E}(2)$ et calculer la p-value. Tracer la fonction F_N associée. Reprendre la même question pour différentes valeurs de p entre]0,1[

- superposer sur un même graphique les différentes fonctions.
- Commenter.
- Que pensez vous de la puissance de ce test.

CHAPITRE 3

Intégration de Monte Carlo

Dans toute cette partie f désigne une densité de probabilité et h une fonction dans $L^1(f)$.

On cherche à estimer des intégrales de la forme

$$I = \int h(x)f(x)\,\mathrm{d}x$$

par des techniques probabilistes qui reposent essentiellement sur la simulation de variables aléatoires.

Exercice 1. Contrôle et visualisation de la convergence des estimateurs de Monte Carlo

Soit X une variable aléatoire distribuée suivant la loi gaussienne standard, on souhaite estimer l'intégrale suivante

$$I = \mathbb{E}(e^X \mathbb{I}_{[-1,1]}(X))$$

par la technique standard de Monte Carlo.

1) Décrire la mise en œuvre de l'estimateur de Monte Carlo, I_n, construit à partir d'un échantillon simulé suivant la loi de X

R 2) Représenter l'estimateur de Monte Carlo en fonction de la taille de l'échantillon.

R 3) Sur la trajectoire simulée à la question précédente, évaluer la variance de l'estimateur de Monte Carlo.

R 4) En déduire les bornes de l'intervalle de confiance de niveau $1 - \alpha = 95\%$ construit en utilisant le Théorème Limite Central. Superposer l'estimateur et ses régions de confiance.

On estime maintenant la variance et les régions de confiance par une méthode non asymptotique. La loi de l'estimateur de Monte Carlo I_n est estimée à n fixé par une méthode de simulation.

R 5) La variance, les quantiles, etc de la loi de I_n sont estimés à partir de N échantillons indépendants de taille n.

La démarche

- On construit une matrice n lignes et N colonnes de nombres aléatoires distribués suivant la loi de densité f. On note A cette matrice
- On calcule la suite des estimateurs $\{I_k,\ k = 1, \ldots, n\}$ sur les N colonnes

```
h = function(x) .... à définir ....
cummean = function(x) cumsum(h(x))/(1 :length(x))
B = apply(A,2,cummean)
```

La k ème ligne de la matrice B contient un échantillon suivant la loi de l'estimateur de Monte Carlo I_k.

Évaluation de la variance de I_n

- Calculer la variance de chacune des lignes

```
V = apply(B,1,var)
```

pour obtenir une estimation de la variance de I_k, $k = 1 \ldots, n$
- Représenter l'estimation de la variance en fonction de la taille de l'échantillon
- Représenter le nuages de points $\{(\log(k), log(V[k]),\ k = 1 \ldots, n\}$. Commenter le résultat et faire le lien avec le résultat $\mathrm{Var}(I_n) = Cn^{-1}$.
- En déduire une estimation de C.

Régions de confiance

- Estimer les quantiles d'ordre $\alpha/2$ et $1 - \alpha/2$ de la loi de I_n. Utiliser la fonction `quantile` et `apply` sur les lignes de la matrice B.
- Représenter sur un même graphique l'estimateur de Monte Carlo et sa région de confiance en fonction de la taille de l'échantillon

Exercice 2. Réduction de la variance : Antithetic sampling

Soit $X_1, \ldots X_n$ n variables aléatoires iid suivant la loi de densité f.
Soit $I_n = \frac{1}{n}\sum_{i=1}^n h(X_i)$ l'estimateur standard de Monte Carlo pour l'intégrale I.
On introduit les variables aléatoires $X_1^*, \ldots X_n^*$ telles que
- pour tout i : $X_i^* = \psi(X_i)$ où ψ est une application mesurable
- les variables aléatoires X_1^* et X_1 ont la même loi.

On pose

$$I_n^* = \frac{1}{n}\sum_{i=1}^n h(X_i^*)$$

et

$$\bar{I}_n = \frac{1}{2}(I_n + I_n^*)$$

1) Calculer la variance de $\bar{I}_n$, puis comparer avec celle de I_{2n}.

   On cherche à construire ψ pour que les variables $h(X_1^*)$ et $h(X_1)$ soient négativement corrélées.

