Université Paris-Dauphine Année 2008-2009
U.F.R. Mathématiques de la décision
L3 - Statistique Mathématique

Partiel

Eléments de correction

Questions de cours : *(3 points)*

Soit $(X_1, \dots, X_n)$ un échantillon de la variable X issu d'une population de taille finie $N \geq n$. On note $\{x_1, \dots, x_N\}$ l'ensemble des valeurs prises par la variable X.

1. Quelle est la moyenne et la variance de X ?
2. Quelle est la loi jointe de l'échantillon $(X_1, \dots, X_n)$ dans le cas d'un tirage avec remise et dans le cas d'un tirage sans remise ?
3. Dans les deux cas que vaut $\mathrm{Cov}(X_1, X_2)$? Justifier votre réponse.

Réponses :

1. La variable X est équidistribuée sur l'ensemble $\{x_1, \dots, x_N\}$ donc
$$\mathbb{E}(X) = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{et} \quad \mathrm{Var}\,(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mathbb{E}(X))^2.$$
2. Dans le cas du tirage avec remise, les X_i sont iid $\sim \mathcal{U}(\{x_1, \dots, x_N\})$ l'uniforme discrète donc $(X_1, \dots, X_n) \sim \mathcal{U}(\{x_1, \dots, x_N\})^{\otimes n}$.
Dans le cas du tirage avec remise, $(X_1, \dots, X_n)$ est équidistribuée sur l'ensemble des n-uplets de $\{x_1, \dots, x_N\}$.
3. Dans le cas du tirage avec remise, par indépendance $\mathrm{Cov}(X_1, X_2) = 0$.
Dans le cas du tirage sans remise, $\mathrm{Cov}(X_1, X_2) = -\sigma^2/(N-1)$.

Exercice 1 : *(5 points)*

Soit $(X_1, ..., X_n)$ un n-échantillon (iid) de loi uniforme $\mathcal{U}([a, b])$ où $a < b$ sont deux réels. On se propose d'estimer la moyenne théorique $m = \mathbb{E}(X_1)$.

1. Calculer m.
2. On considère d'abord l'estimateur usuel $\bar{X}_n$. Déterminer le comportement asymptotique de $\bar{X}_n$.
3. On considère maintenant l'estimateur $T_n = (X_{(1)} + X_{(n)})/2$. Calculer son espérance et sa variance, ainsi que l'erreur quadratique moyenne $\mathbb{E}(T_n - m)^2$ et commenter son

comportement asymptotique.
<u>Indication :</u> on pourra utiliser la fonction β définie par

$$\beta(k,n) = \int_0^1 t^k(1-t)^n dt = \frac{k!n!}{(k+n+1)!} \qquad \forall k,n \in \mathbb{N}.$$

4. Quel estimateur choisir ?

<u>Réponses :</u>

1. $m = \mathbb{E}(X_1) = \int x f_{X_1}(x)dx = \int_a^b x/(b-a)dx = (b^2-a^2)/2(b-a) = (b+a)/2.$
2. Etant donné que les X_i sont iid bornées par b on peut appliquer le TCL :

$$\sqrt{n}\left(\overline{X}_n - \frac{b+a}{2}\right) \to \mathcal{N}(0,\sigma^2) \quad \text{en loi,}$$

où $\sigma^2 = \mathrm{Var}\,(X_1) = (b-a)^2/12.$

3. On considère la transformation affine $(X_i - a)/(b-a) \sim \mathcal{U}([0;1])$ et l'ordre des X_i est respecté. Par linéarité $\mathbb{E}(T_n) = (\mathbb{E}(X_{(1)}) + \mathbb{E}(X_{(n)}))/2$ et

$$\frac{\mathbb{E}(X_{(1)}-a)}{b-a} = \int_0^1 xn(1-x)^{n-1}dx = n\beta(1,n-1) = n\frac{(n-1)!}{(n+1)!} = \frac{1}{n+1},$$
$$\frac{\mathbb{E}(X_{(n)}-a)}{b-a} = \int_0^1 xnx^{n-1}dx = n\beta(0,n) = n\frac{n!}{(n+1)!} = 1 - \frac{1}{n+1}.$$

