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

IS Math314 **Année 2007**

Corrigé du Devoir n° 1

Ce devoir était le sujet intégral du partiel d'I.S. 2006

Ex 1. *Pêche scientifique*

Sur 900 écrevisses pêchées et observées une par une avant d'être relâchées dans la rivière, 240 sont parasitées. On cherche un intervalle de confiance au niveau 98% pour la proportion inconnue p d'écrevisses parasitées dans la rivière.

La procédure suivie s'apparente à un tirage avec remise de boules dans une urne. Le nombre d'écrevisses parasitées dans un échantillon de taille $n = 900$ est donc une variable aléatoire S_n de loi binomiale $\operatorname{Bin}(n, p)$. L'espérance et la variance de S_n valent respectivement np et $np(1-p)$. Le théorème de de Moivre Laplace nous dit que la somme normalisée

$$S_n^* := \frac{S_n - \mathbf{E}S_n}{\sqrt{\operatorname{Var} S_n}} = \frac{S_n - np}{\sqrt{np(1-p)}} = \sqrt{\frac{n}{p(1-p)}}\Big(\frac{S_n}{n} - p\Big)$$

converge en loi quand n tend vers l'infini vers Z gaussienne standard $\mathfrak{N}(0, 1)$. Cette convergence légitime l'approximation de la loi de S_n^* par celle de Z pour n grand (approximation gaussienne). En notant Φ la fonction de répartition de Z, on peut écrire en particulier :

$$\forall t > 0, \quad P(|S_n^*| \leq t) \simeq \Phi(t) - \Phi(-t) = 2\Phi(t) - 1.$$

La résolution numérique de l'équation $2\Phi(t) - 1 = 0{,}98$ à l'aide de la table des valeurs de Φ nous donne la valeur [1] $t = 2{,}33$. En négligeant l'erreur d'approximation gaussienne, on peut donc parier avec probabilité de succès 0,98 sur la réalisation de l'encadrement :

$$|S_n^*| \leq 2{,}33 \quad \Leftrightarrow \quad \frac{S_n}{n} - \frac{2{,}33\sqrt{p(1-p)}}{\sqrt{n}} \leq p \leq \frac{S_n}{n} + \frac{2{,}33\sqrt{p(1-p)}}{\sqrt{n}} \qquad (1)$$

Cet encadrement de la valeur inconnue p a l'inconvénient d'avoir des bornes qui dépendent elles-même de p *via* la quantité $\sqrt{p(1-p)}$. On peut proposer deux façons d'y remédier.

- *Intervalle de confiance avec variance majorée.* On remarque que la fonction $p \mapsto p(1-p)$ atteint son maximum sur $[0, 1]$ pour $p = 1/2$, d'où la majoration $\sqrt{p(1-p)} \leq$

1. Ici il est prudent de prendre la valeur approchée par excès à 10^{-2} qui élargira légèrement l'intervalle de confiance. 0n pourrait obtenir une valeur plus précise de t par *interpolation linéaire*, mais cette précision serait un peu illusoire compte-tenu de l'erreur sur le niveau de confiance réel dûe à l'approximation gaussienne.

$\sqrt{1/4} = 1/2$. On ne diminue pas notre probabilité de gagner notre pari en remplaçant dans l'encadrement ci-dessus $\sqrt{p(1-p)}$ par 1/2. Comme la valeur observée de S_n/n est $\frac{240}{900} = \frac{4}{15} \simeq 0{,}266$, on peut donc proposer l'intervalle de confiance au niveau 98 % :

$$I = \left[0{,}266 - \frac{2{,}33}{2\sqrt{900}}\ ;\ 0{,}266 + \frac{2{,}33}{2\sqrt{900}}\right] = [\,0{,}227\ ;\ 0{,}306\,]$$

