Université Paris Dauphine
U.F.R. Mathématiques de la Décision
Stat. Exploratoire

Année 2010-2011

Enoncés de Travaux Dirigés

Statistique exploratoire

Professeur : Christian ROBERT
xian@ceremade.dauphine.fr

Chargés de TD : Nicolas BOUSQUET
nicolas.bousquet@edf.fr

Youssef CHALDI
chaldi.youssef@gmail.com

Alessandra IACOBUCCI
iacob@ceremade.dauphine.fr

Merlin KELLER
merlinkeller@gmail.com

Julien SOHIER
jusohier@gmail.com

Université Paris Dauphine
U.F.R. Mathématiques de la Décision
Stat. Exploratoire

Année 2010-2011

Feuille de Travaux Dirigés 0

Introduction au logiciel R

Lire attentivement le poly d'introduction au logiciel R.

Exercice 1

Recopier les instructions du poly pages 4 et 5.

Exercice 2 : exercices complémentaires sur les vecteurs

1. En vous aidant du poly, écrire de la façon la plus compacte possible, les vecteurs suivants :
   - nombres de 1 à 3 par pas de 0.1 ;
   - nombres de 3 à 1 par pas de -0.1 ;
   - carrés des 10 premiers entiers ;
   - nombres de la forme $(-1)^n n^2$ pour $n = 1 \dots 10$;
   - dix 0 suivis de dix 1 ;
   - trois 0 suivis de trois 1, suivis de trois 2 .... suivis de trois 9 ;
2. A l'aide de la fonction `sample`, simuler 15 tirages aléatoires de *Pile* ou *Face* dans le cas où :
   - la pièce est équilibrée ;
   - la pièce est truquée : on a 3 fois plus de chance d'obtenir un *Pile* qu'un *Face*.

Exercice 3

Recopier les instructions du poly page 6.

Exercice 4 : exercices complémentaires sur les matrices

1. Ecrire, de la façon la plus compacte possible, les matrices carrées d'ordre 6 suivantes :
   - matrice contenant les entiers de 1 à 36 rangés par colonnes puis par lignes ;
   - matrice dont toutes les lignes sont égales au vecteur des entiers de 1 à 6 ;
   - matrice diagonale dont la diagonale contient les entiers de 1 à 6 (voir la fonction `diag` dans l'aide de R) ;
   - matrice diagonale par blocs, contenant un bloc d'ordre 2 ne contenant que des 2 et un d'ordre 4 contentant les nombres de 1 à 16 rangés en lignes ;
   - matrice $A = ((-1)^{i+j}), i, j = 1 \dots 6$;
   - matrice contenant des "1" sur la diagonale principale, puis des "2" sur les diagonales du dessus et du dessous, puis des "3"....

2. Ecrire la matrice $A = (a_{i,j})$ d'ordre 12 contenant les entiers de 1 à 144 rangés par lignes. Afficher la dimension de la matrice A.

   Extraire de cette matrice les matrices suivantes :
   - coefficients $(a_{i,j})$ pour $i = 1 \dots 6$ et $j = 7 \dots 12$;
   - coefficients $(a_{i,j})$ pour $i + j$ pair ;
   - coefficients $(a_{i,j})$ pour $i, j = 1, 2, 5, 6, 9, 10$.

3. En partant à chaque fois de la matrice A, appliquer les transformations suivantes :
   - supprimer les colonnes 2, 4, 6, 8, 10, 12 ;
   - annuler les termes < 6 de A ;
   - créer une matrice B d'ordre 12 telle que $B_{i,j} = 1$ si $A_{i,j} < 5$, et $B_{i,j} = 0$ sinon ;
   - calculer la somme de A par colonnes, puis par lignes. Calculer la somme de tous les termes de la matrice A (regarder la fonction `apply` dans l'aide).

Exercice 5

Exercice à propos des structures de données : recopier les instructions pages 7 et 8 du poly.

Université Paris Dauphine — Année 2010-2011
U.F.R. Mathématiques de la Décision
Stat. Exploratoire

Feuille de Travaux Dirigés 1

Simulation de variables aléatoires : inversion générique, Box-Müller et algorithme d'Acceptation-Rejet

1 Inversion générique

Principe d'Inversion générique : Rappel

Si U est une variable aléatoire uniforme sur $[0,1)$ et F_X est la fonction de répartition de la variable X, $F_X^{-1}(U)$ a même loi que X

Preuve. On a $P(F_X^{-1}(U) \leq x) = P(U \leq F_X(x)) = F_X(x)$.
Note. Si F_X n'est pas strictement croissante, on prend

$$F_X^{-1}(u) = \inf\{x; F_X(x) \geq u\}$$

Exercice 1 : Application de la méthode d'inversion générique

1. Ecrire une fonction permettant de simuler un échantillon $(X_1, \ldots, X_n)$ de taille n tel que les X_i sont i.i.d. de loi exponentielle de paramètre λ en utilisant la méthode d'inversion générique.
2. Simuler un échantillon de taille 1000 de loi exponentielle à partir de la méthode précédente. Afficher sur un même graphe l'histogramme des réalisations et la fonction de densité d'une loi exponentielle
3. Refaire de même avec une loi de Cauchy.

2 Transformation de Box-Müller

Transformation de Box-Müller : Rappel

Si $U_1, U_2 \sim_{i.i.d.} \mathcal{U}[0,1]$, alors posons :

$$X_1 = \sqrt{-2\log(U_1)}\sin(2\pi U_2) \quad X_2 = \sqrt{-2\log(U_1)}\cos(2\pi U_2)$$

X_1 et X_2 sont i.i.d. de loi normale centrée réduite.

Exercice 2 : Application de la méthode de Box-Müller et loi de Cauchy

Nous considérons une variable aléatoire X suivant une loi de Cauchy $\mathcal{C}(0,1)$ de densité

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$$

1. Montrer que l'on peut simuler des réalisations de X, à l'aide de l'algorithme de Box-Müller de simulation de lois normales en utilisant la propriété suivante : si X_1 , X_2 sont i.i.d de loi $\mathcal{N}(0,1)$, alors $\frac{X_1}{X_2} \sim \mathcal{C}(0,1)$.
2. Etudier l'évolution de la moyenne empirique

$$X_n = \frac{1}{n}\sum_{i=1}^{n} Xi$$

où $X_i \sim_{i.i.d.} \mathcal{C}(0,1)$ quand $n \geq 1$ croît (n allant de 1 à 10000 par exemple). Qu'en pensez vous ? Montrer que la loi $\mathcal{C}(0,1)$ n'a pas de moyenne. Qu'en désuisez-vous ?
3. *Pour aller plus loin (exercice à la maison)* : Construire une expérience de Monte Carlo pour montrer empiriquement que X_n a aussi une loi de Cauchy $\mathcal{C}(0,1)$, $\forall n \geq 1$

3 Algorithme d'acceptation rejet

Algorithme d'Acceptation-Rejet : Rappel

On souhaite simuler une réalisation de la variable aléatoire X de loi caractérisée par la densité f.

