Université des Sciences et Technologies de Lille
U.F.R. de Mathématiques Pures et Appliquées

IS Math314 **Année 2006**

Corrigé de l'examen du 15 juin 2006.
Première partie : exercices 1–3

Ex 1. *Vitesse moyenne (3 points)*

On a enregistré à l'aide d'un radar les vitesses $X_1(\omega) = x_1, \dots, X_{400}(\omega) = x_{400}$ de 400 automobiles dans des conditions permettant de supposer que $X_1, \dots, X_{400}$ sont indépendantes et de même loi. On a obtenu les statistiques suivantes :

$$\sum_{i=1}^{400} x_i = 35\ 200 \text{ km/h}, \quad \sum_{i=1}^{400} x_i^2 = 3\ 107\ 600\ (\text{km/h})^2.$$

On se propose d'estimer $\mathbf{E}X_1$ par un intervalle de confiance au niveau 98%.

Ici la loi des v.a. X_i est inconnue. Il est clair pour des raisons physiques que les X_i sont des v.a. bornées, donc *a fortiori* de carré intégrable. Un estimateur naturel de $\mathbf{E}X_1$ est la moyenne empirique $\overline{X}$ en raison de la loi forte des grands nombres. On va chercher pour $\mathbf{E}X_1$ un intervalle de confiance centré sur $\overline{X}(\omega) = 35\ 200/400 = 88$ km/h.

Pour construire cet intervalle de confiance, on ne peut pas utiliser ici le théorème limite central classique car l'écart-type σ des x_i est inconnu. On va utiliser le TLC avec autonormalisation où l'on remplace σ par S, la racine carrée de la variance empirique

$$S^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbf{E}X_i)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X})^2.$$

Les X_i étant de carré intégrable, le TLC avec autonormalisation nous dit que

$$\sqrt{n}\frac{\overline{X} - \mathbf{E}X_1}{S} \xrightarrow[n\to+\infty]{\text{loi}} Z, \quad Z \sim \mathfrak{N}(0,1). \tag{1}$$

Cette convergence légitime pour n grand, l'approximation

$$P\left(\left|\sqrt{n}\frac{\overline{X} - \mathbf{E}X_1}{S}\right| \leq t\right) \simeq P(|Z| \leq t)) = 2\Phi(t) - 1,$$

où Φ est la f.d.r. de la loi $\mathfrak{N}(0,1)$. La résolution d'inégalité :

$$\left|\sqrt{n}\frac{\overline{X} - \mathbf{E}X_1}{S}\right| \leq t \quad \Leftrightarrow \quad \overline{X} - \frac{tS}{\sqrt{n}} \leq \mathbf{E}X_1 \leq \overline{X} + \frac{tS}{\sqrt{n}},$$

nous donne alors un intervalle de confiance pour $\mathbf{E}X_1$ au niveau $2\Phi(t)-1$. Il s'agit bien d'un intervalle de confiance puisque les bornes $\overline{X} \pm tSn^{-1/2}$ sont calculables à partir des observations sans connaissance de la loi des X_i.

Pour terminer les calculs, on détermine t en résolvant $2\Phi(t)-1=0,98$, ce qui équivaut à $\Phi(t)=0,99$, d'où $t=2,33$ (approximativement). On calcule ensuite $\overline{X}(\omega)$ et $S(\omega)$:

$$\overline{X}(\omega)=\overline{x}=\frac{1}{400}\sum_{i=1}^{400}x_i=\frac{35\ 200}{400}=88 \text{ km/h}.$$

$$S^2(\omega)=s^2=\frac{1}{400}\sum_{i=1}^{400}x_i^2-(\overline{x})^2=\frac{3\ 107\ 600}{400}-88^2=7\ 769-7\ 744=25 \text{ km/h}^2,$$

d'où $S(\omega)=s=5$ km/h. Un intervalle de confiance I au niveau 98% pour $\mathbf{E}X_1$ en km/h est donc

$$I=\left[88-\frac{2,33\times 5}{20};88+\frac{2,33\times 5}{20}\right]=[87{,}41;88{,}59].$$

N.B. On a arrondi la borne gauche par défaut et la borne droite par excès.

