PC 9 : Intervalles de confiance

On corrigera les exercices 3, 5 et 4 en PC.

Exercice 1. (Loi normale) On observe $X_1, \ldots, X_n$ indépendantes et de même loi. On suppose qu'il existe $\theta > 0$ tel que cette loi admet la densité

$$f_\theta(x) = \frac{1}{\sqrt{2\pi\theta^2}} \exp\left(-\frac{x^2}{2\theta^2}\right).$$

Indication : $F^{-1}_{\chi^2_{10}}(0.025) \simeq 3.25$, $F^{-1}_{\chi^2_{10}}(0.975) \simeq 20.48$, $F^{-1}_{\chi^2_{10}}(0.95) \simeq 18.31$, $F_{\chi^2_{10}}(40/3) \simeq 0.79$ et $\Phi^{-1}(0.975) \simeq 2$.

1. On veut estimer $\tau = \theta^2$. Proposer un estimateur $\hat{\tau}$ de τ et étudier sa loi.
2. Construire un intervalle de confiance au niveau $1 - \alpha$ de la forme $[S_1, S_2]$ tel que
$$\mathbb{P}(\tau < S_1) = \mathbb{P}(\tau > S_2) = \alpha/2.$$
3. Donner la loi asymptotique de $\hat{\tau}$ et en déduire un intervalle de confiance asymptotique.
4. Lorsque $n = 10$, $\hat{\tau} = 2$ et $\alpha = 0.05$, comparer l'intervalle de confiance obtenu à la question 2 (non asymptotique) avec celui obtenu à la question 3 (asymptotique).

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

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

1. Justifier que $\sqrt{n} \cdot \frac{\widehat{m}_n - m}{\sigma}$ converge en loi vers limite que l'on identifiera.
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 $\widehat{\sigma}^2_n$ est-il sans biais ($\widehat{\sigma}^2_n$ est sans biais si $\mathbb{E}\left[\widehat{\sigma}^2_n\right] = \sigma^2$) ?

   Indication. On pourra démontrer que $\frac{n-1}{n} \cdot \widehat{\sigma}^2_n = \left(\frac{1}{n}\sum_{k=1}^n X_k^2\right) - \widehat{m}^2_n$.
4. Montrer que
$$\sqrt{n} \cdot \frac{\widehat{m}_n - m}{\widehat{\sigma}_n} \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,1).$$
5. En déduire un intervalle de confiance asymptotique pour m au niveau 95% (en supposant σ inconnu).

Exercice 3. Soient $X_1, \ldots, X_n$ des variables aléatoires réelles i.i.d. dont la loi admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$ donnée par

$$x \in \mathbb{R} \to f(x;\theta) = \frac{2}{\sqrt{\pi\theta}} \exp(-x^2/\theta)\,\mathbb{1}_{\{x>0\}},$$

où $\theta > 0$ est un paramètre inconnu. On observe une réalisation $(x_1, \ldots, x_n)$ du vecteur aléatoire $(X_1, \ldots, X_n)$. On désigne par α un réel donné dans $[0, 1]$ et on note $\hat{m}_2(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n X_i^2$ le moment empirique d'ordre 2.

1. Déterminer l'estimateur du maximum de vraisemblance $\hat{\theta}$ de θ.

2. Déterminer la loi de la variable $X_1/\sqrt{\theta}$. Déduire de ce résultat que la loi de la statistique $\hat{\theta}/\theta$ ne dépend pas de θ, puis donner la loi de $n\hat{\theta}/\theta$.
3. Trouver des réels a et b tels que $[\hat{\theta}/a, \hat{\theta}/b]$ soit un intervalle de confiance de niveau $1-\alpha$ pour θ.

Solution.

1. La fonction de vraisemblance dans ce modèle est donnée par

$$\theta > 0 \to \mathcal{L}(\mathbf{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \frac{2^n}{(\pi\theta)^{n/2}} \exp\left\{-\frac{1}{\theta}\sum_{i=1}^{n} x_i^2\right\}.$$

On en déduit la fonction de log-vraisemblance :

$$\begin{aligned}\ell(\theta) = \log(\mathcal{L}(\mathbf{x};\theta)) &= n\log\frac{2}{\sqrt{\pi}} - \frac{n}{2}\log\theta - \frac{1}{\theta}\sum_{i=1}^{n} x_i^2 \\ &= n\log\frac{2}{\sqrt{\pi}} - \frac{n}{2}\log\theta - \frac{n}{\theta}\hat{m}_2(\mathbf{x}).\end{aligned}$$