– *Intervalle de confiance avec variance estimée*[2]. Une autre possibilité est de remplacer la valeur inconnue p dans $\sqrt{p(1-p)}$ par un *estimateur* de p, en l'occurence S_n/n. En effet par la loi forte des grands nombres, S_n/n converge presque sûrement vers p. On peut donc considérer S_n/n comme une approximation valable de p pour n « grand ». L'avantage de cette méthode est de fournir des intervalles de confiance plus étroits, surtout lorsque p est proche de 0 ou de 1. Le prix à payer est une nouvelle source d'erreur dans l'évaluation de la probabilité de réalisation de l'encadrement obtenu. La justification théorique de cette méthode est fournie par le théorème limite central avec *autonormalisation*, cf. cours d'I.S., chapitre 3. En estimant donc p par $M_n := S_n/n$, on est ainsi conduits compte-tenu de (1), à parier sur l'encadrement :

$$\frac{S_n}{n} - \frac{2{,}33\sqrt{M_n(1-M_n)}}{\sqrt{n}} \leq p \leq \frac{S_n}{n} + \frac{2{,}33\sqrt{M_n(1-M_n)}}{\sqrt{n}}.$$

En négligeant l'erreur d'approximation gaussienne dans le TLC avec autonormalisation, on pourra encore considérer que la probabilité de gagner ce pari sera encore de 0,98. La valeur observée de M_n étant ici 0,266, on obtient ainsi l'intervalle de confiance :

$$\begin{aligned} J &= \left[0{,}266 - \frac{2{,}33\sqrt{0{,}266(1-0{,}266)}}{\sqrt{900}}\ ;\ 0{,}266 + \frac{2{,}33\sqrt{0,266(1-0,266)}}{\sqrt{900}}\right] \\ &= [\,0{,}232\ ;\ 0{,}301\,]. \end{aligned}$$

Ex 2. *Encore une question de dé*

On lance n fois un dé équilibré et on note S_n le nombre de « six » obtenus. Notons A_k l'évènement « obtention du six au k^{e} lancer ». Les A_k sont mutuellement indépendants et de même probabilité $P(A_k) = p = 1/6$. Les variables aléatoires indicatrices $X_k := \mathbf{1}_{A_k}$ sont des v.a. de Bernoulli indépendantes de même paramètre p. Le nombre de « six » obtenus en n lancers est

$$S_n = \sum_{k=1}^{n} \mathbf{1}_{A_k} = \sum_{k=1}^{n} X_k.$$

La v.a. de Bernoulli X_k a pour espérance $p = 1/6$ et variance $p(1-p) = 5/36$. Par additivité de l'espérance et équidistribution[3] des X_k, on a $\mathbf{E}S_n = n\mathbf{E}X_1 = np = n/6$. Par indépendance et équidistribution des X_k, on a $\operatorname{Var} S_n = n \operatorname{Var} X_1 = np(1-p) = 5n/36$.

2. Cette méthode n'était pas exigible, ni au partiel ni au D.M., car elle anticipe sur la suite du cours.
3. « Équidistribution » signifie que les variables ont même loi.

La somme centrée réduite S_n^* du théorème de de Moivre Laplace s'écrit ici

$$S_n^* = \frac{S_n - \mathbf{E}S_n}{\sqrt{\operatorname{Var} S_n}} = \frac{S_n - n/6}{\sqrt{5n/36}} = \frac{6\sqrt{n}}{\sqrt{5}}\Bigl(\frac{S_n}{n} - \frac{1}{6}\Bigr).$$

Ce théorème nous dit que S_n^* converge en loi vers une v.a. gaussienne standard Z lorsque n tend vers l'infini. Ceci légitime pour n grand l'approximation de la loi de S_n^* par la loi $\mathfrak{N}(0,1)$. En particulier pour $t > 0$, on a

$$P(|S_n^*| < t) \simeq P(|Z| < t) = 2\Phi(t) - 1,$$

en notant Φ la f.d.r. de $\mathfrak{N}(0,1)$.

