Petite Classe 8 - MAP 311 - Groupes 1 et 12

Josselin Garnier

19 juin 2017 - Amphi Curie

Exercice 1. (Suite et fin de l'exercice 3 de la PC 7) On observe un échantillon $(X_1, \dots, X_n)$ de n variables aléatoires i.i.d. de loi exponentielle de paramètre $\theta > 0$ inconnu.

1. Déterminer l'estimateur du maximum de vraisemblance $\widehat{\theta}_n = \theta_n(X_1, \dots, X_n)$ de θ et montrer qu'il converge presque sûrement vers θ lorsque $n \to \infty$.
2. Cet estimateur est-il sans biais ?
   Indication. On rappelle que la densité de la loi $\Gamma(a, \lambda)$ est $\frac{1}{\Gamma(a)}\lambda^a x^{a-1} e^{-\lambda x} \mathbf{1}_{x>0}$.
3. Démontrer que $\sqrt{n}(\widehat{\theta}_n - \theta)$ converge en loi vers une loi qu'on déterminera.
   Indication. On pourra utiliser la "méthode Delta".
4. Construire un intervalle de confiance asymptotique à 95% pour θ.

Solution.

3. D'après le théorème central limite,

$$\sqrt{n}\left(\frac{1}{\widehat{\theta}_n} - \frac{1}{\theta}\right) \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0, \frac{1}{\theta^2}).$$

On applique la méthode Delta avec la fonction $g(x) = 1/x$, dérivable au point $1/\theta$. Comme $g'(1/\theta)^2 = \theta^4$, on a

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0, \theta^2).$$

4. D'après le lemme de Slutsky ($\widehat{\theta}_n$ converge presque sûrement vers θ d'après la loi forte des grands nombres), on a

$$\frac{\sqrt{n}}{\widehat{\theta}_n}(\widehat{\theta}_n - \theta) \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0, 1).$$

On note $z_{1-\alpha/2}$ le quantile le quantile d'ordre $1-\alpha/2$ de la loi normale ($\mathbb{P}(Z \geq z_{1-\alpha/2}) = 1 - \mathbb{P}(Z \leq z_{1-\alpha/2}) = \alpha/2$ si Z est une loi normale centrée réduite). On a

$$\mathbb{P}\Big(\frac{\sqrt{n}}{\widehat{\theta}_n}(\widehat{\theta}_n - \theta) \in [-z_{1-\alpha/2}, z_{1-\alpha/2}]\Big) \xrightarrow[n\to\infty]{} \mathbb{P}\big(Z \in [-z_{1-\alpha/2}, z_{1-\alpha/2}]\big) = 1 - \alpha$$

et on peut aussi écrire

$$\mathbb{P}\Big(\frac{\sqrt{n}}{\widehat{\theta}_n}(\widehat{\theta}_n - \theta) \in [-z_{1-\alpha/2}, z_{1-\alpha/2}]\Big) = \mathbb{P}\Big(\theta \in \Big[\widehat{\theta}_n\big(1 - \frac{z_{1-\alpha/2}}{\sqrt{n}}\big), \widehat{\theta}_n\big(1 + \frac{z_{1-\alpha/2}}{\sqrt{n}}\big)\Big]\Big)$$

En prenant $\alpha = 5\%$, on en déduit que

$$\left[\widehat{\theta}_n\left(1 - \frac{1{,}96}{\sqrt{n}}\right), \widehat{\theta}_n\left(1 + \frac{1{,}96}{\sqrt{n}}\right)\right]$$

est un intervalle de confiance asymptotique pour θ au niveau 95% (avec $z_{1-\alpha/2} \simeq 1{,}96$).

Exercice 2. Soit $(X_k)_{k\geq 1}$ des variables aléatoires indépendantes de même loi uniforme sur $[0,\theta]$ (avec θ inconnu).