2) *Résultat préliminaire* Soit g_1 et g_2 deux fonctions monotones, de monotonie opposée et soit X une variable aléatoire.
   Montrer que

$$\mathrm{Cov}(g_1(X), g_2(X)) \leq 0$$

Indications
- on introduit Y une variable aléatoire de même loi que X et indépendante de X.
- vérifier que $(g_1(X)-g_1(Y))(g_2(X)-g_2(Y)) \leq 0$, puis calculer l'espérance du terme de gauche.

3) On note F la fonction de répartition de la loi de densité f. Il existe $U_1, \ldots, U_n$ iid suivant la loi uniforme sur $[0,1]$ tel que $X_i = F^-(U_i)$
On pose
$$X_i^* = F^-(1-U_i)$$
Montrer que si h est monotone alors les variables $h(X_1^*)$ et $h(X_1)$ sont négative corrélées.

4) Montrer que si la loi est symétrique alors $X_i^* = -X_i$

5) En déduire que $\bar{I}_n$ est meilleur que I_{2n} en terme de variance.

R 6) On applique cette méthode dans la situation suivante
- f est la densité de la loi gaussienne
- $h(x) = x/(2^x - 1)$.
- Vérifier que la méthode précédente peut s'appliquer
- Confirmer la réduction de la variance par la simulation. Comparer les variances estimées de $\bar{I}_n$ et I_{2n} en fonction de n (Reprendre la méthode de l'exercice 1 basée sur N trajectoires indépendantes).

Exercice 3. Méthode de Rao Blackwell

Quelques rappels sur la loi de Student
- La densité de la loi de Student à k degrés de liberté (ddl) est égale à
$$f_k(t) = \frac{1}{\sqrt{k\pi}} \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})} \frac{1}{(1+\frac{t^2}{k})^{\frac{k+1}{2}}}$$
- **Proposition** Soient Z une variable aléatoire de loi normale centrée et réduite et U une variable indépendante de Z et distribuée suivant la loi la loi du χ^2 à k ddl. La variable aléatoire $T = \frac{Z}{\sqrt{U/k}}$ suit une loi de Student à k degrés de liberté.

Soit (X, Y) un couple de variables aléatoires dont la loi est définie par
- la loi conditionnelle de X sachant Y est la loi gaussienne centrée et de variance Y^{-1}
- la variable aléatoire Y suit une loi Gamma de paramètres $(\frac{\nu}{2}, \frac{\nu}{2})$

On souhaite estimer $\mathbb{E}(e^{-X^2})$.

1) Montrer que X suit une loi de Student à ν ddl.

2) En déduire une méthode pour simuler un échantillon suivant la loi de Student à ν ddl.

3) Calculer l'espérance conditionnelle $\mathbb{E}(e^{-X^2}|Y)$

4) On simule un échantillon iid suivant la loi du couple (X, Y). Construire deux estimateurs de $\mathbb{E}(e^{-X^2})$ à partir de cet échantillon.

R 5) Représenter les deux estimateurs de Monte Carlo en fonction de n la taille de l'échantillon simulé.

R 6) Comparer les variances estimées des deux estimateurs (reprendre la méthode de l'exercice 1 basée sur N trajectoires indépendantes).

Exercice 4. Importance Sampling (Échantillonnage Pondéré)

On met en œuvre le test suivant

$$H_0 \ : \ \lambda = 2 \qquad H_1 \ : \ \lambda > 2$$

sur le paramètre d'une loi de Poisson à partir d'un échantillon $(X_1, \ldots, X_n)$ de taille $n = 15$.

Si on suppose que la loi de $\sum_{i=1}^{15} X_i$ peut être approchée par une loi gaussienne, alors la région critique du test de niveau α (région où l'on rejette H_0) est de la forme

$$\sum_{i=1}^{n} X_i > n(2 + q(\alpha)\sqrt(2)/\sqrt(n)) := C_n$$

où $q(\alpha)$ est le quantile supérieur de la loi gaussienne standard. On prendra par exemple $\alpha = 1\%$.

On utilise une méthode de Monte Carlo pour évaluer le niveau exact de ce test.

1) Décrire l'estimateur de Monte Carlo standard construit à partir d'un échantillon simulé suivant la loi de Poisson de paramètre $\lambda = 30$

2) Préciser la variance de cet estimateur

R 3) Représenter l'estimateur de Monte Carlo en fonction de la taille de l'échantillon.