On en déduit que $(\mathbb{E}(T_n) - a)/(b-a) = 1/2$ donc $\mathbb{E}(T_n) = (b+a)/2 = m$. On en déduit que la variance et l'erreur quadratique de T_n sont identiques. On a alors

$$\mathrm{Var}\,(T_n) = \mathbb{E}(T_n - m)^2 = \frac{1}{4}\mathbb{E}(X_{(1)} + X_{(n)} - (b+a))^2 = \frac{(b-a)^2}{4}\mathbb{E}\left(\frac{X_{(1)} + X_{(n)} - (b+a)}{b-a}\right)^2.$$

On écrit

$$\begin{aligned}\mathbb{E}\left(\frac{X_{(1)} + X_{(n)} - (b+a)}{b-a}\right)^2 &= \mathbb{E}\left(\frac{X_{(1)}-a}{b-a} + \frac{X_{(n)}-a}{b-a} - 1\right)^2 \\ &= \mathbb{E}\left(\frac{X_{(1)}-a}{b-a}\right)^2 + \mathbb{E}\left(\frac{X_{(n)}-a}{b-a}\right)^2 + 1 - 2\mathbb{E}\left(\frac{X_{(1)}-a}{b-a}\right) - 2\mathbb{E}\left(\frac{X_{(n)}-a}{b-a}\right) \\ &\quad + 2\mathbb{E}\left[\left(\frac{X_{(1)}-a}{b-a}\right)\left(\frac{X_{(n)}-a}{b-a}\right)\right].\end{aligned}$$

Reste à calculer

$$\mathbb{E}\left(\frac{X_{(1)}-a}{b-a}\right)^2 = \int_0^1 x^2 n(1-x)^{n-1}dx = n\beta(2,n-1) = n\frac{2!(n-1)!}{(n+2)!} = \frac{2}{(n+1)(n+2)},$$
$$\mathbb{E}\left(\frac{X_{(n)}-a}{b-a}\right)^2 = \int_0^1 x^2 n x^{n-1}dx = n\beta(0,n+1) = n\frac{(n+1)!}{(n+2)!} = \frac{n}{n+2} \quad \text{et}$$
$$\mathbb{E}\left[\left(\frac{X_{(1)}-a}{b-a}\right)\left(\frac{X_{(n)}-a}{b-a}\right)\right] = \int_0^1\int_0^y xy(n-1)n(y-x)^{n-2}dxdy.$$

Par changement de variable $z = x/y$ pour tout $0 < y \leq 1$ fixé, on trouve

$$\int_0^y x(y-x)^{n-2}dx = y^2 \int_0^1 z(y-yz)^{n-2}dz = y^n \beta(1, n-2).$$

On en déduit que l'intégrale double vaut $(n-1)n\beta(1,n-2)\int_0^1 y^{n+1}dy = 1/(n+2)$. En regroupant tous les termes on obtient

$$\mathrm{Var}\,(T_n) = \frac{(b-a)^2}{4}\left(\frac{2}{(n+1)(n+2)} + \frac{n}{n+2} - 1 + 2 + \frac{2}{n+2}\right) = \frac{(b-a)^2}{2(n+1)(n+2)}.$$

Lorsque n tend vers l'infini, l'erreur quadratique de T_n décroît comme $1/n^2$.

4. Nous choisirons T_n car $\overline{X}_n$ décroît en moyenne quadratique comme $\mathrm{Var}\,(\overline{X}_n) \approx \sigma^2/n$ lorsque n tend vers l'infini d'après la question 2.

Exercice 2 : *(5 points)*

Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires indépendantes de loi uniforme sur $[0,1]$.

