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

IS Math314 **Année 2007**

Corrigé du partiel du 30 mars 2007

Ex 1. *Loi de Cauchy et L.F.G.N.*

Le but de cet exercice est d'étudier la convergence des moyennes arithmétiques d'une suite de v.a. indépendantes et de même loi de Cauchy. Toutes les variables aléatoires intervenant dans l'énoncé sont supposées définies sur le même espace probabilisé $(\Omega, \mathcal{F}, P)$.

1) La variable aléatoire réelle X suivant la loi de Cauchy de densité :

$$f : \mathbb{R} \to \mathbb{R}_+, \quad t \mapsto \frac{1}{\pi(1+t^2)},$$

on a pour tout entier $n \geq 1$:

$$P(X > n) = \int_n^{+\infty} f(t)\,\mathrm{d}t = \int_n^{+\infty} \frac{1}{\pi(1+t^2)}\,\mathrm{d}t.$$

En notant que pour tout $t \in [n, +\infty[$, $t \geq 1$, d'où $1+t^2 \leq 2t^2$, on obtient la minoration :

$$\int_n^{+\infty} \frac{1}{\pi(1+t^2)}\,\mathrm{d}t \geq \int_n^{+\infty} \frac{1}{\pi(2t^2)}\,\mathrm{d}t = \frac{1}{2\pi}\left[\frac{-1}{t}\right]_n^{+\infty} = \frac{1}{2\pi n}.$$

Nous avons ainsi vérifié que

$$\forall n \geq 1, \quad P(X > n) \geq \frac{1}{2\pi n}.$$

2) Soit $(X_k)_{k\geq 1}$ une suite de variables aléatoires indépendantes et de même loi que X. On pose

$$A := \{\omega \in \Omega;\ X_k(\omega) > k \text{ pour une infinité d'indices } k\}.$$

Notons pour tout entier $k \geq 1$, $A_k := \{X_k > n\}$. Les évènements A_k héritent de l'indépendance de la suite des variables aléatoires X_k. D'autre part, $P(A_k) = P(X_k > k) = P(X > k)$ parce que les X_k ont même loi que X. En utilisant la minoration établie à la question précédente, on obtient :

$$\sum_{k=1}^{+\infty} P(A_k) \geq \sum_{k=1}^{+\infty} \frac{1}{2\pi k} = +\infty,$$

en raison de la divergence bien connue de la série harmonique. Par le deuxième lemme de Borel-Cantelli, nous en déduisons que

$$P(\text{réalisation d'une infinité de } A_k) = 1,$$

autrement dit que $P(A) = 1$.

3) Nous allons déduire de ce qui précède que la suite $(X_k)_{k\geq 1}$ ne vérifie pas la loi forte des grands nombres. En posant $S_n := X_1 + \cdots + X_n$, on a

$$\forall \omega \in \Omega, \quad \frac{X_n(\omega)}{n} = \frac{S_n(\omega) - S_{n-1}(\omega)}{n} = \frac{S_n(\omega)}{n} - \frac{n-1}{n} \times \frac{S_{n-1}(\omega)}{n-1}. \tag{1}$$

Supposons maintenant que S_n/n converge presque-sûrement vers une certaine variable aléatoire Y. Notons $C := \{\omega \in \Omega;\ S_n(\omega)/n \xrightarrow[n\to+\infty]{} Y(\omega)\}$. Si ω appartient à C, alors d'après (1) et la définition de C, $X_n(\omega)/n$ doit converger vers 0. Ceci montre que cet ω n'appartient pas à A puisque la convergence vers 0 de la suite $(X_n(\omega)/n)_{n\geq 1}$ interdit qu'une infinité de ses termes soient supérieurs à 1. Nous venons ainsi de vérifier que $A \cap C = \emptyset$. Par conséquent, C est inclus dans le complémentaire de A qui est de probabilité nulle, d'où $P(C) = 0$. Ceci contredit l'hypothèse de convergence presque-sûre de S_n/n vers Y. Comme nous n'avons fait aucune hypothèse sur Y, cela signifie que S_n/n ne peut converger p.s. vers *aucune* v.a. Y. En particulier, la loi forte des grands nombres n'est pas vérifiée dans ce contexte car dans la LFGN, la limite p.s. est une v.a. constante [1].

