Simulations stochastiques - TD 5 (correction)

Modèle de Bernoulli : intervalles de confiance

Exercice 1. ♣ *L'objectif de l'exercice est de comparer numériquement les deux intervalles de confiance étudiés dans le cours pour estimer le paramètre θ d'un modèle de Bernoulli. Toutes les simulations numériques seront réalisées avec des échantillons $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{B}(\theta)$ pour différentes tailles d'échantillon n, différentes valeurs de θ, et différents niveaux α. On répondra notamment aux questions suivantes :*

1. *Rappeler la formule des deux intervalles de confiance vus en cours (celui construit avec l'inégalité de Bienaymé-Tchebychev et celui construit avec le TCL).*

   Pour $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$, l'intervalle de niveau de confiance α construit avec BT s'écrit
   $$\widehat{I}_{n,\alpha} = \left[\widehat{\theta}_n - \frac{1}{\sqrt{4n\alpha}}, \widehat{\theta}_n + \frac{1}{\sqrt{4n\alpha}}\right],$$
   et celui construit avec le TCL s'écrit
   $$\widehat{I}_{n,\alpha} = \left[\widehat{\theta}_n - q_{(1-\alpha/2)}\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}}, \widehat{\theta}_n + q_{(1-\alpha/2)}\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}}\right],$$
   avec $q_{(1-\alpha/2)}$ le quantile d'ordre $1-\alpha/2$ de la gaussienne.

2. *On prend $\theta = 1/2$ et $\alpha = 0.05$. Pour chaque valeur de $n \in \{3, 10, 100, 1000\}$, simuler $N_{simus} = 10^4$ échantillons de taille n et vérifier pour chaque échantillon si les intervalles de confiance obtenus contiennent la vraie valeur de θ. En déduire une estimation $\widehat{\alpha}^{BT}_{n,N_{simus}}$ et $\widehat{\alpha}^{TCL}_{n,N_{simus}}$ du niveau de risque $\mathbb{P}_\theta\big(\widehat{I}_{n,\alpha} \not\ni \theta\big)$ de chacun des deux intervalles $\widehat{I}_{n,\alpha}$. A-t-on bien $\widehat{\alpha}^{BT/TCL}_{n,N_{simus}} \lesssim \alpha$ comme attendu ? Est-ce vrai pour tout n ?*

   En général c'est OK, sauf pour $n = 3$ ou $n = 10$, où on a au contraire $\widehat{\alpha}^{TCL}_{n,N_{\text{simus}}} >> \alpha$, ce qui n'est pas étonnant étant donné que l'intervalle de confiance est un intervalle asymptotique, valable dans le régime $n \to \infty$.

3. *Même question avec $\theta = 1/2$ et $\alpha \in \{0.1, 0.2, 0.3\}$. Commentaires ?*

   Sans commentaires... c'est globalement pareil.

4. *Même question avec $\theta \in \{0.2, 0.1, 0.01\}$ et $\alpha = 0.05$. La valeur de θ influence-t-elle le niveau de risque effectif de l'intervalle de confiance construit avec le TCL ?*

   Lorsque θ se rapproche de 0, le nombre d'observation nécessaire pour que $\widehat{\alpha}^{BT/TCL}_{n,N_{\text{simus}}} \lesssim \alpha$ est de plus en plus grand (pour $\theta = 0.01$, même $n = 1000$ est insuffisant). En pratique, on impose des conditions (recette d'ingénieurs), qui se justifie par des résultats de type Berry-Essen, pour pouvoir utiliser l'intervalle de confiance asymptotique donné par le TCL : par exemple, on prend $n \geqslant 30$, $n\theta \geqslant 5$, $n(1-\theta) \geqslant 5$.

5. *Quel argument mathématique permet de guider le choix de N_{simus} ?*
   On construit un intervalle de confiance (d'un niveau à choisir) pour la probabilité $\mathbb{P}\left[\theta \notin \widehat{I}_{n,\alpha}\right]$, de largeur proportionnelle à $1/\sqrt{N_{\text{simus}}}$.

Exercice 2. *Soit X une variable aléatoire de loi $\mathcal{B}(\theta)$. Nous allons montrer le résultat suivant dû à Hoeffding (lemme d'Hoeffding, 1963) :*

$$\forall \lambda \in \mathbb{R} \ , \qquad \ln \mathbb{E}\big(e^{\lambda(X-\mathbb{E}X)}\big) \leqslant \frac{\lambda^2}{8} \ . \tag{1}$$

Cette inégalité est très utile pour démontrer des inégalités de déviation non-asymptotiques. Dans toute la suite, on pose $\psi(\lambda) = \ln \mathbb{E}\big(e^{\lambda(X-\mathbb{E}X)}\big)$ pour tout $\lambda \in \mathbb{R}$.

