Projet X-MAP-311-2012-CHAFAI-3

Approximations gaussienne et poissonnienne

Proposé par Djalil Chafaï

Second semestre 2011-2012

On s'intéresse en détail aux approximations gaussienne et poissonnienne.

1 Un point de vue statistique

Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires réelles i.i.d. Si elles sont intégrables de moyenne m, la loi forte des grands nombres entraîne que la moyenne empirique

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} = \frac{S_n}{n}$$

est un estimateur presque sûrement convergent et sans biais de m. Si de plus les variables aléatoires $(X_n)_{n\geq 1}$ sont de carré intégrable, et de variance $\sigma^2 > 0$, alors le théorème de la limite centrale affirme que pour tout intervalle I de $\mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sqrt{n}}{\sigma}\left(\frac{X_1 + \cdots + X_n}{n} - m\right) \in I\right) = \frac{1}{\sqrt{2\pi}} \int_I e^{-\frac{u^2}{2}}\, du.$$

En d'autre termes,

$$\frac{\sqrt{n}}{\sigma}\left(\frac{S_n}{n} - m\right) = \frac{S_n - nm}{\sqrt{n\sigma^2}} = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{\sigma^2(S_n)}} \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,1).$$

Par conséquent, pour tout $0 \leq \alpha \leq 1/2$, le théorème de la limite centrale fournit un intervalle de confiance asymptotique pour m de niveau asymptotique α :

$$I_{n,\alpha} = \left[\overline{X}_n - \frac{\sigma q_\alpha}{\sqrt{n}}\,,\, \overline{X}_n + \frac{\sigma q_\alpha}{\sqrt{n}}\right]$$

où q_α est le quantile $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite. On a en effet

$$\mathbb{P}(m \in I_{n,\alpha}) = \mathbb{P}\left(\frac{S_n - nm}{\sqrt{n}\sigma} \in [-q_\alpha, q_\alpha]\right) \underset{n\to\infty}{\longrightarrow} \frac{1}{\sqrt{2\pi}} \int_{-q_\alpha}^{q_\alpha} e^{-\frac{u^2}{2}}\, du = 1 - \alpha.$$

L'intervalle $I_{n,\alpha}$ est de largeur $2\sigma q_\alpha n^{-1/2}$, qui tend vers 0 si $n \to \infty$, et vers ∞ si $\alpha \to 0$. Cette approche repose intégralement sur l'approximation de la loi de $n^{-1/2}(S_n - nm)$ par $\mathcal{N}(0, \sigma^2)$. Malheureusement, cette approximation est parfois très mauvaise.

2 Autour du théorème central limite

Si les (X_n) sont de loi normale $\mathcal{N}(m,\sigma^2)$ alors $(S_n-nm)/\sqrt{n\sigma^2}\sim\mathcal{N}(0,1)$ et cela permet de concevoir le théorème central limite comme une sorte de théorème du point fixe. D'autre part, la vitesse en $\sqrt{n}$ provient de $\sigma(S_n)=\sigma\sqrt{n}$.

Q1 Les moments d'ordre 1 et 2 de $n^{-1/2}(S_n-nm)$ et de $\mathcal{N}(0,1)$ sont identiques, et $(n^{-1/2}(S_n-nm))_{n\geq 1}$ converge en loi vers $\mathcal{N}(0,\sigma^2)$. Proposer une illustration graphique en langage Scilab du comportement des moments d'ordre >2 ;

Q2 Proposer une preuve du théorème central limite par la méthode des moments ;

La preuve classique du théorème central limite est basée sur les fonctions caractéristiques (transformée de Fourier). On se propose d'explorer une nouvelle approche, qui s'avère fructueuse pour certains modèles non-linéaires.

Q3 Démontrer que si $X_1,Y_1,\ldots,X_n,Y_n$ sont des variables aléatoires indépendantes t.q. $\mathbb{E}(|X_k|^3)<\infty$ et $Y_k\sim\mathcal{N}(\mathbb{E}(X_k),\sigma^2(X_k))$ pour tout $1\leq k\leq n$, alors pour toute $f\in\mathcal{C}^3(\mathbb{R},\mathbb{R})$ avec f,f',f'',f''' bornées, en posant $\tau_k^3=\mathbb{E}(|X_k-\mathbb{E}(X_k)|^3)$,

$$|\mathbb{E}(f(X_1+\cdots+X_n))-\mathbb{E}(f(Y_1+\cdots+Y_n))|\leq\frac{(\tau_1^3+\cdots+\tau_n^3)}{2}\left\|f'''\right\|_\infty;$$

Q4 Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires i.i.d. de moyenne m, de variance $\sigma^2>0$, et de moment d'ordre 3 fini. Soit f comme dans la question précédente. Déduire de la question précédente que que pour tout $n\geq 1$,

$$\left|\mathbb{E}\left(f\left(\frac{S_n-nm}{\sqrt{n}\sigma}\right)\right)-\int_{\mathbb{R}}f(u)\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}\,du\right|\leq\frac{\|f'''\|_\infty\,\mathbb{E}(|X_1-m|^3)}{2\sqrt{n}\sigma^3};$$