La première et seconde dérivée de $\ell(\theta)$ sont données par

$$\begin{aligned}\ell'(\theta) &= -\frac{n}{2\theta} + \frac{n}{\theta^2}\hat{m}_2(\mathbf{x}) \\ \ell''(\theta) &= \frac{n}{2\theta^2} - \frac{2n}{\theta^3}\hat{m}_2(\mathbf{x}) = \frac{n}{\theta^2}\left(\frac{1}{2} - \frac{2}{\theta}\hat{m}_2(\mathbf{x})\right).\end{aligned}$$

On cherche les points critiques de $\ell(\theta)$:

$$\ell'(\theta) = 0 \Longleftrightarrow -\frac{n}{2\theta} + \frac{n}{\theta^2}\hat{m}_2(\mathbf{x}) = 0 \Longleftrightarrow \theta = 2\hat{m}_2(\mathbf{x}).$$

On vérifie s'il s'agit d'un point de maximum par la dérivée seconde. En effet, on a

$$\ell''(2\hat{m}_2(\mathbf{x})) = \frac{n}{4\hat{m}_2^2(\mathbf{x})}\left(\frac{1}{2} - 1\right) < 0.$$

L'unique point critique de $\ell(\theta)$ étant un maximum, on conclut qu'il est maximum global. Donc, l'EMV est donné par $\hat{\theta} = 2\hat{m}_2(\mathbf{x})$.

2. Notons $Y = X_1/\sqrt{\theta}$. Calculons la fonction de répartition et la densité de la loi de Y :

$$F_Y(y) = \mathbb{P}(X_1/\sqrt{\theta} \le y) = F_{X_1}(y\sqrt{\theta}) \quad \text{et donc} \quad f_Y(y) = \sqrt{\theta} f_{X_1}(y\sqrt{\theta}) = \frac{2}{\sqrt{\pi}}\,\mathrm{e}^{-y^2}\,\mathbb{1}_{]0,\infty[}(y).$$

On constate que la loi de Y ne dépend pas de θ.
On a

$$\frac{\hat{\theta}}{\theta} = \frac{2\hat{m}_2(\mathbf{x})}{\theta} = \frac{2}{n\theta}\sum_{i=1}^{n} X_i^2 = \frac{2}{n}\sum_{i=1}^{n}\left(\frac{X_i}{\sqrt{\theta}}\right)^2.$$

Avec $Y_i = X_i/\sqrt{\theta}$ pour $i = 1, \ldots, n$, on a alors

$$\frac{\hat{\theta}}{\theta} = \frac{2}{n}\sum_{i=1}^{n} Y_i^2.$$

Il est clair que les variables aléatoires $Y_1, \ldots, Y_n$ sont i.i.d. de même loi que Y, qui ne dépend pas de θ. Par conséquent, la loi de $\hat{\theta}/\theta$ ne dépend pas non plus de θ.
On peut calculer la loi de $n\hat{\theta}/\theta$ explicitement. D'abord, on constate que la variable aléatoire $2Y^2$ suit la loi χ_1^2, car pour tout $t > 0$,

$$F_{2Y^2}(t) = \mathbb{P}(2Y^2 \le t) = \mathbb{P}(Y \le \sqrt{t/2}) = F_Y\left(\sqrt{\frac{t}{2}}\right),$$

et donc

$$f_{2Y^2}(t) = \frac{1}{2\sqrt{2t}} f_Y\left(\sqrt{\frac{t}{2}}\right) = \frac{1}{\sqrt{2t\pi}}\,\mathrm{e}^{-t/2} = \frac{\left(\frac{1}{2}\right)^{1/2}}{\Gamma\left(\frac{1}{2}\right)} t^{1/2-1}\,\mathrm{e}^{-t/2}.$$

En utilisant que la somme de n variables indépendantes de loi χ_1^2 suit la loi χ_n^2, on en déduit que $n\hat{\theta}/\theta$ suit la loi χ_n^2.

3. On cherche a et b tels que

$$1-\alpha = \mathbb{P}_\theta(\theta \in [\hat\theta/a, \hat\theta/b]) = \mathbb{P}_\theta\left(\frac{1}{a} \le \frac{\theta}{\hat\theta} \le \frac{1}{b}\right) = \mathbb{P}_\theta\left(nb \le \frac{n\hat\theta}{\theta} \le na\right)$$
$$= F_{\chi^2_n}(na) - F_{\chi^2_n}(nb).$$

On peut choisir a et b tel que $F_{\chi^2_n}(na) = 1-\alpha/2$ et $F_{\chi^2_n}(nb) = \alpha/2$. On a alors

$$a = q_{1-\alpha/2}(\chi^2_n)/n \qquad b = q_{\alpha/2}(\chi^2_n)/n.$$

Donc on obtient finalement l'intervalle de confiance

$$\left[\frac{n\hat\theta}{q_{1-\alpha/2}(\chi^2_n)}, \frac{n\hat\theta}{q_{\alpha/2}(\chi^2_n)}\right].$$

Exercice 4. Soit la variable aléatoire

$$Y = \mathbb{1}_{\{\theta>\xi\}},$$

où $\theta \in \mathbb{R}$ et ξ est une variable aléatoire de loi $\mathcal{N}(0,1)$. On dispose d'un échantillon $Y_1,\dots,Y_n$ des réalisations i.i.d. de Y.

