

Mise à niveau en statistique

Correction de la feuille de TD 2 : Estimation et intervalles de confiance

M2 Mathématiques de l'Aléatoire – Université Paris-Saclay

Exercice 1 – Loi géométrique

Soient $X_1, \dots, X_n$ i.i.d. de loi géométrique de paramètre $p \in (0, 1)$ sur $\{1, 2, \dots\}$, et $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$.

1. On rappelle que

$$\mathbb{E}[X] = \frac{1}{p}, \quad \text{Var}(X) = \frac{1-p}{p^2}.$$

2. La méthode des moments consiste à résoudre $\bar{X}_n = 1/p$. Elle donne

$$\hat{p}_n = \frac{1}{\bar{X}_n}.$$

3. Par la loi forte des grands nombres, $\bar{X}_n \xrightarrow{\text{p.s.}} 1/p$. La fonction $x \mapsto 1/x$ étant continue en $1/p$, on a $\hat{p}_n \xrightarrow{\text{p.s.}} p$: l'estimateur est fortement consistant, donc consistant.

Le théorème central limite donne

$$\sqrt{n} \left(\bar{X}_n - \frac{1}{p} \right) \xrightarrow{\mathcal{L}} \mathcal{N} \left(0, \frac{1-p}{p^2} \right).$$

Comme la dérivée de $g(x) = 1/x$ vaut $g'(1/p) = -p^2$, la delta-méthode entraîne

$$\sqrt{n}(\hat{p}_n - p) \xrightarrow{\mathcal{L}} \mathcal{N}(0, p^2(1-p)).$$

4. Pour $p \in (0, 1)$, la log-vraisemblance vaut

$$\ell_n(p) = n \log p + \left(\sum_{i=1}^n X_i - n \right) \log(1-p).$$

Si $\bar{X}_n > 1$, elle est strictement concave et son unique maximum est atteint en

$$\hat{p}_n^{\text{MV}} = \frac{1}{\bar{X}_n}.$$

L'estimateur des moments et l'EMV coïncident; les propriétés précédentes s'appliquent donc à l'EMV. Si $X_1 = \dots = X_n = 1$, le supremum sur $(0, 1)$ est atteint lorsque $p \uparrow 1$ mais il n'y a pas de maximum; sur le paramétrage fermé $(0, 1]$, l'EMV vaut 1.

Exercice 2 – Loi de Poisson

Soient $X_1, \dots, X_n$ i.i.d. de loi $\mathcal{P}(\theta)$, avec $\theta > 0$.

1. Comme $\mathbb{E}[X_1] = \theta$, la méthode des moments donne

$$\hat{\theta}_n = \bar{X}_n.$$

2. La loi forte des grands nombres donne $\hat{\theta}_n \xrightarrow{\text{p.s.}} \theta$. L'estimateur est donc fortement consistant, et en particulier consistant. Comme $\text{Var}(X_1) = \theta$, le théorème central limite donne

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \theta).$$

3. La log-vraisemblance est

$$\ell_n(\theta) = -n\theta + \left(\sum_{i=1}^n X_i \right) \log \theta - \sum_{i=1}^n \log(X_i!).$$

Lorsque $\bar{X}_n > 0$, elle est strictement concave et son unique maximum est

$$\hat{\theta}_n^{\text{MV}} = \bar{X}_n.$$

Si toutes les observations sont nulles, le supremum sur $\theta > 0$ est obtenu lorsque $\theta \downarrow 0$; sur le paramétrage $\theta \geq 0$, l'EMV vaut 0.

4. L'EMV coïncide avec l'estimateur des moments. Il est donc fortement consistant et vérifie la normalité asymptotique donnée à la question 2.

Exercice 3 – Loi exponentielle tradatée

On observe $X_i = \theta + Y_i$, où les Y_i sont i.i.d. de loi exponentielle de paramètre 1. On note $X_{(1)} = \min_i X_i$ et $\tilde{X}_n$ une médiane empirique.