R 4) Évaluer la variance de l'estimateur en utilisant N échantillons indépendants

5) Pour réduire la variance de l'estimateur de Monte Carlo, on considère la méthode "Importance Sampling". On utilise comme loi instrumentale une loi de Poisson de paramètre ν. Comment choisir ν pour que l'estimateur I_n^{IS} soit de variance finie.

R 6) Comparer la variance des estimateurs I_n^{IS} pour différentes valeurs de ν. Commenter le choix de ν. (Utiliser N échantillons indépendants pour évaluer les variances des estimateurs I_n^{IS})

CHAPITRE 4

Bootstrap

Exercice 1. Loi uniforme

Soit $X_1, ..., X_n$ n variables aléatoires iid suivant le loi uniforme sur $[0, \theta]$

1) Calculer l'estimateur du maximum de vraisemblance de θ
2) Donner la loi de l'estimateur du MV, sa moyenne et sa variance.
3) On cherche à estimer le biais et la variance de l'estimateur du MV par la technique du bootstrap non paramétrique.
   Calculer la probabilité de l'événement $[\hat{\theta}_n^* = \hat{\theta}_n]$. Commenter

R 4) On met en oeuvre le bootstrap paramétrique et non paramétrique pour estimer le biais et la variance. Les observations sont des données simulées `xdata` suivant la loi uniforme sur $[0, 2]$. Commenter les résultats obtenus.

```
#-----------------------------------------------#
#auteur A. Philippe (U. Nantes ) --------------#
#-----------------------------------------------#
n=50
xdata=runif(n,0,2)
est.theta = max(xdata)
#-----------------------------------------------#


B=500 ;
Mstar=NULL ;

 for ( i  in 1:B)
 {star = max(runif(50,0,est.theta))
  Mstar=c(Mstar,star)
 }

BIAIS = mean(Mstar) - est.theta
VAR   = var(Mstar)

print(c(BIAIS,VAR))
```

```
[1] -0.039047937  0.001458224
#---------------------------------------------#
#une autre facon de programmer: utilisation ---#
#de la fonction boot de la lib. boot----------#
#---------------------------------------------#

  boot.gen <- function(data,emv)
     {     runif(length(data), 0,emv)
      }


  boot.stat=function(x) max(x)

  resultat <- boot(xdata, boot.stat, R=B, sim="parametric",
          ran.gen=boot.gen, mle=max(xdata))


BIAIS = mean(resultat$t) - est.theta
VAR   = var(resultat$t)

> print(c(BIAIS,VAR))
[1] -0.041011839  0.001576235


#---------------------------------------------#
# le bootstrap non parametrique----------------#
#---------------------------------------------#
 boot.stat=function(x,n) max(x[n])
 nonpara= boot(xdata,statistic=boot.stat,R=B)

BIAIS = mean(nonpara$t) - est.theta
VAR   = var(nonpara$t)

print(c(BIAIS,VAR))
[1] -0.0074423526  0.0008635543
> table(nonpara$t)

1.72338678827509 1.77015830529854 1.77376327197999 1.83286639675498
               1                1                3                6
1.91979617066681 1.96550890803337 1.98425709316507 1.98525541741401
              13               37              114              325
> 325/500
[1] 0.65
```

```
> 1-exp(-1)
[1] 0.6321206
```

Exercice 2. Modèle exponentiel

On dispose de 10 données simulées suivant une loi exponentielle

$$f(x) = \frac{1}{\theta} e^{-x/\theta} \mathbb{I}_{\mathbb{R}^+}(x)$$

de paramètre $\theta = 50$

104.91 7.31 4.54 26.24 62.47 116.37 9.97 36.64 109.67 95.96

On note F la fonction de répartition.

R 1) On estime la moyenne $\theta = \theta(F) = \int x \, \mathrm{d}F(x)$, par $T(\underline{X}) = \theta(F_n)$ où F_n est la fonction de répartition empirique.

R 2) Comparer la distribution théorique de $\theta(F_n) - \theta(F)$ avec les approximations obtenues par
- le théorème limite central
- le bootstrap non paramètrique. Discuter du choix de B le nombre d'échantillon simulés suivant F_n pour approcher la loi de $T(\underline{X}^*) - \theta(F_n)$ par une méthode de Monte Carlo.