4) Le résultat obtenu ci-dessus ne contredit pas la loi forte des grands nombres, puisque la condition nécessaire et suffisante pour la convergence p.s. de S_n/n est *l'intégrabilité de* X_1 (dans le cas d'une suite $(X_k)_{k\geq 1}$ i.i.d.). Or la loi de Cauchy n'a pas d'espérance : $\mathbf{E}|X_1| = +\infty$.

Ex 2. *Consommation de carburant*

Dans une étude statistique de la consommation de carburant d'un modèle d'automobile, on a relevé les consommations (en litres aux 100 km) pour un échantillon de 100 conducteurs sur le même trajet (tableau 1). La consommation de carburant est une variable aléatoire X de loi inconnue et on peut interpréter ce tableau comme les observations $X_1(\omega), \ldots, X_{100}(\omega)$ d'une suite de variables aléatoires indépendantes de même loi que X. On s'intéresse à la quantité $p := P(X > 7{,}5)$.

1) Considérons la suite de variables aléatoires $(Y_k)_{k\geq 1}$ définie par $Y_k := \mathbf{1}_{\{X_k > 7{,}5\}}$, où les X_k sont indépendantes et de même loi que X. Les Y_k sont des variables de Bernoulli i.i.d., donc par la loi forte des grands nombres :

$$M_n := \frac{1}{n}\sum_{k=1}^{n} Y_k \xrightarrow[n\to+\infty]{\text{p.s.}} \mathbf{E}Y_1 = \mathbf{E}\mathbf{1}_{\{X_1 > 7{,}5\}} = P(X_1 > 7{,}5) = p.$$

1. En fait, nous avons démontré un petit peu plus que ce que demandait l'énoncé, à savoir que $(Sn/n)_{n\geq 1}$ *ne peut converger sur aucun évènement* C *tel que* $P(C) > 0$. Ceci est en accord avec la loi du zéro-un (non vue en cours) qui implique que l'évènement « asymptotique » $\{S_n/n$ converge$\}$ ne peut avoir pour probabilité que 0 ou 1, pourvu que les X_k soient indépendantes.

7,641	7,329	7,509	7,572	6,910	7,028	7,429	6,931	6,025	6,589
7,542	6,969	7,473	7,779	7,537	7,013	7,848	6,668	7,327	6,172
6,406	7,467	6,980	6,297	6,698	6,478	6,828	6,630	6,767	7,018
6,796	6,934	7,342	6,673	7,666	7,001	6,862	7,028	7,160	6,807
6,469	7,495	7,656	6,616	7,235	5,974	6,947	5,884	6,913	7,384
7,193	7,293	7,582	7,979	7,668	6,885	7,069	6,650	6,096	6,853
7,954	7,196	6,360	7,747	7,824	6,270	6,662	7,434	6,426	6,888
6,494	7,562	7,216	7,868	7,188	7,562	8,107	6,284	6,626	6,915
6,800	7,726	6,613	7,423	6,524	6,280	7,530	5,984	6,664	7,870
6,867	6,799	6,699	6,649	6,274	6,872	7,519	6,833	6,643	6,413

TAB. 1 – Consommation de carburant de 100 conducteurs en litres aux 100 km

Autrement dit, pour n « grand » la fréquence observée $M_n(\omega)$ est proche de p (on considère qu'un évènement de probabilité nulle n'est pas physiquement observable). Ceci nous conduit à proposer *d'estimer* p par la valeur $M_{100}(\omega)$ calculée à partir des données du tableau 1.

Or dans ce tableau, il y a exactement 23 consommations observées supérieures à 7,5 sur 100, donc 23 termes valant 1 dans la somme des $Y_k(\omega)$, les 77 autres valant 0. On estimera donc p par $M_{100}(\omega) = 0{,}23$.

