

Statistiques mathématiques : cours 2

Guillaume Lécué

7 septembre 2017

Références

1. Cours :

- ▶ V. Rivoirard et G. Stoltz, "*Statistiques en action*"
- ▶ P.J. Bickel et K. Doksum, "*Mathematical statistics*"
- ▶ A. Montfort, "*Cours de statistique mathématique*"

2. Exercices :

- ▶ J.J. Daudin, S. Robin et C. Vuillet, "*Statistique inférentielle. Idées, démarches, exemples*"
- ▶ D. Fourdrinier, "*Statistiques inférentielle : cours et exercices corrigés*"
- ▶ B. Cadre et C. Vial, "*Statistique Mathématique Cours et Exercices Corrigés*"

Cours précédent (rappel)

- ▶ Expérience statistique, modèle statistique, échantillonnage
- ▶ Fonction de répartition empirique :

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x), \quad x \in \mathbb{R}$$

et quelques propriétés **asymptotiques** :

$$\widehat{F}_n(x) \xrightarrow{p.s.} F(x), \quad \left\| \widehat{F}_n - F \right\|_{\infty} \xrightarrow{p.s.} 0$$

leurs vitesses de convergence :

$$\sqrt{n}(\widehat{F}_n(x) - F(x)) \xrightarrow{d} \mathcal{N}(0, F(x)(1 - F(x))),$$

$$\sqrt{n} \left\| \widehat{F}_n(x) - F(x) \right\|_{\infty} \xrightarrow{d} K$$

- ▶ propriétés **non-asymptotique** grâce à Tchebychev.

Aujourd'hui

Estimateur "plug-in" et la méthode delta

Quantiles empiriques et applications

Un algorithme "on-line" : Robbins-Monro

- ▶ **Objectif** : estimation d'une caractéristique scalaire $T(F)$ d'une loi inconnue de fonction de répartition F à partir d'un n -échantillon $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} X \sim F$ de cette loi

données : $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} F \rightsquigarrow$ problème : **estimer** $T(F)$

- ▶ Exemples

- ▶ Déjà vu : valeur en un point $T(F) = F(x) = \mathbb{E} I(X \leq x)$
- ▶ Fonctionnelle régulière :

$$T(F) = h \left(\int_{\mathbb{R}} g(x) dF(x) \right) = h(\mathbb{E} g(X))$$

où $g, h : \mathbb{R} \rightarrow \mathbb{R}$ sont **régulières** et $X \sim F$

Exemples de fonctionelles régulières

▶ Moyenne : $T(F) = m(F) = \int_{\mathbb{R}} x dF(x) = \mathbb{E} X$.

▶ Variance :

$$T(F) = \sigma^2(F) = \int_{\mathbb{R}} (x - m(F))^2 dF(x) = \mathbb{E} (X - \mathbb{E} X)^2$$

▶ Asymétrie (skewness) :

$$T(F) = \alpha(F) = \frac{\int_{\mathbb{R}} (x - m(F))^3 dF(x)}{\sigma^3(F)} = \frac{\mathbb{E}(X - \mathbb{E} X)^3}{\sigma^3(F)}$$

▶ Aplatissement (kurtosis) :

$$T(F) = \kappa(F) = \frac{\int_{\mathbb{R}} (x - m(F))^4 dF(x)}{\sigma^4(F)} = \frac{\mathbb{E} (X - \mathbb{E} X)^4}{\sigma^4(F)}$$

Exemples de fonctionnelles non régulières

Définition

Soit X une v.a.r. (de cdf F) et $0 < p < 1$. On appelle **quantile d'ordre p** de X (resp. F) :

$$q_p(F) = \inf\{x \in \mathbb{R} : F(x) \geq p\}$$

- ▶ quand F est **continue et strictement croissante** le **quantile d'ordre p** de la loi F est l'unique solution de

$$F(q_p) = p \quad (\text{càd } q_p = F^{-1}(p)).$$

- ▶ la **médiane** = $\text{med}(F) = q_{1/2}(F)$
- ▶ les **quartiles** = $\{q_{1/4}(F), \text{med}(F), q_{3/4}(F)\}$

Estimateur "plug-in"

Définition

On appelle estimateur "plug-in" (càd "par substitution") de $T(F)$ l'estimateur $T(\hat{F}_n)$.

