ENSAE
ECOLE NATIONALE
DE LA STATISTIQUE
ET DE L'ADMINISTRATION
ECONOMIQUE

Cursus Intégré
2004-2005

Rappels de statistique mathématique
Enoncé des travaux dirigés n° 8

Guillaume Lacôte
Bureau **E03**
✉ Guillaume.Lacote@ensae.fr
☞ http://ensae.no-ip.com/SE222/

Version du 20050421-15h38, révisée le 21 avril 2005

Enoncé de l'exercice 1

Soit un échantillon $X_1, \ldots, X_n$ de variables aléatoires i.i.d telles que :
– X_i a une probabilité α de valoir a, et
– une probabilité $(1-\alpha)$ de suivre une loi normale $\mathcal{N}(0,1)$,
ce qui s'écrit encore :

$$X_1 = a \cdot \mathbb{1}_{Z_1=1} + Y_1 \cdot \mathbb{1}_{Z_1=0}$$

où $Z_1 \rightsquigarrow \mathcal{B}(1,\alpha)$ est indépendante de $Y_1 \rightsquigarrow \mathcal{N}(0,1)$.

☞ Q1 Montrer que $(\delta_a + \lambda)^{\otimes n}$ est une mesure dominante pour le modèle considéré.

☞ Q2 Montrer que :

$$\frac{dP_{X_1,\ldots,X_n}(x_1,\ldots,x_n)}{d(\delta_a+\lambda)^{\otimes n}} = \prod_{i=1}^{n} (\alpha \mathbb{1}_a(x_i) + (1-\alpha)(1-\mathbb{1}_a(x_i))\phi(x_i))$$

où ϕ est la densité d'une loi normale $\mathcal{N}(0,1)$.

Enoncé de l'exercice 2

Cet exercice s'inspire librement des travaux de Dov Samet, 2003.[1]

On s'intéresse pour tout couple (X,Y) de variables aléatoires réelles à la propriété $\mathcal{P}$ suivant

$$(\mathcal{P}) \quad \forall \mathcal{A} \text{ mesurable}, \begin{cases} \mathbb{P}(X < Y \text{ et } X \in \mathcal{A}) = \mathbb{P}(X > Y \text{ et } X \in \mathcal{A}) = \frac{1}{2}\mathbb{P}(X \in \mathcal{A}) \\ \mathbb{P}(Y < X \text{ et } Y \in \mathcal{A}) = \mathbb{P}(Y > X \text{ et } Y \in \mathcal{A}) = \frac{1}{2}\mathbb{P}(Y \in \mathcal{A}) \end{cases}$$

☞ Q1 Comment s'interprête la propriété $\mathcal{P}$?

■ Partie 1 **"*Devinez quel est le plus grand !*"**

On s'intéresse au problème suivant, présenté originellement par Blackwell (1951) :

> Deux nombres réels sont tirés aléatoirement, et chacun est placé dans une enveloppe. L'une d'elle est (indépendamment) tirée au hasard et vous est présentée. Vous devez deviner, au vu du nombre qu'elle contient, s'il s'agit du plus grand ou le plus petit des deux. Sauriez-vous deviner juste plus d'une fois sur deux en moyenne ?

[1] Résultats présentés à la 14ième conférence internationale de Théorie des jeux, Stony Brook 2003.
Voir `Http ://www.sunysb.edu/gametheory/Conf03/twopuzzles.pdf` pour un exposé plus complet.

Pour simplifier la présentation, supposons que vous gagnez $+1$ si vous avez deviné juste et -1 sinon ; ce problème revient à savoir si vous pouvez vous garantir un gain espéré strictement positif.

☞ Q1 Soient X et Y les deux nombres tirés aléatoirement, et supposons que (X, Y) vérifie $\mathcal{P}$. Quel est votre meilleur gain espéré ? Interprêter.

☞ Q2 Cherchons néanmoins à construire une stratégie gagnante.

Pour ce faire, donnons-nous $s \in \mathbb{R}$ et soit $\sigma_s : \begin{pmatrix} \mathbb{R} & \to & \{-1,1\} \\ x & \mapsto & \begin{cases} 1 & \text{si } x > s \\ -1 & \text{sinon} \end{cases} \end{pmatrix}$ la *stratégie de seuil* associée à s.

(a) Soit $(x \neq y)$ une réalisation du tirage ; le nombre qui vous est présenté est soit x, soit y. Quel est votre gain espéré en jouant selon σ_s sachant que $x < s$ et $y < s$?