2) On peut obtenir un intervalle de confiance pour p en commençant par remarquer que le théorème limite central appliqué à la suite des v.a. Y_k nous donne :

$$M_n^* := \sqrt{\frac{n}{p(1-p)}}(M_n - p) \xrightarrow[n\to+\infty]{\text{loi}} Z, \quad \text{où } Z \sim \mathfrak{N}(0,1).$$

En particulier pour tout réel $t > 0$, $P(|M_n^*| \leq t)$ converge quand n tend vers $+\infty$ vers $P(|Z| \leq t) = 2\Phi(t) - 1$, où Φ désigne la fonction de répartition de la loi normale $\mathfrak{N}(0,1)$. La résolution numérique de l'équation $2\Phi(t) - 1 = 0,95$ à l'aide de la table des valeurs de Φ nous donne $t = 1{,}96$. En négligeant l'erreur d'approximation gaussienne, on peut donc dire que si on parie sur l'encadrement $|M_n^*| \leq 1{,}96$, on a une probabilité de succès de 0,95. Cet encadrement peut se réécrire sous la forme :

$$M_n - 1{,}96\sqrt{\frac{p(1-p)}{n}} \leq p \leq M_n + 1{,}96\sqrt{\frac{p(1-p)}{n}}. \tag{2}$$

L'inconvénient de cet encadrement de p est que ses bornes dépendent du paramètre inconnu p, via le produit $p(1-p)$ qui est la variance de Y_1. Il est bien connu que le maximum de ce produit lorsque p décrit $[0,1]$ est atteint en $p = 1/2$, ce qui nous donne la majoration $p(1-p) \leq 1/4$ valable pour tout $p \in [0,1]$. Comme on ne diminue pas la probabilité de réalisation de (2) en élargissant l'intervalle d'encadrement, le remplacement de $p(1-p)$ par $1/4$ dans (2) nous fournit l'encadrement

$$M_n - \frac{1{,}96}{2\sqrt{n}} \leq p \leq M_n + \frac{1{,}96}{2\sqrt{n}}, \tag{3}$$

sur lequel on peut parier avec une probabilité de succès *d'au moins* 0,95, en négligeant encore l'erreur d'approximation gaussienne. En appliquant ceci avec $n = 100$ et la valeur

observée $M_{100}(\omega) = 0{,}23$, on obtient finalement pour p l'intervalle de confiance

$$I = \big[0{,}132;\ 0{,}328\big] \quad \text{au niveau } 0{,}95.$$

Ex 3. *Quel candidat est en tête ?*

1) Pour tout $\omega \in \Omega$, $\mathbf{1}_A(\omega)\mathbf{1}_B(\omega)$ vaut 0 ou 1 comme produit de deux facteurs valant chacun 0 ou 1. Ce produit vaut 1 si et seulement si *ses deux facteurs* valent 1, c'est-à- dire $\omega \in A$ et $\omega \in B$, soit encore $\omega \in A \cap B$. Ainsi

$$\forall \omega \in \Omega, \quad \mathbf{1}_A(\omega)\mathbf{1}_B(\omega) = \begin{cases} 1 & \text{si } \omega \in A \cap B \\ 0 & \text{sinon} \end{cases} = \mathbf{1}_{A\cap B}(\omega),$$

ce qui justifie la formule $\mathbf{1}_A\mathbf{1}_B = \mathbf{1}_{A\cap B}$.

2) Soient A et B deux évènements disjoints ($A \cap B = \emptyset$). On note $p_1 := P(A)$, $p_2 := P(B)$ et on définit la variable aléatoire

$$X := \mathbf{1}_A - \mathbf{1}_B.$$

On suppose de plus que p_1 et p_2 sont strictement positifs. Notons C le complémentaire de $A \cup B$. Comme A, B, C sont disjoints et de réunion Ω, on a $P(C) = 1 - p_1 - p_2$ et

$$\forall \omega \in \Omega, \quad X(\omega) = \begin{cases} 1 & \text{si } \omega \in A, \\ -1 & \text{si } \omega \in B, \\ 0 & \text{si } \omega \in C. \end{cases}$$