- ▶ quand $T(F) = h(\mathbb{E} g(X))$ alors l'estimateur *plug-in* de $T(F)$ est :

$$T(\hat{F}_n) = h\left(\frac{1}{n} \sum_{i=1}^n g(X_i)\right)$$

- ▶ quand $T(F) = q_p(F) = \inf\{x \in \mathbb{R} : F(x) \geq p\}$, l'estimateur *plug-in* est le **quantile empirique** :

$$T(\hat{F}_n) = \inf\{x \in \mathbb{R} : \hat{F}_n(x) \geq p\}$$

Performances asymptotiques de l'estimateur plug-in pour l'estimation de fonctionnelles régulières de la forme $T(F) = h(\mathbb{E} g(X))$

Convergence (consistance) : si $g, h : \mathbb{R} \rightarrow \mathbb{R}$, h continue et $\mathbb{E} |g(X)| < \infty$, alors $T(\widehat{F}_n) \xrightarrow{\text{p.s.}} T(F)$ (LFGN + continuous map theorem).

Vitesse de convergence (normalité asymptotique) :

1. TCL :

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n g(X_i) - \mathbb{E} g(X) \right) \xrightarrow{d} \mathcal{N}(0, \text{Var}[g(X)])$$

où $\text{Var}[g(X)] = \mathbb{E} [(g(X) - \mathbb{E} g(X))^2]$

2. On a un résultat du type $\sqrt{n}(Z_n - c_1) \xrightarrow{d} \mathcal{N}(0, c_2)$. Comment transférer ce résultat à $\sqrt{n}(h(Z_n) - h(c_1)) \xrightarrow{d} ?$

Vitesse de convergence de $T(\widehat{F}_n)$ vers $T(F) = h(\mathbb{E} g(X))$

Théorème (Méthode « delta »)

Soit (Z_n) une suite de v.a.r. et V une v.a.r. telles que

$$a_n(Z_n - c_0) \xrightarrow{d} V$$

où (a_n) est une suite de réels positifs tendant vers $+\infty$ et c_0 est une constante. Soit $h : \mathbb{R} \rightarrow \mathbb{R}$ une fonction **continue et dérivable en c_0** . Alors

$$a_n(h(Z_n) - h(c_0)) \xrightarrow{d} h'(c_0)V$$

Méthode Delta

1. si $\sqrt{n}(Z_n - c_1) \xrightarrow{d} \mathcal{N}(0, c_2)$ et h dérivable en c_1 alors

$$\sqrt{n}(h(Z_n) - h(c_1)) \xrightarrow{d} \mathcal{N}(0, c_2[h'(c_1)]^2)$$

2. si $V \sim \mathcal{N}(\mu, \nu)$ et $a \in \mathbb{R}$ alors $aV \sim \mathcal{N}(a\mu, a^2\nu)$.
3. l'idée centrale de la preuve de la méthode Delta est un développement limité de h en c_0 : quand $n \rightarrow \infty$

$$a_n(h(Z_n) - h(c_0)) \approx h'(c_0)[a_n(Z_n - c_0)] \approx h'(c_0)V$$

Conclusion : normalité asymptotique de l'estimateur plug-in dans le cas de fonctionnelles régulières $T(F) = h(\mathbb{E}g(X))$

Proposition

Si $\mathbb{E}[g(X)^2] < +\infty$ et h est une fonction continue et dérivable en $\mathbb{E}g(X)$, alors

$$\sqrt{n}(T(\hat{F}_n) - T(F)) \xrightarrow{d} \mathcal{N}(0, v(F)),$$

où $v(F) = h'(\mathbb{E}[g(X)])^2 \text{Var}[g(X)]$.