1. Soit Φ la fonction de répartition de la loi $\mathcal{N}(0,1)$. Montrer que $\hat\theta_n = \Phi^{-1}(\frac{1}{n}\sum_{i=1}^n Y_i)$ est l'estimateur de maximum de vraisemblance de θ. Cet estimateur est-il consistant ?
2. Chercher la loi limite de l'estimateur $\hat\theta_n$ quand $n\to\infty$.
3. Soit $0<\alpha<1$. Proposer un intervalle de confiance asymptotique de niveau $1-\alpha$.

Solution. 1. Comme Y prend ses valeurs dans $\{0,1\}$, Y suit une loi de Bernoulli avec paramètre $\mathbb{P}(Y=1) = \mathbb{P}(\theta>\xi) = \Phi(\theta)$. La fonction de vraisemblance s'écrit

$$\mathcal{L}(\theta; Y_1,\dots,Y_n) = \prod_{i=1}^n (\Phi(\theta))^{Y_i}(1-\Phi(\theta))^{1-Y_i} = (\Phi(\theta))^{\sum_{i=1}^n Y_i}(1-\Phi(\theta))^{n-\sum_{i=1}^n Y_i},$$

et la log-vraisemblance est donnée par

$$\ell(\theta) = \sum_{i=1}^n Y_i \log(\Phi(\theta)) + (n - \sum_{i=1}^n Y_i)\log(1-\Phi(\theta)).$$

Notons φ la densité de la loi $\mathcal{N}(0,1)$. On dérive

$$\begin{aligned}\ell'(\theta) &= \frac{\sum_{i=1}^n Y_i}{\Phi(\theta)}\varphi(\theta) - \frac{n-\sum_{i=1}^n Y_i}{1-\Phi(\theta)}\varphi(\theta)\\ &= \frac{(1-\Phi(\theta))\sum_{i=1}^n Y_i - \Phi(\theta)(n-\sum_{i=1}^n Y_i)}{\Phi(\theta)(1-\Phi(\theta))}\varphi(\theta)\\ &= \frac{\sum_{i=1}^n Y_i - n\Phi(\theta)}{\Phi(\theta)(1-\Phi(\theta))}\varphi(\theta).\end{aligned}$$

Puisque $\varphi>0$ et la fonction de répartition $\Phi:\mathbb{R}\to(0,1)$ est bijective, la réciproque Φ^{-1} est bien définie. Ainsi,

$$\ell'(\theta) = 0 \Leftrightarrow \sum_{i=1}^n Y_i - n\Phi(\theta) = 0 \Leftrightarrow \frac{1}{n}\sum_{i=1}^n Y_i = \Phi(\theta) \Leftrightarrow \theta = \Phi^{-1}\left(\frac{1}{n}\sum_{i=1}^n Y_i\right).$$

On vérifie que $\ell(\theta)$ atteint un maximum en $\frac{1}{n}\sum_{i=1}^n Y_i$. En effet, d'une part on a $\varphi(\theta)/(\Phi(\theta)(1-\Phi(\theta)))>0$ pour tout $\theta\in\mathbb{R}$. D'autre part, $\sum_{i=1}^n Y_i - n\Phi(\theta) < 0$ si et seulement si $\theta > \Phi^{-1}\left(\frac{1}{n}\sum_{i=1}^n Y_i\right)$. Donc, la dérivée vérifie $\ell'(\theta)<0$ si et seulement si $\theta > \Phi^{-1}\left(\frac{1}{n}\sum_{i=1}^n Y_i\right)$. On déduit que $\ell(\theta)$ atteint son maximum en $\frac{1}{n}\sum_{i=1}^n Y_i$, et donc l'estimateur de maximum de vraisemblance est $\hat\theta_n = \Phi^{-1}(\frac{1}{n}\sum_{i=1}^n Y_i)$.

Puisque $\frac{1}{n}\sum_{i=1}^n Y_i \to \mathbb{E}[Y_1] = \Phi(\theta)$ *p.s.* et Φ^{-1} est une fonction continue, on a $\hat\theta_n = \Phi^{-1}(\frac{1}{n}\sum_{i=1}^n Y_i) \to \Phi^{-1}(\Phi(\theta)) = \theta$ *p.s.*. Donc $\hat\theta_n$ est consistant pour θ.