1. *Calculer $\psi(\lambda)$ et $\psi'(\lambda)$ puis montrer que $\psi(0) = \psi'(0) = 0$.*
   Remarquons que
   $$\mathbb{E}\big(e^{\lambda(X-\mathbb{E}X)}\big) = \mathbb{E}\big(e^{\lambda(X-\theta)}\big) = \theta e^{\lambda(1-\theta)} + (1-\theta)e^{-\lambda\theta}.$$
   On obtient alors que
   $$\psi(\lambda) = \ln\left(\theta e^{\lambda(1-\theta)} + (1-\theta)e^{-\lambda\theta}\right) = \ln\left(e^{-\lambda\theta}(\theta e^{\lambda} + (1-\theta))\right) = -\lambda\theta + \ln\left(\theta e^{\lambda} + (1-\theta)\right).$$
   En dérivant, on obtient
   $$\psi'(\lambda) = -\theta + \frac{\theta e^{\lambda}}{\theta e^{\lambda} + 1 - \theta} = -\theta + \frac{1}{1 + \frac{1-\theta}{\theta}e^{-\lambda}},$$
   d'où $\psi(0) = \psi'(0) = 0$.
2. *Montrer que, pour tout $\lambda \in \mathbb{R}$,*
   $$\psi''(\lambda) = \frac{\frac{1-\theta}{\theta}e^{-\lambda}}{\left(1 + \frac{1-\theta}{\theta}e^{-\lambda}\right)^2} \ .$$
   On dérive l'inverse de $1 + \frac{1-\theta}{\theta}e^{-\lambda}$ avec les formules de la dérivée d'un inverse, et on obtient directement le résultat.
3. *Etudier la fonction $u \mapsto u/(1+u)^2$ sur $\mathbb{R}_+$ et montrer que $u/(1+u)^2 \leqslant 1/4$ pour tout $u \geqslant 0$.*
   Ici , on est dans la même configuration que lorsque l'on veut prouver que $u(1-u) \leqslant 1/4$. La dérivée de $u \mapsto u/(1+u)^2$ s'écrit
   $$\frac{(1+u)^2 - 2u(1+u)}{(1+u)^4} = \frac{(1+u)(1-u)}{(1+u)^4},$$
   qui est positif sur $[0,1]$ et négatif sur $[1,+\infty[$, ce qui fait que la fonction $u/(1+u)^2$ est croissante jusqu'en 1 puis décroissante. On en déduite que
   $$u/(1+u)^2 \leqslant 1/(1+1)^2 = \frac{1}{4}.$$
4. *En déduire que $\psi''(\lambda) \leqslant 1/4$ pour tout $\lambda \in \mathbb{R}$, puis en déduire l'inégalité* (1).
   Indications : on commencera d'abord par le cas où $\lambda \geqslant 0$ en montrant que $\psi'(\lambda) \leqslant \lambda/4$, puis $\psi(\lambda) \leqslant \lambda^2/8$. On traitera enfin le cas $\lambda < 0$ (attention aux signes !).

Pour majorer ψ' sur $[0,+\infty[$, on peut utiliser l'égalité

$$\psi'(\lambda)=\psi'(0)+\int_0^\lambda \psi''(t)dt,$$

qui est facilement majorable en utilisant les 3 questions précédentes. En effet, $\psi'(0)=0$ et $\psi''(t)\leqslant 1/4$, donc pour $\lambda\geqslant 0$, on a

$$\psi'(\lambda)\leqslant 0+\int_0^\lambda \frac{1}{4}dt=\lambda/4.$$

Une fois qu'on a obtenu $\psi'(\lambda)\leqslant\lambda/4$ sur $[0,+\infty[$, on écrit cette fois

$$\psi(\lambda)=\psi(0)+\int_0^\lambda \psi'(t)dt\leqslant 0+\int_0^\lambda \frac{t}{4}dt=\lambda^2/8,$$

qui est l'inégalité de Hoeffding lorsque $\lambda\geqslant 0$.
Pour $\lambda<0$, il faut faire attention à remettre les intégrales dans le bon sens, pour avoir des bornes dans le bon sens :

$$\psi'(\lambda)=\psi'(0)+\int_0^\lambda \psi''(t)dt=\psi'(0)-\int_\lambda^0 \psi''(t)dt,$$

qu'on minore par

$$0-\int_\lambda^0 \frac{1}{4}dt=\lambda/4,$$

et on écrit ensuite

$$\psi(\lambda)=\psi(0)-\int_\lambda^0 \psi'(t)dt\leqslant 0-\int_\lambda^0 \frac{t}{4}dt=\lambda^2/8,$$

qui est l'inégalité de Hoeffding lorsque $\lambda\geqslant 0$.