Pour construire un **intervalle de confiance**, on aimerait remplacer $v(F)$ par $v(\hat{F}_n)$: quand h est \mathcal{C}^1 , on montre que $v(\hat{F}_n) \xrightarrow{\mathbb{P}} v(F)$ et, via le lemme de Slutsky,

$$\sqrt{n} \frac{T(\hat{F}_n) - T(F)}{v(\hat{F}_n)^{1/2}} \xrightarrow{d} \mathcal{N}(0, 1)$$

Autre application de la méthode Delta : stabilisation de la variance

Soit X_1, \dots, X_n un n -échantillon de loi Exponentielle de paramètre $\theta \in [0, 1]$.

- ▶ densité $f(\theta, x) = \theta \exp(-\theta x)I(x > 0)$, moyenne $\mathbb{E}_\theta X = 1/\theta$, variance $\text{Var}_\theta X = 1/\theta^2$
- ▶ TCL : $\sqrt{n}(\bar{X}_n - 1/\theta) \xrightarrow{d} \mathcal{N}(0, 1/\theta^2)$
- ▶ **Pb.** : La variance asymptotique dépend du paramètre inconnu θ
- ▶ Méthode Delta : si g est \mathcal{C}^1 alors :

$$\sqrt{n}(g(\bar{X}_n) - g(1/\theta)) \xrightarrow{d} \mathcal{N}(0, (g'(1/\theta))^2/\theta^2)$$

- ▶ en particulier pour $g(\theta) = \log(\theta)$, on a

$$\sqrt{n}(g(\bar{X}_n) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, 1)$$

Application : stabilisation de la variance (Bernoulli)

Soit X_1, \dots, X_n un n -échantillon dans le modèle de Bernoulli de paramètre $\theta \in [0, 1]$.

- ▶ TCL : $\sqrt{n}(\bar{X}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \theta(1 - \theta))$
- ▶ La variance asymptotique dépend du paramètre inconnu θ
- ▶ Méthode Delta : si g est \mathcal{C}^1 alors :

$$\sqrt{n}(g(\bar{X}_n) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, (g'(\theta))^2 \theta(1 - \theta))$$

- ▶ en particulier pour $g(\theta) = 2\arcsin(\sqrt{\theta})$, on a

$$\sqrt{n}(g(\bar{X}_n) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, 1)$$

En dimension $k > 1$

- ▶ Il s'agit de fonctionnelles de la forme

$$T(F) = h(\mathbb{E} g_1(X), \dots, \mathbb{E} g_k(X))$$

où $h : \mathbb{R}^k \rightarrow \mathbb{R}$ est \mathcal{C}^1 .

- ▶ **Exemple** : le coefficient d'asymétrie

$$T(F) = \frac{\mathbb{E}(X - \mathbb{E} X)^3}{\sigma^3} = h(\mathbb{E} X, \mathbb{E} X^2, \mathbb{E} X^3)$$

où σ est l'écart-type de X .

- ▶ **Outil** : Versions multidimensionnelles
 1. du TCL
 2. de la « méthode delta ».

TCL et méthode « delta » multidimensionnelle

- ▶ **TCL multidimensionnel** : $(\mathbf{X}_n)_{n \geq 1}$ vecteurs aléatoires dans \mathbb{R}^k , i.i.d., de moyenne $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}_1]$ et de matrice de variance-covariance $\Sigma = \mathbb{E}[(\mathbf{X}_1 - \boldsymbol{\mu})(\mathbf{X}_1 - \boldsymbol{\mu})^\top]$. Alors $\bar{\mathbf{X}}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i$ vérifie :

$$\sqrt{n}(\bar{\mathbf{X}}_n - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

- ▶ **Méthode « delta » multidimensionnelle** : Si, de plus, $h : \mathbb{R}^k \rightarrow \mathbb{R}^d$ continûment différentiable, alors

$$\sqrt{n}(h(\bar{\mathbf{X}}_n) - h(\boldsymbol{\mu})) \xrightarrow{d} \mathcal{N}\left(0, \nabla h(\boldsymbol{\mu})^\top \Sigma \nabla h(\boldsymbol{\mu})\right).$$

rem. : si $A \in \mathbb{R}^{k \times d}$ et $G \sim \mathcal{N}_k(\boldsymbol{\mu}, \Sigma)$ alors $A^\top G \sim \mathcal{N}_d(A^\top \boldsymbol{\mu}, A^\top \Sigma A)$

