Université Paris I, Panthéon - Sorbonne

Première Année Master M.A.E.F. 2016 – 2017

Statistiques II

Contrôle continu n°1, mars 2017

Examen de 1h30. Tout document ou calculatrice est interdit.

1. On considère une suite $(X_i)_{i\in\mathbb{N}}$ de variables aléatoires indépendantes et identiquement distribuées suivant une loi discrète à valeurs dans $\{-1,0,1\}$ telle que $\Pr(X_0=0)=p_0>0$ and $p_1=\Pr(X_0=1)>\Pr(X_0=-1)>0$. On définit également:
$$Y_n=\sum_{i=1}^{n} X_i, \quad \text{pour tout } n\in\mathbb{N}^*.$$
   (a) Déterminer $\mathbb{E}(X_0)$ en fonction de p_0 et p_1 et montrer que $\mathbb{E}(X_0)=m>0$ **(1pt)**. Déterminer en fonction de p_0 et p_1, $\sigma^2=\mathrm{var}(X_0)$ **(1pt)**.
   (b) Montrer que Y_n est une variable aléatoire discrète et préciser l'ensemble de ses valeurs possibles **(0.5pts)**.
   (c) Déterminer $\mathbb{E}(Y_n)$ puis $\mathrm{var}(Y_n)$ pour tout $n\in\mathbf{N}^*$ **(1pt)**. (Y_n) est-il un processus stationnaire **(0.5pts)**?
   (d) Montrer que $Y_n \xrightarrow[n\to+\infty]{p.s.} +\infty$ **(1.5pts)**.
   (e) Montrer que $(Y_n)_{n\in\mathbf{N}}$ peut se mettre sous la forme
$$Y_n=m\,n+\sqrt{n}\,\varepsilon_n \quad \text{pour tout } n\in\mathbb{N}^*,$$
   où $(\varepsilon_n)_{n\in\mathbf{N}^*}$ est une suite de variables aléatoires centrées de même variance (à préciser) **(1pt)**. Quelle est la tendance de $(Y_n)_{n\in\mathbf{N}}$ **(0.5pts)**? Montrer que la loi de ε_n converge vers une loi à préciser lorsque $n\to\infty$ **(1pt)**. En déduire que les variables $(\varepsilon_n)_{n\in\mathbf{N}^*}$ ne forment pas un processus stationnaire **(1.5pts)**.
   (f) Déterminer $\mathrm{cov}(Y_i,Y_j)$ puis $\mathrm{cov}(\varepsilon_i,\varepsilon_j)$ pour $i,j\in\mathbf{N}^*$ **(2pts)**. Que devient cette quantité lorsque $|i-j|$ est "grand" **(1pt)**?
   (g) On suppose maintenant que $(Y_1,\cdots,Y_N)$ est un échantillon observé de $(Y_n)_{n\in\mathbf{N}^*}$. On suppose que tous les paramètres m, p_0, p_1 sont inconnus. Montrer que l'estimateur $\widehat{m}_N$ de m par régression linéaire par moindres carrés est défini par
$$\widehat{m}_N=\frac{6}{N(N+1)(2N+1)}\sum_{i=1}^{N} i\,Y_i \qquad \textbf{(3pts)}.$$
   Calculer $\mathbb{E}(\widehat{m}_N)$ et $\mathrm{var}(\widehat{m}_N)$ **(2pts)**. L'estimateur $\widehat{m}_N$ est-il un estimateur convergent de m **(1pt)**?
   (h) Montrer que les estimateurs $\widehat{p}_0=\frac{1}{N}\sum_{i=0}^{N-1}\mathbb{I}_{Y_{i+1}-Y_i=0}$ et $\widehat{p}_1=\frac{1}{N}\sum_{i=0}^{N-1}\mathbb{I}_{Y_{i+1}-Y_i=1}$ (on pose par convention $Y_0=0$) sont les estimateurs de p_0 et p_1 par maximum de vraisemblance (on pourra écrire la vraisemblance à l'aide d'un produit de probabilités conditionnelles) **(3pts)**. En justifiant, donner les théorèmes de la limite centrale vérifiés par $\widehat{p}_0$ et $\widehat{p}_1$ **(2pts)**.
   (i) En utilisant l'expression de m en fonction de p_0 et p_1, en déduire un autre estimateur $\tilde{m}_N$ non biaisé de m **(1pt)**. Converge-t-il plus vite que $\widehat{m}_N$ vers m (au sens du risque quadratique) **(2pts)**?