1. On a

$$\mathbb{E}[X_i] = \theta + 1, \quad \text{Var}(X_i) = 1.$$

2. La méthode des moments donne

$$\hat{\theta}_n = \bar{X}_n - 1.$$

Cet estimateur est sans biais. Par la loi forte des grands nombres, il est fortement consistant, donc consistant. Enfin,

$$\sqrt{n}(\hat{\theta}_n - \theta) = \sqrt{n}(\bar{X}_n - (\theta + 1)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

3. Comme l'estimateur est sans biais,

$$\mathbb{E}[(\hat{\theta}_n - \theta)^2] = \text{Var}(\hat{\theta}_n) = \frac{1}{n}.$$

4. La vraisemblance s'écrit

$$L_n(\theta) = \exp \left(- \sum_{i=1}^n X_i + n\theta \right) \mathbf{1}_{\{\theta \leq X_{(1)}\}}.$$

Elle est croissante en θ sur $(-\infty, X_{(1)}]$. Ainsi,

$$\hat{\theta}_n^{\text{MV}} = X_{(1)}.$$

5. Pour tout $x \geq 0$,

$$\begin{aligned} \mathbb{P}(n(\hat{\theta}_n^{\text{MV}} - \theta) > x) &= \mathbb{P}\left(Y_1 > \frac{x}{n}, \dots, Y_n > \frac{x}{n}\right) \\ &= e^{-x}. \end{aligned}$$

Par conséquent, cette identité étant exacte pour tout n ,

$$n(\hat{\theta}_n^{\text{MV}} - \theta) \sim \text{Exp}(1).$$

En particulier, l'EMV est fortement consistant et son biais vaut $1/n$.

6. La médiane m de X_1 vérifie $1 - e^{-(m-\theta)} = 1/2$, donc

$$m = \theta + \log 2.$$

On propose alors

$$\hat{\theta}_n^{\text{med}} = \tilde{X}_n - \log 2.$$

La consistance forte des quantiles empiriques donne $\hat{\theta}_n^{\text{med}} \xrightarrow{\text{p.s.}} \theta$. De plus, la densité de X_1 en sa médiane vaut $f_\theta(m) = 1/2$. Le théorème central limite pour les quantiles entraîne

$$\sqrt{n}(\hat{\theta}_n^{\text{med}} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{(1/2)(1 - 1/2)}{f_\theta(m)^2}\right) = \mathcal{N}(0, 1).$$

7. Les estimateurs des moments et par la médiane ont une erreur d'ordre $n^{-1/2}$, tandis que l'EMV a une erreur d'ordre n^{-1} et une loi exacte simple. Dans ce modèle correctement spécifié, on choisira donc naturellement $X_{(1)}$. Il est biaisé, mais sa version $X_{(1)} - 1/n$ est sans biais et conserve la vitesse n^{-1} .

Exercice 4 – Loi de Laplace

1. Pour le modèle d'échelle

$$f_\theta(x) = \frac{1}{2\theta} e^{-|x|/\theta}, \quad \theta > 0,$$

la log-vraisemblance vaut

$$\ell_n(\theta) = -n \log(2\theta) - \frac{1}{\theta} \sum_{i=1}^n |X_i|.$$

Elle croît jusqu'à $n^{-1} \sum_i |X_i|$, puis décroît. Comme $\sum_i |X_i| > 0$ presque sûrement,

$$\hat{\theta}_n^{\text{MV}} = \frac{1}{n} \sum_{i=1}^n |X_i|$$

est presque sûrement l'unique EMV. Or $|X_i| \sim \text{Exp}(1/\theta)$, avec $\mathbb{E}|X_i| = \theta$ et $\text{Var}(|X_i|) = \theta^2$. Ainsi,

$$\hat{\theta}_n^{\text{MV}} \xrightarrow{\text{p.s.}} \theta, \quad \sqrt{n}(\hat{\theta}_n^{\text{MV}} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \theta^2).$$

2. Dans le modèle de position

$$f_\mu(x) = \frac{1}{2} e^{-|x-\mu|}, \quad \mu \in \mathbb{R},$$

maximiser la vraisemblance revient à minimiser $\sum_i |X_i - \mu|$. L'ensemble des EMV est donc l'ensemble des médianes empiriques :

$$\arg \max_{\mu \in \mathbb{R}} L_n(\mu) = \begin{cases} \{X_{((n+1)/2)}\}, & n \text{ impair}, \\ [X_{(n/2)}, X_{(n/2+1)}], & n \text{ pair}. \end{cases}$$

L'EMV est presque sûrement unique lorsque n est impair et ne l'est pas lorsque n est pair. Toute sélection usuelle de cet intervalle est fortement consistante. Comme la densité en la médiane μ vaut $f_\mu(\mu) = 1/2$,

$$\sqrt{n}(\hat{\mu}_n^{\text{MV}} - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Exercice 5 – Propriétés des moyenne et médiane empiriques

1. (a) Pour tout $t \in \mathbb{R}$,

$$\mathbb{E}[(X_1 - t)^2] = \text{Var}(X_1) + (t - \mathbb{E}[X_1])^2.$$

Le minimum est donc unique et atteint en $t = \mathbb{E}[X_1]$.

