X2016 – MAP 311
PC 8 – 19 juin 2017 – Intervalles de confiance

Igor Kortchemski – igor.kortchemski@cmap.polytechnique.fr

Corrigé des exercices non traités sur http://www.normalesup.org/~kortchem/MAP311 *un peu après la PC.*

1 Intervalles de confiance

Exercice 1. **(Monte Carlo et intervalle de confiance exact pour π)** Soit $((X_i, Y_i))_{i\geq 1}$ une suite de vecteurs aléatoires indépendants tels que pour tout $i \geq 1$, X_i et Y_i sont des variables aléatoires indépendantes de loi uniforme sur $[0,1]$. Posons $Z_i = 1$ si $X_i^2 + Y_i^2 \leq 1$ et $Z_i = 0$ sinon. Finalement, on pose

$$S_n = \frac{4}{n} \cdot \sum_{i=1}^{n} Z_i,$$

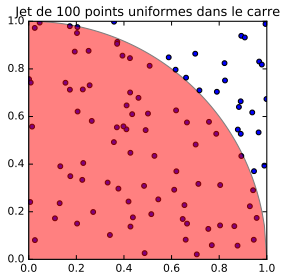

FIGURE 1 – Dans ce exemple, $S_{100} = \frac{4}{100} \cdot 77$ (il y a 77 points dans la région pleine).

(1) Identifier la loi de Z_1.

(2) Montrer que S_n converge presque sûrement vers π.

(3) Montrer que $\mathbb{P}(|S_n - \pi| \geq x) \leq \frac{4}{x^2 n}$ pour tout $x > 0$.
Indication. On pourra utiliser l'inégalité de Bienaymé–Tchebychev.

(4) Expliquer comment simuler un intervalle de confiance de π d'amplitude au plus 10^{-2} à 99%

Exercice 2. **(Intervalles de confiance asymptotiques avec le TCL)** Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires indépendantes et de même loi. On suppose que X_1 est de carré intégrable, de moyenne m et de variance $\sigma^2 > 0$. On pose

$$\widehat{m}_n = \frac{X_1 + \cdots + X_n}{n}, \qquad \widehat{\sigma}_n^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \widehat{m}_n)^2.$$

Pour des questions, demande d'explications etc., n'hésitez pas à m'envoyer un mail.

(1) Justifier que $\sqrt{n} \cdot \frac{\widehat{m}_n - m}{\sigma}$ converge en loi vers $\mathcal{N}(0,1)$, une gaussienne centrée réduite.

(2) En déduire un intervalle de confiance asymptotique pour m au niveau 95% (en supposant σ connu).

(3) Montrer que $\widehat{\sigma}_n$ converge presque sûrement vers σ. L'estimateur $\widehat{\sigma}_n^2$ est-il sans biais ?
Indication. On pourra démontrer que $\frac{n-1}{n} \cdot \widehat{\sigma}_n^2 = \left(\frac{1}{n}\sum_{k=1}^n X_k^2\right) - \widehat{m}_n^2$.

(4) Montrer que
$$\sqrt{n} \cdot \frac{\widehat{m}_n - m}{\widehat{\sigma}_n} \quad \xrightarrow[n\to\infty]{\text{loi}} \quad \mathcal{N}(0,1).$$

(5) En déduire un intervalle de confiance asymptotique pour m au niveau 95% (en supposant σ inconnu).

Exercice 3. **(Intervalles de confiance asymptotiques)** Le but de cet exercice est d'estimer le temps d'attente du RER B, qui suit une loi exponentielle de paramètre $\lambda > 0$ inconnu. Soit $(E_i)_{i\geq 1}$ une suite de variables aléatoires indépendantes de même loi exponentielle de paramètre $\lambda > 0$ inconnu. On pose $\widehat{\lambda}_n = \frac{E_1 + \cdots + E_n}{n}$ et

$$\widehat{\sigma}_n^2 = \frac{1}{n-1}\sum_{k=1}^n (E_k - \widehat{\lambda}_n)^2.$$

(1) Donner un intervalle de confiance asymptotique pour $\frac{1}{\lambda}$ au niveau 95%.

(2) Comment obtenir un intervalle de confiance asymptotique pour λ au niveau 95% ?

2 À chercher pour la prochaine fois

Exercice 4. **(Référendum)** Avant un référendum, on effectue une enquête pour estimer la proportion p de personnes votant oui. On interroge un échantillon représentatif de n personnes et on note $\widehat{F}_n$ la proportion du nombre de réponses oui dans l'échantillon.

(1) Quelle est la loi de $n\widehat{F}_n$ (en faisant des hypothèses raisonnables) ?

(2) Démontrer que
$$\frac{\sqrt{n}}{\sqrt{\widehat{F}_n(1-\widehat{F}_n)}}\left(\widehat{F}_n - p\right) \quad \xrightarrow[n\to\infty]{\text{loi}} \quad \mathcal{N}(0,1).$$
En déduire un intervalle de confiance asymptotique $\widehat{I}_n$ à 95% pour p.

(3) En utilisant l'inégalité de Bienaymé-Tchebychev, obtenir un intervalle de confiance exact $\widehat{J}_n$ à 95%.

(4) On rappelle l'inégalité de concentration (vue en Amphi 3) pour la loi binomiale :
$$\mathbb{P}\left(|\mathsf{Binom}(n,p) - np| > r\right) \leq 2\exp\left(-2\frac{r^2}{n}\right).$$
Utiliser cette inégalité pour obtenir un intervalle de confiance exact $\widehat{K}_n$ à 95%.