En raison de l'équivalence

$$\Bigl|\frac{S_n}{n} - \frac{1}{6}\Bigr| < 0{,}01 \;\Leftrightarrow\; |S_n^*| < \frac{0{,}06\sqrt{n}}{\sqrt{5}},$$

on est ramené à chercher — en négligeant l'erreur d'approximation gaussienne — le plus petit entier n tel qu'avec $t = 0{,}06 \times 5^{-1/2} n^{1/2}$, $2\Phi(t) - 1 \geq 0{,}9$. Comme Φ est croissante et continue et t est une fonction croissante de n, cela revient à chercher le plus petit entier n tel que $\Phi(t) \geq 0{,}95$. La lecture inverse de la table nous donne $\Phi^{-1}(0{,}95) = 1{,}645$. L'entier n cherché est donc la plus petite solution entière de l'inéquation :

$$\frac{0{,}06\sqrt{n}}{\sqrt{5}} \geq 1{,}645,$$

qui équivaut à

$$n \geq \frac{5 \times 1{,}645^2}{0{,}06^2} = 3\,758{,}37.$$

La valeur cherchée est donc $n = 3\,759$.

Ex 3. *Différence de sommes*

On suppose que les deux suites de variables aléatoires $(X_i)_{i\geq 1}$ et $(Y_i)_{i\geq 1}$ sont définies sur le même espace probabilisé $(\Omega, \mathcal{F}, P)$ et vérifient les hypothèses suivantes :

a) chacune de ces suites est i.i.d. ;

b) X_i et Y_i sont de carré intégrable, $\mathbf{E}X_i =: m_1$, $\mathbf{E}Y_i =: m_2$, $\operatorname{Var} X_i =: \sigma_1^2 > 0$, $\operatorname{Var} Y_i =: \sigma_2^2 > 0$;

c) les suites $(X_i)_{i\geq 1}$ et $(Y_i)_{i\geq 1}$ sont indépendantes l'une de l'autre, ce qui implique notamment que pour toute fonction mesurable $g : \mathbb{R}^2 \to \mathbb{R}$, $(g(X_i, Y_i))_{i\geq 1}$ est une suite de v.a. indépendantes et aussi que pour tout $i \geq 1$, X_i et Y_i sont indépendantes.

On pose pour tout $n \geq 1$,

$$S_n := \sum_{i=1}^n X_i, \quad T_n := \sum_{i=1}^n Y_i.$$

1) On suppose que $m_2 > m_1$. On va montrer que

$$W_n := \frac{1}{\sqrt{n}}\bigl(T_n - S_n\bigr) \xrightarrow[n\to+\infty]{\text{p.s.}} +\infty.$$

On remarque que W_n peut s'écrire

$$W_n = \sqrt{n}V_n, \quad \text{avec} \quad V_n := \frac{T_n}{n} - \frac{S_n}{n}.$$

Les suites $(X_i)_{i\geq 1}$ et $(Y_i)_{i\geq 1}$ sont des suites de variables aléatoires i.i.d. et intégrables. Elles vérifent ainsi les hypothèses de la loi forte des grands nombres, de sorte que les évènements

$$\Omega_1 := \Bigl\{\frac{S_n}{n} \xrightarrow[n\to+\infty]{} m_1\Bigr\}, \quad \Omega_2 := \Bigl\{\frac{T_n}{n} \xrightarrow[n\to+\infty]{} m_2\Bigr\}$$

ont chacun pour probabilité 1, ainsi que leur intersection Ω'. Pour tout $\omega \in \Omega'$, $V_n(\omega)$ converge vers $m_2 - m_1$ quand n tend vers $+\infty$. Comme cette limite est strictement positive, le produit $\sqrt{n}V_n(\omega)$ tend vers $+\infty$. Ceci étant vrai pour tout ω de l'évènement Ω' de probabilité 1, on peut conclure à la convergence presque-sûre de W_n vers $+\infty$.

2) Bien entendu le raisonnement ci-dessus n'est plus valable si $m_1 = m_2$, puisqu'alors le produit $\sqrt{n}V_n(\omega)$ se présente pour $\omega \in \Omega'$ comme une forme indéterminée « $+\infty \times 0$ » lorsque n tend vers l'infini.