1. Montrer que l'estimateur du maximum de vraisemblance $\widehat{\theta}_n = \widehat{\theta}_n(X_1,\dots,X_n)$ de θ est $\widehat{\theta}_n = \max(X_1,\dots,X_n)$.
2. Montrer que $W_n = n\big(1-\frac{\widehat{\theta}_n}{\theta}\big)$ converge en loi quand $n\to\infty$ et déterminer sa limite.
3. Construire un intervalle de confiance asymptotique à 95% pour θ.

Solution.

1. Une densité f_θ de $(X_1,\dots,X_n)$ au point $(x_1,\dots,x_n)$ est
$$f_\theta(x_1,\dots,x_n) = \frac{1}{\theta^n}\prod_{i=1}^n \mathbf{1}_{0\leq x_i\leq\theta} = \frac{\mathbf{1}_{\max(x_1,\dots,x_n)\leq\theta}}{\theta^n},$$
qui, en θ, est bien maximale au point $\max(x_1,\dots,x_n)$.
2. Pour $x<0$, on a $\mathbb{P}(W_n\leq x)$=0. Pour $x\geq 0$, on a
$$\mathbb{P}\big(W_n\leq x\big) = \mathbb{P}\Big(\widehat{\theta}_n \geq \theta - \frac{\theta}{nx}\Big) = 1-\Big(1-\frac{1}{nx}\Big)^n \quad \underset{n\to\infty}{\longrightarrow} \quad 1-e^{-x}.$$
Donc W_n converge en loi vers une loi exponentielle $\mathcal{E}$ de paramètre 1.
3. Soit c tel que $\mathbb{P}(\mathcal{E}>c)=5\%$ (ainsi $c\simeq 3$). On a donc
$$\mathbb{P}\Big(0\leq n\big(1-\frac{\widehat{\theta}_n}{\theta}\big)\leq c\Big) \quad \underset{n\to\infty}{\longrightarrow} \quad 95\%,$$
ou encore
$$\mathbb{P}\Big(\widehat{\theta}_n \leq\theta\leq \frac{\widehat{\theta}_n}{1-\frac{c}{n}}\Big) \quad \underset{n\to\infty}{\longrightarrow} \quad 95\%.$$
Un intervalle de confiance asymptotique à 95% pour θ est donc $\Big[\widehat{\theta}_n, \frac{\widehat{\theta}_n}{1-\frac{c}{n}}\Big]$.

Exercice 3. (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
$$\overline{X}_n = \frac{X_1+\cdots+X_n}{n}, \qquad \widehat{\sigma}_n^2 = \frac{1}{n-1}\sum_{k=1}^n (X_k-\overline{X}_n)^2.$$

1. Justifier que $\sqrt{n}\frac{\overline{X}_n-m}{\sigma}$ 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 ?
4. Montrer que
$$\sqrt{n}\frac{\overline{X}_n - m}{\widehat{\sigma}_n} \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,1).$$
5. En déduire un intervalle de confiance asymptotique pour m au niveau 95% (en supposant σ inconnu).

Solution.