Proof. (a) On a $\mathbb{E}X_0=2p_1+p_0-1>0$ (car $p_{-1}=1-p_0-p_1<p_1$) et $\mathrm{var}\,X_0=1-p_0+(2p_1+p_0-1)^2$.
(b) Pour tout $n\in\mathbf{N}^*$, Y_n est une somme de variables discrètes donc Y_n est aussi une variable discrète est ses valeurs appartiennent à $\{-n,-(n-1),\cdots,n-1,n\}$.
(c) On a $\mathbb{E}Y_n=n\,m$ et $\mathrm{var}\,Y_n=n\,\sigma^2$ comme somme de variables i.i.d. L'espérance dépendant de n ce ne peut être un processus stationnaire.
(d) D'après la Loi des Grands Nombres forte comme les X_i sont des v.a.i.i.d. telles que $\mathbb{E}|X_i|<\infty$ alors $\frac{1}{n}Y_n \xrightarrow[n\to+\infty]{p.s.} m$ avec $m>0$. Ainsi $\mathbb{P}(\lim_{n\to\infty}Y_n>\frac{1}{2}m\,n)=1$; ceci entraîne que $Y_n \xrightarrow[n\to+\infty]{p.s.} +\infty$.

(e) Pour tout $n \in \mathbb{N}^*$, $Y_n = \mathbb{E}Y_n + (Y_n - \mathbb{E}Y_n) = m\,n + \sqrt{n}\frac{(Y_n - \mathbb{E}Y_n)}{\sqrt{n}}$. Soit $\varepsilon_n = \frac{(Y_n - \mathbb{E}Y_n)}{\sqrt{n}}$. Alors $\mathbb{E}\varepsilon_n = 0$ et $\operatorname{var}\varepsilon_n = \frac{1}{n}\operatorname{var} Y_n = \sigma^2$: les (ε_n) sont bien centrées et de même variance. D'après le Théorème de la Limite Centrale (que l'on peut utiliser car les X_i sont des v.a.i.i.d. de variance σ^2 finie), $\varepsilon_n = \sqrt{n}\big(\frac{1}{n}Y_n - m\big) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,\sigma^2)$. En conséquence, comme à n fini les ε_n ont une loi discrète, leur loi dépend de n (la loi normale n'est pas discrète): elles ne forment une suite stationnaire.
(f) On a $\operatorname{cov}(Y_i, Y_j) = \sum_{k=1}^{i}\sum_{\ell=1}^{j}\operatorname{cov}(X_k, X_\ell)$. Or pour $k \neq \ell$ alors $\operatorname{cov}(X_k, X_\ell) = 0$. Donc $\operatorname{cov}(Y_i, Y_j) = \sum_{k=1}^{i}\sigma^2 = \min(i,j)\sigma^2$. En raison de la définition de ε_n, on en déduit que $\operatorname{cov}(\varepsilon_i, \varepsilon_j) = \frac{\min(i,j)}{\sqrt{ij}}\sigma^2$. Enfin, si $|j-i| \to \infty$ avec par exemple $j/i \to \infty$ alors $\operatorname{cov}(\varepsilon_i, \varepsilon_j) \to 0$.
(g) On pose $Z = {}^t(1, \cdots, N)$ car $Y = Z\,(m) + U$ avec $\mathbb{E}U = 0$, d'où ${}^t Z\,Z = \sum_{i=1}^{N} i^2 = \frac{1}{6}N(N+1)(2N+1)$. En conséquence $\widehat{m}_N = ({}^t Z\,Z)^{-1}{}^t Z\,Y = \frac{6}{N(N+1)(2N+1)}\sum_{i=1}^{N} i\,Y_i$. On sait que $\widehat{m}_N$ est sans biais soit $\mathbb{E}\widehat{m}_N = m$ car $\mathbb{E}U = 0$. Pour le calcul de la variance, la matrice de covariance de Y n'étant pas diagonale,

$$
\begin{aligned}
\operatorname{var}\widehat{m}_N &= \Big(\frac{6}{N(N+1)(2N+1)}\Big)^2 \sum_{i,j=1}^{N} i\,j\,\operatorname{cov}(Y_i, Y_j)\\
&= \sigma^2\Big(\frac{6}{N(N+1)(2N+1)}\Big)^2\Big(\sum_{i=1}^{N} i^3 + 2\sum_{1\le i<j\le N} i^2 j\Big)\\
&= \sigma^2\Big(\frac{6}{N(N+1)(2N+1)}\Big)^2\Big(\frac{1}{4}(N^4+2N^3+N^2) + \sum_{1\le i\le N-1} i^2(N(N+1) - i(i+1))\Big)\\
&= \sigma^2\Big(\frac{6}{N(N+1)(2N+1)}\Big)^2\Big(\frac{1}{4}(N^4+2N^3+N^2) + \frac{1}{6}N^2(N^2-1)(4N^2-1) - \sum_{1\le i\le N-1} i^4 + i^3\Big)\\
&= \sigma^2\Big(\frac{6}{N(N+1)(2N+1)}\Big)^2\Big(\frac{1}{4}(N^4+2N^3+N^2) + \frac{1}{6}N^2(N^2-1)(4N^2-1) - \frac{1}{30}\big(6(N-1)^5 + 15(N-1)^4 + 10N^3 - (N-1)\big)\\
&\qquad\qquad -\frac{1}{4}((N-1)^4 + 2(N-1)^3 + (N-1)^2)\Big)\\
&\underset{N\to+\infty}{\sim} \frac{21}{5}\,\sigma^2\,\frac{1}{N}.
\end{aligned}
$$

