X2016 – MAP 311

PC 8 – 19 juin 2017 – Intervalles de confiance

Corrigé des questions non abordées en PC

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

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

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 ?

Corrigé : Commençons par une inégalité qui sera utile dans toutes les questions. Notons $\widehat{P}_n$ la proportion observée de personnes présentant des complications dans un échantillon de n personnes. D'après l'inégalité de Bienamyé-Tchebychev, on a

$$\mathbb{P}\left(|\widehat{P}_n - p| > r\right) \leq \frac{p(1-p)}{r^2 n} \leq \frac{1}{4r^2 n}$$

pour tout $r > 0$.

(1) En prenant r tel que $\frac{1}{4r^2 n} = 0.05$, on trouve $r \simeq \frac{2.24}{\sqrt{n}}$, et donc, avec $n = 400$ et $\widehat{P}_n = 0.1$, l'intervalle de confiance obtenu est

$$\left[0.1 - \frac{2.24}{\sqrt{400}}, 0.1 + \frac{2.24}{\sqrt{400}}\right] = [-0.01, 0.21].$$

(2) Dans ce cas, on veut $r = 0.005$ et $\frac{1}{4r^2 n} = 0.05$. On trouve $n = 20 \cdot 10^6$.

(3) Dans ce cas, on a $r = 0.005$, $n = 400$ et on trouve que $\frac{1}{4r^2 n} = 25 > 1$. L'inégalité de Bienaymé-Tchebychev donne donc une inégalité triviale, et on ne peut donc pas estimer le risque avec un effectif $n = 400$ un intervalle de demi-largeur 0.005 en utilisant l'inégalité de Bienaymé-Tchebychev.

□

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.

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

(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 $\lim_{n\to\infty} \mathbb{P}\left(\theta \in \widehat{I}_{n,\alpha}\right) = 1-\alpha$.

Corrigé :

(1) La loi forte des grands nombres nous dit que $\overline{X}_n$ converge presque sûrement vers $\mathbb{E}[X_1] = \theta$. Le théorème central limite nous dit que $\sqrt{n}(\overline{X}_n - \theta)$ converge en loi vers une loi normale $\mathcal{N}(0, \theta(1-\theta))$.

(2) On essaye d'appliquer la méthode delta : si g est une fonction dérivable en θ telle que $g'(\theta) \neq 0$, alors

$$\sqrt{n}\left(g(\overline{X}_n) - g(\theta)\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \theta(1-\theta)g'(\theta)^2).$$

On cherche donc g telle que $\theta(1-\theta)g'(\theta)^2 = 1$, ou encore $g'(\theta) = \frac{1}{\sqrt{\theta(1-\theta)}}$ pour tout $\theta \in (0,1)$. On trouve ainsi (à constante additive près) $g(\theta) = 2\arcsin(\sqrt{\theta})$.

(3) On a

$$\mathbb{P}\left(\theta \in \widehat{I}_{n,\alpha}\right) = \mathbb{P}\left(\left|\sqrt{n}(g(\overline{X}_n) - g(\theta))\right| \leq z_\alpha\right) \xrightarrow[n\to\infty]{} \mathbb{P}(-z_\alpha \leq Z \leq z_\alpha) = 1-\alpha$$

avec Z une loi normale centrée réduite. Ainsi, on prend

$$\widehat{I}_{n,\alpha} = \left[\sin\left(\arcsin((\overline{X}_n)^{1/2}) - \frac{z_\alpha}{2\sqrt{n}}\right)^2, \sin\left(\arcsin((\overline{X}_n)^{1/2}) + \frac{z_\alpha}{2\sqrt{n}}\right)^2\right].$$

□

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 $d_1, \ldots, d_p$ 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 $(Y_1, \ldots, Y_p)$ 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 $x = (x_1, x_2, \ldots, x_n)$ 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 $U = (U_1, U_2, \ldots, U_n)$ sont des variables aléatoires gaussiennes centrées réduites indépendantes. On note $Y = (Y_1, Y_2, \ldots, Y_n)$,

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

(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 $(e_1, \ldots, e_n)$ de $\mathbb{R}^n$. Notons P la projection orthogonale sur $E_1 = \mathrm{Vect}(e_1)$ et Q la projection orthogonale sur $E_2 = \mathrm{Vect}(e_2, \ldots, e_n)$.

(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.