Notations : gradient (1/2)

$$h : \begin{cases} \mathbb{R}^k & \rightarrow \\ x & \mapsto \end{cases} \begin{pmatrix} \mathbb{R}^d \\ h_1(x) \\ \vdots \\ h_d(x) \end{pmatrix}$$

alors $\nabla h(x) = (\nabla h_1(x) \quad \nabla h_2(x) \quad \cdots \quad \nabla h_d(x)) \in \mathbb{R}^{k \times d}$

$$\text{où } \nabla h_j(x) = \begin{pmatrix} \partial_{x_1} h_j(x) \\ \vdots \\ \partial_{x_k} h_j(x) \end{pmatrix} \quad j = 1, \dots, d$$

tel que $h_1(x + v) \approx h_1(x) + \langle \nabla h_1(x), v \rangle = h_1(x) + \nabla h_1(x)^\top v$ et de même

$$h(x + v) \approx h(x) + \langle \nabla h(x), v \rangle = h(x) + \nabla h(x)^\top v$$

Notations : gradient (2/2)

Par exemple :

1. pour $h(x) = Ax$ où $A \in \mathbb{R}^{d \times k}$, on a :

$$\nabla h(x) = A^\top$$

2. pour $h(x) = \|Ax\|_2^2$, on a :

$$\nabla h(x) = 2A^\top Ax$$

3. pour $h(x) = \|y - Ax\|_2^2$, on a :

$$\nabla h(x) = -2A^\top (y - Ax)$$

Application : coefficient d'asymétrie

- ▶ **Coefficient d'asymétrie** : on a

$$T(F) = h\left(\mathbb{E} X, \mathbb{E} X^2, \mathbb{E} X^3\right)$$

avec

$$h(\alpha, \beta, \gamma) = \frac{\gamma - 3\alpha\beta + 2\alpha^3}{(\beta - \alpha^2)^{3/2}}.$$

- ▶ **l'estimateur plug-in** est

$$T(\widehat{F}_n) = h\left(\frac{1}{n} \sum_{i=1}^n X_i, \frac{1}{n} \sum_{i=1}^n X_i^2, \frac{1}{n} \sum_{i=1}^n X_i^3\right).$$

- ▶ On applique le **TCL multidimensionnel** avec $\mathbf{X}_i = (X_i, X_i^2, X_i^3)^\top$ et $\boldsymbol{\mu} = (\mathbb{E} X, \mathbb{E} X^2, \mathbb{E} X^3)^\top$, puis la **méthode « delta »** avec h .

Quantiles théoriques et empiriques

Quantile "théorique" d'ordre p :

$$T(F) = q_p(F) = \inf\{x \in \mathbb{R} : F(x) \geq p\}$$

Quantile empirique d'ordre p :

$$T(\hat{F}_n) = \hat{q}_{n,p} = \inf\{x \in \mathbb{R} : \hat{F}_n(x) \geq p\}$$

Question : Quelles sont les propriétés statistiques d'estimation de $q_p(F)$ par $\hat{q}_{n,p}$? (Pb. : on n'est plus dans le cas régulier)

Définition

Soit X_1, \dots, X_n un n -échantillon de v.a.r.. On appelle *statistiques d'ordre* les n statistiques $X_{(1)}, \dots, X_{(n)}$ construites telles que

$$X_{(1)} \leq \dots \leq X_{(n)}$$

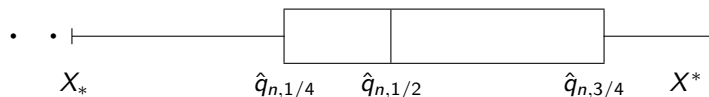
1. pour le quantile d'ordre $0 < p < 1$:

$$\hat{q}_{n,p} = X_{(k)} = X_{(\lceil np \rceil)} \text{ quand } \frac{k-1}{n} < p \leq \frac{k}{n}$$

2. en particulier, la médiane empirique vérifie :

$$\hat{q}_{n,1/2} = \text{med}(\hat{F}_n) = X_{(\lceil n/2 \rceil)} \text{ où } \lceil t \rceil = \min(n \in \mathbb{N} : n \geq t)$$