Comme $\operatorname{var}\widehat{m}_N \underset{N\to+\infty}{\longrightarrow} 0$ et $\widehat{m}_N$ sans biais on en déduit que $\widehat{m}_N$ est convergent.
(h) La vraisemblance est
$\mathbb{P}\big((Y_1,\cdots,Y_N) = (y_1,\cdots,y_N)\big) = \mathbb{P}\big(Y_N = y_N | (Y_1,\cdots,Y_{N-1}) = (y_1,\cdots,y_{N-1}))\big) \times \mathbb{P}\big(Y_{N-1} = y_{N-1} | (Y_1,\cdots,Y_{N-2}) = (y_1,\cdots,y_{N-2}))\big) \times \cdots \times \mathbb{P}(Y_1 = y_1)$. Comme on a des chaînes de Markov on peut simplifier et
$\mathbb{P}\big((Y_1,\cdots,Y_N) = (y_1,\cdots,y_N)\big) = \mathbb{P}\big(Y_N = y_N | Y_{N-1} = y_{N-1}\big) \times \mathbb{P}\big(Y_{N-1} = y_{N-1} | Y_{N-2} = y_{N-2}\big) \times \cdots \times \mathbb{P}(Y_1 = y_1)$,
soit $\mathbb{P}\big((Y_1,\cdots,Y_N) = (y_1,\cdots,y_N)\big) = p_1^{N\widehat{p}_1} p_0^{N\widehat{p}_0}(1-p_1-p_0)^{N-N\widehat{p}_1 - N\widehat{p}_0}$. On peut alors chercher le maximum en dérivant par rapport à p_0 et p_1 et en annulant les dérivées: on arrive ainsi au fait que $\widehat{p}_0$ et $\widehat{p}_1$ sont les estimateurs de p_0 et p_1 par maximum de vraisemblance. Comme les $Y_{i+1} - Y_i = X_{i+1}$ sont des v.a.i.i.d., il en est de même pour les $\mathbb{I}_{Y_{i+1}-Y_i=c}$ avec $c = 1$ ou 0, et on peut appliquer un Théorème de la Limite Centrale (la variance des $\mathbb{I}_{Y_{i+1}-Y_i=c}$ existe car ce sont des variables de Bernoulli d'espérance $\mathbb{P}(X_i = c)$). Ainsi on obtient:

$$\sqrt{N}\big(\widehat{p}_0 - p_0\big) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}\big(0, p_0(1-p_0)\big) \quad \text{et} \quad \sqrt{N}\big(\widehat{p}_1 - p_1\big) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}\big(0, p_1(1-p_1)\big).$$

(i) On peut donc considérer $\tilde{m}_N = 2\widehat{p}_1 + \widehat{p}_0 - 1$. On peut calculer $\operatorname{cov}(\mathbb{I}_{X_i=0}, \mathbb{I}_{X_i=1}) = \mathbb{E}\mathbb{I}_{X_i=0,\mathbb{I}_{X_i=1}} - p_0 p_1 = -p_0 p_1$. Donc $\operatorname{cov}(\widehat{p}_0, \widehat{p}_1) = \frac{1}{N^2}\sum_{i=1}^{N}(-p_0 p_1) = -p_0 p_1 \frac{1}{N}$. En conséquence, $\operatorname{var}\tilde{m}_N = \frac{1}{N}\big(4p_1(1-p_1) + p_0(1-p_0) - 4p_0 p_1\big) = \frac{1}{N}\big(4p_1 + p_0 - (2p_1+p_0)^2\big)$. Il est possible alors de comparer les variances et il suffit alors de comparer $4p_1 + p_0 - (2p_1+p_0)^2$ et $\frac{21}{5}\,(1 - p_0 - (2p_1 + p_0 - 1)^2)$. On montre après quelques calculs que $\operatorname{var}\widehat{m}_N \geq \operatorname{var}\tilde{m}_N$: il vaut mieux utiliser $\tilde{m}_N$ pour estimer m (ce qui est normal car c'est aussi l'estimateur du maximum de vraisemblance de m). □