Exercice 3 (Intervalle de confiance avec le lemme d'Hoeffding)**.** *Dans toute la suite, on suppose qu'on a accès à un échantillon $X_1,X_2,\dots,X_n$ de v.a. indépendantes et de loi de Bernoulli de paramètre $\theta\in(0,1)$ inconnu. On note $\widehat{\theta}_n=\frac{1}{n}\sum_{i=1}^n X_i$ la moyenne empirique.*

1. *Montrer que $\mathbb{E}\big(\exp\big(\sum_{i=1}^n \frac{\lambda}{n}(X_i-\theta)\big)\big)=\big[\mathbb{E}\big(\exp\big(\frac{\lambda}{n}(X_1-\theta)\big)\big)\big]^n$, puis utiliser l'inégalité (1) prouvée à l'exercice 2 (le lemme d'Hoeffding), pour montrer que :*

$$\forall\lambda\in\mathbb{R}\ ,\qquad \ln\mathbb{E}\left(e^{\lambda(\widehat{\theta}_n-\theta)}\right)\leqslant\frac{\lambda^2}{8n}\ .$$

Par l'indépendance des X_i, on a

$$\begin{aligned}\mathbb{E}\left(\exp\left(\sum_{i=1}^n\frac{\lambda}{n}(X_i-\theta)\right)\right)&=\mathbb{E}\left(\prod_{i=1}^n\exp\left(\frac{\lambda}{n}(X_i-\theta)\right)\right)\\&=\prod_{i=1}^n\mathbb{E}\left(\exp\left(\frac{\lambda}{n}(X_i-\theta)\right)\right)=\left[\mathbb{E}\left(\exp\left(\frac{\lambda}{n}(X_1-\theta)\right)\right)\right]^n.\end{aligned}$$

On en déduit que

$$\ln\mathbb{E}\left(\exp\left(\sum_{i=1}^n\frac{\lambda}{n}(X_i-\theta)\right)\right)=n\ln\left[\mathbb{E}\left(\exp\left(\frac{\lambda}{n}(X_1-\theta)\right)\right)\right].$$

Mais l'inégalité de Hoeffding nous dit que

$$\ln\left[\mathbb{E}\left(\exp\left(\frac{\lambda}{n}(X_1-\theta)\right)\right)\right] = \ln\left[\mathbb{E}\left(\exp\left(\frac{\lambda}{n}(X_1-\mathbb{E}[X_1])\right)\right)\right] \leqslant \frac{(\lambda/n)^2}{8},$$

ce qui permet de conclure

$$\ln\mathbb{E}\left(e^{\lambda(\widehat{\theta}_n-\theta)}\right) = \ln\mathbb{E}\left(\exp\left(\sum_{i=1}^{n}\frac{\lambda}{n}(X_i-\theta)\right)\right) = n\ln\left[\mathbb{E}\left(\exp\left(\frac{\lambda}{n}(X_1-\theta)\right)\right)\right] \leqslant \frac{\lambda^2}{8n}$$

2. *A l'aide de l'inégalité de Markov, montrer que, pour tout $x > 0$,*

$$\mathbb{P}\big(\widehat{\theta}_n > \theta + x\big) \leqslant e^{\lambda^2/(8n)}e^{-\lambda x} \qquad \text{pour tout } \lambda > 0\ .$$

En optimisant en λ, en déduire que $\mathbb{P}\big(\widehat{\theta}_n > \theta + x\big) \leqslant e^{-2nx^2}$ pour tout $x > 0$.
Ici, il y a une astuce fréquemment utilisé en probabilité, qui consiste à appliquer l'inégalité de Markov à la variable aléatoire $e^{\lambda(\widehat{\theta}_n-\theta)}$. Voilà le calcul

$$\begin{aligned}\mathbb{P}\left[\widehat{\theta}_n > \theta + x\right] &= \mathbb{P}\left[\lambda(\widehat{\theta}_n - \theta) > \lambda x\right] \\ &= \mathbb{P}\left[e^{\lambda(\widehat{\theta}_n-\theta)} > e^{\lambda x}\right] \\ &\leqslant \frac{\mathbb{E}[e^{\lambda(\widehat{\theta}_n-\theta)}]}{e^{\lambda x}} \quad \text{(Markov)} \\ &\leqslant \frac{e^{\lambda^2/8n}}{e^{\lambda x}} = \exp(\lambda^2/8n - \lambda x) \quad \text{(question 1)}\end{aligned}$$