1. Trouver une densité g facilement simulable telle que $\sup_x \frac{f(x)}{g(x)} = M$. ($M \in]1, < \infty[$)
2. Générer
$$Y_1, Y_2, \ldots \sim_{i.i.d.} g\,, \qquad U_1, U_2, \ldots \sim_{i.i.d.} \mathcal{U}([0,1])$$
3. Prendre $X = Y_k$ où
$$k = \inf\{n\,; U_n \leq f(Y_n)/Mg(Y_n)\}$$

La variable produite par la régle d'arrêt ci-dessus est distribuée suivant la loi f_X

Exercice 3 : Application de l'algorithme d'Acceptation-Rejet

1. A l'aide de la méthode Acceptation-Rejet, générer une réalisation d'une variable aléatoire normale centré réduite. *On pourra utiliser la fonction* `rcauchy`.
   (a) Montrer que la constante M vaut $\sqrt{2\pi}e^{-1/2}$
   (b) Illustrer graphiquement la pertinence de votre algorithme de génération.
   (c) Faire varier la constante M et regarder son influence sur le temps d'attente nécéssaire à la génération de la variable aléatoire.
2. A l'aide de la méthode Acceptation-Rejet, générer une réalisation d'une variable aléatoire X ayant pour densité de probabilité

$$f(x) = \frac{2}{5}(2 + \cos(x))e^{-x}$$

On pourra utiliser la fonction `rexp`. Vérifier graphiquement la pertinence de votre algorithme de génération.

Exercice 4 : Pour aller plus loin : génération de variables aléatoires tronquées
Nous considérons une variable aléatoire gaussienne X centrée réduite contrainte au support $[a, b]$ avec $0 < b$

1. Donner la densité de cette variable aléatoire en précisant la constante de normalisation
2. Tracer avec le logiciel R la probabilité $\mathbb{P}(Y \in [0, b])$ lorsque $Y \sim \mathcal{N}(0, 1)$ et a et b varient.
3. Discuter de la pertinence de l'algorithme qui consiste à simuler $Y \sim \mathcal{N}(0, 1)$ jusqu'à ce que $Y \in [a, b]$
4. On prend $a = 0$. Proposer une méthode d'acceptation-rejet fondée sur une loi exponentielle $\mathcal{E}(\lambda)$. Optimiser en λ et écrire une fonction R adaptée.

Université Paris Dauphine — Année 2010-2011
U.F.R. Mathématiques de la Décision
Stat. Exploratoire

Feuille de Travaux Dirigés 2

Méthodes de Monte Carlo

1 Intégration par méthode de Monte Carlo

Intégration par Monte Carlo : rappels

Soit X une variable aléatoire de densité f et h une fonction définie sur le support de X, telle que $\int |h(x)|f(x)dx < \infty$. Nous cherchons à évaluer

$$\mathfrak{I} = \int h(x)f(x)dx = \mathbb{E}_f[h(X)]$$

Dans de nombreuses situations, ce calcul ne peut être fait de façon explicite. L'intégration par méthode de Monte Carlo en fournit une approximation.
Principe : D'après la loi des grands nombres, si $X_1, \ldots, X_n$ sont indépendantes et identiquement distribuées de loi de densité f alors,

$$\lim_{n\to\infty} \hat{\mathfrak{I}}_n = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} h(X_i) = \mathfrak{I}, \quad \text{p.s.}$$

En pratique : Il suffit de simuler un n-échantillon $X_1, \ldots, X_n$ de loi f et on approchera $\mathfrak{I}$ par $\frac{1}{n}\sum_{i=1}^{n} h(X_i)$.
Vitesse de convergence : Posons $\hat{\sigma}_n^2(h(X)) = \frac{1}{n}\sum_{i=1}^{n}(h(X_i) - \hat{\mathfrak{I}})^2$. Supposons que $\int |h(x)|^2 f(x)dx < \infty$. Alors, d'après le théorème Central Limit

$$\lim_{n\to\infty} \sqrt{n}\frac{\hat{\mathfrak{I}}_n - \mathfrak{I}}{\hat{\sigma}_n(h(X))} = \mathcal{N}(0,1) \quad (\mathcal{L})$$

Conséquence : Intervalle de confiance asymptotique pour $\mathfrak{I}$. Pour n grand, $\hat{\mathfrak{I}}_n \sim \mathcal{N}\left(\mathfrak{I}, \frac{1}{n}\hat{\sigma}_n^2(h(X))\right)$. Ainsi, soit $q_{1-\alpha}$ $(1-\alpha)$-quantile d'une loi $\mathcal{N}(0,1)$. Alors

$$\lim_{n\to} \mathbb{P}\left(\mathfrak{I} \in \left[\hat{\mathfrak{I}}_n - q_{1-\alpha}\frac{1}{\sqrt{n}}\hat{\sigma}_n(h(X)),\, \hat{\mathfrak{I}}_n + q_{1-\alpha}\frac{1}{\sqrt{n}}\hat{\sigma}_n(h(X))\right]\right) = (1-\alpha)\%$$

Remarque : Attention à ne pas confondre la variance de X estimée par $\hat{\sigma}_n^2(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ avec la variance de $\mathfrak{I}$ estimée par $\frac{\hat{\sigma}_n^2(h(X))}{n}$.

Exercice 1 : Application de la méthode de Monte Carlo
Nous considérons une variable aléatoire X de loi $\mathcal{G}amma(a, b)$ de densité de probabilité :

$$f_{a,b}(x) = \frac{b^a x^{a-1}}{\Gamma(a)} \exp(-bx)\mathbb{I}_{x>0}$$

Dans la suite, on utilisera $a = 4$ et $b = 1$.

1. Simuler un échantillon de 1000 réalisations de X.
2. Déterminer par la méthode de Monte-Carlo des approximations de l'espérance et de la variance de X, puis donner la variance de l'estimateur de l'espérance.
3. Calculer par une méthode de simulation des approximations de la fonction de répartition de X aux points 2 et 5.
4. Donner des approximations des quantiles à 85, 90 et 95% de la loi de X .

Exercice 2 : Application de la méthode de Monte Carlo (2)
Nous considérons une variable aléatoire X dont la densité de probabilité est proportionnelle à la fonction suivante :

$$(2 + \sin^2(x)) \exp\left(-\left(2 + \cos^3(3x) + \sin^3(2x)\right) x\right) \mathbf{1}_{\mathbb{R}^+}(x).$$

1. Vérifier que, pour tout $x \in [-\pi, \pi]$, $\cos^3(3x) + \sin^3(2x) > -\frac{7}{4}$ et construire un algorithme de génération de réalisations de X .
2. Déterminer par une méthode de simulation des approximations de l'espérance et de la variance de X .
3. Calculer par une méthode de simulation des approximations de la fonction de répartition de X aux points 0.5, 1, 1.5, 5, 10, 15 et des approximations des quantiles à 85, 90 et 95% de la loi de X .

2 Méthode de Monte Carlo avec échantillonnage d'importance

Importance Sampling : rappels

Nous cherchons toujours à donner une valeur approchée de $\mathfrak{I} = \int h(x)f(x)dx$. Nous introduisons la représentation alternative suivante :

$$\mathfrak{I} = \int h(x)f(x)dx = \int h(x)\frac{f(x)}{g(x)}g(x)dx$$

où g est telle que $\int \left|h(x)\frac{f(x)}{g(x)}\right| g(x)dx < \infty$
Conséquence : Si $Y_1, \ldots, Y_n$ i.i.d. de loi g, par la loi des grands nombres

$$\tilde{\mathfrak{I}}_n = \frac{1}{n}\sum_{i=1}^{n} h(Y_i)\frac{f(Y_i)}{g(Y_i)} \longrightarrow \mathfrak{I} \quad \text{p.s.}$$

En pratique : Il suffit de simuler un n-échantillon $Y_1, \ldots, Y_n$ de loi g et on approchera $\mathcal{I}$ par $\frac{1}{n}\sum_{i=1}^{n} h(Y_i)\frac{f(Y_i)}{g(Y_i)}$

Avantages :
- Fonctionne pour tout choix de g tel que $\text{supp}(g) \supset \text{supp}(f)$
- Amélioration possible de la variance de l'estimateur de $\mathcal{I}$
- Recyclage de simulations $Y_i \sim g$ pour d'autres densités f
- Utilisation de lois simples g