Q5 En approchant les indicatrices d'intervalle par des fonctions régulières, déduire de la question précédente le théorème central limite lorsque les variables aléatoires $(X_n)_{n\geq 1}$ ont un moment d'ordre 3 fini ;

Le théorème central limite suggère que la loi de S_n est proche de $\mathcal{N}(nm,n\sigma^2)$ lorsque n est grand, et les questions **4**-**5** permettent de quantifier cette proximité pour certaines « fonctions test ». Le théorème de Berry-Esseen, dans le même esprit, affirme que pour tout $t\in\mathbb{R}$ et $n\geq 1$, en notant $\tau^3=\mathbb{E}(|X_1-\mathbb{E}(X_1)|^3)$,

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\frac{S_n-nm}{\sqrt{n}\sigma}\leq t\right)-\int_{-\infty}^{t}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\,dx\right|\leq\frac{\tau^3}{\sqrt{n}\sigma^3}.$$

Q6 Trouver une preuve courte du théorème de Berry-Essen.

3 Cas des variables de Bernoulli

Si $(X_n)_{n\geq 1}$ sont des variables de Bernoulli sur $\{0,1\}$ de paramètre $p\in]0,1[$, on a $m=p$, $\sigma^2=pq$ et $\tau^3=pq(1-2pq)$ où $q=1-p$, et le théorème de Berry-Esseen donne

$$\sup_{t\in\mathbb{R}}\left(\mathbb{P}\left(S_n\leq\sqrt{npq}t+np\right)-\int_{-\infty}^{t}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\,dx\right)\leq\frac{1-2pq}{\sqrt{npq}}.$$

Cette approximation de la loi binomiale par la loi normale est d'autant plus bonne que $(1-2p(1-p))/\sqrt{np(1-p)}$ est petit. À n fixé, cette borne est minimale pour $p=1/2$ mais explose quand p se rapproche de 0 ou de 1.

Q7 Illustrer le comportement et la justesse de la borne de Berry-Esseen pour le cas Bernoulli en fonction de p au moyen d'un programme en langage Scilab ;

Il est possible de quantifier la proximité de la loi binomiale à la loi normale en utilisant la densité plutôt que la fonction de répartition.

Q8 En utilisant une formule de Stirling rafinée, montrer que si $-\infty < a < b < +\infty$ alors la convergence uniforme suivante a lieu :

$$\lim_{n\to\infty} \sqrt{n} \sup_{k\in I_n(a,b)} \left| \mathbb{P}(S_n = k) - \frac{1}{\sqrt{2\pi npq}} \exp\left(-\frac{(k-np)^2}{2npq}\right)\right| = 0$$

où

$$I_n(a,b) = \left\{ 0 \le k \le n : \frac{k-np}{\sqrt{npq}} \in [a,b] \right\};$$

Q9 En déduire que

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - np}{\sqrt{npq}} \in [a,b]\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}}\, dx.$$

On a $S_n \sim \mathrm{Binom}(n,p)$. Lorsque p est trop proche de 0 ou de 1, l'approximation gaussienne du théorème central limite n'est pas bonne. Pour tout $k \in \mathbb{N}$, on a,

$$\mathbb{P}(S_n = k) - e^{-np}\frac{(np)^k}{k!} = \left(\frac{n}{n(1-p)} \cdots \frac{n-k+1}{n(1-p)}(1-p)^n - e^{-np}\right)\frac{(np)^k}{k!}.$$

Si p dépend de n avec $\lim_{n\to\infty} np = \lambda$ alors la loi de S_n tend vers la loi de Poisson $\mathrm{Poi}(\lambda)$. Il est naturel de chercher à quantifier cette convergence vers la loi de Poisson dans l'esprit de l'inégalité de Berry-Esseen. Dans cette direction, le but de la section suivante est d'établir l'inégalité de poissonisation suivante, utile si np^2 est petit :

$$\sum_{k=0}^{\infty} \left| \mathbb{P}(S_n = k) - e^{-np}\frac{(np)^k}{k!} \right| \le 2np^2.$$

4 Appoximation poissonnienne de la loi binomiale

Dans toute cette section, E est un ensemble au plus dénombrable.