R 3) Construire et comparer les intervalles de confiance obtenus par les méthodes suivantes
- la loi exacte de $\theta(F_n)$
- le théorème limite central
- le bootstrap non paramètrique
  - méthode de base
  - percentile
  - t-bootstrap

`Indication :` La fonction `boot.ci` de la librairie `boot` calcule les différentes régions de confiance

Exercice 3. Coefficient de corrélation

Soit $(X_1, \ldots, X_n)$ n vecteurs aléatoires de $\mathbb{R}^2$ indépendants et identiquement distribués. On estime le coefficient de corrélation ρ entre les coordonnées de X_1.

Situation A On suppose que les vecteurs sont gaussiens de moyenne $\begin{pmatrix} 0 \\ 2 \end{pmatrix}$ et de variance $\begin{pmatrix} 4 & 1.5 \\ 1.5 & 1 \end{pmatrix}$

R-A1) Écrire une fonction qui retourne un échantillon de taille n suivant la loi de X_1

R-A2) Pour des échantillons de tailles $n = 10, 50, 100, 500$, comparer la probabilité de couverture des intervalles de confiance obtenus par les méthodes suivantes
- théorème limite central
- bootstrap de base (B=2000)

– bootstrap percentile (B=2000)

`Indications` La probabilité de couverture des intervalles est estimée par une méthode de Monte Carlo. On simule $N = 1000$ échantillons suivant la loi de X_1, on calcule les intervalles de confiance et on évalue la proportion des intervalles qui contiennent la vraie valeur du paramètre.

Situation B La loi du vecteur $X_1 = \begin{pmatrix} X_1(1) \\ X_1(2) \end{pmatrix}$ est définie par $X_1(1)$ suit une loi gaussienne standard et $X_1(2) = X_1(1)^2$

R-B1) Vérifier que les coordonnées de X_1 sont non corrélées $\rho = 0$

R-B2) Reprendre les questions de la partie A pour ce modèle.

Exercice 4

Les données `cars` (`data(cars)`) donnent des distances de freinage en fonction de la vitesse. On modélise la distance par le modèle linéaire suivant

$$\text{distance} = a + b\text{vitesse} + \varepsilon$$

où les variables aléatoires ε_i sont iid centrées et de variance σ^2

On souhaite comparer les performances de deux techniques de bootstrap [bootstrap sur les données] /[bootstrap sur les résidus]

1) Donner une estimation des paramètres (a, b, σ^2)
2) Tester l'hypothèse "les résidus sont iid suivant une loi gaussienne $\mathcal{N}(0, \sigma^2)$

le bootstrap

3) En utilisant la fonction `lm.boot` de la librairie `simpleboot` ou la fonction `boot` de la librairie `boot`, mettre en oeuvre les deux techniques de bootstrap.
4) Donner une estimation du biais et de la variance des estimateurs de (a, b, σ^2). Comparer les deux méthodes de bootstrap.
5) Tracer l'histogramme et le boxplot des valeurs $\hat{a}_n^{*i}$, $\hat{b}_n^{*i}$, et $\hat{\sigma}_n^{*i}$ pour $i = 1, ..., n$. Comparer les deux méthodes de bootstrap
6) Tracer sur un même graphique les B droites $y = \hat{a}^* + \hat{b}^* x$ construites sur les échantillons bootstrapés. Commenter.
7) Reprendre les questions précédentes pour le bootstrap paramétrique sur les résidus, c'est à dire
   – on simule les ε^* iid suivant la loi normale $\mathcal{N}(0, \hat{\sigma}_\varepsilon^2)$.

CHAPITRE 5

Statistique Bayésienne

Exercice 1. Calcul des lois a priori conjuguées

Donner une famille de lois conjuguées pour les modèles $\{P_\theta, \theta \in \Theta\}$ suivants

1) Poisson : $\{\mathcal{P}(\theta), \theta \in \mathbb{R}^+\}$
   Indication : loi gamma
2) Gaussien de variance connue : $\{\mathcal{N}(\theta, \sigma^2), \theta \in \mathbb{R}\}$
   Indication : loi gaussienne
3) Gaussien de moyenne connue : $\left\{\mathcal{N}(\mu, \frac{1}{\theta}), \theta \in \mathbb{R}_+\right\}$
   Indication : loi gamma

Exercice 2. Calcul des lois de Jeffrey

Calculer la loi non informative de Jeffrey pour les modèles suivants :

1) Poisson : $\{\mathcal{P}(\theta), \theta \in \mathbb{R}^+\}$
2) Gaussien : $\{\mathcal{N}(\mu, \sigma^2), (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+\}$ dans les deux situations suivantes
   a. $\theta = \mu$ et σ^2 est connu.
   b. $\theta = \sigma$ et μ est connu.
   c. les deux paramètres sont inconnus : $\theta = (\mu, \sigma)$.

