Inégalités de concentration pour les statistiques d'ordre

Méthode entropique et représentation de Rényi

Maud Thomas[1]
en collaboration avec Stéphane Boucheron[1]

[1]LPMA Université Paris-Diderot

Colloque Jeunes Probabilistes et Statisticiens
Forges-les-Eaux, 6-11 avril 2014

Statistiques d'ordre

- $X_1, \ldots, X_n \sim_{\text{i.i.d.}} F$.

Statistiques d'ordre

$X_{(1)} \geq \ldots \geq X_{(n)}$ réarrangement décroissant de $X_1, \ldots, X_n$.

- $X_{(1)}$: maximum.
- $X_{(n/2)}$: médiane empirique.
- $\mathbb{P}\{X_{(k)} \leq t\} = \sum_{i=k}^{n} \binom{n}{i} F^i(t)(1 - F(t))^{n-i}$.
- Statistique classique et théorie des valeurs extrêmes :
  - Distributions asymptotiques.
  - Convergence des moments.

Objectif

Obtenir des bornes simples, non asymptotiques de la variance et des queues de probabilités des statistiques d'ordre.

Concentration

Concentration de la mesure

Une fonction de plusieurs variables aléatoires indépendantes qui ne dépend pas trop de chacune d'entre-elles est concentrée autour de son espérance.

Exemple : Concentration gaussienne

- X vecteur aléatoire gaussien standard et $Z = f(X)$.
- Inégalité de Poincaré : $\mathrm{Var}[Z] \leq \mathbb{E}\|\nabla f\|^2$.
- Inegalité de Sobolev logarithmique de Gross : $\mathrm{Ent}[Z^2] \leq 2\mathbb{E}\|\nabla f\|^2$.
- Inégalité de Cirelson : $\mathbb{P}\{Z \geq \mathbb{E}Z + t\} \leq \exp(-t^2/(2L^2))$ si $\|\nabla f\| \leq L$.

Statistiques d'ordre et inégalité de Poincaré

- $f(X_1, \ldots, X_n) = X_{(k)}$
  la k^e statistique d'ordre d'un échantillon quelconque est une fonction régulière de n variables aléatoires indépendantes.
- $||\nabla f|| = 1$.
- X_i v.a gaussiennes standard.
  - Inégalité de Poincaré $\Rightarrow \mathrm{Var}[X_{(k)}] \leq 1$.
  - Théorie des valeurs extrêmes $\Rightarrow \mathrm{Var}[X_{(1)}] = O(1/\log n)$.
  - Théorie classique des statistiques $\Rightarrow \mathrm{Var}[X_{(n/2)}] = O(1/n)$.

Problème

On comprend mal en quel sens les statistiques d'ordre sont des fonctions régulières de l'échantillon.

Statistiques d'ordre et espacements

Proposition (Boucheron, T. (2012))

Pour $0 < k \leq n/2$ *et* $\lambda \in \mathbb{R}$

- $$\mathrm{Var}[X_{(k)}] \leq k\mathbb{E}\left[\left(X_{(k)} - X_{(k+1)}\right)^2\right] = k\mathbb{E}[\Delta_k{}^2]\,.$$

- $$\begin{aligned}\mathrm{Ent}\left[e^{\lambda X_{(k)}}\right] &:= \lambda\mathbb{E}[X_{(k)}e^{\lambda X_{(k)}}] - \mathbb{E}[e^{\lambda X_{(k)}}]\log\mathbb{E}[e^{\lambda X_{(k)}}] \\ &\leq k\mathbb{E}\left[e^{\lambda X_{(k+1)}}\psi(\lambda(X_{(k)} - X_{(k+1)}))\right] \\ &= k\mathbb{E}\left[e^{\lambda X_{(k+1)}}\psi(\lambda\Delta_k)\right]\end{aligned}$$

avec $\psi(x) = 1 + (x-1)e^x$.

