Comportement asymptotique de la variance empirique.

Soient $(X_1,\ldots,X_n)$ un n-échantillon d'une variable aléatoire X, telle que $\mathbb{E}[X^4] < \infty$. On notera $\mu = \mathbb{E}[X]$, $\sigma^2 = \mathbb{V}(X)$ et $\mathbb{V} = \mathbb{V}((X-\mu)^2)$. On considère les statistiques suivantes :

$$\begin{aligned} \overline{X}_n &= \frac{1}{n}\sum_{i=1}^n X_i \\ V_n &= \frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2 \\ S_n^2 &= \frac{1}{n}\sum_{i=1}^n (X_i-\overline{X}_n)^2 \end{aligned}$$

On notera que V_n est la variance empirique dans le cas où l'espérance est connue mais est incalculable. Le but de l'exercice est de montrer que V_n et S_n^2 ont le même comportement asymptotique.

1°. Montrer que

$$\sqrt{n}\left(V_n-\sigma^2\right) \rightsquigarrow \mathcal{N}(0,\mathbb{V})$$

Les $(X_i-\mu)_i$ sont des vaiid et par conséquent les $\left((X_i-\mu)^2\right)_i$ le sont aussi. Comme les X_i ont un moment d'ordre 4, les $X_i-\mu$ sont de carré intégrable et l'on peut donc appliquer le TLC. $\sqrt{n}\left(V_n-\sigma^2\right)$ est asymptotiquement normal et la loi limite a pour variance $\mathbb{V}$.

2°. Montrer que $S_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2$

Il s'agit de la formule de König-Huygens. On développe l'expression :

$$\begin{aligned} S_n^2 &= \frac{1}{n}\sum_{i=1}^n (X_i-\overline{X}_n)^2 \\ &= \frac{1}{n}\sum_{i=1}^n (X_i^2 - 2X_i\overline{X}_n + \overline{X}_n^2) \\ &= \frac{1}{n}\sum_{i=1}^n X_i^2 - 2\overline{X}_n^2 + \overline{X}_n^2 \\ &= \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2 \end{aligned}$$

3°. Montrer que $S_n^2 \xrightarrow[n\to\infty]{\mathbb{P}} \sigma^2$.

D'après la loi forte des grands nombres, $\overline{X_n^2}$ converge presque sûrement vers $\mathbb{E}[X_i^2] = \sigma^2+\mu^2$ et $\overline{X}_n$ converge presque sûrement vers $\mathbb{E}[X_i] = \mu$ (les va sont iid et ont un moment d'ordre 1). D'après le théorème de l'application continue, on en déduit que $\overline{X}_n^2$ converge presque sûrement vers $\mathbb{E}[X_i]^2 = \mu^2$. Enfin, la convergence presque sûre est stable par addition et par conséquent $S_n^2 = \overline{X_n^2} - \overline{X}_n^2$ converge presque sûrement vers $\sigma^2+\mu^2-\mu^2 = \sigma^2$.

La convergence presque sûre implique la convergence en probabilité, d'où la réponse.

4°. On suppose $\mathbb{E}[X] = 0$, quitte à travailler sur les variables centrées. Montrer, grâce au théorème de Slutsky, que

$$\sqrt{n}\left(S_n^2-\sigma^2\right) \rightsquigarrow \mathcal{N}(0,\mathbb{V})$$

Nous avons vu que $V_n-\sigma^2$ converge vers 0 presque sûrement. En factorisant à partir de leur expression on voit facilement que $S_n^2 - V_n = -(\overline{X}_n-\mu)^2$. Ainsi :

$$\begin{aligned} \sqrt{n}(S_n^2-\sigma^2) &= \sqrt{n}(S_n^2-V_n) + \sqrt{n}(V_n-\sigma^2) && (1) \\ &= -\sqrt{n}(\overline{X}_n-\mu)^2 + \sqrt{n}(V_n-\sigma^2) && (2) \\ &= -\sqrt{n}(\overline{X}_n-\mu)(\overline{X}_n-\mu) + \sqrt{n}(V_n-\sigma^2) && (3) \end{aligned}$$