La loi de la variable aléatoire discrète X est caractérisée par le tableau :

k	-1	0	1
$P(X=k)$	p_2	$1-p_1-p_2$	p_1

On en déduit immédiatement que :

$$\mathbf{E}X = p_2 \times (-1) + (1 - p_1 - p_2) \times 0 + p_1 \times 1 = p_1 - p_2.$$

On aurait pu obtenir ce résultat sans passer par la loi de X en utilisant la linéarité de l'espérance :

$$\mathbf{E}X = \mathbf{E}\big(\mathbf{1}_A - \mathbf{1}_B\big) = \mathbf{E}\mathbf{1}_A - \mathbf{E}\mathbf{1}_B = P(A) - P(B) = p_1 - p_2.$$

De même, pour le calcul de $\mathbf{E}X^2$, on peut soit utiliser la loi de X :

$$\mathbf{E}X^2 = p_2 \times (-1)^2 + (1 - p_1 - p_2) \times 0^2 + p_1 \times 1^2 = p_1 + p_2,$$

soit développer $(\mathbf{1}_A - \mathbf{1}_B)^2$ en notant que $\mathbf{1}_A^2 = \mathbf{1}_A$ car $0^2 = 0$ et $1^2 = 1$ et en utilisant la formule $\mathbf{1}_A\mathbf{1}_B = \mathbf{1}_{A\cap B}$:

$$\mathbf{E}X^2 = \mathbf{E}(\mathbf{1}_A - \mathbf{1}_B)^2 = \mathbf{E}(\mathbf{1}_A + \mathbf{1}_B - 2\mathbf{1}_{A\cap B}) = P(A) + P(B) - 2P(A \cap B) = p_1 + p_2.$$

La variance σ^2 de X vaut donc

$$\sigma^2 = \mathbf{E}X^2 - (\mathbf{E}X)^2 = p_1 + p_2 - (p_1 - p_2)^2 = p_1(1-p_1) + p_2(1-p_2) + 2p_1p_2. \qquad (4)$$

Sous cette forme, il est facile de voir que la variance de X vérifie $0 < \operatorname{Var} X \leq 1$. En effet, $p_1 + p_2 = P(A \cup B) \leq 1$, donc $p_2 \leq 1 - p_1$. Comme le maximum de $p(1-p)$ lorsque p décrit $[0,1]$ est atteint pour $p = 1/2$ et vaut $1/4$, on a $p_i(1-p_i) \leq 1/4$ $(i = 1, 2)$ et $p_1p_2 \leq p_1(1-p_1) \leq 1/4$ d'où

$$\operatorname{Var} X \leq \frac{1}{4} + \frac{1}{4} + \frac{2}{4} = 1$$

De plus comme $p_1 > 0$ et $0 < p_2 \leq 1 - p_1$, on a la minoration $\operatorname{Var} X \geq p_1(1-p_1) > 0$.

3) Soit $(X_k)_{k\geq 1}$ une suite de variables aléatoires indépendantes de même loi que X et pour tout $n \geq 1$, $S_n := X_1 + \cdots + X_n$. Les v.a. X_k étant bornées ($|X_k| \leq 1$) sont évidemment de carré intégrable. Toutes les hypothèses du théorème limite central (cas i.i.d.) étant vérifiées, on a :

$$S_n^* := \frac{S_n - \mathbf{E}S_n}{\sqrt{\operatorname{Var} S_n}} \xrightarrow[n\to+\infty]{\text{loi}} Z, \quad Z \sim \mathfrak{N}(0,1). \qquad (5)$$

Par linéarité de l'espérance et équidistribution des X_k, $\mathbf{E}S_n = n\mathbf{E}X = n(p_1 - p_2)$. Par indépendance et équidistribution des X_k, $\operatorname{Var} S_n = n \operatorname{Var} X = n\sigma^2$, avec σ^2 donnée par (4). Ceci nous permet de réécrire S_n^* sous la forme :

$$S_n^* = \frac{S_n - n(p_1 - p_2)}{\sigma\sqrt{n}} = \frac{\sqrt{n}}{\sigma}\Big(\frac{S_n}{n} - (p_1 - p_2)\Big).$$