1. Ceci provient du théorème central limite car $\sqrt{n}\frac{\overline{X}_n-m}{\sigma} = \frac{X_1+\cdots+X_n-nm}{\sigma\sqrt{n}}$.
2. Notons $z_{1-\alpha/2}$ le quantile le quantile d'ordre $1-\alpha/2$ de la loi normale ($\mathbb{P}(Z \geq z_{1-\alpha/2}) = \alpha/2$ si Z est une loi normale centrée réduite). On a alors
$$\mathbb{P}(-z_{1-\alpha/2} \leq \sqrt{n}\frac{\overline{X}_n - m}{\sigma} \leq z_{1-\alpha/2}) \underset{n\to\infty}{\longrightarrow} \mathbb{P}(|Z| \leq z_{1-\alpha/2}) = \alpha.$$
L'intervalle
$$\widehat{I}_n = \left[\overline{X}_n - \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}}, \overline{X}_n + \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}}\right]$$
est donc un intervalle de confiance asymptotique au niveau $1-\alpha$ car
$$m \in \left[\overline{X}_n - \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}}, \overline{X}_n + \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}}\right] \iff -z_{1-\alpha/2} \leq \sqrt{n}\frac{\overline{X}_n - m}{\sigma} \leq z_{1-\alpha/2}.$$
En prenant $\alpha = 0{,}05$ et en utilisant le fait que $z_{1-\alpha/2} \simeq 1{,}96$, l'intervalle
$$\widehat{I}_n = \left[\overline{X}_n - \frac{1{,}96\sigma}{\sqrt{n}}, \overline{X}_n + \frac{1{,}96\sigma}{\sqrt{n}}\right]$$
est donc un intervalle de confiance asymptotique au niveau 95%.
3. Tout d'abord, on a
$$\frac{n-1}{n}\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k^2 - 2\overline{X}_n X_k + \overline{X}_n^2) = \left(\frac{1}{n}\sum_{k=1}^{n} X_k^2\right) - 2\overline{X}_n^2 + \overline{X}_n^2 = \left(\frac{1}{n}\sum_{k=1}^{n} X_k^2\right) - \overline{X}_n^2.$$
D'après la loi forte des grands nombres, $\frac{1}{n}\sum_{k=1}^{n} X_k^2$ converge presque sûrement vers $\mathbb{E}(X_1^2) = \sigma^2 + m^2$ et $\overline{X}_n$ converge presque sûrement vers m. Donc $\widehat{\sigma}_n^2$ converge presque sûrement vers σ^2. Donc $\widehat{\sigma}_n$ converge presque sûrement vers σ.
Pour étudier le biais de $\widehat{\sigma}_n^2$, calculons d'abord $\mathbb{E}(\overline{X}_n^2)$ en utilisant l'indépendance des variables aléatoires:
$$\mathbb{E}(\overline{X}_n^2) = \mathrm{Var}(\overline{X}_n) + \mathbb{E}(\overline{X}_n)^2 = \frac{n\sigma^2}{n^2} + m^2 = \frac{\sigma^2}{n} + m^2.$$

On en déduit que

$$\mathbb{E}(\frac{n-1}{n}\widehat{\sigma}_n^2) = \sigma^2 + m^2 - \left(m^2 + \frac{\sigma^2}{n}\right) = \frac{n-1}{n}\sigma^2.$$

Donc $\mathbb{E}(\widehat{\sigma}_n^2) = \sigma^2$ et $\widehat{\sigma}_n^2$ est un estimateur sans biais de la variance.

4. Comme $\widehat{\sigma}_n$ converge en probabilité vers σ, d'après le théorème de Slutsky, la convergence

$$\left(\sqrt{n}(\overline{X}_n - m), \widehat{\sigma}_n\right) \underset{n\to\infty}{\longrightarrow} (\sigma\mathcal{N}(0,1), \sigma)$$

a lieu conjointement en loi. En composant ceci par la fonction $f(x,y) = x/y$, continue sur $\mathbb{R} \times \mathbb{R}^*$, on obtient le résultat voulu.

5. Comme dans la première question, l'intervalle

$$\widehat{I}_n = \left[\overline{X}_n - \frac{\widehat{\sigma}_n z_{1-\alpha/2}}{\sqrt{n}}, \overline{X}_n + \frac{\widehat{\sigma}_n z_{1-\alpha/2}}{\sqrt{n}}\right]$$

est donc un intervalle de confiance asymptotique au niveau $1-\alpha$ et l'intervalle

$$\widehat{I}_n = \left[\overline{X}_n - \frac{1{,}96\widehat{\sigma}_n}{\sqrt{n}}, \overline{X}_n + \frac{1{,}96\widehat{\sigma}_n}{\sqrt{n}}\right]$$

est donc un intervalle de confiance asymptotique au niveau 95%.

Exercice 4. 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% ?

Solution.