Eléments de preuve

Inégalité Efron-Stein (Efron, Stein (1981) ; Steele (1986))

Soit $f : \mathbb{R}^n \to \mathbb{R}$ mesurable, et soit $Z = f(X_1, \ldots, X_n)$.
Soit $Z_i = f_i(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ où $f_i : \mathbb{R}^{n-1} \to \mathbb{R}$ est une fonction mesurable arbitraire.
Si Z est de carré intégrable, alors :

$$\mathrm{Var}[Z] \leq \mathbb{E}\left[\sum_{i=1}^{n} (Z - Z_i)^2\right] .$$

Inégalité Sobolev logarithmique modifiée (Wu(2000) ; Massart (2000))

Soit $\tau(x) = e^x - x - 1$. Avec les mêmes notations, pour tout $\lambda \in \mathbb{R}$,

$$\mathrm{Ent}\left[e^{\lambda Z}\right] \leq \mathbb{E}\left[\sum_{i=1}^{n} e^{\lambda Z} \tau\left(-\lambda(Z - Z_i)\right)\right] .$$

Représentation de Rényi

- Les statistiques d'ordre d'un échantillon exponentiel sont distribuées comme des sommes partielles de variables exponentielles indépendantes.

Représentation de Rényi (Rényi (1953))

Soient $Y_{(1)} \geq Y_{(2)} \geq \ldots \geq Y_{(n)}$ les statistiques d'ordre d'un échantillon exponentiel, alors

$$\left(Y_{(n)}, \ldots, Y_{(i)}, \ldots, Y_{(1)}\right) \sim \left(\frac{E_n}{n}, \ldots, \sum_{k=i}^{n} \frac{E_k}{k}, \ldots, \sum_{k=1}^{n} \frac{E_k}{k}\right)$$

où $E_1, \ldots, E_n$ sont des variables aléatoires exponentielles indépendantes.

Transformation quantile

Définition (Fonction quantile)

$$F^{\leftarrow}(p) = \inf\{x : F(x) \geq p\}, p \in (0,1) \ .$$

Notation

$$U(t) = F^{\leftarrow}(1 - 1/t), t \in (1, \infty).$$

Représentation pour les statistiques d'ordre

Si $Y_{(1)} \geq \ldots \geq Y_{(n)}$ sont les statistiques d'ordre d'un échantillon exponentiel, alors

$$(U \circ \exp)(Y_{(1)}) \geq \ldots \geq (U \circ \exp)(Y_{(n)})$$

sont distribuées comme les statistiques d'un échantillon tiré selon F.

Taux de hasard

Définition (Taux de hasard)

Le taux de hasard h d'une fonction de répartition F différentiable est définie par :

$$h = F'/\overline{F} = F'/(1-F) \ .$$

Lemme

Le taux de hasard de la fonction de répartition F est croissant, ssi $U \circ \exp$ est concave.

En effet,

$$(U \circ \exp)' = \frac{1}{h\,(U \circ \exp)} \ .$$

Majorer la variance

Théorème (Boucheron, T. (2012))

Si F a un taux de hasard croissant h, alors pour $1 \leq k \leq n/2$,

$$\mathrm{Var}\left[X_{(k)}\right] \leq \frac{2}{k}\mathbb{E}\left[\left(\frac{1}{h(X_{(\boldsymbol{k}+\mathbf{1})})}\right)^2\right] .$$

Proposition (Cas gaussien (Boucheron, T. (2012)))

Soit $n \geq 3$, soit $X_{(k)}$ la k^e statistique d'ordre d'un échantillon distribué comme les valeurs absolues de n v.a gaussiennes.

$$\mathrm{Var}[X_{(k)}] \leq \frac{1}{k \log 2} \frac{8}{\log\left(\frac{2n}{k}\right) - \log(1 + \frac{4}{k}\log\log\left(\frac{2n}{k}\right))} .$$

Inégalité Efron-Stein exponentielle

Association négative

X et Y sont négativement associées si pour toutes fonctions croissantes f, g

$$\mathbb{E}[f(X)g(Y)] \leq \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)] .$$

Lemme

Si F a un taux de hasard croissant, alors $X_{(k+1)}$ et $\Delta_k = X_{(k)} - X_{(k+1)}$ sont négativement associées.

Théorème (Boucheron, T. (2012))

Si F a un taux de hasard croissant h,
alors pour $\lambda \geq 0$, et $1 \leq k \leq n/2$,

$$\log \mathbb{E} e^{\lambda(X_{(k)} - \mathbb{E}X_{(k)})} = \lambda \frac{k}{2} \mathbb{E}\left[\sqrt{\frac{V_k}{k}}\left(e^{\lambda\sqrt{V_k/k}} - 1\right)\right] .$$

Un peu de théorie des valeurs extrêmes (1)

- Etant donné $X_1, \ldots, X_n$ i.i.d $\sim F$, quel est le comportement asymptotique de $\max(X_1, \ldots, X_n)$?