Corrigé :

(1) Le vecteur U étant constitué de variables aléatoires gaussiennes indépendantes, c'est un vecteur gaussien. Donc $\widehat{A}_n$ est une gaussienne, et il reste à déterminer son espérance et sa variance :

$$\mathbb{E}\left[\widehat{A}_n\right] = \frac{\sum_{i=1}^n x_i \mathbb{E}[Y_i]}{\sum_{i=1}^n x_i^2} = \frac{\sum_{i=1}^n a x_i^2}{\sum_{i=1}^n x_i^2} = a, \qquad \mathrm{Var}(\widehat{A}_n) = \frac{1}{\left(\sum_{i=1}^n x_i^2\right)^2} \sum_{i=1}^n x_i^2 \mathrm{Var}(Y_i) = \frac{b}{\|x\|}.$$

Comme $\mathbb{E}\left[(\widehat{A}_n - a)^2\right] = \mathrm{Var}(\widehat{A}_n)$, on voit que $\mathbb{E}\left[(\widehat{A}_n - a)^2\right] \to 0$ si et seulement si $\|x\| \to \infty$ lorsque $n \to \infty$.

(2) Tout d'abord, on remarque que $\widehat{A}_n = \frac{1}{\|x\|^2}\langle Y, x\rangle$.

Alors

$$PY = \langle Y, e_1\rangle e_1 = \langle Y, \frac{x}{\|x\|}\rangle e_1 = \widehat{A}_n x, \qquad QY = Y - PY = Y - \widehat{A}_n x$$

et on remarque d'une part que le vecteur gaussien QY est centré et que $(n-1)\widehat{B}_n = \|QY\|^2$.

D'après le théorème de Cochran, d'une part PY et QY sont indépendants et donc $\widehat{A}_n$ et $\widehat{B}_n$ sont indépendants, et d'autre part d'après le théorème de Cochran $\|QY\|^2/b$ suit une loi $\chi^2(n-1)$. En particulier, en écrivant $\widehat{B}_n = \frac{b}{n-1} \cdot \frac{\|QY\|^2}{b}$ et en utilisant le fait qu'une loi du χ^2 à k degrés de liberté a pour espérance k et variance $2k$,

$$\mathbb{E}\left[\widehat{B}_n\right] = \frac{b}{n-1}\mathbb{E}\left[\frac{\|QY\|^2}{b}\right] = b, \qquad \mathrm{Var}(\widehat{B}_n) = \frac{b^2}{(n-1)^2}\mathrm{Var}\left(\frac{\|QY\|^2}{b}\right) = \frac{2b^2}{n-1}.$$

(3) Comme $\widehat{A}_n - a$ suit une loi $\mathcal{N}(0, \frac{b}{\|x\|})$, on en déduit que

$$\sqrt{\frac{\|x\|}{n-1}} \cdot \frac{\widehat{A}_n - a}{\sqrt{\widehat{B}_n}} = \frac{\sqrt{\frac{\|x\|}{b}}\left(\widehat{A}_n - a\right)}{\sqrt{\frac{(n-1)\widehat{B}_n}{b}}}$$

suit une loi de Student à $n-1$ degrés de liberté.

(4) Si le paramètre b est connu, comme $\widehat{A}_n - a$ suit une loi $\mathcal{N}(0, \frac{b}{\|x\|})$, en notant $z_{\alpha/2}$ le quantile le quantile d'ordre $1-\alpha/2$ de la loi normale ($\mathbb{P}(Z \geq z_{\alpha/2}) = \alpha/2$ si Z est une loi normale centrée réduite), en prenant $\alpha = 5\%$, un intervalle de confiance pour a au niveau de (exactement) 95% est

$$\left[\widehat{A}_n - z_{\alpha/2}\sqrt{\frac{b}{\|x\|}}, \widehat{A}_n + z_{\alpha/2}\sqrt{\frac{b}{\|x\|}}\right].$$

Si le paramètre b est inconnu, notons $t_{n-1,1-\alpha/2}$ le quantile d'orde $1-\alpha/2$ de la loi de Student à $n-1$ degrés de liberté avec $\alpha = 5\%$. Alors un intervalle de confiance pour a au niveau de (exactement) 95% est

$$\left[\widehat{A}_n - t_{n-1,1-\alpha/2}\sqrt{\frac{(n-1)\widehat{B}_n}{\|x\|}}, \widehat{A}_n + t_{n-1,1-\alpha/2}\sqrt{\frac{(n-1)\widehat{B}_n}{\|x\|}}\right].$$

□