- (b) De même,

$$\frac{1}{n} \sum_{i=1}^n (X_i - t)^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 + (t - \bar{X}_n)^2.$$

Ainsi, l'unique minimiseur est $\bar{X}_n$.

2. (a) Pour toute variable aléatoire positive Z , le théorème de Fubini donne

$$\mathbb{E}[Z] = \int_0^\infty \mathbb{P}(Z > z) dz.$$

Appliqué à $Z = |X_1 - t|$, ceci conduit à

$$\mathbb{E}|X_1 - t| = \int_t^\infty (1 - F(u)) du + \int_{-\infty}^t F(u) du.$$

Comme F est continue, la dérivée par rapport à t vaut

$$2F(t) - 1.$$

Elle est négative à gauche de $x_{1/2}$ et positive à droite, sous les hypothèses de l'énoncé. Par conséquent, $x_{1/2}$ est l'unique minimiseur de $t \mapsto \mathbb{E}|X_1 - t|$.

- (b) Ordonnons l'échantillon : $X_{(1)} \leq \dots \leq X_{(n)}$. La fonction $g(t) = n^{-1} \sum_i |X_i - t|$ est convexe et affine par morceaux. Entre deux statistiques d'ordre successives, sa pente est égale au nombre d'observations à gauche de t moins le nombre d'observations à droite de t , divisé par n . Il s'ensuit que

$$\arg \min_{t \in \mathbb{R}} g(t) = \begin{cases} \{X_{((n+1)/2)}\}, & n \text{ impair}, \\ [X_{(n/2)}, X_{(n/2+1)}], & n \text{ pair}. \end{cases}$$

Exercice 6 – Intervalle de confiance et loi normale

On a $X_i \sim \mathcal{N}(0, \theta^2)$ et l'on pose $\tau = \theta^2$.

1. Un estimateur naturel, qui est aussi l'estimateur des moments et l'EMV, est

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^n X_i^2.$$

Comme $X_i^2/\tau \sim \chi_1^2$ et que les observations sont indépendantes,

$$\frac{n\hat{\tau}}{\tau} \sim \chi_n^2.$$

En particulier, $\mathbb{E}[\hat{\tau}] = \tau$ et $\text{Var}(\hat{\tau}) = 2\tau^2/n$.

2. Notons $q_{\nu, \gamma} = F_{\chi_\nu^2}^{-1}(\gamma)$. Puisque

$$\mathbb{P}\left(q_{n, \alpha/2} \leq \frac{n\hat{\tau}}{\tau} \leq q_{n, 1-\alpha/2}\right) = 1 - \alpha,$$

on obtient l'intervalle exact

$$\left[\frac{n\hat{\tau}}{q_{n, 1-\alpha/2}}, \frac{n\hat{\tau}}{q_{n, \alpha/2}} \right].$$

Ses probabilités d'erreur inférieure et supérieure valent chacune $\alpha/2$.

3. Le théorème central limite appliqué à X_i^2 donne

$$\sqrt{n} \frac{\hat{\tau} - \tau}{\sqrt{2} \tau} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

En notant $z_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ et en inversant le pivot asymptotique, on obtient, lorsque $z_{1-\alpha/2} \sqrt{2/n} < 1$,

$$\left[\frac{\hat{\tau}}{1 + z_{1-\alpha/2} \sqrt{2/n}}, \frac{\hat{\tau}}{1 - z_{1-\alpha/2} \sqrt{2/n}} \right].$$

4. Pour $n = 10$, $\hat{\tau} = 2$ et $\alpha = 0.05$, l'intervalle exact est

$$\left[\frac{20}{20.48}, \frac{20}{3.25} \right] \simeq [0.98, 6.15].$$

Avec $z_{0.975} \simeq 1.96$, l'intervalle asymptotique précédent vaut environ

$$[1.07, 16.2].$$

Pour ce petit échantillon, l'approximation est médiocre et produit ici une borne supérieure très grande. L'intervalle exact est à privilégier.