Ex 2. *Estimation par maximum de vraisemblance (3 points)*

On note $X_1,\dots,X_n$ un échantillon associé au modèle statistique $(\Omega,\mathcal{F},(P_\theta)_{\theta\in]0,+\infty[})$, les X_i étant des v.a. à valeurs dans $]0,1[$ ayant pour densité sous P_θ la fonction :

$$t\mapsto f(t,\theta):=\theta t^{\theta-1}\mathbf{1}_{]0,1[}(t).$$

On se propose d'estimer θ par maximum de vraisemblance.

1) La fonction de vraisemblance s'écrit ici

$$L(x_1,\dots,x_n,\theta)=\prod_{i=1}^n f(x_i,\theta)=\theta^n(x_1\cdots x_n)^{\theta-1}\mathbf{1}_{]0,1[}(x_1)\cdots\mathbf{1}_{]0,1[}(x_n).$$

Comme les X_i sont à valeurs dans $]0,1[$, toutes les observations x_i sont forcément dans cet intervalle, donc toutes les indicatrices ci-dessus valent 1, d'où

$$L(x_1,\dots,x_n,\theta)=\theta^n(x_1\cdots x_n)^{\theta-1}.$$

Il est commode ici de considérer la log-vraisemblance (de même sens de variation que L) :

$$\ln L(x_1,\dots,x_n,\theta)=n\ln\theta+(\theta-1)\ln(x_1\cdots x_n).$$

La dérivée par rapport à θ est

$$\frac{\partial}{\partial\theta}\ln L(x_1,\dots,x_n,\theta)=\frac{n}{\theta}+\ln(x_1\cdots x_n).$$

En notant que le réel $x_1\cdots x_n$ appartient à $]0,1[$ et a donc un logarithme *négatif*, on voit que cette dérivée est strictement positive pour $0<\theta<-n/\ln(x_1\cdots x_n)$ et strictement négative pour $\theta > -n/\ln(x_1\cdots x_n)$. La log-vraisemblance et la vraisemblance ont donc un unique maximum absolu atteint au point

$$\widehat{\theta} = \frac{-n}{\ln(x_1\cdots x_n)}.$$

On en déduit que l'estimateur T_n de θ par maximum de vraisemblance est la statistique

$$T_n := \frac{-n}{\ln(X_1\cdots X_n)} = \frac{-n}{\ln X_1 + \cdots + \ln X_n}. \tag{2}$$

2) Pour justifier l'existence de $\mathbf{E}_\theta(\ln X_1)$ il nous faut montrer que $\mathbf{E}_\theta|\ln X_1| < +\infty$. Or X_1 ayant sous chaque P_θ une loi à densité $f(t,\theta)$, cette espérance s'écrit

$$\mathbf{E}_\theta|\ln X_1| = \int_{-\infty}^{+\infty} |\ln t| f(t,\theta)\,\mathrm{d}t = \int_{-\infty}^{+\infty} |\ln t|\theta t^{\theta-1}\mathbf{1}_{]0,1[}(t) = \int_0^1 |\ln t|\theta t^{\theta-1}\,\mathrm{d}t.$$

Pour $\theta > 1$, la fonction $g :]0,1] \to \mathbb{R}$, $t \mapsto |\ln t|\theta t^{\theta-1}$ se prolonge par continuité en 0 en posant $g(0) := 0$ et l'intégrale ci-dessus se réduit à une intégrale de Riemann ordinaire d'une fonction continue sur un intervalle fermé borné. Elle est donc finie. Par contre si $0 < \theta \leq 1$, g tend vers $+\infty$ à droite en 0 et on a une vraie intégrale généralisée. Pour montrer sa convergence, le plus simple est de la calculer en intégrant par parties, ce calcul valable pour tout $\theta > 0$, pouvant être recyclé dans le calcul de $\mathbf{E}_\theta(\ln X_1)$. Considérons donc pour tout $\varepsilon \in]0,1[$, l'intégrale

$$\int_\varepsilon^1 |\ln t|\theta t^{\theta-1}\,\mathrm{d}t = \int_\varepsilon^1 (-\ln t)\theta t^{\theta-1}\,\mathrm{d}t,$$

en notant que $\ln t$ est négatif pour tout $t \in]0,1[$. L'intégration par parties permet de faire disparaître le logarithme par dérivation. On pose donc ici

$$\begin{aligned} u(t) &= -\ln t & v'(t) &= \theta t^{\theta-1} \\ u'(t) &= \frac{-1}{t} & v(t) &= t^\theta. \end{aligned}$$