Exercice 3 : Application de la méthode d'Importance Sampling
On cherche à évaluer la valeur de

$$I = \int_2^{\infty} \frac{1}{\pi(1+x^2)}dx$$

1. Calculer analytiquement la valeur de I.
2. Calculer par une méthode de simulation directe une approximation de I, $\widehat{I}_{1,n}$ où $\widehat{I}_n$ est obtenu à partir d'un échantillon de n simulations. Donner l'intervalle de confiance à 95% pour I correspondant.
3. Montrer que $I = \frac{1}{2} - \int_0^2 \frac{1}{\pi(1+x^2)}dx$ et proposer une nouvelle approximation de I, $\widehat{I}_{2,n}$. Donner l'intervalle de confiance à 95% pour I correspondant.
4. Montrer que $I = \int_0^{1/2} \frac{y^{-2}}{\pi(1+y^{-2})}dy$ et proposer une nouvelle approximation de I, $\widehat{I}_{3,n}$.Donner l'intervalle de confiance à 95% pour I correspondant.
5. Tracer $\widehat{I}_{1,n}$ en fonction de n (n entre 1 et 10000). Ajouter les courbes correspondant à $\widehat{I}_{2,n}$ et $\widehat{I}_{3,n}$ en fonction de n.

Exercice 4 : Application de la méthode d'Importance Sampling (2)
Nous considérons une variable aléatoire X dont la densité de probabilité est proportionnelle à la fonction suivante :

$$(2 + \sin^2(x)) \exp\left(-\left(3 + \cos^3(3x)\right)x\right) \mathbf{1}_{\mathbb{R}^+}(x).$$

La densité de X n'est connue qu'à un facteur multiplicatif près. Déterminer par une méthode de simulation une approximation de ce facteur.

Université Paris Dauphine — Année 2010-2011
U.F.R. Mathématiques de la Décision
Stat. Exploratoire

Feuille de Travaux Dirigés 3

Etude de la fonction de répartition

1 Définition de la fonction de répartition empirique

Rappels : Fonction de répartition empirique

Définition : Soit $(X_1, X_2, ..., X_n)$ un n-échantillon de v.a. indépendantes de fonction de répartition F. Sans aucune hypothèse supplémentaire sur F, cette fonction peut être estimée en tout point t par la fonction de répartition empirique $\widehat{F_n}$:

$$\widehat{F_n}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{X_i \leq t\}}$$

Remarque : $\widehat{F_n}(t)$ est un estimateur non-paramétrique de $F(t)$.
$\widehat{F_n}(t)$ est un estimateur sans biais de $F(t)$. $\widehat{F_n}$ est une fonction en escaliers :

$$\hat{F}_n(t) = \begin{cases} 0 & \text{si } t < X_{(1)} \\ \frac{1}{n} & \text{si } X_{(1)} \leq t < X_{(2)} \\ \vdots & \\ \frac{i}{n} & \text{si } X_{(i)} \leq t < X_{(i+1)} \\ \vdots & \\ 1 & \text{si } t \geq X_{(n)} \end{cases}$$

où $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ correspond aux valeurs de X ordonnées dans le sens croissant.

Dans les exercices 1 à 3, $\boldsymbol{X}$ est un n-échantillon de loi $\mathcal{N}(0,1)$.

Exercice 1 : Calcul et représentation de $\widehat{F_n}$

1. Créer une fonction qui, à partir du n-échantillon $\boldsymbol{X}$, calcule $\widehat{F_n}(t)$.
2. Tracer le graphe de $\widehat{F_n}$ (prendre $n = 100$). Superposer la courbe de F.

A la maison : Même travail avec un échantillon de loi $\mathcal{E}xp(1)$.

2 Comportement asymptotique de la fonction de répartition empirique

Rappels : Loi forte des grands nombres et théorème central limit

Loi forte des grands nombres (LFGN) : En tout point t, $\widehat{F_n}(t)$ est la proportion d'observations inférieures à t, autrement dit un estimateur de $P(X \leq t)$. Une conséquence de la LFGN est que

$$\forall\ t,\ \widehat{F_n}(t) \xrightarrow{ps} F(t), n \to \infty$$

Théorème central limit (TCL) : En remarquant que $\mathbb{I}_{\{X \leq t\}} \sim \mathcal{B}er(F(t))$, on obtient facilement que $\mathbb{V}(\widehat{F_n}(t)) = \frac{1}{n}F(t)(1-F(t))$. L'application du TCL nous donne que :

$$\sqrt{n}(\widehat{F_n}(t) - F(t)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, F(t)(1-F(t))),\ n \to \infty$$

Conséquence : En utilisant le TCL et la LFGN, On peut construire un intervalle de confiance pour $F(t)$. Soit $q_{1-\alpha/2}$ le $(1-\alpha/2)$-quantile d'une loi $\mathcal{N}(0,1)$. Alors

$$\lim_{n\to\infty} \mathbb{P}\left(F(t) \in \left[\widehat{F_n}(t) \pm q_{1-\alpha/2}\frac{1}{\sqrt{n}}\sqrt{\widehat{F_n}(t)(1-\widehat{F_n}(t))}\right]\right) = (1-\alpha)\%$$

Exercice 2 : Vérification de la LFGN et du TCL

1. LFGN : Etudier la convergence de $\widehat{F_n}$ vers F (Utiliser pour n les valeurs $30, 50, 100, 500$).
2. TCL : Vérifier graphiquement que $\sqrt{n}(\widehat{F_n}(t) - F(t)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, F(t)(1-F(t)))$ (Utiliser les mêmes valeurs de n que précédemment, $t = 0$). Donner les $IC_{.95}$ correspondants.

A la maison : Même travail avec un échantillon de loi $\mathcal{E}xp(1)$ et $t = 2$.

3 Précision de l'estimateur de la fonction de répartition

Rappels : Précision de l'estimateur

Définition : Pour un intervalle de confiance de la forme

$$IC_{1-\alpha}(I) = \left(\widehat{I} \pm q_{1-\alpha/2}\sigma_n(\widehat{I})\right)$$

la précision correspond à la longueur de l'intervalle, soit $p = 2q_{1-\alpha/2}\sigma_n(\widehat{I})$.

Conséquence : Avec une approximation de $\sigma_n(\widehat{I})$ en fonction de n, on peut trouver la taille d'échantillon nécessaire pour obtenir une précision donnée.

Exemple d'approximation : Pour une loi symétrique en a, $F(a) = 1/2$, donc $F(t)(1-F(t)) \approx \frac{1}{4}$ pour t proche de a. Par conséquent, pour t proche de a, la variance de $\widehat{F_n}(t)$, peut être approchée par $\frac{1}{4n}$ ($\sigma_n(\widehat{I}) \approx \frac{1}{2\sqrt{n}}$).

Exercice 3 : Détermination de la taille d'échantillon nécessaire pour obtenir une précision donnée

1. Toujours en utilisant un échantillon $\boldsymbol{X}$ de loi normale centrée réduite, déterminer une approximation de la variance de $\widehat{F_n}(t)$ quand $t \approx 0$.
2. En déduire la taille d'échantillon n^* nécessaire pour obtenir une précision de $10e-3$. Vérifier par expérience de Monte-Carlo que la précision obtenue est suffisante.

A la maison : Construire un $IC_{.95}$ de $F(t)$ pour $t = 2$ basé sur un échantillon de taille n^*. La précision de cet estimateur est-elle inférieure ou supérieure à $10e-3$? Pourquoi ?