La convergence en loi (5) peut donc se réécrire :

$$\frac{\sqrt{n}}{\sigma}\Big(\frac{S_n}{n} - (p_1 - p_2)\Big) \xrightarrow[n\to+\infty]{\text{loi}} Z, \quad Z \sim \mathfrak{N}(0,1). \qquad (6)$$

4) Pour une élection, 5 candidats A, B, C, D, E sont en concurrence. Dans un sondage sur 1 000 électeurs choisis au hasard dans la population totale, 370 déclarent voter pour A, 340 pour B, 154 pour C, 122 pour D et 14 pour E. On note $p_1, \ldots, p_5$, les proportions inconnues respectives d'électeurs de $A, \ldots, E$ dans la population totale. Notons A_k l'évènement « le k^{e} sondé déclare voter pour A », et définissons de même, B_k, C_k, D_k, E_k. Posons $X_k = \mathbf{1}_{A_k} - \mathbf{1}_{B_k}$. Autrement dit, X_k vaut 1 si le k^{e} sondé vote A, -1 s'il vote B et 0 dans tous les autres cas. Alors les X_k sont i.i.d. de même loi que X et S_n est exactement la différence des intentions de vote pour A et de celles pour B parmi les n sondés. Ainsi pour l'échantillon sondé dont les réponses nous fournissent les valeurs de $X_1(\omega), \ldots, X_{1\,000}(\omega)$ (ce codage occasionnant une perte d'information pour les intentions de votes en faveur de C, D, E), la somme $S_{1\,000}(\omega)$ comporte 370 termes $+1$, 340 termes -1 et 290 termes nuls, d'où $S_{1\,000}(\omega) = 370 - 340 = 30$.

La convergence en loi (6) légitime pour n grand l'approximation gaussienne $P(|S_n^*| \leq t) \simeq P(|Z| \leq t) = 2\Phi(t) - 1$. La résolution approchée de l'équation $2\Phi(t) - 1 = 0{,}98$ à

l'aide de la table nous donne $t = 2{,}33$. En négligeant l'erreur d'approximation gaussienne, on peut donc parier avec une probabilité 0,98 sur la réalisation de l'encadrement :

$$\frac{S_n}{n} - \frac{2{,}33\sigma}{\sqrt{n}} \leq p_1 - p_2 \leq \frac{S_n}{n} + \frac{2{,}33\sigma}{\sqrt{n}}.$$

L'écart-type σ dépendant de p_1 et p_2 est inconnu, mais comme il est majoré par 1, on peut parier sur l'encadrement

$$\frac{S_n}{n} - \frac{2{,}33}{\sqrt{n}} \leq p_1 - p_2 \leq \frac{S_n}{n} + \frac{2{,}33}{\sqrt{n}}.$$

avec une probabilité de succès *d'au moins* 0,98. En appliquant ceci à l'échantillon observé, avec $n = 1\ 000$ et $S_n(\omega)/n = 30/1\ 000 = 0{,}03$, on obtient pour $p_1 - p_2$ l'intervalle de confiance :

$$I = [-0{,}044;\ 0{,}104] \quad \text{au niveau } 98\%.$$

Comme la borne inférieure de cet intervalle est strictement négative, on ne peut pas en conclure qu'il y a 98% de chances que A devance B le jour de l'élection.