Il vient alors

$$\int_\varepsilon^1 (-\ln t)\theta t^{\theta-1}\,\mathrm{d}t = \big[(-\ln t)t^\theta\big]_\varepsilon^1 - \int_\varepsilon^1 \frac{-1}{t}t^\theta\,\mathrm{d}t = \varepsilon^\theta \ln\varepsilon + \Big(\frac{1^\theta}{\theta} - \frac{\varepsilon^\theta}{\theta}\Big).$$

En faisant tendre ε vers 0 et en rappelant que θ est strictement positif, on voit que ceci converge vers $1/\theta$. On a donc montré la finitude de $\mathbf{E}_\theta|\ln X_1|$. L'existence de $\mathbf{E}_\theta(\ln X_1)$ est ainsi justifiée. Pour calculer cette dernière espérance, on remarque simplement que, puisque $X_1(\omega) \in]0,1[$ pour tout $\omega \in \Omega$, $\ln X_1 = -|\ln X_1|$ et on déduit immédiatement du calcul ci-dessus que

$$\mathbf{E}_\theta(\ln X_1) = \frac{-1}{\theta}.$$

Il n'est pas surprenant de trouver une valeur négative puisque $\ln X_1$ ne prend que des valeurs négatives.

3) En revenant à (2), on voit que T_n peut aussi s'écrire

$$T_n = \frac{-1}{\frac{1}{n}(\ln X_1 + \cdots + \ln X_n)},$$

ce qui suggère d'appliquer la loi forte des grands nombres à la suite des v.a. $Y_i := \ln X_i$. La suite $(Y_i)_{i\geq 1}$ hérite de l'indépendance et de l'équidistribution de $(X_i)_{i\geq 1}$. D'après la question 2, $\mathbf{E}_\theta |Y_1| < +\infty$. Les hypothèses de la loi forte des grands nombres sont donc bien vérifiées par $(Y_i)_{i\geq 1}$, d'où pour tout $\theta \in]0, +\infty[$,

$$\frac{1}{n}(Y_1 + \cdots + Y_n) = \frac{1}{n}(\ln X_1 + \cdots + \ln X_n) \xrightarrow[n\to+\infty]{P_\theta-\text{p.s.}} \mathbf{E}_\theta(\ln X_1) = \frac{-1}{\theta}.$$

La fonction $h : t \mapsto -1/t$ étant continue sur $\mathbb{R}^*$, donc en particulier au point $-1/\theta$, on en déduit que $T_n = h\big(n^{-1}(Y_1 + \cdots + Y_n)\big)$ converge $P_\theta -$ p.s. vers $h(-1/\theta) = \theta$. Ainsi T_n est un estimateur fortement consistant de θ.

Ex 3. *Convergence en loi (4 points)*

Soit $(X_i)_{i\geq 1}$ une suite de variables aléatoires définies sur le même espace probabilisé $(\Omega, \mathcal{F}, P)$, indépendantes et de même loi telle que $\mathbf{E}X_1^2 = 1$ et $\mathbf{E}X_1 = 0$. On suppose de plus que pour tout i, $X_i(\omega)$ n'est jamais nul. On définit alors sur $(\Omega, \mathcal{F}, P)$ les suites de variables aléatoires $(Y_n)_{n\geq 1}$ et $(Z_n)_{n\geq 1}$ par

$$Y_n := \sqrt{n}\frac{X_1 + \cdots + X_n}{X_1^2 + \cdots + X_n^2}, \quad Z_n := \frac{X_1 + \cdots + X_n}{(X_1^2 + \cdots + X_n^2)^{1/2}}.$$

Montrons que $(Y_n)_{n\geq 1}$ et $(Z_n)_{n\geq 1}$ convergent en loi quand n tend vers l'infini vers une v.a. gaussienne de loi $\mathfrak{N}(0, 1)$.