Le boxplot : représentation synthétique de la dispersion de données réelles



fin de la “moustache” (whiskers) :

$$X_* = \min\{X_i : |X_i - \hat{q}_{n,1/4}| \leq 1,5\mathcal{I}_n\},$$

$$X^* = \max\{X_i : |X_i - \hat{q}_{n,3/4}| \leq 1,5\mathcal{I}_n\}.$$

Intervalle interquartile :

$$\mathcal{I}_n = \hat{q}_{n,3/4} - \hat{q}_{n,1/4}.$$

Les données au-delà des whiskers sont considérées comme *outliers*.
(Il existe d'autres variantes)

Exemple d'application du boxplot

http://localhost:8888/notebooks/box_qqplots.ipynb Box-plot

Le qq-plot : test d'adéquation à une loi

Etant donné un n -échantillon X_1, \dots, X_n et une cdf F_{ref} , on veut tester si l'hypothèse suivante est acceptable :

$$(H_0) \quad \text{“Les } X_i \text{ sont distribués selon } F_{ref}\text{”}$$

Pour “accepter ou refuser visuellement” cette hypothèse, on peut tracer le qq-plot : c'est le **nuage de points**

$$\left(q_{i/n}(F_{ref}), \hat{q}_{n,i/n} \right)_{i=1}^n = \left(q_{i/n}(F_{ref}), X_{(i)} \right)_{i=1}^n$$

1. si le nuage de points est “approximativement” aligné avec la droite $y = x$ alors l'hypothèse est acceptée (on trace aussi la droite $y = x$ sur un qq-plot)
2. si les points sont “approximativement” alignés avec une droite affine alors l'hypothèse est vraie à une transformation de centrage et scaling près (généralement, on normalise les données)

convergence des quantiles empiriques

Théorème

Soit X une v.a.r. (on note par F sa cdf) admettant une densité f_X par rapport à la mesure de Lebesgue. On suppose que f_X est strictement positive p.s. sur un intervalle $I \subset \mathbb{R}$ et nulle en dehors. Soit $0 < p < 1$. On a

$$\widehat{q}_{n,p} \xrightarrow{p.s.} q_p(F) = q_p$$

Si de plus la densité f_X de X admet une version continue en q_p alors $\widehat{q}_{n,p}$ est asymptotiquement Gaussien :

$$\sqrt{n}(\widehat{q}_{n,p} - q_p) \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{f_X(q_p)^2}\right)$$

Convergence des quantiles empiriques

La variance asymptotique de $\hat{q}_{n,p}$ est

$$\frac{p(1-p)}{f_X(q_p)^2}$$

La quantité $f_X(q_p)$ est inconnue.

- ▶ Comme $\hat{q}_{n,p}$ est **fortement consistant** et f_X est continue en q_p ,

$$f_X(\hat{q}_{n,p}) \xrightarrow{p.s.} f_X(q_p)$$

On peut donc "remplacer" q_p par $\hat{q}_{n,p}$ grâce à Slutsky :

$$\frac{\sqrt{n}f_X(\hat{q}_{n,p})}{\sqrt{p(1-p)}}(\hat{q}_{n,p} - q_p) \xrightarrow{d} \mathcal{N}(0, 1)$$

- ▶ Mais $f_X(\hat{q}_{n,p})$ est aussi inconnue ! (problème d'estimation de densité)

Limites de l'approche "plug-in"

L'estimation de $T(F)$ par $T(\hat{F}_n)$ n'est pas toujours **possible** :

- ▶ Exemple : si F admet une densité f continue par rapport à la mesure de Lebesgue qu'on souhaite estimer en un x_0 donné :

$$T(F) = f(x_0) = F'(x_0),$$

on ne **peut pas prendre** comme estimateur $\hat{F}'_n(x_0)$ car \hat{F}_n est constante par morceaux.

L'estimation de $T(F)$ par $T(\hat{F}_n)$ n'est pas toujours **souhaitable** :

- ▶ Souvent on dispose d'information **a priori** supplémentaire : F appartient à une sous-classe particulière de distributions (**le modèle**) et il y a des choix plus judicieux que l'estimateur par plug-in (cf. cours suivants).