Indication : Regarder le tableau de variation de la fonction $g(x) = x(1-x)$ sur l'intervalle $[0, 1]$.

4 Génération d'un n-échantillon de fonction de répartition $\widehat{F_n}$

Rappels : Génération d'un m-échantillon de fonction de répartition $\widehat{F_n}$

Soit $X_1, \cdots, X_n$ un n-échantillon de fonction de répartition F^X. A partir de cet échantillon, nous sommes capables de calculer la fonction de répartition empirique $\widehat{F_n^X}$. Cette fonction de répartition $\widehat{F_n^X}$ définit une nouvelle loi de probabilité ayant pour support $X_1, \cdots, X_n$ uniquement.

But : On cherche à générer un m-échantillon $Y_1, \cdots, Y_m$ de fonction de répartition $F^Y = \widehat{F_n^X}$.
Conséquence : $Y_i, i = 1, \cdots, m$ peut prendre les valeurs $X_1, \cdots, X_n$ uniquement et

$$\mathbb{P}(Y_i = X_k) = \frac{1}{n} \quad \forall k = 1 \dots n, \forall i = 1 \dots m$$

En pratique, on fera un tirage équiprobable dans l'échantillon $X_1, \cdots, X_n$ en utilisant la fonction `sample` de R.

Attention : La fonction de répartition de Y est $\widehat{F_n^X}$, sa fonction de répartition empirique $\widehat{F_m^Y} = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}_{\{Y_i \leq t\}}$ et on a seulement

$$\forall\ t,\ \widehat{F_m^Y}(t) \xrightarrow{ps} F_n^X(t), m \to \infty$$

Exercice 4 : Simulation d'un échantillon connaissant sa fonction de répartition

Soit X un n-échantillon de $\mathcal{N}(0,1)$, $n = 30$.

1. Représenter la fonction de répartition empirique de X, $\widehat{F_n^X}$.
2. Simuler un échantillon $Y_1, \cdots, Y_m$ suivant la fonction de répartition $\widehat{F_n^X}$.
3. Représenter la fonction de répartition empirique de Y $\widehat{F_m^Y}$.
4. Vérifier graphiquement que $\widehat{F_m^Y}$ se rapproche de $\widehat{F_n^X}$ quand m devient grand (utiliser pour m les valeurs $30, 50, 100, 500$).

Université Paris Dauphine — Année 2010-2011
U.F.R. Mathématiques de la Décision
Stat. Exploratoire

Feuille de Travaux Dirigés 4

Bootstrap

L'objectif de ce TP est de présenter la méthode de réchantillonnage Bootstrap. Cette méthode permet, lorsque les méthodes classiques de statistique ne s'appliquent pas, de résoudre des probèmes usuels de statistique inférentielle (biais, variance, erreur quadratique moyenne d'un estimateur, intervalles de confiance, tests d'hypothèses...).

Le bootstrap est une technique d'inférence statistique basée sur une succession de rééchantillonnages. Soit X une variable aléatoire réelle de fonction de répartition F inconnue : $F(x) = P(X \leq x)$. Soit $(X_1, ..., X_n)$ un échantillon de la loi de X et F_n la fonction de répartition empirique associée : $F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{X_i \leq x}$.

On s'intéresse à un paramètre θ de la loi de X. θ peut s'écrire comme une fonctionnelle de F soit $\theta = t(F)$. Un estimateur naturel de $\theta = t(F)$ est donné par $\widehat{\theta} = t(\widehat{F}_n) = T(X_1, \ldots, X_n)$.

Exemples :

1. Si $\theta = E[h(X)] = \int h(x)dF(x)$, où h est une fonction de $\mathbb{R}$ dans $\mathbb{R}$,
   $\hat{\theta} = \int h(x)d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n h(X_i)$
2. Si $\theta = Var[h(X)] = E[(h(X) - E[h(X)])^2] = \int h(x)^2 dF(x) - \left(\int h(x)dF(x)\right)^2$ où h est une fonction de $\mathbb{R}$ dans $\mathbb{R}$,
   $\hat{\theta} = \int h(x)^2 d\widehat{F}_n(x) - \left(\int h(x)d\widehat{F}_n(x)\right)^2 = \frac{1}{n}\sum_{i=1}^n h(X_i)^2 - \left[\frac{1}{n}\sum_{i=1}^n h(X_i)\right]^2$
3. Si θ est la médiane de X, $\hat{\theta} = (X_{(n/2)} + X_{(n/2+1)})/2$ si n est pair, $X_{(n+1)/2}$ si n est impair, où $(X_{(1)}, ..., X_{(n)})$ est la statistique d'ordre associée à $X_1, ..., X_n$ (i.e. l'échantillon ordonné par ordre croissant) .
4. Si θ est le quantile de niveau $1 - \alpha$ de la loi de X ,$\hat{\theta} = X_{([(1-\alpha)n]+1)}$.

On souhaite estimer le biais, la variance ou l'erreur quadratique d'un tel estimateur $\hat{\theta} = T(X_1, ..., X_n)$, obtenir des intervalles de confiance sur θ etc...

1 Estimation du biais de $\hat{\theta} = T(X_1, ..., X_n)$.

On appelle biais de $\hat{\theta}$ la quantité $E[\hat{\theta}] - \theta$. Ce biais est en général inconnu puisqu'il dépend de F, elle-même inconnue. On souhaite l'estimer sur la base d'une observation $(X_1, \ldots X_n)$. Il existe plusieurs situations possibles.
Tout d'abord supposons que F **et** θ **sont connus** (ce qui est seulement un cas d'école puisque si tel était le cas, nous n'aurions pas besoin d'estimer θ !). Alors :

- soit on peut calculer $E[\hat{\theta}] - \theta$ analytiquement et le probléme est réglé,
- soit on ne peut pas calculer $E[\hat{\theta}] - \theta$ analytiquement et on a recours à une procédure de Monte Carlo. Plus précisément,
  1. On simule B n-échantillons $(X_1^l, ..., X_n^l)$ sous la loi F.
  2. Pour chaque échantillon, on calcule $\hat{\theta}^l = T(X_1^l, ..., X_n^l)$.
  3. On obtient finalement une estimation du biais $E[\hat{\theta}] - \theta$ par : $\frac{1}{B}\sum_{l=1}^{B} \hat{\theta}^l - \theta$

Mais, dans la réalité, F et θ ne sont pas connus. La procédure précédente est donc inapplicable.

La méthode du Bootstrap consiste à remplacer dans la procédure de Monte Carlo précédente F par $\widehat{F}_n$ et θ par $\hat{\theta}$.

Rappel : Au cours du TP 3, nous avons appris comment simuler un n-échantillon sous la loi $\widehat{F}_n$: cela revient é tirer au hasard avec remise n variables dans les observations $X_1, \dots X_n$

Finalement la procédure Bootstrap d'estimation du biais s'écrit :

Procédure Bootstrap d'estimation du biais

1. Calcul de $\hat{\theta}$ à partir des observations $X_1 \dots X_n$.
2. Pour $l = 1 \dots B$,
   (a) On simule un n-échantillon $(X_1^{*l}, ..., X_n^{*l})$ sous la loi $\widehat{F}_n$ i.e. on tire au hasard avec remise n variables dans les observations $(X_1, ..., X_n)$: sample(X,n,replace=TRUE).
   (b) Pour chaque nouvel échantillon, on calcule $\hat{\theta}^{*l} = T(X_1^{*l}, ..., X_n^{*l})$
3. On obtient finalement une estimation du biais $E[\hat{\theta}] - \theta$ par : $\frac{1}{B}\sum_{l=1}^{B} \hat{\theta}^{*l} - \hat{\theta}$

Exercice 1

On s'intéresse à $\widehat{\sigma_n^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ l'estimateur naturel de la variance $V(X) = \sigma^2$.