(b) Quel est votre gain espéré sachant que $x < s < y$?

(c) En déduire que si votre gain espéré (sur tous les tirages possibles) est strictement positif.

☞ Q3 En déduire qu'aucune paire de variables aléatoires réelles (X, Y) ne peut vérifier $\mathcal{P}$.

Partie 2 ***Le jeu des enveloppes***

On s'intéresse désormais au célèbre jeu dit *des enveloppes*, introduit par Kraitchik (1953) sous la forme suivante :

> On considère deux enveloppes contenant une certaine somme d'argent, mais l'une contenant le double de l'autre ; chacune a autant de chances que l'autre de contenir la plus grosse somme. L'une d'elles vous est donnée au hasard, et vous devez choisir entre la prendre, ou prendre l'autre.
> A supposer qu'elle contienne x, l'autre a une chance sur deux de contenir $2x$ et une chance sur deux de contenir $\frac{x}{2}$; donc le contenu espéré de l'autre enveloppe est $\frac{5}{4}x > x$, de sorte que vous avez intérêt à changer d'enveloppe. Mais le même raisonnement s'applique aussi à la nouvelle enveloppe, de sorte que vous avez à nouveau intérêt à changer d'enveloppe, et ainsi de suite ...
> Que faire dans cette situation ?

Résoudre ce paradoxe a été le prétexte à une littérature foisonnante ; voici une solution.

Notons X la somme (variable aléatoire réelle) contenue dans l'enveloppe qui vous est présentée, et Y celle contenue dans l'autre.

☞ Q1 Quel sens donner à l'expression $\mathcal{H}_x$: "l'autre a une chance sur deux de contenir $2x$ et une chance sur deux de contenir $\frac{x}{2}$" ?

☞ Q2 En déduire que si cette hypothèse est vraie, alors la distribution de (X, Y) vérifie $\mathcal{P}$.

☞ Q3 Conclure.

Enoncé de l'exercice 3

Soient $X_1, \ldots, X_n$ n variables aléatoires i.i.d dans $\mathbb{R}^k$, de loi paramétrique $\mathcal{P}_\theta$ de densité

$$\frac{d\mathcal{P}_\theta(x)}{d\lambda} = f(x - \theta)$$

$\theta \in \mathbb{R}^k$ est appelé *paramètre de position*.

On cherche à étudier les estimateurs T du paramètre θ. On se restreint pour cela à la classe d estimateurs *équivariants*, c'est-à-dire des estimateurs T vérifiant :

$$\forall c \in \mathbb{R}^k, T(X_1 + c, \ldots, X_n + c) = T(X_1, \ldots, X_n) + c$$

Cette restriction est naturelle : en effet, pour tout $c \in \mathbb{R}^k$, $X_i + c \rightsquigarrow \mathcal{P}_{\theta+c}$, et on ne peut admet qu'un simple changement d'échelle puisse mener à une estimation différente.

☞ Q1 Montrer que :

$$\text{T équivariant} \Leftrightarrow \begin{array}{l} \exists T_1 : (\mathbb{R}^{k-1} \to \mathbb{R}^k) \ / \ \forall (x_1, \ldots, x_n) \in \mathbb{R}^k \\ T(x_1, \ldots, x_n) = T_1(x_2 - x_1, \ldots, x_n - x_1) + x_1 \end{array}$$

☞ Q2 Soit $W(x, y) = w(x - y)$ une fonction de perte. Montrer que, si un estimateur T est équivaria le risque associé, pour la perte W, ne dépend pas de θ.

Un estimateur équivariant T qui minimise le risque $R_w(T, 0)$ (c'est-à-dire au point $\theta = 0$) alors un estimateur optimal (parmi les équivariants) pour la fonction de perte W.

☞ Q3 Soit la fonction

$$\psi(x) = \int_{\mathbb{R}^k} w(x - u) \prod_{i=1}^{n} f(X_i - u) du$$

et l'estimateur T^* défini par

$$\psi(T^*) = \inf_{x \in \mathbb{R}^k} \psi(x)$$

T^* est appelé *l'estimateur de Pitman* (on suppose ici que w est telle que l'équation $\psi(x^*)$ $\inf_{x \in \mathbb{R}^k} \psi(x)$ admette une solution et une seule).

(a) Montrer que T^* est équivariant.