En posant $Z_i := Y_i - X_i$ pour tout $i \geq 1$, on voit que la suite $(Z_i)_{i\geq 1}$ est i.i.d., de carré intégrable et que $\mathbf{E}Z_1 = m_2 - m_1 = 0$. On peut calculer la variance de Z_1 combinaison linéaire de deux v.a. indépendantes[4] :

$$\operatorname{Var} Z_1 = \operatorname{Var} Y_1 + (-1)^2 \operatorname{Var} X_1 = \sigma_1^2 + \sigma_2^2 =: \sigma^2.$$

Notons σ la racine carrée strictement positive de σ^2. Posons $R_n := Z_1 + \dots + Z_n$, et notons que $\mathbf{E}R_n = n\mathbf{E}Z_1 = 0$ et $\operatorname{Var} R_n = n\sigma^2$. On peut alors écrire

$$W_n = \sigma\frac{R_n}{\sigma\sqrt{n}} = \sigma R_n^* \quad \text{où} \quad R_n^* := \frac{R_n - \mathbf{E}R_n}{\sqrt{\operatorname{Var} R_n}}.$$

Par le théorème limite central, R_n^* converge en loi quand n tend vers l'infini vers Z gaussienne standard (de loi $\mathfrak{N}(0,1)$). On en déduit que σR_n^* converge en loi vers σZ parce que la convergence en loi est préservée par multiplication par une constante (l'application $x \mapsto \sigma x$ est continue). La variable aléatoire σZ est gaussienne $\mathfrak{N}(0,\sigma)$.

4. Attention à l'« ânerie : $\operatorname{Var}(Y_1 - X_1) = \operatorname{Var} Y_1 - \operatorname{Var} X_1$ » rencontrée dans de trop nombreuses copies. Elle pourrait notamment donner une variance négative ! La formule correcte est $\operatorname{Var}(aX_1 + bY_1) = a^2 \operatorname{Var} X_1 + 2ab \operatorname{Cov}(X_1, Y_1) + b^2 \operatorname{Var} Y_1$ et lorsque les deux v.a. X_1 et Y_1 sont indépendantes, la covariance est nulle.

Ex 4. *Convergences p.s. de produits*

1) Soit U une variable aléatoire à valeurs dans $]0,1[$, de loi uniforme sur $]0,1[$ et $X := -\ln U$. La fonction exponentielle et sa réciproque la fonction logarithme étant croissantes strictement, on a l'équivalence :

$$\forall \omega \in \Omega, \quad -\ln U > x \Leftrightarrow U < \mathrm{e}^{-x}.$$

Cette équivalence se traduit par l'égalité d'évènements $\{-\ln U > x\} = \{U < \mathrm{e}^{-x}\}$. Comme U suit la loi uniforme sur $]0,1[$, sa fonction de répartition F est *continue* et

$$P(U < t) = P(U \leq t) = F(t) = \begin{cases} t & \text{si } 0 < t < 1, \\ 1 & \text{si } t \geq 1. \end{cases}$$

Pour $x \geq 0$, e^{-x} est dans $]0,1[$, donc d'après ce qui précède, $P(-\ln U > x) = P(U < \mathrm{e}^{-x}) = F(\mathrm{e}^{-x}) = \mathrm{e}^{-x}$. D'autre part, comme U est à valeurs dans $]0,1[$, $-\ln U$ est une v.a. positive, donc $P(-\ln U > x)$ vaut 1 pour $x < 0$. Finalement la fonction de survie G de X est donnée par

$$G(x) = \begin{cases} 1 & \text{si } x < 0, \\ \exp(-x) & \text{si } x \geq 0. \end{cases}$$

On reconnaît ici la fonction de survie de la loi exponentielle de paramètre 1. En particulier, X a pour espérance 1.

2) Soit $(U_i)_{i\geq 1}$ une suite de variables aléatoires définies sur le même espace probabilisé $(\Omega, \mathcal{F}, P)$, à valeurs dans $]0,1[$, indépendantes et de même loi uniforme sur $]0,1[$. Pour tout $n \geq 1$, on note M_n la *moyenne géométrique* de $U_1, \dots, U_n$, définie par

$$M_n := \left(\prod_{i=1}^{n} U_i \right)^{1/n}.$$

Par construction, M_n est une v.a. à valeurs strictement positives, donc son logarithme est bien défini. On remarque que

$$\ln M_n = -\frac{1}{n} \sum_{i=1}^{n} X_i, \quad X_i := -\ln U_i.$$

Les X_i sont indépendantes de même loi exponentielle de paramètre 1, donc intégrables et d'espérance 1. Par la loi forte des grands nombres, $n^{-1}(X_1 + \dots + X_n)$ tend presque-sûrement vers 1, d'où :

$$\ln M_n \xrightarrow[n\to+\infty]{\text{p.s.}} -1.$$

La convergence presque-sûre étant conservée par les applications continues, on en déduit en prenant l'exponentielle que M_n converge presque-sûrement vers e^{-1} quand n tend vers l'infini.