1. Montrer que $\bar{X}_n = \frac{1}{n}\sum_{k=1}^n X_k$ converge en probabilité vers $\frac{1}{2}$. Cette convergence a-t-elle lieu presque sûrement ?
2. Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction continue et bornée, montrer que

$$\lim_{n\to\infty} \int_{[0,1]^n} f\left(\frac{x_1 + \ldots + x_n}{n}\right) dx_1 \ldots dx_n = f\left(\frac{1}{2}\right).$$

On considère maintenant une suite $(Y_n)_{n\geq 1}$ de variables aléatoires indépendantes de loi de Poisson de paramètre 1.

3. Déterminer la loi de $Z_n = \sum_{k=1}^n Y_k$. En déduire que

$$P\left(\frac{Z_n - n}{\sqrt{n}} \leq 0\right) = \sum_{k=0}^n e^{-n}\frac{n^k}{k!}.$$

4. Montrer que la suite $\frac{Z_n - n}{\sqrt{n}}$ converge en loi et décrire sa limite.
5. En déduire que $\lim_{n\to\infty} e^{-n}(1 + n + \frac{n^2}{2!} + ... + \frac{n^n}{n!}) = \frac{1}{2}$.

Réponses :

1. En appliquant la LFGN sur les X_i iid et bornées par 1 on obtient presque sûrement

$$\overline{X}_n \to \mathbb{E}(X_1) = \frac{1}{2}.$$

2. La convergence presque sûre impliquant la convergence en loi de $\overline{X}_n$ vers la Dirac $\delta_{1/2}$, par définition de celle-ci on obtient que pour toute fonction f continue bornée

$$\mathbb{E}(f(\overline{X}_n)) \to \mathbb{E}(f(Y)) \text{ quand } n \to \infty \text{ avec } Y \sim \delta_{1/2}.$$

On en déduit le résultat souhaité car $\mathbb{E}(f(\overline{X}_n)) = \int_{[0,1]^n} f\left((x_1 + \ldots + x_n)/n\right) dx_1 \ldots dx_n$ par indépendance et $\mathbb{E}(f(Y)) = f(1/2)$ par définition de la mesure de Dirac.

3. Par indépendance des Y_i, la loi de Z_n est une loi de Poisson de paramètre n. on en déduit que
$$P\left(Z_n \leq n\right) = \sum_{k=0}^{n} P(Z_n = k) = \sum_{k=0}^{n} e^{-n}\frac{n^k}{k!}$$
d'où le résultat souhaité.
4. Etant donné que par définition $(Z_n - n)/\sqrt{n} = \sqrt{n}(\overline{Y}_n - 1)$ on applique le TCL sur les Y_i iid tels que $\mathbb{E}(Y_1^2) = 1 < \infty$ et on obtient la convergence en loi de $(Z_n - n)/\sqrt{n}$ vers une $\mathcal{N}(0,1)$.
5. La convergence en loi est équivalente à la convergence simple des fonctions de répartition et en particulier $P((Z_n - n)/\sqrt{n} \leq 0) \to P(N \leq 0)$ où $N \sim \mathcal{N}(0,1)$. Par symétrie de la loi normale centrée, on a $P(N \leq 0) = 1/2$ et la propriété est prouvée en utilisant l'écriture de $P((Z_n - n)/\sqrt{n} \leq 0)$ trouvée en 3.

Problème : *(7 points)*

On considère une suite $(\varepsilon_i)_{i\in\mathbb{N}}$ de variables aléatoires indépendantes et identiquement distribuées suivant une loi $\mathcal{N}(0,\sigma^2)$, où $\sigma^2 > 0$ est un paramètre inconnu. Pour tout $n \in \mathbb{N}^*$, on définit :
$$X_n = \varepsilon_n - \alpha\varepsilon_{n-1},$$
où $\alpha \in \mathbb{R}$ est inconnu.