(b) Montrer que, pour toute fonction φ telle que $\mathbb{E}\left(\|\varphi(X_1, \ldots, X_n)\|\right) < +\infty$, on a :

$$E_\theta[\varphi(X_1, ..., X_n) | X_2 - X_1, ..., X_n - X_1] = \int_{\mathbb{R}^k} \varphi(X_1 + \theta - u, ..., X_n + \theta - u) \frac{\prod_{i=1}^n f(X_i - u)}{\int_{\mathbb{R}^k} \prod_{i=1}^n f(X_i - v) dv} du$$

(c) Montrer finalement que T^* est optimal pour w.

☞ Q4 Donner l'expression de T^* lorsque $w(x - y) = \|x - y\|^2$.

☞ Q5 Dans ce cas, que vaut T^* lorsque $X_i \rightsquigarrow \mathcal{N}(\theta, 1)$?

Et lorsque $X_i \rightsquigarrow \mathcal{U}_{[-\frac{1}{2}+\theta, \frac{1}{2}+\theta]}$?

Enoncé de l'exercice 4

On considère une loi multinomiale à K modalités ($K \geq 3$) de probabilités $p_1, p_2, \ldots, p_K$. On dispose de n observations indépendantes issues de cette loi. On notera N_k le nombre d'observations de la modalité $k \in [\![1, K]\!]$.

On veut tester l'hypothèse : $\underline{H_0 :}$ "$p_1 + p_2 = \frac{1}{2}$"

☞ Q1 Donner les estimateurs du maximum de vraisemblance contraints $(\widehat{p_k^0})_k$ et non contraints $(\widehat{p_k})_k$ de $p_1, \ldots, p_K$.

☞ Q2 Calculer la statistique ξ^W du test de Wald de l'hypothèse H_0.

☞ Q3 (a) En constatant que $\widehat{p_3^0}$ est asymptotiquement efficace sous H_0, montrer que

$$\mathbb{C}ov_{as}\left(\widehat{p_3^0}, \widehat{p_3} - \widehat{p_3^0}\right) = 0$$

et en déduire que

$$\mathbb{V}_{as}\left(\widehat{p_3} - \widehat{p_3^0}\right) = \mathbb{V}_{as}\left(\widehat{p_3}\right) - \mathbb{V}_{as}\left(\widehat{p_3^0}\right)$$

(b) Calculer la loi limite de $\left(\widehat{p_3^0} - p_3\right)$, et en déduire $\mathbb{V}_{as}\left(\widehat{p_3^0}\right)$.

(c) Donner un estimateur $\widehat{\mathbb{V}_{as}\left(\widehat{p_3} - \widehat{p_3^0}\right)}$ convergent sous H_0 de $\mathbb{V}_{as}\left(\widehat{p_3} - \widehat{p_3^0}\right)$.
En déduire l'expression de la statistique du test d'Hausman de H_0

$$\xi^H = n \frac{\left(\widehat{p_3} - \widehat{p_3^0}\right)^2}{\widehat{\mathbb{V}_{as}\left(\widehat{p_3} - \widehat{p_3^0}\right)}}$$

(d) Vérifier que le test d'Hausman est convergent.

(e) Montrer que les statistiques ξ^W et ξ^H sont asymptotiquement équivalentes sous H_0.

Enoncé de l'exercice 5

La mise en œuvre de nombreuses techniques d'analyse, et notamment la quasi-totalité des te
niques d'estimation et de test de modèles statistiques, nécessitent d'*engendrer* des réalisati
x d'une variable aléatoire réelle X de loi $\mathcal{L}$ donnée.

L'objet de cet exercice est de présenter quelques techniques permettant à un ordinateur, mach
essentiellement déterministe, d'engendrer de telles variables. Il s'agit donc de construire p
toute loi $\mathcal{L}$ une fonction récursive $f^{\mathcal{L}} : \mathbb{N} \to I$ qui énumère des réalisations d'un variable $X \rightsquigarrow$
i.e. telle que la distribution empirique des $\left(f^{\mathcal{L}}(n)\right)_{n\in\mathbb{N}}$ converge vers la vraie distribution F

☞ Q1 Cherchons tout d'abord à engendrer une variable aléatoire **entière** uniforme.
Soit $N \in \mathbb{N}^*$ et considérons $\mathcal{L} = \mathcal{U}_{[\![0,N-1]\!]}$.

(a) Définissons pour $\phi : [\![0, N-1]\!] \to [\![0, N-1]\!]$ et $s \in [\![0, N-1]\!]$

$$f_{\phi,s} : \begin{pmatrix} [\![0, N-1]\!] & \to & [\![0, N-1]\!] \\ t & \mapsto & \phi^t(s) \end{pmatrix}$$

Montrer que $f_{\phi,s}$ est périodique à partir d'un certain rang ; on note $T_{\phi,s}$ sa période.
En quel sens peut-on dire que $f_{\phi,s}$ "génère" une variable uniformément distribuée $[\![0, N-1]\!]$?