$(\overline{X}_n-\mu)$ tend presque-sûrement vers 0, donc également en probabilité. $\sqrt{n}(\overline{X}_n-\mu)$ converge en loi vers une loi normale centrée (d'après le TLC). D'après le lemme de Slutsky, le produit des deux termes converge en loi vers le produit des limites, à savoir 0. Cette convergence en loi a lieu vers une constante, elle est donc équivalente à une convergence en probabilité.

En appliquant à nouveau le lemme de Slutsky à la somme des deux termes de (3), comme $\sqrt{n}(S_n^2-V_n)$ converge vers 0 en probabilité et $\sqrt{n}(V_n-\sigma^2)$ converge vers une loi normale, la somme des deux converge en loi vers la limite de $\sqrt{n}(V_n-\sigma^2)$ qui est une loi $N(0,\mathbb{V})$.

5°. En travaillant sur le TCL appliqué au couple aléatoire $(\overline{X}_n,\overline{X_n^2})$ et en utilisant la delta-méthode, retrouver le résultat précédent.

Les vecteurs $(X_i,X_i^2)^T$ sont des vaiid de carré intégrable (car les X_i ont des moments d'ordre 4 par hypothèse). On peut donc appliquer la LFGN et le théorème de la limite centrale (en dimension 2) à leur moyenne empirique $(\overline{X}_n,\overline{X_n^2})^T$:

$$\left(\begin{pmatrix} \overline{X}_n \\ \overline{X_n^2} \end{pmatrix} - \begin{pmatrix} \mu \\ \mu^2+\sigma^2 \end{pmatrix}\right) \rightsquigarrow \mathcal{N}((),\mathbb{C})$$

où $\mathbb{C}$ est la matrice de covariance du couple (X_i,X_i^2). Posons maintenant $\psi(x,y) = y - x^2$. Il s'agit d'une fonction polynomiale, donc en particulier de classe C^1 dans $\mathbb{R}^2$. La valeur absolue de son déterminant jacobien est $2|x|$ et l'on peut appliquer la méthode delta à cette fonction. On a alors $S_n^2 = \overline{X_n^2} - \overline{X}_n^2 = \psi(\overline{X}_n,\overline{X_n^2})$ et

$$\left(\psi(\overline{X}_n,\overline{X_n^2}) - \psi\left(\mu,\mu^2+\sigma^2\right)\right) \rightsquigarrow \mathcal{N}\left(O,J\mathbb{C}J^T\right)$$

où J est la matrice jacobienne de ψ. Nous n'avons même pas à calculer ce produit matriciel puisque l'on veut juste retrouver le résultat de la question 4° : la limite est une variable gaussienne univariée de moyenne $\psi(\mu,\mu^2+\sigma^2) = \mu^2+\sigma^2-\mu^2 = \sigma^2$. Dans le cours, vous avez une autre version corrigée de cette question.

6°. Construire un intervalle de confiance à 95% pour σ^2.

Remarques

• Il est important de bien voir que l'on ne peut pas utiliser directement la LFGN ou le TLC à S_n^2 car les variables $(X_i - \overline{X})^2$ ne sont pas indépendantes.

• Bien se souvenir des propriétés de la convergence presque sûre : elle est stable par toutes les opérations usuelles et par application d'une fonction continue.

• On rappelle que le lemme de Slutsky assure que si $(X_n)_n$ et $(Y_n)_n$ sont deux suites de va convergeant respectivement en loi vers X et c où c est une constante, alors $(X_n + Y_n) \rightsquigarrow (X + c)$, $X_nY_n \rightsquigarrow cX$.

• Enfin, la converge en loi vers une constante est équivalente à la convergence en probabilité vers une constante.