Exercice 3

R 1) Récupérer le fichier de données de taille $n = 500$
   `http ://www.math.sciences.univ-nantes.fr/~philippe/lecture/Piecesdef.txt`.
   La i$^{\text{e}}$ observation N_i est égale au nombre de pièces défectueuses dans le i$^{\text{e}}$ lot. Un lot est constitué de 50 pièces. On veut estimer la probabilité de produire une pièce défectueuse.
2) Décrire le modèle statistique.
3) Montrer que la famille des lois β est une famille de lois conjuguées pour ce modèle.
4) Le service qualité fournit les informations suivantes
   – la proportion de pièces défectueuses est en moyenne 0.15
   – la proportion de pièces $X_1, ..., X_n$ défectueuses appartient à l intervalle $[0.1, 0.2]$ avec une probabilité de 95%

   Utiliser ces informations pour fixer les paramètres de la loi a priori conjuguée.

5) Calculer la loi a posteriori et l'estimateur de Bayes associé au coût quadratique.

6) **Situation non informative :** Calculer la loi a priori de Jeffrey pour ce modèle.

7) Calculer la loi a posteriori et l'estimateur de Bayes.

R 8) Représenter les deux estimateurs de Bayes en fonction du nombre d'observations n.

R 9) Tracer sur un même graphique les deux lois a posteriori calculées sur les K premières observations. Faire varier K par exemple $5, 10, 15, 20$.

Exercice 4. Régions de confiance bayésienne

On dispose de n observations $X_1, ..., X_n$ iid suivant une loi gaussienne $\mathcal{N}(\theta, 1)$.
On choisit comme loi a priori sur θ la loi gaussienne $\mathcal{N}(0, \tau^{-2})$, $\tau > 0$

1) Montrer que la loi a posteriori est une loi Gaussienne

$$\mathcal{N}(\frac{\bar{X}_n}{1+\tau^2/n}, \frac{1}{n+\tau^2})$$

2) Montrer que les régions HPD de niveau $1-\alpha(=.95)$ sont de la forme

$$\theta \in \left[\frac{\bar{X}_n}{1+\tau^2/n} - \frac{u_{1-\alpha/2}}{\sqrt{n+\tau^2}}; \frac{\bar{X}_n}{1+\tau^2/n} + \frac{u_{1-\alpha/2}}{\sqrt{n+\tau^2}}\right] = I^{HPD}(\tau, \bar{X}_n)$$

où u_α est le quartile d'ordre α de la loi gaussienne standard.

R 3) Tracer les bornes des régions de confiance en fonction de τ. On prend par exemple $\theta = 0$.

R 4) Ajouter sur ce graphique les bornes de la région de confiance classique de niveau $1-\alpha$.

5) Montrer que

$$P_\theta(\theta \in I^{HPD}(\tau, \bar{X}_n)) = F\left(\frac{\theta\tau^2}{\sqrt{n}} + u_{1-\alpha/2}\sqrt{\frac{n+\tau^2}{n}}\right) - F\left(\frac{\theta\tau^2}{\sqrt{n}} - u_{1-\alpha/2}\sqrt{\frac{n+\tau^2}{n}}\right)$$

où F est la fonction de répartition de la loi gaussienne standard.

R 6) Tracer cette probabilité en fonction de τ et θ. Commenter.

7) Comment choisir la loi a priori pour que les régions HPD de niveau $1-\alpha$ soient aussi des régions de confiance au sens classique de niveau $1-\alpha$. Vérifier que la loi de Jeffrey satisfait cette propriété.