On cherche le lambda qui minimise $\exp(\lambda^2/8n-\lambda x)$. Autrement dit, on cherche le minimum du polynôme de second degré $\lambda^2/8n-\lambda x$, atteint en $\lambda = -b/2a = 4nx$. On a donc, en utilisant l'inégalité pour $\lambda = 4nx$, que

$$\mathbb{P}\left[\widehat{\theta}_n > \theta + x\right] \leqslant \exp(-2nx^2).$$

3. *Montrer que, de même, $\mathbb{P}\big(\widehat{\theta}_n < \theta - x\big) \leqslant e^{-2nx^2}$, puis que*

$$\mathbb{P}\big(\big|\widehat{\theta}_n - \theta\big| > x\big) \leqslant 2e^{-2nx^2} \qquad \text{pour tout } x > 0.$$

Comment interpréter cette inégalité ?
Pour l'inégalité dans l'autre sens, il faut utiliser la même chose avec $-\lambda$.

$$\begin{aligned}\mathbb{P}\left[\widehat{\theta}_n < \theta - x\right] &= \mathbb{P}\left[-\lambda(\widehat{\theta}_n - \theta) > \lambda x\right] \\ &= \mathbb{P}\left[e^{-\lambda(\widehat{\theta}_n-\theta)} > e^{-\lambda x}\right] \\ &\leqslant \frac{\mathbb{E}[e^{\lambda(\widehat{\theta}_n-\theta)}]}{e^{\lambda x}} \quad \text{(Markov)} \\ &\leqslant \frac{e^{\lambda^2/8n}}{e^{\lambda x}} = \exp(\lambda^2/8n - \lambda x) \quad \text{(question 1).}\end{aligned}$$

De même que précédemment, on prend $\lambda = 4nx$ pour obtenir $\mathbb{P}\big(\widehat{\theta}_n < \theta - x\big) \leqslant e^{-2nx^2}$. D'où

$$\mathbb{P}\big(\big|\widehat{\theta}_n - \theta\big| > x\big) = \mathbb{P}\left[\widehat{\theta}_n < \theta - x\right] + \mathbb{P}\left[\widehat{\theta}_n > \theta + x\right] \leqslant 2e^{-2nx^2}.$$

C'est un contrôle de la déviation de $\widehat{\theta}_n$ à sa moyenne, donc un contrôle de l'erreur de l'estimateur $\widehat{\theta}_n$ de θ.

4. *Utiliser l'inégalité précédente pour montrer que l'intervalle aléatoire donné par* $\widehat{I}_{n,\alpha} = \left[\widehat{\theta}_n - \sqrt{\frac{\ln(2/\alpha)}{2n}}\,;\widehat{\theta}_n + \sqrt{\frac{\ln(2/\alpha)}{2n}}\,\right]$ *est un intervalle de confiance pour* θ *de niveau (de risque)* α. *Attention, le niveau doit être garanti pour tout* $n \in \mathbb{N}^*$, *contrairement aux intervalles de confiance de niveau asymptotique* α *pour lesquels le niveau n'est garanti qu'approximativement lorsque* $n \to +\infty$. *Dans notre cas, on parlera donc d'intervalle de confiance* non-asymptotique.
Pour avoir un niveau α, il faut avoir $2e^{-2nx^2} = \alpha$ dans l'inégalité de la question précédente. Ca correspond à $x = \sqrt{\frac{\ln(2/\alpha)}{2n}}$. On a donc

$$\mathbb{P}\big(\theta \notin \widehat{I}_{n,\alpha}\big) = \mathbb{P}\left(\big|\widehat{\theta}_n - \theta\big| > \sqrt{\frac{\ln(2/\alpha)}{2n}}\right) =\leqslant \alpha.$$

5. *On pose* $\varphi(x) = e^{-x^2/2}/\sqrt{2\pi}$ *pour* $x > 0$ *et* $\bar{\Phi}(u) = \displaystyle\int_u^{+\infty} \varphi(x)dx$ *pour* $u > 0$. *Montrer que* $\frac{d}{dx}\big(-x^{-1}\varphi(x)\big)= \big(1+x^{-2}\big)\varphi(x)$, *puis* $u^{-1}\varphi(u) \leqslant \big(1+u^{-2}\big)\bar{\Phi}(u)$ *pour tout* $u > 0$, *et en déduire que*

$$\forall u > 0\ , \qquad \bar{\Phi}(u) \geqslant \frac{\varphi(u)}{u + u^{-1}}\ .$$