5. Pour tester $H_0 : \tau \leq 3$ contre $H_1 : \tau > 3$, les grandes valeurs de $\hat{\tau}$ sont défavorables à H_0 . Sous $\tau = 3$,

$$\frac{n\hat{\tau}}{3} \sim \chi_n^2.$$

La probabilité de rejet est croissante en τ sur H_0 ; le cas le moins favorable est donc $\tau = 3$. Un test exact de niveau α rejette lorsque

$$\hat{\tau} > \frac{3}{n} q_{n, 1-\alpha}.$$

Pour $n = 10$ et $\alpha = 0.05$, la valeur critique est

$$\frac{3}{10} q_{10, 0.95} \simeq \frac{3}{10} \times 18.31 = 5.49.$$

Comme $\hat{\tau} = 4 < 5.49$, on ne rejette pas H_0 . La p -valeur vaut

$$1 - F_{\chi_{10}^2} \left(\frac{10 \times 4}{3} \right) = 1 - F_{\chi_{10}^2} (40/3) \simeq 0.21.$$

Exercice bonus – Des tanks nazis

On suppose que $(X_1, \dots, X_n)$ est un échantillon uniforme sans remise dans $\{1, \dots, N\}$, et l'on note $M_n = X_{(n)} = \max_i X_i$.

1. Chaque sous-ensemble de n numéros est équiprobable parmi les $\binom{N}{n}$ sous-ensembles possibles. L'événement $\{M_n \leq m\}$ signifie que les n numéros observés appartiennent à $\{1, \dots, m\}$. Pour $m = n, \dots, N$,

$$\mathbb{P}_N(M_n \leq m) = \frac{\binom{m}{n}}{\binom{N}{n}}.$$

De même, si $M_n = m$, les $n - 1$ autres numéros doivent être choisis parmi $\{1, \dots, m - 1\}$, d'où

$$\mathbb{P}_N(M_n = m) = \frac{\binom{m-1}{n-1}}{\binom{N}{n}}.$$

En utilisant $m \binom{m-1}{n-1} = n \binom{m}{n}$ puis l'identité du hockey-stick,

$$\mathbb{E}_N[M_n] = \frac{1}{\binom{N}{n}} \sum_{m=n}^N m \binom{m-1}{n-1} = \frac{n}{\binom{N}{n}} \sum_{m=n}^N \binom{m}{n} = \frac{n}{\binom{N}{n}} \binom{N+1}{n+1} = \frac{n(N+1)}{n+1}.$$

Par échangeabilité, chaque X_i est marginalement uniforme sur $\{1, \dots, N\}$. L'indépendance n'est pas nécessaire pour calculer l'espérance de la moyenne :

$$\mathbb{E}_N[\bar{X}_n] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_N[X_i] = \frac{1}{N} \sum_{k=1}^N k = \frac{N+1}{2}.$$

Deux estimateurs sans biais de N sont ainsi

$$\hat{N}_{\max} = \frac{n+1}{n} M_n - 1, \quad \hat{N}_{\text{moy}} = 2\bar{X}_n - 1.$$

2. Construisons deux intervalles de confiance de niveau au moins $1 - \alpha$.

Intervalle fondé sur le maximum. La borne inférieure $N \geq M_n$ est certaine. Pour construire une borne supérieure, remarquons que, pour $m \leq N$,

$$\mathbb{P}_N(M_n \leq m) = \prod_{j=0}^{n-1} \frac{m-j}{N-j} \leq \left(\frac{m+1}{N+1} \right)^n.$$

En effet, $(m-j)/(N-j) \leq (m+1)/(N+1)$ pour chaque $j = 0, \dots, n-1$. Fixons $p \in (0, 1)$. L'événement

$$N > \frac{M_n + 1}{p} - 1$$

implique $(M_n + 1)/(N + 1) < p$; sa probabilité est donc au plus p^n . En choisissant $p = \alpha^{1/n}$, on obtient l'intervalle conservatif

$$N \in \left[M_n, \frac{M_n + 1}{\alpha^{1/n}} - 1 \right]$$

avec probabilité au moins $1 - \alpha$. Pour tenir compte du caractère entier de N , on prend la partie entière de la borne supérieure. Une version un peu moins fine, mais plus immédiate, s'obtient avec $\mathbb{P}_N(M_n \leq m) \leq (m/N)^n$ et donne $[M_n, M_n/\alpha^{1/n}]$.