On montre maintenant par deux méthodes la convergence p.s. vers 0 *de*

$$T_n := \prod_{i=1}^{n} U_i = M_n^n.$$

3) Par la question 2, l'évènement $\Omega' := \left\{ M_n \xrightarrow[n\to+\infty]{} \mathrm{e}^{-1} \right\}$ a pour probabilité 1. Soit ω un élément quelconque de Ω'. Comme la limite e^{-1} de $M_n(\omega)$ est *strictement* inférieure à $2/\mathrm{e}$, il existe un rang $k(\omega)$ tel que pour tout $n \geq k(\omega)$, $M_n(\omega) < 2/\mathrm{e}$. On en déduit que :

$$\forall n \geq k(\omega), \quad 0 < M_n(\omega)^n \leq \left(\frac{2}{\mathrm{e}} \right)^n.$$

Comme $0 < 2/\mathrm{e} < 1$, ce majorant de $M_n(\omega)^n$ tend vers 0 quand n tend vers $+\infty$ et il en va donc de même pour $M_n(\omega)^n$. Ceci étant vrai pour tout ω de l'évènement Ω' de probabilité 1, on a établi la convergence presque sûre vers 0 de $T_n = M_n^n$.

4) Voici une deuxième méthode qui n'utilise pas la question 2.

a) Fixons ω quelconque dans Ω et regardons le comportement la suite de nombres réels $(T_n(\omega))_{n\geq 1}$. Comme les réels $U_1(\omega), \dots, U_n(\omega)$ sont tous dans $]0,1[$, il en est de même pour leur produit $T_n(\omega)$. On note ensuite que $(T_n(\omega))_{n\geq 1}$ est strictement décroissante car pour tout $n \geq 1$, $T_{n+1}(\omega) = T_n(\omega)U_{n+1}(\omega)$ et $0 < U_{n+1}(\omega) < 1$. Ainsi la suite de réels $(T_n(\omega))_{n\geq 1}$ est décroissante et minorée par 0. Elle converge donc dans $[0,1[$. En posant

$$\forall \omega \in \Omega, \quad T(\omega) := \lim_{n\to+\infty} T_n(\omega),$$

on définit une application $T : \Omega \to [0,1[$. Cette application hérite de la mesurabilité des variables aléatoires T_n (résultat admis en cours), c'est donc bien une variable aléatoire.

b) Les variables aléatoires $U_1, \dots, U_n$ sont intégrables (car bornées) *et* indépendantes. Leur produit T_n est donc intégrable et

$$\mathbf{E}T_n = \mathbf{E}U_1 \dots \mathbf{E}U_n = (\mathbf{E}U_1)^n = \left(\frac{1}{2} \right)^n = 2^{-n}.$$

Ainsi $\mathbf{E}T_n$ tend vers 0 quand n tend vers l'infini. La v.a. T_n étant positive, cette convergence vers 0 est aussi celle de $\mathbf{E}|T_n - 0|$, autrement dit T_n tend vers 0 au sens L^1 donc aussi en probabilité[5].

c) On a vu au a) que $T_n(\omega)$ converge vers $T(\omega)$ pour tout $\omega \in \Omega$. Ceci implique bien évidemment la convergence presque-sûre de T_n vers T, donc aussi sa convergence en probabilité. Mais T_n converge aussi en probabilité vers 0 d'après le b). Par unicité de la limite en probabilité modulo l'égalité presque-sûre, on en déduit que la v.a. T est nulle presque-sûrement. En revenant à la convergence p.s. de T_n vers T, on conclut que T_n converge p.s. vers 0.

5. Le fait que la convergence L^1 implique la convergence en probabilité avait été vu en cours, ou pouvait être redémontré immédiatement par l'inégalité de Markov.