Question subsidiaire : *En utilisant l'approximation gaussienne de la loi de S_n^*, que pouvez vous proposer comme estimation de la probabilité que A devance B le jour de l'élection, au vu des résultats du sondage ?*

Ex 4. *Quel candidat est en tête ? par le TLC vectoriel (4 points)*

1) Pour tout $n \geq 1$, soit $N_n = (N_{n,1}, \ldots, N_{n,d})$ un vecteur aléatoire suivant la loi multinomiale de paramètres n et $p = (p_1, \ldots, p_d)$. Le théorème limite central vectoriel appliqué au vecteur aléatoire N_n s'écrit :

$$N_n^* := \frac{1}{\sqrt{n}}(N_n - np) \xrightarrow[n\to+\infty]{\text{loi}} W, \tag{7}$$

où W est le vecteur gaussien d'espérance le vecteur nul de $\mathbb{R}^d$ et de matrice de covariance $K = [K_{i,j}]$, avec $K_{i,j} = \delta_{i,j}p_i - p_ip_j$. Grâce à la conservation de la convergence en loi par images continues, on en déduit que pour toute $h : \mathbb{R}^d \to \mathbb{R}$ continue, $h(N_n^*)$ converge en loi vers $h(W)$. En particulier en prenant pour h la forme linéaire $h : (x_1, \ldots, x_d) \mapsto ax_1 + bx_2$ et en notant que :

$$N_n^* = \sqrt{n}\Big(\frac{1}{n}N_n - p\Big),$$

on obtient la convergence en loi de

$$Y_n := \sqrt{n}\Big(a\frac{N_{n,1}}{n} + b\frac{N_{n,2}}{n} - (ap_1 + bp_2)\Big) = h(N_n^*)$$

vers $h(W)$ qui est gaussienne comme combinaison linéaire d'un vecteur gaussien. Les paramètres de la loi gaussienne de la v.a. $h(W)$ sont $\mathbf{E}h(W) = h(\mathbf{E}W) = h(0) = 0$

(l'espérance des vecteurs aléatoires commute avec les applications linéaires) et l'écart-type obtenu à partir de la variance $\operatorname{Var} h(W) = \operatorname{Var}(aW_1 + bW_2)$. Cette variance se calcule à l'aide de la matrice K :

$$\begin{aligned}\operatorname{Var}(aW_1 + bW_2) &= a^2 \operatorname{Var} W_1 + b^2 \operatorname{Var} W_2 + 2ab \operatorname{Cov}(W_1, W_2) \\ &= a^2 K_{1,1} + b^2 K_{2,2} + 2ab K_{1,2} \\ &= a^2 p_1(1-p_1) + b^2 p_2(1-p_2) - 2ab p_1 p_2.\end{aligned}$$

2) Cette approche vectorielle permet de retrouver les résultats de l'exercice précédent. En effet, avec $d = 5$, posons $N_{n,1} = \sum_{k=1}^n \mathbf{1}_{A_k}, \ldots, N_{n,5} = \sum_{k=1}^n \mathbf{1}_{E_k}$. Le vecteur aléatoire $N_n = (N_{n,1}, \ldots, N_{n,5})$ ainsi formé suit la loi multinomiale de paramètres n et $p = (p_1, \ldots, p_5)$. On remarque que $N_{n,1} - N_{n,2}$ est exactement la variable aléatoire S_n de l'exercice précédent. En choisissant $a = 1/\sigma$ et $b = -a$, on voit que $Y_n = S_n^*$. Avec ce choix, la v.a. gaussienne $h(W)$ a pour espérance 0 et pour variance :

$$\operatorname{Var} h(W) = \frac{1}{\sigma^2}\big(p_1(1-p_1) + p_2(1-p_2) + 2p_1p_2\big) = 1,$$

en rappelant (4). La convergence en loi de Y_n est alors exactement celle de S_n^* donnée par (6). L'application au sondage en découle.

Remarque : Ici on a choisi un vecteur multinomial de dimension 5 pour coller au plus près au sondage présenté à l'exercice précédent. En réalité, n'importe quel vecteur multinomial de dimension $d \geq 3$ dont la loi de la projection sur les 2 premières composantes est la même que celle de $(N_{n,1}, N_{n,2})$ aurait tout aussi bien fait l'affaire. En particulier, on aurait pu prendre un vecteur multinomial de dimension 3 et de paramètres n et (p_1, p_2, q_3), avec $q_3 = 1 - (p_1 + p_2)$, ce qui revient à regrouper dans une même classe tous les sondés ne votant pour aucun des candidats A et B.