2. En vertu du TCL (car $\mathbb{E}[Y_1^2] < \infty$), on a $\sqrt{n}(\frac{1}{n}\sum_{i=1}^n Y_i - \Phi(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathrm{Var}(Y_1)) = \mathcal{N}(0, \Phi(\theta)(1-\Phi(\theta)))$. La fonction $\Phi^{-1}(\theta)$ est continument dérivable avec dérivée $(\Phi^{-1})'(\theta) = 1/\varphi(\Phi^{-1}(\theta))$. On obtient par la delta-méthode

$$\sqrt{n}(\hat{\theta}_n - \theta) = \sqrt{n}\left(\Phi^{-1}\left(\frac{1}{n}\sum_{i=1}^n Y_i\right) - \Phi^{-1}(\Phi(\theta))\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, ((\Phi^{-1})'(\Phi(\theta)))^2\Phi(\theta)(1-\Phi(\theta)))$$
$$= \mathcal{N}\left(0, \frac{\Phi(\theta)(1-\Phi(\theta))}{\varphi^2(\theta)}\right).$$

3. On a, d'après la question précédente,

$$1-\alpha = \lim_{n\to\infty} \mathbb{P}_\theta\left(\sqrt{n}(\hat{\theta}_n - \theta) \in \left[\sqrt{\frac{\Phi(\theta)(1-\Phi(\theta))}{\varphi^2(\theta)}}q^N_{\alpha/2}, -\sqrt{\frac{\Phi(\theta)(1-\Phi(\theta))}{\varphi^2(\theta)}}q^N_{\alpha/2}\right]\right)$$
$$= \lim_{n\to\infty} \mathbb{P}_\theta\left(\hat{\theta}_n - \sqrt{\frac{\Phi(\theta)(1-\Phi(\theta))}{n\varphi^2(\theta)}}q^N_{\alpha/2} \le \theta \le \hat{\theta}_n + \sqrt{\frac{\Phi(\theta)(1-\Phi(\theta))}{n\varphi^2(\theta)}}q^N_{\alpha/2}\right).$$

Un intervalle de confiance asymptotique de niveau $1-\alpha$ est donc $\left[\hat{\theta}_n - \sqrt{\frac{\Phi(\theta)(1-\Phi(\theta))}{n\varphi^2(\theta)}}q^N_{\alpha/2}, \hat{\theta}_n + \sqrt{\frac{\Phi(\theta)(1-\Phi(\theta))}{n\varphi^2(\theta)}}q^N_{\alpha/2}\right]$.
Alternative : En vertu du TCL et du théorème de Slutsky, on a

$$\sqrt{n}\frac{\bar{Y}_n - \Phi(\theta)}{\sqrt{S_n^2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

où $S_n^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$. Ce qui implique

$$1-\alpha = \lim_{n\to\infty} \mathbb{P}_\theta\left(\sqrt{n}\frac{\bar{Y}_n - \Phi(\theta)}{\sqrt{S_n^2}} \in [q^N_{\alpha/2}, -q^N_{\alpha/2}]\right)$$
$$= \lim_{n\to\infty} \mathbb{P}_\theta\left(\bar{Y}_n - q^N_{\alpha/2}\sqrt{\frac{S_n^2}{n}} \le \Phi(\theta) \le \bar{Y}_n + q^N_{\alpha/2}\sqrt{\frac{S_n^2}{n}}\right)$$
$$= \lim_{n\to\infty} \mathbb{P}_\theta\left(\Phi^{-1}\left(\bar{Y}_n - q^N_{\alpha/2}\sqrt{\frac{S_n^2}{n}}\right) \le \theta \le \Phi^{-1}\left(\bar{Y}_n + q^N_{\alpha/2}\sqrt{\frac{S_n^2}{n}}\right)\right),$$

car la fonction Φ est strictement croissante. On a donc montré que $\left[\Phi^{-1}\left(\bar{Y}_n - q^N_{\alpha/2}\sqrt{\frac{S_n^2}{n}}\right), \Phi^{-1}\left(\bar{Y}_n + q^N_{\alpha/2}\sqrt{\frac{S_n^2}{n}}\right)\right]$ est un intervalle de confiance asymptotique de niveau $1-\alpha$ pour θ.

Exercice 5 (Loi exponentielle)**.** Soit $(E_i)_{i\ge1}$ 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 :

$$\widehat{I}_n = \left[\widehat{\lambda}_n - \frac{1.96\cdot\widehat{\sigma}_n}{\sqrt{n}}, \widehat{\lambda}_n + \frac{1.96\cdot\widehat{\sigma}_n}{\sqrt{n}}\right]$$

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

$$\widehat{J}_n = \left[\frac{1}{\widehat{\lambda}_n + \frac{1.96 \cdot \widehat{\sigma}_n}{\sqrt{n}}}, \frac{1}{\widehat{\lambda}_n - \frac{1.96 \cdot \widehat{\sigma}_n}{\sqrt{n}}} \right]$$

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

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

En appliquant le lemme de Slutsky, on obtient que

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

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 6. 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 ?