Calculons

$$\frac{d}{dx}\big(-x^{-1}\varphi(x)\big)= \frac{x^2 e^{-x^2/2}}{\sqrt{2\pi} + e^{-x^2/2}}\sqrt{2\pi}x^2 = (1 + x^{-2})\varphi(x),$$

d'où

$$\begin{aligned} u^{-1}\varphi(u) &= -\int_u^{+\infty} \frac{d}{dx}\big(-x^{-1}\varphi(x)\big)dx \\ &= \int_u^{+\infty} (1 + x^{-2})\varphi(x)dx \leqslant (1 + u^{-2}) \int_u^{+\infty} \varphi(x)dx = (1 + u^{-2})\bar{\Phi}(u), \end{aligned}$$

ce qui entraîne $\bar{\Phi}(u) \geqslant \frac{\varphi(u)}{u+u^{-1}}$.

6. *Montrer que le niveau asymptotique est égal à* $\lim\limits_{n\to+\infty} \mathbb{P}_\theta\big(\widehat{I}_{n,\alpha} \not\ni \theta\big) = 2\bar{\Phi}\Big(\sqrt{\frac{\ln(2/\alpha)}{2\theta(1-\theta)}}\Big)$. *Utiliser ensuite l'inégalité de la question précédente pour montrer que ce niveau asymptotique est plus proche de* α *que celui de l'intervalle de confiance construit avec l'inégalité de Bienaymé-Tchebychev. Est-ce logique ? souhaitable ?*
On a d'après le TCL

$$\mathbb{P}_\theta\big(\widehat{I}_{n,\alpha} \not\ni \theta\big) = \mathbb{P}\left(\big|\sqrt{n}\frac{\widehat{\theta}_n - \theta}{\sqrt{\theta(1-\theta)}}\big| > \sqrt{\frac{\ln(2/\alpha)}{2\theta(1-\theta)}}\right) \to_{n\to\infty} 2\bar{\Phi}\left(\sqrt{\frac{\ln(2/\alpha)}{2\theta(1-\theta)}}\right).$$

En appelant $u_\alpha = \sqrt{\frac{\ln(2/\alpha)}{2\theta(1-\theta)}}$, on a donc

$$\lim_n \mathbb{P}_\theta\big(\widehat{I}_{n,\alpha} \not\ni \theta\big) = 2\bar{\Phi}(u_\alpha) \geqslant \frac{2\varphi(u_\alpha)}{u_\alpha + u_\alpha^{-1}}$$

Or $\theta(1-\theta) \leqslant 1/4$, donc $u_\alpha \geqslant \sqrt{2\ln(2/\alpha)} \geqslant \sqrt{2\ln(2)} \geqslant 1$. D'où

$$u_\alpha + u_\alpha^{-1} \leqslant u_\alpha + 1 \leqslant 2u_\alpha.$$

On a donc
$$\frac{2\varphi(u_\alpha)}{u_\alpha + u_\alpha^{-1}} \geqslant \frac{2\varphi(u_\alpha)}{2u_\alpha} = \frac{(\alpha/2)^{\frac{1}{4\theta(1-\theta)}}}{\sqrt{\frac{\ln(2/\alpha)}{2\theta(1-\theta)}}}.$$
Quand α proche de 0, ce terme est bien supérieur à $\exp(-1/2\alpha)$ obtenu pour l'intervalle de confiance construit avec BT.

Exercice 4. Comparer les intervalles de confiance étudiés dans l'exercice 1 avec l'intervalle de confiance non-asymptotique construit dans l'exercice 3. Toutes les simulations numériques seront réalisées avec des échantillons $X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{B}(\theta)$ pour différentes tailles d'échantillon n, différentes valeurs de θ, et différents niveaux α. On répondra notamment aux questions suivantes :

1. ♣ Comment vérifier numériquement que l'intervalle de confiance de l'exercice 3 est bien de niveau α ? Est-ce vrai pour tout $n \in \mathbb{N}^*$?
2. Quel intervalle de confiance non-asymptotique est le moins large entre celui construit avec l'inégalité de Bienaymé-Tchebychev (cf. cours) et celui construit avec le lemme d'Hoeffding (exercice 3) ? La réponse est-elle vraie pour tout $\alpha \in (0, 1)$ ou dépend-elle de la valeur de α ?
3. ♣ Vérifier numériquement votre réponse à la question précédente.
4. ♣ En pratique, que prendriez-vous comme intervalle de confiance ?