Q10 Montrer que l'ensemble des lois sur E est un espace métrique complet pour

$$d(\mu,\nu) := \sup_{A\subset E} |\mu(A) - \nu(A)|;$$

Q11 Montrer que si μ et ν sont des lois sur E alors

$$d(\mu,\nu) = \frac{1}{2} \sup_{f:E\to[-1,1]} \left| \int f\, d\mu - \int f\, d\nu \right| = \frac{1}{2}\sum_{x\in E} |\mu(x) - \nu(x)|.$$

De plus, le supremum dans la définition de d est atteint pour l'ensemble

$$A = A_* := \{x \in E : \mu(x) \ge \nu(x)\}$$

tandis que dans l'expression variationnelle fonctionnelle, il est atteint pour

$$f = f_* := \mathbf{1}_{A_*} - \mathbf{1}_{A_*^c};$$

Q12 Montrer que si $(X_n)_{n\geq 1}$ sont des variables aléatoires sur E et si μ_n désigne la loi de X_n, alors pour toute loi μ sur E, les propriétés suivantes sont équivalentes :

(a) $\lim_{n\to\infty} \int f\, d\mu_n = \int f\, d\mu$ pour toute fonction bornée $f : E \to \mathbb{R}$;

(b) $\lim_{n\to\infty} \mu_n(x) = \mu(x)$ pour tout $x \in E$;

(c) $\lim_{n\to\infty} d(\mu_n, \mu) = 0$.

Q13 Montrer que si (μ_n) sont des lois sur E et $\mu(x) := \lim_{n\to\infty} \mu_n(x)$ alors μ n'est pas forcément une loi, sauf si E est fini ;

Q14 Montrer que si μ et ν sont des lois sur E alors

$$d(\mu, \nu) = 1 - \sum_{x\in E} (\mu(x) \wedge \nu(x)).$$

En particulier, $d(\mu, \nu) = 1$ si et seulement si μ et ν ont des support disjoints ;

Q15 Montrer que si μ et ν sont des lois sur E alors

$$d(\mu, \nu) = \inf_{(X,Y)} \mathbb{P}(X \neq Y)$$

où l'infimim porte sur les couples de v.a. sur $E \times E$ de lois marginales μ et ν. Montrer de plus que l'infimum est atteint ;

Q16 Montrer que si $\alpha_1, \ldots, \alpha_n$ et $\beta_1, \ldots, \beta_n$ sont des lois de probabilité sur $\mathbb{N}$ alors

$$d(\alpha_1 * \cdots * \alpha_n, \beta_1 * \cdots * \beta_n) \leq d(\alpha_1, \beta_1) + \cdots + d(\alpha_n, \beta_n);$$

Q17 Montrer que $d(\mathrm{Bin}(1, p), \mathrm{Poi}(p)) \leq p^2$;

Q18 Déduire des deux questions précédentes que si $X_1, \ldots, X_n$ sont des variables aléatoires indépendantes de lois de Bernoulli $\mathrm{Bin}(1, p_1), \ldots, \mathrm{Bin}(1, p_n)$, si μ_n est la loi de $S_n = X_1 + \cdots + X_n$, et si $\nu_n = \mathrm{Poi}(p_1 + \ldots + p_n)$ est la loi de Poisson de même moyenne que S_n, alors on a

$$d(\mu_n, \nu_n) \leq p_1^2 + \cdots + p_n^2;$$

Q19 En déduire que si $(X_{n,k})_{1\leq k\leq n}$ est un tableau triangulaire de v.a.r. indépendantes de lois de Bernoulli avec $X_{n,k} \sim \mathrm{Bin}(1, p_{n,k})$ pour tous $n \geq k \geq 1$, et si

$$\lim_{n\to\infty} p_{n,1} + \cdots + p_{n,n} = \lambda \quad \text{et} \quad \lim_{n\to\infty} \max_{1\leq k\leq n} p_{n,k} = 0,$$

alors $X_{n,1} + \cdots + X_{n,n}$ converge en loi vers $\mathrm{Poi}(\lambda)$ quand $n \to \infty$;

Q20 Illustrer le résultat précédent au moyen d'un programme Scilab.

Il est possible d'obtenir une meilleure borne :

$$d(\mathrm{Bin}(1, p_1) * \cdots \mathrm{Bin}(1, p_n), \mathrm{Poi}(p_1 + \cdots + p_n)) \leq (1 - e^{-(p_1+\cdots+p_n)}) \frac{p_1^2 + \cdots + p_n^2}{p_1 + \cdots + p_n}.$$

Dans certains cas, cette borne permet d'établir une proximité asymptotique à une loi de Poisson dont le paramètre ne converge pas, de manière plus souple que la borne en $p_1^2 + \cdots + p_n^2$. D'autre part, on peut voir d comme une distance de couplage de Wasserstein : comme $\mathbb{P}(X \neq Y) = \mathbb{E}(d(X, Y))$ pour la distance atomique $d(x, y) = \delta_{x\neq y}$, on a, en notant $\Pi(\mu, \nu)$ l'ensemble des lois sur $E \times E$ de marginales μ et ν,

$$d(\mu, \nu) = \min_{\pi\in\Pi(\mu,\nu)} \int_{E\times E} d(x, y)\, d\pi(x, y).$$