(b) En quel sens peut-on dire que $f_{\phi,s}$ est d'autant meilleure que $T_{\phi,s}$ est grande ?

(c) Soient $a \in [\![2, N-1]\!]$, $c \in [\![0, N-1]\!]$ et $p \in [\![2, N-1]\!]$ et définissons $\psi_{a,c,}$
$\begin{pmatrix} \mathbb{N} & \to & [\![0, N-1]\!] \\ n & \mapsto & (an+c) \mod p \end{pmatrix}$.
A quelles conditions $f_{\psi_{a,c,p},s}$ est-elle un bon générateur uniforme sur $[\![0, N-1]\!]$?

☞ Q2 Cherchons à engendrer une variable aléatoire réelle uniforme sur $[0, 1]$.

(a) Soit $(x_n)_{n\in\mathbb{N}} \in [0, 1]^{\mathbb{N}}$, et définissons pour $f : [0, 1] \to \mathbb{R}$ continue et $n \in \mathbb{N}^*$

$$S_n^x(f) = \frac{1}{n} \sum_{k=1}^{n} f(x_k)$$

On dit de x qu'elle vérifie le *critère de Weyl* si pour toute f continue

$$S_n^x(f) \xrightarrow[n\to+\infty]{} \int_0^1 f$$

Comment s'interprète ce critère ?

(b) Soit $r \in \mathbb{R}$, et définissons $x = (frac(rk))_{k\in\mathbb{N}}$ où $frac(u) = u - E(u) \in [0, 1[$ désigne
partie fractionnaire de $u \in \mathbb{R}$.
Montrer que x vérifie le critère de Weyl *ssi* r est irrationnel.

☞ Q3 On cherche désormais à engendrer des réalisations d'une variable aléatoire de loi quelconqu

(a) Soit $\mathcal{L}$ la loi de densité $f_{\mathcal{L}} = \frac{1}{3}\delta_1 + \frac{1}{4}\delta_2 + \frac{5}{12}\delta_3$.
Tracer la fonction de répartition $F_{\mathcal{L}}$ de $\mathcal{L}$.
En déduire son inverse généralisée

$$F^{-1} : \begin{pmatrix} [0,1] & \to & \mathbb{R} \\ s & \mapsto & \inf\{x \in \mathbb{R} \,/\, F_{\mathcal{L}}(x) \geq s\} \end{pmatrix}$$

Soit $U \rightsquigarrow \mathcal{U}_{[0,1]}$ et définissons $X = F_{\mathcal{L}}^{-1}(U)$; quelle est la loi de X ?

(b) Soit alors $F_{\mathcal{L}}$ la fonction de répartition supposée continue et strictement croissante d'une loi $\mathcal{L}$ quelconque.
On définit de même $X = F_{\mathcal{L}}^{-1}(U)$ pour $U \rightsquigarrow \mathcal{U}_{[0,1]}$; quelle est la loi de X ?

(c) Proposer un algorithme qui génère des réalisations successives d'une variable X de loi exponentielle $\mathcal{E}(\lambda)$ de densité $f_{\mathcal{E}_\lambda}(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,+\infty[}(x)$ pour $\lambda > 0$.

(d) Soit $(X, Y) \rightsquigarrow \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 I_2\right)$.
Soient $R^2 = X^2 + Y^2$ et $\theta = \arctan \frac{Y}{X}$.
Montrer que R^2 et θ sont indépendantes, et donner leurs lois.
En déduire un algorithme de génération d'une variable gaussienne.

(e) Même question si X suit une loi de Weibull $\mathcal{W}(\alpha, \beta)$ de densité $f_{\mathcal{W}(\alpha,\beta)}(x) = \alpha\beta x^{\beta-1} e^{-\alpha x^\beta} \mathbb{1}_{x \geq 0}$.

(f) Même question si X suit une loi de Poisson $\mathcal{P}(\lambda)$ de densité $f_{\mathcal{P}(\lambda)}(k) = \frac{\lambda^k}{k!} e^{-\lambda}$.
On pourra comparer $x_n = u_1 \times \cdots \times u_n$, où chaque u_n est une réalisation de $U \rightsquigarrow U_{[0,1]}$, à $e^{-\lambda}$.