Théorème (Fisher and Tippet (1928) and Gnedenko (1943))

S'il existe $a_n > 0$ et b_n et une distribution non dégénérée G telles que :

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G(x)$$

en tout point de continuité x de G, alors G est de la forme :

$$G(x) = \exp\left(-(1+\gamma x)^{-1/\gamma}\right), \quad 1+\gamma x > 0$$

pour $\gamma = 0$, $G(x) = \exp(-e^{-x})$.

- Dans ce cas, on dit que F appartient au max-domaine d'attraction de $G = G_\gamma$, noté $F \in \mathrm{MDA}(G_\gamma)$.
- γ est appelé indice de valeurs extrêmes.

Un peu de théorie des valeurs extrêmes (2)

- On distingue trois domaines
  - Domaine de Fréchet ($\gamma > 0$) : lois à queue lourde. Ex : Pareto, Cauchy.
  - Domaine de Gumbel ($\gamma = 0$) : lois à queue plutôt fine. Ex : Exponentielle, Gaussienne.
  - Domaine de Weibull ($\gamma < 0$) : lois à queue finies à droite. Ex : Uniforme, Beta.

Problème : Estimation de γ

- Hill (1975) : pour $\gamma > 0$.
- Pickands (1975) : pour $\gamma \in \mathbb{R}$.
- Estimateur des moments : pour $\gamma \in \mathbb{R}$.
- Estimateur du maximum de vraisemblance : pour $\gamma > -1/2$.
- ...

Estimateur de Hill $\widehat{\gamma}$ ($\gamma > 0$)

- $X_1, \ldots, X_n \in \mathrm{MDA}(D_\gamma)$, $\gamma > 0$.
- $X_{(1)} \geq \ldots \geq X_{(n)}$: statistiques d'ordre associées.
- Pour $1 \leq k \leq n$, on définit :

$$\widehat{\gamma}(k) = \frac{1}{k} \sum_{i=1}^{k} \log \frac{X_{(i)}}{X_{(k+1)}} \quad .$$

- $\overline{F}(t) = (1 - F)(t) \approx t^{-1/\gamma}$ pour t grand.
- Estimateur de Hill $\widehat{\gamma}$ = estimateur du maximum de vraisemblance de γ si $\overline{F} = t^{-1/\gamma}$.
- Choix de k = compromis biais-variance.

Concentration pour l'estimateur de Hill

- J_n un entier.
- $Y_{(k+1)}$ la $(k+1)^{\text{e}}$ statistique d'ordre d'un échantillon exponentiel.
- $\overline{\eta}$ est une fonction décroissante tendant vers 0 en ∞.

Proposition

Sous certaines conditions, si $F \in \mathrm{MDA}(\gamma), \gamma > 0$,
Pour $1 \leq k \leq J_n$

$$-\frac{3\gamma}{k}\mathbb{E}\left[\overline{\eta}\left(e^{Y_{(k+1)}}\right)\right] \leq \mathrm{Var}[\widehat{\gamma}(k)] - \frac{\gamma^2}{k} \leq \frac{3\gamma}{k}\mathbb{E}\left[\overline{\eta}\left(e^{Y_{(k+1)}}\right)\right] + \frac{3}{k}\mathbb{E}\left[\overline{\eta}\left(e^{Y_{(k+1)}}\right)^2\right].$$

Soit $\tau \geq 1$, *alors pour tout* $t > 0$ *sur* $\{\overline{F}(X_{(J_n+1)}) \leq \frac{1}{\tau}\}$,

$$\mathbb{P}\left\{|\widehat{\gamma}(k) - \mathbb{E}\left[\widehat{\gamma}(k) \mid X_{(J_n+1)}\right]| \geq 2\frac{\gamma + 2\overline{\eta}(\tau)}{\sqrt{k}}\left(\sqrt{4t} + \frac{t}{\sqrt{k}}\right) \mid X_{(J_n+1)}\right\} \leq 2e^{-t}.$$

Merci de votre attention !