Intervalle fondé sur la moyenne : justification de Hoeffding. Les X_i ne sont pas indépendantes, et l'on ne peut donc pas appliquer directement la preuve usuelle. Introduisons $Y_1, \dots, Y_n$, tirées *avec* remise dans $\{1, \dots, N\}$. Les Y_i sont i.i.d. et ont la même loi marginale que les X_i .

Le théorème de comparaison de Hoeffding pour les tirages sans remise affirme que, pour toute fonction convexe φ pour laquelle les espérances existent,

$$\mathbb{E}_N \left[\varphi \left(\sum_{i=1}^n X_i \right) \right] \leq \mathbb{E}_N \left[\varphi \left(\sum_{i=1}^n Y_i \right) \right].$$

Le tirage sans remise est donc plus concentré que le tirage avec remise au sens de l'ordre convexe. En particulier, cette comparaison s'applique à $\varphi(u) = e^{\lambda u}$, ce qui permet de transférer la méthode de Chernoff.

Posons $\mu = (N + 1)/2$. Pour $\lambda, t > 0$, l'inégalité de Markov donne

$$\begin{aligned}\mathbb{P}_N(\bar{X}_n - \mu \geq t) &\leq e^{-\lambda nt} \mathbb{E}_N \left[e^{\lambda \sum_{i=1}^n (X_i - \mu)} \right] \\ &\leq e^{-\lambda nt} \mathbb{E}_N \left[e^{\lambda \sum_{i=1}^n (Y_i - \mu)} \right].\end{aligned}$$

Or $Y_i - \mu$ est centrée et $Y_i \in [1, N]$. Le lemme de Hoeffding donne

$$\mathbb{E}_N[e^{\lambda(Y_i - \mu)}] \leq \exp\left(\frac{\lambda^2(N-1)^2}{8}\right).$$

En utilisant l'indépendance des Y_i , puis en optimisant en $\lambda = 4t/(N-1)^2$, il vient

$$\mathbb{P}_N(\bar{X}_n - \mu \geq t) \leq \exp\left(-\frac{2nt^2}{(N-1)^2}\right).$$

La même démonstration appliquée à $-X_i$, puis l'union des deux événements, donnent

$$\mathbb{P}_N\left(\left|\bar{X}_n - \frac{N+1}{2}\right| \geq t\right) \leq 2 \exp\left(-\frac{2nt^2}{(N-1)^2}\right).$$

L'astuce faisant intervenir $N + 1$. La borne précédente dépend encore du paramètre inconnu par $N - 1$. On utilise alors

$$N - 1 \leq N + 1 = 2\mu.$$

En prenant $t = x\mu = x(N + 1)/2$, on obtient

$$\begin{aligned}\mathbb{P}_N\left(\left|\frac{\bar{X}_n}{\mu} - 1\right| \geq x\right) &\leq 2 \exp\left(-\frac{nx^2}{2} \frac{(N+1)^2}{(N-1)^2}\right) \\ &\leq 2e^{-nx^2/2}.\end{aligned}$$

Cette dernière borne ne dépend plus de N . Posons

$$x_{n,\alpha} = \sqrt{\frac{2\log(2/\alpha)}{n}}.$$

Avec probabilité au moins $1 - \alpha$,

$$(1 - x_{n,\alpha})\mu \leq \bar{X}_n \leq (1 + x_{n,\alpha})\mu.$$

Si $x_{n,\alpha} < 1$, on isole $\mu = (N + 1)/2$ et l'on obtient

$$N \in \left[\frac{2\bar{X}_n}{1 + x_{n,\alpha}} - 1, \frac{2\bar{X}_n}{1 - x_{n,\alpha}} - 1 \right].$$

Cet intervalle a un niveau au moins $1 - \alpha$. On peut renforcer sa borne inférieure à l'aide de la contrainte certaine $N \geq M_n$, puis arrondir les deux bornes de manière appropriée puisque N est entier.