(5) Quel intervalle choisiriez-vous ?

3 Plus appliqué

Exercice 5. **(Enquête)** On effectue une enquête, durant une épidémie de grippe, dans le but de connaître la proportion p de personnes présentant ensuite des complications graves. On observe un échantillon représentatif de 400 personnes et pour un tel échantillon 40 personnes ont présenté des complications.

(1) Donner un intervalle de confiance pour p au risque 5%.

(2) On désire que la valeur estimée $\widehat{p}$ diffère de la proportion inconnue exacte p de moins de 0.005 avec une probabilité égale à 95%. Quel sera l'effectif d'un tel échantillon ?

(3) Quel devrait être le risque pour obtenir le même intervalle qu'à la question précédente en conservant l'effectif $n = 400$? Quelle conclusion peut-on en tirer ?

4 Pour aller plus loin

Exercice 6. **(Stabilisation de la variance)** On dispose d'un échantillon $X_1, \dots, X_n$ de variables aléatoires indépendantes de même loi de Bernoulli de paramètre $0 < \theta < 1$.

(1) On note $\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$ la moyenne empirique des X_i. Que donne la loi forte des grands et le TCL ?

(2) Trouver une fonction g telle que $\sqrt{n}(g(\overline{X}_n) - g(\theta))$ converge en loi vers une loi loi gaussienne centrée réduite.

(3) On note z_α le quantile d'ordre $1 - \alpha/2$ de la loi normale ($\mathbb{P}(Z \geq z_\alpha) = \alpha/2$ si Z est une loi normale centrée réduite). En déduire un intervalle de confiance asymptotique $\widehat{I}_{n,\alpha}$ (qui dépend de z_α, n et $\overline{X}_n$) tel que $\lim_{n\to\infty} \mathbb{P}\left(\theta \in \widehat{I}_{n,\alpha}\right) = 1 - \alpha$.

Rappel (théorème de Cochran, extension de la proposition 7.2.4 du poly). Soit X un vecteur colonne aléatoire de $\mathbb{R}^n$ de loi $\mathcal{N}(m, \sigma^2 I_n)$ (avec $m \in \mathbb{R}^n$, $\sigma > 0$) et $\mathbb{R}^n = E_1 \oplus \cdots \oplus E_p$ une décomposition de $\mathbb{R}^n$ en somme directe de p sous-espaces vectoriels orthogonaux de dimensions $d_1, \dots, d_p$ avec $d_1 + \cdots + d_p = n$. Soit $\mathbf{P}_k$ la matrice du projecteur orthogonal sur E_k et $Y_k = \mathbf{P}_k X$ la projection orthogonale de X sur E_k. Alors :

(1) les vecteurs aléatoires $(Y_1, \dots, Y_p)$ sont indépendants et Y_k suit la loi $\mathcal{N}(\mathbf{P}_k m, \sigma^2 \mathbf{P}_k)$;

(2) les variables aléatoires réelles $(\|Y_i - \mathbf{P}_i m\|^2)_{1 \leq i \leq p}$ sont indépendantes et $\|Y_k - \mathbf{P}_k m\|^2/\sigma^2$ suit la loi $\chi^2(d_k)$.

Exercice 7. **(Étalonnage)** On considère que la réponse d'un appareil de mesure à un signal déterministe ξ est égale à $a\xi$ plus un bruit gaussien centré de variance b, où $(a,b) \in \mathbb{R} \times \mathbb{R}_+^*$. On se propose d'étalonner l'appareil (c'est-à-dire estimer les valeurs de a et b) en envoyant une suite $x = (x_1, x_2, \dots, x_n)$ de signaux connus. On note $Y_i = ax_i + \sqrt{b}U_i$ la réponse au i-ième signal où on suppose que les coordonnées du vecteur $U = (U_1, U_2, \dots, U_n)$ sont des variables aléatoires gaussiennes centrées réduites indépendantes. On note $Y = (Y_1, Y_2, \dots, Y_n)$,

$$\widehat{A}_n = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2} \qquad \text{et} \qquad \widehat{B}_n = \frac{\sum_{i=1}^n (Y_i - x_i \widehat{A}_n)^2}{n-1}.$$

(1) Donner la loi de $\widehat{A}_n$. À quelle condition sur la suite $(x_i)_{i\geq 1}$ a-t-on $\mathbb{E}\left[(\widehat{A}_n - a)^2\right] \to 0$ lorsque $n \to \infty$?

On complète $e_1 = \frac{x}{\|x\|}$ en une base orthonormée $(e_1, \dots, e_n)$ de $\mathbb{R}^n$. Notons P la projection orthogonale sur $E_1 = \mathrm{Vect}(e_1)$ et Q la projection orthogonale sur $E_2 = \mathrm{Vect}(e_2, \dots, e_n)$.

(2) Déterminer PY et QY. En déduire que $\widehat{A}_n$ et $\widehat{B}_n$ sont des variables aléatoires indépendantes. Donner l'espérance et la variance de $\widehat{B}_n$.

(3) Montrer qu'il existe une constante c (dépendant de x) telle que la variable aléatoire $c\frac{\widehat{A}_n - a}{\sqrt{\widehat{B}_n}}$ suive une loi de Student.

(4) Donner un intervalle de confiance à 95% pour le paramètre a, suivant que l'on connaît la valeur de b ou non.