La méthode pour prouver ces convergences en loi de $(Y_n)_{n\geq 1}$ et $(Z_n)_{n\geq 1}$ est la même que celle vue en cours pour établir le théorème limite central avec autonormalisation. On met en facteur S_n^*, la somme centrée réduite du TLC classique et on montre que l'autre facteur converge en probabilité vers une constante. Ceci permet de conclure *via* le lemme de Slutsky.

Commençons par calculer $S_n^* = (\operatorname{Var} S_n)^{-1/2}(S_n - \mathbf{E}S_n)$, où $S_n := X_1 + \cdots + X_n$. Comme les X_i ont une espérance nulle, $\mathbf{E}S_n = \sum_{i=1}^n \mathbf{E}X_i = 0$. De plus par indépendance des v.a. de carré intégrable X_i,

$$\operatorname{Var} S_n = \sum_{i=1}^n \operatorname{Var} X_i = n \operatorname{Var} X_1 = n\mathbf{E}(X_1^2) = n,$$

où la deuxième égalité repose sur l'équidistribution des X_i et la troisième égalité utilise la nullité de $\mathbf{E}X_1$. Par conséquent,

$$S_n^* = \frac{X_1 + \cdots + X_n}{\sqrt{n}}. \tag{3}$$

Les X_i étant i.i.d. et de carré intégrable, le théorème limite central nous donne la convergence en loi :

$$\frac{X_1 + \cdots + X_n}{\sqrt{n}} \xrightarrow[n\to+\infty]{\text{loi}} Z, \quad Z \sim \mathfrak{N}(0,1). \tag{4}$$

Au vu de (3), la factorisation de Y_n par S_n^* s'écrit :

$$Y_n = S_n^* \frac{n}{X_1^2 + \cdots + X_n^2} = S_n^* \Big(\frac{X_1^2 + \cdots + X_n^2}{n} \Big)^{-1}. \tag{5}$$

Les v.a. X_i^2 sont i.i.d. et intégrables puisque les X_i sont i.i.d. de carré intégrable. La loi forte des grands nombres appliquée aux X_i^2 nous donne donc

$$\frac{X_1^2 + \cdots + X_n^2}{n} \xrightarrow[n\to+\infty]{\text{p.s.}} \mathbf{E}X_1^2 = 1.$$

L'application $g : x \mapsto 1/x$ étant continue sur $\mathbb{R}^*$, donc en particulier au point 1, on en déduit que

$$V_n := \Big(\frac{X_1^2 + \cdots + X_n^2}{n} \Big)^{-1} = g\Big(\frac{X_1^2 + \cdots + X_n^2}{n} \Big) \xrightarrow[n\to+\infty]{\text{p.s.}} 1.$$

La convergence p.s. implique la convergence en probabilité, d'où

$$\Big(\frac{X_1^2 + \cdots + X_n^2}{n} \Big)^{-1} \xrightarrow[n\to+\infty]{\text{Pr}} 1. \tag{6}$$

La factorisation (5) nous a donc permis de représenter Y_n comme le produit de $S_n^* V_n$ où S_n^* converge en loi vers Z et V_n converge en probabilité vers 1. Par le lemme de Slutsky, le *vecteur aléatoire* (S_n^*, V_n) converge en loi vers $(Z, 1)$. En prenant son image par le fonction continue $h : \mathbb{R}^2 \to \mathbb{R}$, $(x, y) \mapsto xy$, on en déduit la convergence en loi de $Y_n = h(S_n^*, V_n)$ vers $g(Z, 1) = Z$.

Pour établir la convergence en loi de Z_n vers Z, on remplace (5) par la factorisation

$$Z_n = S_n^* \Big(\frac{n}{X_1^2 + \cdots + X_n^2} \Big)^{1/2} = S_n^* \Big(\frac{X_1^2 + \cdots + X_n^2}{n} \Big)^{-1/2}. \tag{7}$$

Ensuite on utilise le même raisonnement que ci-dessus en prenant pour g l'application $x \mapsto x^{-1/2}$.