Un algorithme "on-line" : Robbins-Monro

"Batch" vs "on-line"

Il existe principalement deux manières de générer/recevoir des données :

- ▶ **"batch"** : les données sont toutes obtenues en une seule fois (ex. : jeux de données)
- ▶ **"on-line"** : les données sont obtenues les unes à la suite des autres (ex. : données en temps réel)

Remarque :

1. \hat{F}_n et $\hat{q}_{n,\alpha}$ sont des estimateurs "batch"
2. on peut regarder les données "batch" comme des données "on-line" (cf. vovpal wabbit)

Estimation "on-line" des quantiles

Question : ebay souhaite connaître le 95-ième pourcentile des montants de transaction sur son site.

Deux stratégies :

1. "batch" : on reprend tous les achats passés sur eBay depuis sa création et on calcul $\hat{q}_{n,95/100}$. Problème : n est très grand !
2. "on-line" : à chaque nouvel achat, on actualise un estimateur (en temps réel).

Rem. : De nombreux estimateurs on-line sont adaptés d'algorithmes d'optimisation convexe itératifs comme la **descente de gradient**.

Descente de gradient / méthode de Newton

Problème : trouver un zéro d'une fonction f croissante et \mathcal{C}^1 : trouver x tel que

$$f(x) = 0$$

La méthode de Newton est une méthode itérative :

Init : $x_0 \in \mathbb{R}$ **while** *stopping criteria* **do**

1. on fait une DL de f en x_k :

$$f(x) \approx f(x_k) + f'(x_k)(x - x_k)$$

2. on résoud $f(x_k) + f'(x_k)(x - x_k) = 0$ (au lieu de $f(x) = 0$) :

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

end

Descente de gradient / méthode de Newton

1. Critère d'arrêt (pour ϵ donné) :

$$|f(x_k)| \leq \epsilon \text{ ou } |x_{k+1} - x_k| \leq \epsilon$$

2. Quand la fonction n'est pas dérivable ou que la dérivée est difficile à calculer, on remplace $f'(x_k)$ par η_k^{-1} (step size)
3. chercher le minimum d'une fonction convexe h c'est chercher un zéro d'une fonction croissante h' : méthode de Newton = descente de gradient

$$x_{k+1} = x_k - \frac{h'(x_k)}{h''(x_k)}$$

(et si h'' n'existe pas ou difficile à calculer : $h''(x_k) \leftrightarrow \eta_k^{-1}$)

Estimation "on-line" des quantiles par Robbins-Monro (1/2)

Soit X une v.a.r. admettant une densité f strictement positive sur une intervalle $I \subset \mathbb{R}$ et nulle en dehors de cet intervalle. On note par F la cdf de X .

1. F est dérivable sur \mathbb{R} : $F' = f$ p.p.
2. F est strictement croissante sur I
3. soit $p \in (0, 1)$, le quantile d'ordre p de X est l'unique solution de

$$F(x) - p = 0$$

On est donc amené à trouver le zéro d'une fonction dérivable strictement croissante : on peut utiliser la méthode de Newton

L'algorithme de Newton est

$$x_{k+1} = x_k - \frac{F(x_k) - p}{f(x_k)}$$

Problèmes :

1. f est inconnu : $f(x_k) \leftrightarrow \eta_k^{-1}$ (step size)
2. F est inconnue : on écrit $F(x_k) = \mathbb{E} I(X \leq x_k)$ et on "estime" $F(x_k)$ par $I(X_{k+1} \leq x_k)$ grâce à la nouvelle donnée X_{k+1}

On obtient l'algorithme de Robbins-Monro (1954) :

$$x_{k+1} = x_k - \eta_k (I(X_{k+1} \leq x_k) - p)$$

Robbins-Monro / descente de gradient stochastique

L'**algorithme de Robbins-Monro** (RM) pour l'estimation du quantile d'ordre $p \in (0, 1)$ est le suivant :

Data: X_1, \dots, X_n v.a.r.i.i.d.