1. On suppose que X suit une loi $\mathcal{N}(0,1)$.
   (a) Simuler un 100 - échantillon $(X_1, \dots, X_{100})$ et le stocker dans le vecteur XX.
   (b) Calculer analytiquement le biais $b = E[\widehat{\sigma_n^2}] - \sigma^2$.
   (c) Evaluer ce biais par la méthode de Monte Carlo. Tracer un graphe faisant apparaître les variations de cette approximation en fonction du nombre d'itérations B. Ajouter à ce graphe une droite horizontale d'ordonnée b.
2. On reprend $(X_1, \dots, X_{100})$ stockées dans le vecteur XX mais on suppose désormais qu'on ne connait pas la loi des observations. Estimer le biais par une procédure Bootstrap. Comme précédemment, tracer un graphe faisant apparaître les variations de cette approximation en fonction du nombre d'itérations B.
3. Comparer les trois méthodes.

Remarque : La même procédure peut être utilisée pour estimer les variance, erreur moyenne quadratique d'un estimateur...

2 Calcul des intervalles de confiance par Bootstrap

Soit $\hat{\theta}$ un estimateur de θ . On cherche à avoir un intervalle de confiance sur θ i.e on cherche $q_1(\hat{\theta})$ et $q_2(\hat{\theta})$ tels que $\mathcal{P}\left[q_1(\hat{\theta}) \leq \theta \leq q_2(\hat{\theta})\right] = 1-\alpha$ (α fixé).

Rappels sur les intervalles de confiance (voir cours de Stat de L2) :

1. Si $(X_1, \ldots, X_n)$ un échantillon de loi normale $\mathcal{N}(\theta, \sigma^2)$, alors $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n$ est l'estimateur naturel de θ .
   (a) Si σ^2 est connu, soit Φ la fonction de répartition d'une loi normale centrée réduite et q tel que $\Phi(q) = 0.975$. Alors $I_c = [\hat{\theta} - q\frac{\sigma}{\sqrt{n}}, \hat{\theta} + q\frac{\sigma}{\sqrt{n}}]$ est un intervalle de confiance exact de θ de niveau 95%.
   (b) Si σ^2 est inconnu, sous l'hypothèse de normalité des données, posons $\widehat{S}_n = \frac{1}{n-1}\sum_{i-1}^{n}(X_i - \overline{X}_n)^2$ un estimateur de la variance. Alors $\frac{\overline{X}_n - \theta}{\sqrt{\widehat{S}_n/n}} \sim \mathcal{T}(n-1)$ Ainsi, soit Φ_T la fonction de répartition d'une loi de Student à $n-1$ degrés de liberté et q tel que $\Phi_T(q) = 0.975$ alors $I_c = \left[\overline{X}_n - q\sqrt{\widehat{S}_n/n}, \overline{X}_n + q\sqrt{\widehat{S}_n/n},\right]$ est un intervalle de confiance de θ de niveau 95%.
2. Maintenant, supposons que $(X_1, \ldots, X_n)$ est un n-échantillon de X de loi inconnue telle que $E[X] = \theta$. Comme précédemment, $\hat{\theta} = \overline{X}_n$ est un estimateur naturel de θ.
   (a) Si σ^2 est connu, le théorème Central-Limit permet de donner un intervalle de confiance asymptotique.
   (b) Si σ^2 est inconnu, soit $\widehat{S}_n$ un estimateur consistant de σ^2. Grâce au lemme de Slutsky et au théorème Central Limit, on a convergence en loi de $\frac{\overline{X}_n - \theta}{\sqrt{\widehat{S}_n/n}}$ vers une loi normale centrée réduite. Nous obtenons alors facilement des intervalles de confiance asymptotiques.

Problèmes :

- Les résultats précédents s'appliquent si on a un théorème du type Central Limit.
- Ces résultats sont asymptotiques donc valables seulement si on a un grand nombre d'observations.
- Ils necéssitent d'avoir un estimateur consistant de la variance.

La procédure des percentiles Bootstrap permet de résoudre ce probléme. Son principe est d'approcher la fonction de répartition de l'estimateur $\hat{\theta} = T(X_1, ..., X_n)$ par sa fonction de répartition empirique obtenue grâce à un échantillonnage Bootstrap. Les bornes de l'intervalle de confiance sont alors obtenues à partir de cette fonction de répartition empirique.

Procédure des percentiles Bootstrap

Soit $(X_1, \ldots, X_n)$ un n-échantillon observé.

1. Pour $l = 1 \ldots B$,
   (a) On simule un n-échantillon $(X_1^{*l}, ..., X_n^{*l})$ sous la loi $\widehat{F}_n$ i.e. on tire au hasard avec remise n variables dans les observations $(X_1, ..., X_n)$:

   sample(X,n,replace=TRUE).

   (b) On calcule $\hat{\theta}^{*l} = T(X_1^{*l}, ..., X_n^{*l})$
2. Les $(\hat{\theta}^{*l})_{l=1...B}$ fournissent une approximation de la fonction de répartition de $\hat{\theta}$. On calcule q_1 et q_2 en utilisant la fonction quantile.

Exercice 2

On s'intéresse à l'espérance μ d'un échantillon d'une loi normale.

1. Simuler un n-échantillon $(X_1, ..., X_n)$ pour $n = 10$ avec une loi normale $\mathcal{N}(5, 2)$. On suppose désormais μ et σ^2 inconnus.
2. Calculer un intervalle de confiance à 95% pour μ à l'aide de la loi de Student.
3. Calculer un intervalle de confiance asymptotique à 95% pour μ à l'aide du théorème Central Limit
4. Calculer un intervalle de confiance asymptotique à 95% pour μ à l'aide de la procédure des percentiles Bootstrap
5. Comparer les méthodes
6. Reprendre l'énoncé en faisant varier n.

3 Tests d'hypothèses par méthode Bootstrap

Problème

Dans de nombreux problèmes pratiques, on modélise fructueusement la relation existant entre deux quantités Y et X par une forme affine. Supposons que l'on dispose de n valeurs de X fixées, notéees x_i, et que pour chaque x_i on observe une réalisation d'une variable aléatoire Y , notée y_i. Le modèle de régression linéaire simple consiste à postuler que :

$$Y_i = \alpha + \beta x_i + E_i$$

où les E_i sont des variables aléatoires i.i.d. d'espérance nulle et de variance égale à σ^2. On peut alors estimer les paramètres inconnus α et β grâce au critère des moindres carrés. Cela consiste à déterminer les quantités $\hat{\alpha}$ et $\hat{\beta}$ minimisant : $\sum_{i=1}^{n}(Y_i - \alpha - \beta x_i)^2$.
Posons $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{Y} = \sum_{i=1}^{n} Y_i$, $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$ et $S_{YY} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$, $S_{xY} = \sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})$. Les estimateurs des moindres carrés de α et β et s'écrivent :

$$\hat{\beta} = \frac{S_{xY}}{S_{xx}} \quad \hat{\alpha} = \overline{Y} - \hat{\beta}\overline{x}$$

1. Pour $n = 30$, $\sigma^2 = 0.5$, $\alpha = 2$, $\beta = 1$ et $x = seq(0, 10, length = n)$, générer des réalisations de Y_i selon le modèle de régression linéaire simple dans le cas où les résidus E_i suivent $\gamma\mathcal{T}(5)$ où γ est une constante à déterminer.
2. Déterminer une estimation bootstrap de α et β
3. Déterminer, par échantillonnage bootstrap, un intervalle de confiance à 95% pour les paramètres α et β.
4. Posons

$$T = \sqrt{(n-2)S_{xx}}\,\frac{\hat{\beta} - 1}{\sum_{i=1}^{n}(Y_I - \hat{\alpha} - \hat{\beta}x_i)2}$$

Sur l'échantillon généré à la question 1, nous souhaitons tester l'hypothèse nulle H_0 selon laquelle $\beta = 1$ contre l'hypothèse alternative H_1 selon laquelle $\beta = 1.5$. Nous proposons d'utiliser la règle de décision consistant à rejeter l'hypothèse H_0 si $T > F^{-1}_{St(28)}(0.95)$, stratégie optimale lorsque les résidus sont gaussiens. Déterminer, par échantillonnage bootstrap, le taux d'erreur du test précédent (i.e. probabilité de rejeter H_0 en commettant une erreur, probabilité égale à 5% dans le cas gaussien).