1. Montrer que $\mathrm{Var}\left(\varepsilon_i^2\right) = 2\sigma^4$.
2. Pour tout $i \in \mathbb{N}^*$, déterminer $\mathbb{E}(X_i)$ et $\mathrm{Var}(X_i)$. Montrer que $(X_i)_i$ est une suite de variables identiquement distribuées dont vous préciserez la loi.
3. Montrer que $\mathrm{Cov}(X_i, X_j) = (1+\alpha^2)\sigma^2$ si $i = j$, $\mathrm{Cov}(X_i, X_j) = -\alpha\sigma^2$ si $|i-j| = 1$ et $\mathrm{Cov}(X_i, X_j) = 0$ sinon.
4. Pour n fixé, en déduire la loi du vecteur $(X_1, \ldots, X_n)$.
5. Soit $\widehat{\sigma}_n^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2$. Montrer que $\mathbb{E}(\widehat{\sigma}_n^2) = (1+\alpha^2)\sigma^2$. Soit
$$Z_n^{(1)} = \frac{1}{n}\sum_{k=1}^{[n/2]} X_{2k}^2 \quad \text{et} \quad Z_n^{(2)} = \frac{1}{n}\sum_{k=1}^{[(n+1)/2]} X_{2k-1}^2,$$
avec $[x]$ la partie entière de x. Montrer que les suites de variables aléatoires $(Z_n^{(1)})_n$ et $(Z_n^{(2)})_n$ convergent presque sûrement et préciser leurs limites.
6. Montrer que si deux suites de variables aléatoires convergent presque sûrement alors leur somme converge presque sûrement. En déduire que $(\widehat{\sigma}_n^2)_n$ converge presque sûrement vers $(1+\alpha^2)\sigma^2$.
7. Soit $\widehat{\rho}_n = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n-1} X_i X_{i+1}$. En utilisant le même type d'argument que précédemment, montrer que $(\widehat{\rho}_n)_n$ converge presque sûrement vers $-\alpha\sigma^2$. En déduire une suite de variables aléatoires convergeant presque sûrement vers α.

Réponses :