Init : $x_0 \in (0, 1)$, $(\eta_k)_k$ une suite de nombre réels positifs

for $k = 0, \dots, n$ **do**

$$x_{k+1} = x_k - \eta_k (I(X_{k+1} \leq x_k) - p)$$

end

1. écriture en **pseudo-code**
2. algorithme **itératif**
3. $(\eta_k)_k$ est appelé le **step size**. Par exemple :

$$\eta_k = k^{-a}, \text{ où } a \in (1/2, 1] \text{ (ou "line search")}$$

4. x_0 **starting point** (cf. "warm start")

Théorème

Soit $p \in (0, 1)$ et X une v.a.r. dont la cdf F vérifie :

1. F est continue
2. il existe un unique $q_p \in \mathbb{R}$ tel que pour tout $x \neq q_p$,

$$(x - q_p)(F(x) - p) > 0$$

Soit $(X_k)_k \stackrel{i.i.d.}{\sim} X$. Alors, la suite itérative de RM $(x_k)_k$ où $x_0 \in \mathbb{R}$ et $x_{k+1} = x_k - \eta_k(I(X_{k+1} \leq x_k) - p)$ **converge presque sûrement vers q_p** quand le step size $(\eta_k)_k$ vérifie :

$$\sum_k \eta_k = +\infty \text{ et } \sum_k \eta_k^2 < +\infty$$

Vitesse de convergence de RM

Théorème

Si de plus F est \mathcal{C}^2 alors pour $f = F'$ (densité de X) et $\sigma^2 = p(1-p)$, quand $n \rightarrow \infty$:

1. si $f(q_p) > 1/2$ alors

$$\sqrt{n}(x_n - q_p) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2}{2f(q_p) - 1}\right)$$

2. si $f(q_p) = 1/2$ alors

$$\sqrt{\frac{n}{\log n}}(x_n - q_p) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

3. si $0 < f(q_p) < 1/2$ alors $n^{f(q_p)}(x_n - q_p) \xrightarrow{d} Z$ où Z est une variable aléatoire bornée p.s..

Comparaison d'estimateurs (1/2)

Problème : Dans le cadre "batch", on a construit deux estimateurs du quantile $q_p(F)$: $\hat{q}_{n,p}$ et x_n (RM) lequel choisir ?

1) critères théoriques (asymptotique) :

- ▶ les deux estimateurs sont fortement **consistants**
- ▶ la **vitesse de convergence** de $\hat{q}_{n,p}$ est toujours en $1/\sqrt{n}$ alors que celle de x_n se dégrade quand $f(q_p) \leq 1/2 \Rightarrow$

$\hat{q}_{n,p}$ est préférable à x_n quand $f(q_p) \leq 1/2$

- ▶ quand $1/2 < f(q_p)$, $\hat{q}_{n,p}$ et x_n sont tous les deux asymptotiquement normaux de vitesse de convergence en $1/\sqrt{n}$ mais leurs **variances asymptotiques** sont

- ★ pour $\hat{q}_{n,p}$: $\sigma^2/f(q_p)^2$

- ★ pour x_n : $\sigma^2/(2f(q_p) - 1)$

or $\sigma^2/f(q_p)^2 \leq \sigma^2/(2f(q_p) - 1)$ donc

$\hat{q}_{n,p}$ est préférable à x_n quand $1/2 < f(q_p)$

D'un point de vue théorique, $\hat{q}_{n,p}$ est préférable à x_n

Comparaison d'estimateurs (2/2)

2) critères empiriques :

- ▶ **coût de calcul** : la construction de $\hat{q}_{n,p}$ nécessite le tri des données X_1, \dots, X_n (qui peuvent être distribuée quand n est grand) contrairement à x_n qui est on-line \Rightarrow
 x_n est préférable à $\hat{q}_{n,p}$ quand n est grand
- ▶ **Etude de la convergence sur des données simulées** : l'intérêt des données simulées est qu'on connaît la valeur de l'objet à estimer.
http://localhost:8888/notebooks/rm_quantile.ipynb
Robbins-Monro
- ▶ **Etude des estimateurs sur des données réelles** : cohérence des résultats; échantillon test.