Université Paris Dauphine
UFR de mathématiques de la Décision / MD3
Mat 6-3- NOISE

Année 2010-2011

Feuille de Travaux Dirigés 5

Statistique non paramétrique: estimation de la densité.

Dans ce TP, on s'intéresse à l'estimation de fonctions de densité par une méthode non- paramétrique.

Soit X une variable aléatoire réelle de fonction de densité f inconnue (notons F la fonction de répartition correspondante). Soit $(X_1, ..., X_n)$ un échantillon issu de X.

But: On cherche à estimer f à partir de l'échantillon observé.

Une idée qui ne marche pas: nous savons que $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{X_{(i)}\leq x}(x)$ est un estimateur convergent et sans biais de F. Or $f(x) = \frac{dF}{dx}(x)$, on peut donc songer à prendre la dérivée de la fonction de répartition empirique. Or cette fonction n'est pas dérivable, *il faut donc estimer directement f.*

1 Estimation par histogramme

1.1 Principe

L'idée naturelle est d'approcher la densité f par l'histogramme des observations. Ceci revient à proposer comme estimateur de f une fonction en escaliers $\widehat{f}_n(x) = \sum_{k=1}^{K} \omega_k \mathbb{I}_{[a_k, a_{k+1}[}(x)$, où les $(a_k)_{k=1...K+1}$ et $(\omega_k)_{k=1...K}$ vérifient les trois propriétés suivantes:

- $a_1 < \cdots < a_{K+1}$.
- $\sum_{k=1}^{K} \omega_k(a_{k+1} - a_k) = 1$ (aire sous la courbe).
- La quantité $\omega_k(a_{k+1} - a_k)$ est un estimateur de la $\mathbb{P}_F(X \in [a_k, a_{k+1}[)$. i.e., par exemple:

$$\omega_k(a_{k+1} - a_k) = \frac{\text{Nombre d'observations dans l'intervalle } [a_k, a_{k+1}[}{n} = \widehat{F}_n(a_{k+1}) - \widehat{F}_n(a_k)$$

$\hookrightarrow$ **Sous R**, `hist(x)$density` donne les poids $(\omega_k)_{k=1...K}$ et `hist(x)$breaks` donne les $(a_k)_{k=1...K+1}$.

Interprétation probabiliste: L'estimation $\widehat{f}_n(x)$ en escaliers revient en fait à estimer la loi des observations par un mélange de lois uniformes. En effet:

$$\widehat{f}_n(x) = \sum_{k=1}^{K} \omega_i \mathbb{I}_{[a_k, a_{k+1}[}(x)) = \sum_{k=1}^{K} \omega_i(a_{k+1} - a_k) \underbrace{\frac{\mathbb{I}_{[a_k, a_{k+1}[}(x)}{a_{k+1} - a_k}}_{\text{densité de } \mathcal{U}_{[a_k, a_{k+1}[}}$$

Remarque: Il est possible d'obtenir un estimateur de la fonction de répartition F à partir de $\widehat{f}_n$ en intégrant cette fonction ($F(x) = \int_{-\infty}^{x} f(t)dt$). Le nouvel estimateur de F obtenu est une fonction linéaire par morceaux donc différente de la fonction de répartition empirique.

1.2 Défauts de la méthode et solutions

La méthode précédente repose sur le choix des $(a_k)_{k=1\dots K+1}$ (dépendant en général des observations). Leur choix est délicat. En effet:

- Nécéssairement, $a_1 > -\infty$ et $a_{K+1} < \infty$. Sinon $\int_{\mathbb{R}} \widehat{f}_n(x)dx = \infty$. Or, en général, $Supp(f) = \mathbb{R}$. a_1 et a_{K+1} doivent donc être choisi de façon à approcher au mieux le support de f.
- On cherche à ce que $\lim_{n\to\infty} \widehat{f}_n(x) = f(x)$ i.e. plus la taille de l'échantillon est grande, meilleur est l'estimateur. Ainsi $(a_k)_{k=1\dots K+1}$ et $(\omega_k)_{k=1\dots K+1}$ doivent dépendre de n. Mais de quelle façon? Si $(a_{k+1}(n) - a_k(n))$ décroit est trop vite vers 0, le nombre d'observations tombant dans l'intervalle $[a_k, a_{k+1}[$ est trop petit pour avoir une bonne estimation.

  $\Rightarrow$ **Fenêtres de Scott** : *choix optimal de la largeur des classes* $[a_k, a_{k+1}[$
  Un choix optimal de largeur de classe est donné par

  $$a_{k+1}(n) - a_k(n) = h_n = 3.5\widehat{\sigma}n^{-1/3}.$$

  où $\widehat{\sigma}$ est un estimateur du l'écart type des observations. De plus, cette largeur de classe assure la convergence de $\widehat{f}_n$ vers f quand n tend vers l'infini.

$\hookrightarrow$ **Sous R**, on utilisera la fonction `hist` en spécifiant le nombre de classes par `nclass=(max(`x`)-min(`x`))/`h_n

Exercice

On s'intéresse à la densité d'un mélange de lois normales: $\frac{1}{3}\mathcal{N}(0, 1.5^2) + \frac{2}{3}\mathcal{N}(4, 1.5^2)$.

1. *Simulation des données*: donner et tracer la densité f de cette loi. Ecrire une fonction `rmelange.m` permettant de simuler un échantillon de taille n issu de f et simuler un échantillon de taille $n = 450$. Stocker les valeurs dans un vecteur `XX`.

   Rappel : Pour simuler une v.a. X de densité $p\ f_0 + (1-p)\ f_1$ où f_0 et f_1 sont les composantes du mélange, les 2 étapes sont :

   (a) Choix de la composante : $c \sim \mathcal{B}(p)$
   (b) Tirage de $X|c \sim f_c$ où c est la composante choisie à l'etape précédente.

   On cherche maintenant à retrouver f à partir des observations `XX`.

2. *Estimation paramétrique de f*: On cherche d'abord à estimer f par une gaussienne i.e. on cherche un estimateur sous la forme:

   $$g_\theta(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{(x-\mu)^2}{2\sigma^2}$$

   Donner un estimateur $\widehat{\theta}$ de $\theta = (\mu, \sigma^2)$. En déduire un estimateur $g_{\widehat{\theta}}$ de f. Tracer sur un même graphe f et son estimateur paramétrique $g_{\widehat{\theta}}$. Conclure

3. *Estimation par histogramme de f*:

   (a) Utiliser la fonction `hist` pour donner un estimateur de f.
   (b) Faire varier le nombre de classes K. Comparer aux fenêtres optimales de Scott. On tracera les estimations obtenues pour différents K sur plusieurs graphiques dans une même fenêtre (utiliser `par(mfrow=c(2,2))`).