1. Comme $\mathbb{E}(E_1) = \frac{1}{\lambda}$, nous pouvons appliquer l'exercice précédent:

$$\widehat{I}_n = \left[\widehat{\lambda}_n - \frac{1{,}96\widehat{\sigma}_n}{\sqrt{n}}, \widehat{\lambda}_n + \frac{1{,}96\widehat{\sigma}_n}{\sqrt{n}}\right]$$

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

2. Dire que $\frac{1}{\lambda} \in [A, B]$ est équivalent au fait que $\frac{1}{B} \leq \lambda \leq \frac{1}{A}$. On en déduit que

$$\widehat{J}_n = \left[\frac{1}{\widehat{\lambda}_n + \frac{1{,}96\widehat{\sigma}_n}{\sqrt{n}}}, \frac{1}{\widehat{\lambda}_n - \frac{1{,}96\widehat{\sigma}_n}{\sqrt{n}}} \right]$$

est un intervalle de confiance asymptotique pour λ au niveau 95%.

Alternativement, on aurait pu appliquer la méthode Delta : en notant $\sigma^2 = \frac{1}{\lambda^2}$ la variance de E_1, en appliquant la méthode Delta avec la fonction $f(x) = 1/x$ (dérivable en $1/\lambda$ avec $f'(1/\lambda) = -\lambda^2 \neq 0$),

$$\sqrt{n} \left(f(\widehat{\lambda}_n) - f(1/\lambda) \right) \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0, (f'(1/\lambda))^2 \sigma^2)$$

Ainsi, $\sqrt{n} \left(\frac{1}{\widehat{\lambda}_n} - \lambda \right)$ converge en loi vers $\mathcal{N}(0, (f'(1/\lambda))^2\sigma^2) = \mathcal{N}(0, \lambda^2) = \mathcal{N}(0, 1/\sigma^2)$. Ainsi,

$$\sigma\sqrt{n} \left(\frac{1}{\widehat{\lambda}_n} - \lambda \right) \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0, 1).$$

En appliquant le lemme de Slutsky, on obtient que

$$\widehat{\sigma}_n\sqrt{n} \left(\frac{1}{\widehat{\lambda}_n} - \lambda \right) \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0, 1).$$

On en déduit que

$$\left[\frac{1}{\widehat{\lambda}_n} - \frac{1{,}96}{\widehat{\sigma}_n\sqrt{n}}, \frac{1}{\widehat{\lambda}_n} + \frac{1{,}96}{\widehat{\sigma}_n\sqrt{n}} \right]$$

est un intervalle de confiance asymptotique pour λ au niveau 95%.

Exercice 5. (Stabilisation de la variance) On dispose d'un échantillon $X_1, \ldots, X_n$ 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 donnent la loi forte des grands nombres et le TCL?
2. Trouver une fonction g telle que $\sqrt{n}\big(g(\overline{X}_n) - g(\theta)\big)$ converge en loi vers une loi gaussienne centrée réduite.
3. On note $z_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi normale ($\mathbb{P}(Z \geq z_{1-\alpha/2}) = \alpha/2$ si Z est une loi normale centrée réduite). En déduire un intervalle de confiance asymptotique $\widehat{I}_{n,\alpha}$ (qui dépend de $z_{1-\alpha/2}$, n et $\overline{X}_n$) tel que $\lim_{n\to\infty} \mathbb{P}(\theta \in \widehat{I}_{n,\alpha}) = 1 - \alpha$.

Solution.

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 on a la convergence en loi suivante

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

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

$$\begin{aligned} \mathbb{P}(\theta \in \widehat{I}_{n,\alpha}) &= \mathbb{P}\Big(\left|\sqrt{n}(g(\overline{X}_n) - g(\theta))\right| \leq z_{1-\alpha/2}\Big) \\ &\overset{n\to\infty}{\longrightarrow} \mathbb{P}(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1-\alpha \end{aligned}$$

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

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

Exercice 6. Soient $(X_i)_{1\leq i\leq n}$ des variables aléatoires réelles indépendantes de même loi, de carré intégrable, d'espérance m et de variance σ^2. On suppose que la moyenne empirique $\overline{X}_n$ et la variance empirique V_n, définies par

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i, \qquad V_n = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2,$$

sont indépendantes. Le but de cet exercice est de démontrer que la loi de X_i est une loi gaussienne $\mathcal{N}(m, \sigma^2)$.