1. Par définition $\varepsilon_i^2 \sim \sigma^2\chi(1)$. Or on a $\chi(1) = \gamma(1/2, 1/2)$ donc la variance de cette loi vaut $1/2/(1/2)^2 = 2$ d'où le résultat souhaité.
2. Par linéarité $\mathbb{E}(X_i) = \mathbb{E}(\varepsilon_i) - \alpha\mathbb{E}(\varepsilon_{i-1}) = 0$ et par indépendance des ε_i on a $\mathrm{Var}\,(X_i) = \mathrm{Var}\,(\varepsilon_i) + \alpha^2\mathrm{Var}\,(\varepsilon_{i-1}) = (1+\alpha^2)\sigma^2$. Etant donné que la somme de deux gaussiennes est une gaussienne, les X_i sont tous de même loi $\mathcal{N}(0, (1+\alpha^2)\sigma^2)$.
3. Pour tout i, $\mathrm{Cov}(X_i, X_i) = \mathrm{Var}\,(X_i)$ et $\mathrm{Cov}(X_i, X_{i-1}) = \mathbb{E}(X_iX_{i-1}) = \mathbb{E}((\varepsilon_i - \alpha\varepsilon_{i-1})(\varepsilon_{i-1} - \alpha\varepsilon_{i-2}) = -\alpha\mathbb{E}(\varepsilon_{i-1}^2) = -\alpha\sigma^2$. Enfin, lorsque $|i-j| > 1$ alors X_i et X_j sont fonctions des couples $(\varepsilon_i, \varepsilon_{i-1})$ et $(\varepsilon_j, \varepsilon_{j-1})$. Etant donné que $\{i, i-1\} \cap \{j, j-1\} = \emptyset$ les couples sont indépendants et il en est de même de X_i et X_j. Leur covariance est donc nulle.
4. Soit $X = (X_1, \ldots, X_n)'$, soit $\varepsilon = (\varepsilon_n, \ldots, \varepsilon_0)'$ et soit M la matrice de taille $n \times (n+1)$
$$M = \begin{pmatrix} 1 & -\alpha & 0 & \ldots & 0 \\ 0 & 1 & -\alpha & 0 & \vdots \\ \vdots & 0 & 1 & -\alpha & 0 \\ 0 & \ldots & 0 & 1 & -\alpha \end{pmatrix}.$$
Alors par indépendance $\varepsilon \sim \mathcal{N}_{n+1}(0, I)$ et $X = M\varepsilon$ est aussi un vecteur gaussien. Sa moyenne vaut $M0 = 0$ et sa matrice de variance covariance $\Sigma = MIM' = MM'$ est donnée par
$$\Sigma = \begin{pmatrix} (1+\alpha^2)\sigma^2 & -\alpha\sigma^2 & 0 & \ldots & 0 \\ -\alpha\sigma^2 & (1+\alpha^2)\sigma^2 & -\alpha\sigma^2 & 0 & \vdots \\ 0 & -\alpha\sigma^2 & (1+\alpha^2)\sigma^2 & -\alpha\sigma^2 & 0 \\ \vdots & 0 & -\alpha\sigma^2 & (1+\alpha^2)\sigma^2 & -\alpha\sigma^2 \\ 0 & 0 & 0 & -\alpha\sigma^2 & (1+\alpha^2)\sigma^2 \end{pmatrix}.$$
5. Par linéarité $\mathbb{E}(\widehat{\sigma}_n^2) = \mathbb{E}(X_i^2) = \mathrm{Var}(X_i) = (1+\alpha^2)\sigma^2$. On a par définition $Z_{2n}^{(1)} = \sum_{k=1}^n X_{2k}^2/(2n)$, la moitié de la moyenne empirique des $(X_2^2, X_4^2, \ldots, X_{2n}^2)$. On applique la loi forte des grands nombres aux X_{2k}^2 iid tels que $\mathbb{E}(X_{2k}^2) = (1+\alpha^2)\sigma^2$ et on obtient la convergence presque sûre de $Z_{2n}^{(1)}$ vers $(1+\alpha^2)\sigma^2/2$. Il en est de même pour $Z_n^{(1)}$ et pour $Z_n^{(2)}$.
6. Soient $V_n \to V$ et $W_n \to W$ deux suites de variables aléatoires convergeant presque sûrement. Alors par définition $P(\{\lim V_n = V\}^c) = 0$ et $P(\{\lim W_n = W\}^c) = 0$ donc $P(\{\lim V_n = V\}^c \cup \{\lim W_n = W\}^c) = 0$ par sous additivité. Par passage au complémentaire $P(\{\lim V_n = V\} \cap \{\lim W_n = W\}) = 1$ mais $\{\lim(V_n + W_n) = V + W\} \supseteq \{\lim V_n = V\} \cap \{\lim W_n = W\}$ d'où $P(\{\lim(V_n + W_n) = V + W\}) = 1$. On a montré que $Z_n^{(1)}$ et $Z_n^{(2)}$ convergent presque sûrement vers $(1+\alpha^2)\sigma^2/2$ donc leur somme $\widehat{\sigma}_n^2$ converge aussi presque sûrement (bien que $Z_n^{(1)}$ et $Z_n^{(2)}$ soient dépendants).
7. On définit
$$Y_n^{(1)} = \frac{1}{n-1}\sum_{k=1}^{[(n-1)/3]} X_{3k}X_{3k+1} \quad , \quad Y_n^{(2)} = \frac{1}{n-1}\sum_{k=1}^{[(n-1)/3]} X_{3k-1}X_{3k} \quad \text{et}$$
$$Y_n^{(3)} = \frac{1}{n-1}\sum_{k=1}^{[(n-1)/3]} X_{3k-2}X_{3k-1}.$$

Alors pour tout $i = 1, 2, 3$ la variable $Y_n^{(i)}$ est équivalente au tiers de la moyenne empirique de variables aléatoires iid d'espérance $-\alpha\sigma^2$. En appliquant la LFGN les suites $Y_n^{(i)}$ convergent presque sûrement vers $-\alpha\sigma^2/3$ et leur somme $\widehat{\rho}_n$ converge elle-aussi presque sûrement vers $-\alpha\sigma^2$.