2 Estimation par noyau

2.1 Principe

Nous avons vu que l'estimateur par histogramme approche f par un mélange de lois uniformes. Nous cherchons maintenant à généraliser cette méthode à d'autres mélanges. Revenons à la définition de f comme dérivée de la fonction de répartition empirique:

$$f(x) = \frac{dF}{dx}(x) = \lim_{h\to 0} \frac{F(x+h) - F(x-h)}{2h} \simeq \frac{F(x+h) - F(x-h)}{2h} \qquad \text{pour } h \text{ petit} \tag{2.1}$$

De (**??**), on déduit un nouvel estimateur $\widehat{f}_n$ de f (pour h petit) :

$$\widehat{f}_n(x) = \frac{\widehat{F}_n(x+h) - \widehat{F}_n(x-h)}{2h} = \frac{1}{2hn}\sum_{i=1}^{n} \mathbb{I}_{[-h,h]}(x - X_i) = \frac{1}{hn}\sum_{i=1}^{n} \frac{1}{2}\mathbb{I}_{[-1,1]}\left(\frac{x - X_i}{h}\right)$$

Cet estimateur est un cas particulier de l'estimation par histogramme avec des $(a_k)_{k=1\dots K+1}$ de la forme $X_i \pm h$. On parle d'*estimation par noyau uniforme.*

Exercice (Suite)

4 *Estimation par noyau uniforme de f*:

(a) Ecrire une fonction estimant f par des noyaux uniformes pour h fixé. On utilisera la fonction `density` avec l'argument `kernel="rectangular"`.

(b) Faire varier h et comparer graphiquement f et ses estimations (on fera un graphe pour chaque h).

Au lieu de considérer une approximation uniforme autour de chaque observation X_i, on peut utiliser une distribution plus lisse de la forme :

$$\widehat{f}_n(x) = \frac{1}{hn}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{2.2}$$

où K est une densité de probabilité (aussi appelée "noyau") et h est un facteur d'échelle ou largeur de fenêtre (assez petit).

2.2 Choix du noyau

En théorie, toutes les densités K sont acceptables cependant certaines sont privilégiées en pratique et implémentées sous R. Plus précisément, on trouve:

- le noyau normal $K(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)$ $\hookrightarrow$Utiliser la fonction `density` avec l'option `kernel="g"`
- le noyau d'Epanechnikov $K(y) = C(1-y^2)^2\mathbb{I}_{[-1,1]}(y)$ $\hookrightarrow$ `kernel="epanechnikov"` ou `"e"`
- le noyau triangulaire $K(y) = (1+y)\mathbb{I}_{[-1,0]}(y) + (1-y)\mathbb{I}_{[0,1]}(y)$ $\hookrightarrow$ `kernel="triangular"` ou `"t"`

Exercice (Suite)

5 *Estimation de f par d'autres noyaux* : Estimer f par tous les noyaux proposés ci-dessus (uniforme y compris) et comparer les réponses obtenues. Conclure.

2.3 Propriétés de l'estimateurs à noyaux et du paramètre d'échelle h

Le choix de h est crucial puisqu'il conditionne la qualité de l'estimateur. En effet, si h est trop grand, beaucoup d'observation contribuent à l'estimation de $f(x)$ et la fonction obtenue est trop lisse (on parle d'"over-smoothing"). Si h est trop petit, alors trop peu d'observations X_i contribuent à l'estimation de $f(x)$: on parle d'"under-smoothing". Il faut donc trouver un compromis.
Pour cela, définissons l'Erreur Moyenne Intégrée (EMI):

$$d(f, \widehat{f}_h) = \mathbb{E}\left[\int \left(f(x) - \widehat{f}_h(x)\right)^2 dx\right].$$

On cherche le paramètre d'échelle dépendant de n (noté dans la suite h_n) minimisant l'EMI i.e. minimisant la distance entre f et son estimateur $\widehat{f}_{h_n}$:

$$h_n^* = \text{argmin}_{h_n} d(f, \widehat{f}_{h_n}).$$

On peut facilement montrer que l'EMI se décompose en un terme de biais et un terme de variance *(à faire en exercice)*:

$$d(f, \widehat{f}_{h_n}) = \int \left\{ \underbrace{\left(\mathbb{E}\left[\widehat{f}_{h_n}(x)\right] - f(x)\right)^2}_{Biais^2} + \underbrace{\mathbb{E}\left[\widehat{f}^2_{h_n}(x)\right] - \left(\mathbb{E}\left[\widehat{f}^2_{h_n}(x)\right]\right)^2}_{Var} \right\} dx$$

Déterminons le comportement du biais et de la variance en fonction de h_n et déduisons h_n^*.

- **Calcul du Biais**: D'après la définition de l'estimateur $\widehat{f}_h(x)$ (**??**) nous avons:

$$\begin{aligned} \text{Biais}\left(\widehat{f}_h(x)\right) &= \mathbb{E}_f\left(\widehat{f}_{h_n}(x)\right) - f(x) \\ &= \frac{1}{h_n n}\sum_{i=1}^{n} \mathbb{E}_f\left[K\left(\frac{x - X_i}{h_n}\right)\right] - f(x) = \frac{n}{nh_n}\mathbb{E}_f\left[K\left(\frac{x - X}{h_n}\right)\right] - f(x) \\ &= \frac{1}{h_n}\int K\left(\frac{x-s}{h_n}\right) f(s)ds - f(x) = \int K(y)\, f(x - h_n y)dy - f(x) \end{aligned} \tag{2.3}$$

  D'où, d'après (**??**), $\lim_{h_n \to 0} \mathbb{E}_f\left(\widehat{f}_{h_n}(x)\right) - f(x) = 0$. $\widehat{f}_{h_n}(x)$ est donc un estimateur sans biais de f si $h_n \to 0$. Il est possible d'avoir un résultat plus fin si on suppose f de classe $\mathcal{C}^2$. En effet , un développement de Taylor du second ordre du terme $f(x - h_n y)$ autour de x donne alors:

$$\text{Biais}\left(\widehat{f}_{h_n}(x)\right) = \frac{h_n^2}{2} f''(x)\mu_2(K) + o(h_n^2), \quad h_n \to 0$$

  où $\mu_2(K) = \int y^2 K(y)dy$. Par conséquent, le biais de $\widehat{f}_{h_n}(x)$ est quadratique en h_n et dépend de la courbure de f (Biais $\left(\widehat{f}_h(x)\right) \propto f''(x)$). De plus, le biais diminue si h_n diminue.

- **Variance de l'estimateur à noyau**

  Posons $\|K\|_2^2 = \int K^2(y)dy$. Alors, on peut montrer que: $\text{Var}\left(\widehat{f}_{h_n}(x)\right) = \frac{1}{nh_n} f(x)\|K\|_2^2 + o\left(\frac{1}{nh_n}\right)$ où, $nh_n \to \infty$. Ainsi, la variance augmente si h_n diminue.

- **Finalement**, nous obtenons:

$$d(f, \widehat{f}_{h_n}) = \underbrace{\frac{h_n^4}{4}\|f''\|_2^2 \mu_2^2(K)}_{\searrow \text{ si } h_n \searrow} + \underbrace{\frac{1}{nh_n}\|K\|_2^2}_{\nearrow \text{ si } h_n \searrow} + o\left(\frac{1}{nh_n}\right) + o(h_n^4) \tag{2.4}$$

  Cette dernière expression met en évidence la nécéssite d'un paramètre d'échelle h_n permettant un compromis entre le biais et la variance.