1. On suppose dans les questions 1 à 4 que $m = 0$. Calculer $\mathbb{E}((n-1)V_n)$ en fonction de σ^2. Montrer que pour tout $t \in \mathbb{R}$:

$$\mathbb{E}\big((n-1)V_n e^{itn\overline{X}_n}\big) = (n-1)\psi(t)^n\sigma^2,$$

où on note $\psi(t)$ la fonction caractéristique de X_1.

2. En développant V_n dans l'égalité précédente, vérifier que :

$$\mathbb{E}\big((n-1)V_n e^{itn\overline{X}_n}\big) = -(n-1)\psi''(t)\psi(t)^{n-1} + (n-1)\psi'(t)^2\psi(t)^{n-2}.$$

3. En déduire que, sur un voisinage ouvert de 0, ψ est solution de l'équation différentielle :

$$\frac{\psi''}{\psi} - \left(\frac{\psi'}{\psi}\right)^2 = -\sigma^2, \qquad \psi(0) = 1, \qquad \psi'(0) = 0.$$

4. En déduire que la loi des variables X_i est la loi gaussienne $\mathcal{N}(0, \sigma^2)$.
Indication : Considérer $\phi = \psi'/\psi$.

5. Que peut-on dire si l'on ne suppose plus $m = 0$?

Solution.

1. On a

$$\mathbb{E}((n-1)V_n) = \mathbb{E}\Big(\sum_{j=1}^n \Big(X_j^2 - 2\overline{X}_n X_j + \overline{X}_n^2\Big)\Big) = n\sigma^2 - 2n\frac{\sigma^2}{n} + n^2\frac{\sigma^2}{n^2} = (n-1)\sigma^2.$$

   Comme V_n et $\overline{X}_n$ sont indépendants, on a :

$$\mathbb{E}\big((n-1)V_n e^{itn\overline{X}_n}\big) = \mathbb{E}((n-1)V_n)\mathbb{E}\big(e^{itn\overline{X}_n}\big) = (n-1)\sigma^2\psi(t)^n.$$

2. En développant, il vient :

$$\begin{aligned}
\mathbb{E}\big((n-1)V_n e^{itn\overline{X}_n}\big) &= \mathbb{E}\Big(\Big((1-\frac{1}{n})\sum_{j=1}^n X_j^2 - \frac{2}{n}\sum_{1\le j<k\le n} X_j X_k\Big)e^{it\sum_{j=1}^n X_j}\Big) \\
&= (1-\frac{1}{n})\sum_{j=1}^n \mathbb{E}(X_j^2 e^{itX_j})\mathbb{E}(e^{itX_1})^{n-1} \\
&\qquad -\frac{2}{n}\sum_{1\le j<k\le n}\mathbb{E}(X_j e^{itX_j})\mathbb{E}(X_k e^{itX_k})\mathbb{E}(e^{itX_1})^{n-2} \\
&= -(1-\frac{1}{n})n\psi''(t)\psi(t)^{n-1} - \frac{2}{n}\cdot\frac{n(n-1)}{2}(-i\psi'(t))^2\psi(t)^{n-2} \\
&= -(n-1)\psi''(t)\psi(t)^{n-1} + (n-1)\psi'(t)^2\psi(t)^{n-2}.
\end{aligned}$$

3. Soit O la composante connexe de l'ouvert $\{t \in \mathbb{R} : \psi(t) \neq 0\}$ contenant 0. On en déduit de la question précédente que ψ est solution de l'équation différentielle

$$\frac{\psi''}{\psi} - \left(\frac{\psi'}{\psi}\right)^2 = -\sigma^2 \text{ sur } O, \qquad \psi(0) = 1, \qquad \psi'(0) = 0.$$