Classification bayésienne

Exercice 5

Récupérer les fichiers

`http://www.math.sciences.univ-nantes.fr/~philippe/lecture/mix.simple.R`
`http://www.math.sciences.univ-nantes.fr/~philippe/lecture/mix.expo.R`

- `rmix.gauss` est un générateur de nombres aléatoires suivant un mélange de lois gaussiennes
$$p\mathcal{N}(m_1, \sigma_1^2) + (1-p)\mathcal{N}(m_2, \sigma_2^2)$$
Méthode utilisée : On simule $Z \sim B(p)$, puis $X \sim \mathcal{N}(m_1, \sigma_1^2)$ si $Z = 1$ et $X \sim \mathcal{N}(m_2, \sigma_2^2)$ si $Z = 0$.
La fonction retourne les X_i et les Z_i.
- `rmix.expo` est un générateur de nombres aléatoires suivant un mélange de lois exponentielles
$$p\mathcal{E}(m_1) + (1-p)\mathcal{E}(m_2)$$
- `mcmc.mix.gauss` est un générateur de nombres aléatoires suivant la loi a posteriori pour un mélange gaussien. On estime le paramètre p et on suppose que les paramètres m_i, σ_i sont connus.
La fonction retourne
`list(z.sim=z,p.sim= p)`
  - `z.sim` est une matrice dont la i ème colonne contient un échantillon simulé suivant la loi a posteriori de la donnée manquante Z_i.
  - `p.sim` est un vecteur qui contient un échantillon simulé suivant la loi a posteriori du paramètre p

  Méthode utilisée c'est un algorithme de Monte Carlo par chaîne de Markov, l'algorithme de Gibbs.
- `mcmc.mix.expo` est un générateur de nombres aléatoires suivant la loi a posteriori pour un mélange exponentiel. On estime les paramètres $p; m_1, m_2$ sous la contrainte $m_1 > m_2$

R 1) Simuler un échantillon de taille $n = 150$ iid suivant la loi de mélange
$$p\mathcal{N}(m_1, \sigma_1^2) + (1-p)\mathcal{N}(m_2, \sigma_2^2)$$
- $p = 1/2\,;\ m_1 = 0\,;\ \sigma_1 = 1\,;\ m_2 = 5\,;\ \sigma_2 = 1$
- $p = 1/4\,;\ m_1 = 0\,;\ \sigma_1 = 1\,;\ m_2 = 5\,;\ \sigma_2 = 1$
- $p = 1/2\,;\ m_1 = 0\,;\ \sigma_1 = 1\,;\ m_2 = 1\,;\ \sigma_2 = 1$
- $p = 1/2\,;\ m_1 = 0\,;\ \sigma_1 = 1\,;\ m_2 = 0\,;\ \sigma_2 = 3$

R 2) En utilisant `mcmc.mix.gauss`, donner une approximation de l'estimateur de Bayes du paramètre p

R 3) Donner une approximation des probabilités a posteriori $P(Z_i = 1 | X_1 \dots X_n)$ pour $i = 1, \dots, n$.

R 4) En déduire une classification des observations.

R 5) Estimer le pourcentage de variables bien classées.

R 6) Simuler un échantillon de taille $n = 150$ iid suivant la loi de mélange

$$p\mathcal{E}(m_1) + (1-p)\mathcal{E}(m_2)$$

Choix des paramètres

- $p = 1/2$; $m_1 = 3$; $m_2 = 1$
- $p = 1/4$; $m_1 = 5$; $m_2 = 1$

R 7) Tracer sur un même graphique
- l'estimation de la loi des observations (histogramme)
- la densité de la loi de mélange $p\mathcal{E}(m_1) + (1-p)\mathcal{E}(m_2)$
- les densités des lois exponentielles $\mathcal{E}(m_1)$ et $\mathcal{E}(m_2)$

R 8) Tracer sur une même page `(par(mfrow=c(2,1))`, l'histogramme des observations issues de la première composante et celui des observations issues de la seconde composante.

R 9) En utilisant `mcmc.mix.expo`, donner une approximation des estimateurs de Bayes des paramètres p, m_1, m_2

R 10) Donner une estimation des probabilités a posteriori $P(Z_i = 1|X_1 \dots X_n)$ pour $i = 1, \dots, n$.

R 11) En déduire une classification des observations.

R 12) Estimer le pourcentage de variables bien classées.

R 13) Tracer l'histogramme des observations classées dans la première composante, puis celui des observations classées dans la seconde composante. Comparer avec les résultats obtenus à ceux de la question 8