- **Optimisation du paramètre d'échelle** Dans le cas du noyau gaussien, on peut montrer que le *paramètre d'échelle optimal théorique* est $h_n^* = \left(n\sqrt{2\pi}\int (f''(x))^2 dx\right)^{-1/5}$ (il suffit de minimiser l'expression (**??**) en h_n par annulation de la dérivée). Le h_n^* dépend de f et est donc impossible à calculer en pratique. On propose donc un *paramètre d'échelle optimal empirique* :

$$\widehat{h}_n^* = \frac{0.9\,\min(\widehat{\sigma}, \widehat{q}_{0.75} - \widehat{q}_{0.25})}{(1.34n)^{1/5}}$$

Remarque: Les constantes 0.9 et 1.34 sont propres au noyau gaussien.

Exercice (Suite)

6 *Estimation de f par noyau gaussien* : Dans l'estimation par noyau gaussien, faire varier h_n. Comparer au paramètre d'échelle optimal empirique.

Université Paris Dauphine
UFR de mathématiques de la Décision / MD3
Mat 6-3- NOISE

Année 2010-2011

Feuille de Travaux Dirigés 6

Test de Kolmogorov-Smirnov.

Dans ce TP, *on s'intéresse au jeu de données* faithful *inclus dans R.*

Etude préliminaire des données

1. Télécharger le jeu de données et en faire une étude sommaire (type de données, taille de l'échantillon, caractères étudiées, moyennes, variances... etc)

   Penser à utiliser data(faithful), help(faithful), summary(faithful)

2. Tracer une représentation graphique permettant de d'écrire sommairement la distribution de probabilité des données faithful

3. Estimer par la méthode du noyau uniforme la densité de probabilité des données faithful.

4. Estimer par la méthode du noyau gaussien la densité de probabilité des données faithful

5. Etudier l'influence de la taille de la fenêtre sur l'estimateur à noyau gaussien.

1 Introduction au test de Kolmogorov-Smirnov

1.1 Test d'adéquation à une loi donnée

Soit $(X_1, \ldots, X_n)$ un n-échantillon d'une loi inconnue P. Soit P_0 une loi connue, fixée. On cherche à tester l'hypothèse

$$\mathcal{H}_0\text{: "les données } X_1, \ldots, X_n \text{ sont distribuées selon la loi } P_0\text{"}$$

contre

$$\mathcal{H}_1\text{: les données } X_1, \ldots, X_n \text{ ne sont pas distribuées selon la loi } P_0.$$

Principe du test : Le test de Kolmogorov-Smirnov permet de répondre à cette problématique. L'idée est la suivante : si l'hypothèse $\mathcal{H}_0$ est correcte, alors la fonction de répartition empirique $\widehat{F}_n$ de l'échantillon doit être proche de F_0, la fonction de répartition conrrespondant à la loi P_0.

Statistique de test : On mesure l'adéquation de la fonction de répartition empirique à la fonction F_0 par la distance de Kolmogorov-Smirnov, qui est la distance de la norme uniforme entre fonctions de répartitions. Pour la calculer, il suffit d'évaluer la différence entre $\widehat{F}_n$ et F_0 aux points $X_{(i)}$.

$$D_{KS}(F_0, \widehat{F}_n) = \max_{i=1,\ldots,n} \left\{ \left| F_0(X_{(i)}) - \frac{i}{n} \right| , \left| F_0(X_{(i)}) - \frac{i-1}{n} \right| \right\} .$$

Construction du test: On va rejetter l'hypothèse $\mathcal{H}_0$ si la distance entre $\widehat{F}_n$ et F_0 est grande, i.e. si $D_{KS}(F_0, \widehat{F}_n)$ dépasse un certain seuil q_α à définir.

A propos du seuil: On choisit le seuil q_α tel que, si l'hypothèse $\mathcal{H}_0$ est vraie, la probabilité de se rejeter $\mathcal{H}_0$ est petite (typiquement $\alpha = 5\%$)

$$\mathbb{P}_{X_i \sim F_0}\left(D_{KS}(F_0, \widehat{F}_n) > q_\alpha\right) = \alpha$$

Pour obtenir ce seuil, il faut connaître la loi de $D_{KS}(F_0, \widehat{F}_n)$ dans le cas où les X_i sont distribués sous la loi F_0. Or on peut montrer que sous l'hypothèse $\mathcal{H}_0$, la loi de la statistique $D_{KS}(F_0, \widehat{F})$ ne dépend pas de F_0. Cependant, la fonction de répartition de $D_{KS}(F_0, \widehat{F})$ n'a pas d'expression explicite simple et doit être calculée numériquement. Cette loi a été tabulée.

$\hookrightarrow$ **Sous R**,

- le test d'adéquation à une loi est implémenté dans la fonction `ks.test`
- la réponse de cette fonction est une liste d'objet dont la p-value. La p-value est le α minimal auquel on aurait rejeté $\mathcal{H}_0$. Si la p-value est inférieure à 5% on rejette l'hypothèse $\mathcal{H}_0$ au niveau 5%

Exercice. Test d'adéquation à une loi

On s'intéresse au temps d'éruptions dépassant 3 minutes.

1. Créer un vecteur long contenant les temps d'éruptions dépassant 3 minutes.
2. Tester l'hypothèse selon laquelle les temps d'éruption observés dépassant 3 minutes suivent une loi $\mathcal{N}(4, 0.1)$.

Remarque: Le test de Kolmogorov-Smirnov s'étend à la comparaison de deux fonctions de répartition empiriques, et permet alors de tester l'hypothèse que deux échantillons sont issus de la même loi. Pour cela on utilise la fonction `ks.test` dans laquelle les seuils sont corrigés.

1.2 Test d'adéquation à une famille de loi

Soit $X_1 \dots X_n$ un n-échantillon de loi inconnue. Soit $\mathcal{F}_\theta$ une famille paramétrique de lois. Par exemple, $\mathcal{F}_\theta = \left\{\mathcal{N}(\mu, \sigma^2), \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^{*+}\right\}$. On cherche maintenant à tester l'hypothèse: $\mathcal{H}_0$: "La loi des données $X_1, \dots, X_n$ appartient à la famille de distributions $\mathcal{F}_\theta$" contre $\mathcal{H}_1$: "La loi des données $X_1, \dots, X_n$ n'appartient pas à la famille de distributions $\mathcal{F}_\theta$"

Méthode: Soit $\widehat{\theta}$ l'estimateur par maximum de vraisemblance du paramètre θ. Comme précédemment, on rejette $\mathcal{H}_0$ si

$$D_{KS}(F_{\widehat{\theta}}, \widehat{F}_n) > q'_\alpha$$

A propos du seuil: Comme précédemment, on calcule le seuil de façon à minimiser la probabilité de rejetter $\mathcal{H}_0$ à tort. Pour cela il faut avoir la loi de la statistique de test $D_{KS}(F_{\widehat{\theta}}, \widehat{F}_n) > q'_\alpha$.

- *Attention*: Comme $F_{\widehat{\theta}}$ dépend de l'échantillon, la loi de la statistique de test n'est pas la même que dans le cas du test d'adéquation à une loi. La fonction ks.test ne réalise pas le test d'ajustement à une famille de loi, mais le test d'ajustement à une loi connue. On ne peut pas l'utiliser dans ce contexte.
- La fonction R permettant de réaliser le test de Kolmogorov-Smirnov d'ajustement à la famille gaussienne est lillie.test(x), du package nortest.

Exercice. Test d'adéquation à une famille loi

Tester l'éventuelle normalité de la distribution de probabilité des temps d'éruption observés dépassant 3 minutes en utilisant la fonction lillie.test(x).