4. On pose $\phi = \psi'/\psi$ sur O. On obtient que $\phi'(t) = -\sigma^2$ et $\phi(0) = 0$. On en déduit que $\phi(t) = -\sigma^2 t$. Donc $\psi'(t) = -\sigma^2 t\psi(t)$. Une solution de cette équation sur O est $\psi(t) = e^{-t^2\sigma^2/2}$. On cherche alors les solutions sous la forme $\psi(t) = e^{-t^2\sigma^2/2}h(t)$ (méthode de la variation de la constante). On en déduit que $h' = 0$. La condition $\psi(0) = 1$ implique que $e^{-t^2\sigma^2/2}$ est la seule solution de l'équation différentielle considérée et $O = \mathbb{R}$. La loi de X_i est donc la loi gaussienne $\mathcal{N}(0, \sigma^2)$.

5. Si $m \neq 0$, on applique ce qui précède à $X_i^* = X_i - m$. On trouve alors que la loi de X_i est la loi gaussienne $\mathcal{N}(m, \sigma^2)$.

Exercice 7. 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 ?

Solution. 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 Bienaymé-Chebychev, on a

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

pour tout $r > 0$.

1. En prenant r tel que $\frac{1}{4r^2n} = 0{,}05$, 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 $\frac{1}{4r^2n} = 0{,}95$. On trouve $n = 10527$.
3. Dans ce cas, on a $r = 0{,}005$, $n = 400$ et on trouve que $\frac{1}{4r^2n} = 25 > 1$. L'inégalité de Bienaymé-Chebychev 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-Chebychev.

Pour aller plus loin

Rappel (théorème de Cochran, extension de la proposition 6.2.8 du poly). Soit $\boldsymbol{X}$ un vecteur colonne aléatoire de $\mathbb{R}^n$ de loi $\mathcal{N}(\boldsymbol{m}, \sigma^2 \mathbf{I}_n)$ (avec $\boldsymbol{m} \in \mathbb{R}^n$, $\sigma > 0$) 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 $\boldsymbol{Y}_k = \mathbf{P}_k \boldsymbol{X}$ la projection orthogonale de $\boldsymbol{X}$ sur E_k. Alors:

1. Les vecteurs aléatoires $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_p)$ sont indépendants et $\boldsymbol{Y}_k$ suit la loi $\mathcal{N}(\mathbf{P}_k \boldsymbol{m}, \sigma^2 \mathbf{P}_k)$.
2. Les variables aléatoires réelles $(\|\boldsymbol{Y}_i - \mathbf{P}_i \boldsymbol{m}\|^2)_{1 \leq i \leq p}$ sont indépendantes et $\|\boldsymbol{Y}_k - \mathbf{P}_k \boldsymbol{m}\|^2/\sigma^2$ suit la loi $\chi^2_{d_k}$.

Exercice 8. 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 $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ de signaux connus. On note $Y_i = a x_i + \sqrt{b} U_i$ la réponse au i-ième signal où on suppose que les coordonnées du vecteur $\boldsymbol{U} = (U_1, U_2, \ldots, U_n)$ sont des variables aléatoires gaussiennes centrées réduites indépendantes. On note $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$,

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

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

On complète $\boldsymbol{e}_1 = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$ en une base orthonormée $(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n)$ de $\mathbb{R}^n$. Notons $\mathbf{P}$ la projection orthogonale sur $E_1 = \mathrm{Vect}(\boldsymbol{e}_1)$ et $\mathbf{Q}$ la projection orthogonale sur $E_2 = \mathrm{Vect}(\boldsymbol{e}_2, \dots, \boldsymbol{e}_n)$.

2. Déterminer $\mathbf{P}\boldsymbol{Y}$ et $\mathbf{Q}\boldsymbol{Y}$. 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 $\boldsymbol{x}$) 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.

Solution.

1. Le vecteur $\boldsymbol{U}$ 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:
$$\mathbb{E}(\widehat{A}_n) = \frac{\sum_{i=1}^n x_i \mathbb{E}(Y_i)}{\sum_{i=1}^n x_i^2} = \frac{\sum_{i=1}^n a x_i^2}{\sum_{i=1}^n x_i^2} = a,$$
$$\mathrm{Var}(\widehat{A}_n) = \frac{1}{\left(\sum_{i=1}^n x_i^2\right)^2} \sum_{i=1}^n x_i^2 \mathrm{Var}(Y_i) = \frac{b}{\|\boldsymbol{x}\|^2}.$$
Comme $\mathbb{E}((\widehat{A}_n - a)^2) = \mathrm{Var}(\widehat{A}_n)$, on voit que $\mathbb{E}((\widehat{A}_n - a)^2) \to 0$ si et seulement si $\|\boldsymbol{x}\| \to \infty$ lorsque $n \to \infty$.

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

   Alors
$$\mathbf{P}\boldsymbol{Y} = \langle \boldsymbol{Y}, \boldsymbol{e}_1\rangle \boldsymbol{e}_1 = \langle \boldsymbol{Y}, \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\rangle \boldsymbol{e}_1 = \widehat{A}_n \boldsymbol{x}, \qquad \mathbf{Q}\boldsymbol{Y} = \boldsymbol{Y} - \mathbf{P}\boldsymbol{Y} = \boldsymbol{Y} - \widehat{A}_n \boldsymbol{x}$$
   et on remarque d'une part que le vecteur gaussien $\mathbf{Q}\boldsymbol{Y}$ est centré et que $(n-1)\widehat{B}_n = \|\mathbf{Q}\boldsymbol{Y}\|^2$.

   D'après le théorème de Cochran, d'une part $\mathbf{P}\boldsymbol{Y}$ et $\mathbf{Q}\boldsymbol{Y}$ sont indépendants et donc $\widehat{A}_n$ et $\widehat{B}_n$ sont indépendants, et d'autre part $\|\mathbf{Q}\boldsymbol{Y}\|^2/b$ suit une loi χ^2_{n-1}. En particulier, en écrivant $\widehat{B}_n = \frac{b}{n-1}\frac{\|\mathbf{Q}\boldsymbol{Y}\|^2}{b}$ et en utilisant le fait qu'une loi du χ^2 à k degrés de liberté a pour espérance k et variance $2k$,
$$\mathbb{E}(\widehat{B}_n) = \frac{b}{n-1}\mathbb{E}\Big(\frac{\|\mathbf{Q}\boldsymbol{Y}\|^2}{b}\Big) = b, \qquad \mathrm{Var}(\widehat{B}_n) = \frac{b^2}{(n-1)^2}\mathrm{Var}\left(\frac{\|\mathbf{Q}\boldsymbol{Y}\|^2}{b}\right) = \frac{2b^2}{n-1}.$$

3. Comme $\widehat{A}_n - a$ suit une loi $\mathcal{N}(0, \frac{b}{\|\boldsymbol{x}\|^2})$, on en déduit que
$$\frac{\|\boldsymbol{x}\|}{\sqrt{n-1}} \frac{\widehat{A}_n - a}{\sqrt{\widehat{B}_n}} = \frac{\frac{\|\boldsymbol{x}\|}{\sqrt{b}}\left(\widehat{A}_n - a\right)}{\sqrt{\frac{(n-1)\widehat{B}_n}{b}}}$$
   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 $\mathcal{N}(0, \frac{b}{\|\boldsymbol{x}\|^2})$, en notant $z_{1-\alpha/2}$ le quantile le quantile d'ordre $1-\alpha/2$ de la loi normale ($\mathbb{P}(Z \geq z_{1-\alpha/2}) = \alpha/2$ si Z est une loi normale centrée réduite), en prenant $\alpha = 5\%$, un intervalle de confiance pour a au niveau de (exactement) 95% est

$$\left[\widehat{A}_n - z_{1-\alpha/2}\frac{\sqrt{b}}{\|\boldsymbol{x}\|}, \widehat{A}_n + z_{1-\alpha/2}\frac{\sqrt{b}}{\|\boldsymbol{x}\|}\right].$$

Si le paramètre b est inconnu, notons $t_{n-1,1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ 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

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