Rappels

On rappelle les éléments suivants du cours de Probabilités, considérés comme acquis par la suite:

- **Loi (forte) des grands nombres**: Si $X_1, \dots, X_n$ sont des v.a. i.i.d. sur $\mathbb{R}$ satisfaisant $\mathbb{E}|X_1| < +\infty$, alors
$$\frac{1}{n}\sum_{i=1}^n X_i \underset{n\to+\infty}{\overset{\text{p.s.}}{\longrightarrow}} \mathbb{E}(X_1).$$
- **Théorème central limite**: Si $X_1, \dots, X_n$ sont des v.a. i.i.d. sur $\mathbb{R}$, satisfaisant $\mathbb{E}|X_1|^2 < +\infty$, alors
$$\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mathbb{E}(X_i)) \underset{n\to+\infty}{\rightsquigarrow} \mathcal{N}(0, \mathrm{Var}(X_1)).$$
On rappelle aussi les définitions équivalentes de la convergence en loi (notée $\rightsquigarrow$): $X_n \rightsquigarrow X$ ssi, de manière équivalente
    1. $\forall f \in C_b(\mathbb{R})$ $\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$.
    2. Si F_X est continue en t, $F_{X_n}(t) \to F_X(t)$ (fonctions de répartition).
    3. Pour tout $t \in \mathbb{R}$, $\varphi_{X_n}(t) \to \varphi_X(t)$ (fonctions caractéristiques).
- **Inégalité de Markov**: Si X est positive, alors, pour tout $t > 0$,
$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t}.$$

Exercices: outils de base en concentration

Exercice 1 (Autour de Byenaymé Tchebychev). — Soit X une variable réelle satisfaisant $\mathbb{E}(X^2) < +\infty$.

1. Montrer l'inégalité de Byenaymé Tchebychev: $\forall t > 0 \quad \mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}$.
2. Montrer l'inégalité de Cantelli: $\forall t > 0 \quad \mathbb{P}(X \geq \mathbb{E}(X) + t) \leq \frac{\mathrm{Var}(X)}{\mathrm{Var}(X)+t^2}$.
3. Comparer ces deux inégalités.
4. Montrer l'inégalité de Paley-Zygmund: en supposant de plus $X \geq 0$,
$$\forall t \in [0,1] \quad \mathbb{P}(X \geq t\mathbb{E}(X)) \geq (1-t)^2 \frac{\mathbb{E}(X)^2}{\mathbb{E}(X^2)}.$$

Solution 1. — 1. Pour $t > 0$, on a
$$\begin{aligned}\mathbb{P}(|X - \mathbb{E}(X)| \geq t) &= \mathbb{P}\left((X - \mathbb{E}(X))^2 \geq t^2\right) \\ &\leq \frac{\mathbb{E}\left((X - \mathbb{E}(X))^2\right)}{t^2} \qquad \text{(Markov).}\end{aligned}$$

2. Soit $t > 0$. On a, pour tout $u \geq 0$,

$$\begin{aligned}
\mathbb{P}\left(X - \mathbb{E}(X) \geq t\right) &= \mathbb{P}\left(X - \mathbb{E}(X) + u \geq t + u\right) \\
&\leq \mathbb{P}\left((X - \mathbb{E}(X) + u)^2 \geq (t+u)^2\right) \qquad (t + u \geq 0) \\
&\leq \frac{\mathbb{E}\left((X - \mathbb{E}(X) + u)^2\right)}{(t+u)^2} \qquad \text{(Markov)} \\
&\leq \frac{\mathrm{Var}(X) + u^2}{(t+u)^2} := f(u).
\end{aligned}$$

On vérifie que $f'(u) = 0 \Leftrightarrow u = \mathrm{Var}(X)/t$, et on obtient le résultat en choisissant cette valeur de u.

3. Pour $t > 0$, la majoration donnée par Cantelli de $\mathbb{P}(X \geq \mathbb{E}(X) + t)$ sera toujours meilleure que celle de Byenaymé Tchebychev (et a la bonne propriété d'être toujours majorée par 1). En revanche, si on veut borner la déviation des deux côtés en utilisant l'inégalité de Cantelli, on a d'une part

$$\begin{aligned}
\mathbb{P}\left(X \leq \mathbb{E}(X) - t\right) &= \mathbb{P}\left((-X) \geq \mathbb{E}(-X) + t\right) \\
&\leq \frac{\mathrm{Var}(X)}{\mathrm{Var}(X) + t^2},
\end{aligned}$$

et donc

$$\mathbb{P}\left(|X - \mathbb{E}(X)| \geq t\right) \leq 2\frac{\mathrm{Var}(X)}{\mathrm{Var}(X) + t^2}.$$

On remarque alors que

$$2\frac{\mathrm{Var}(X)}{\mathrm{Var}(X) + t^2} \leq \frac{\mathrm{Var}(X)}{t^2} \Leftrightarrow t^2 \leq \mathrm{Var}(X).$$

Or, pour $t^2 \leq \mathrm{Var}(X)$, les deux bornes sont supérieures à 1 (donc triviales). On en déduit que l'inégalité de Byenaymé Tchebychev est toujours meilleure que celle de Cantelli pour borner les déviations bilatères de manière non triviale.

4. On part de

$$\begin{aligned}
\mathbb{E}(X) &= \mathbb{E}(X\mathbb{1}_{X \leq t\mathbb{E}(X)}) + \mathbb{E}(X\mathbb{1}_{X > t\mathbb{E}(X)}) \\
&\leq t\mathbb{E}(X) + \mathbb{E}(X\mathbb{1}_{X > t\mathbb{E}(X)}).
\end{aligned}$$

On borne alors le deuxième terme via

$$\mathbb{E}(X\mathbb{1}_{X > t\mathbb{E}(X)}) \leq \mathbb{E}(X^2)^{\frac{1}{2}}(\mathbb{P}(X > t\mathbb{E}(X)))^{\frac{1}{2}} \qquad \text{(Cauchy Schwarz)},$$

ce qui permet de conclure.

Exercice 2 (Inégalité de Hoeffding). — On va commencer par prouver le Lemme de Hoeffding: si Z est une v.a. telle que $Z \in [a, b]$ p.s. et $\mathbb{E}(Z) = 0$, alors

$$\forall t \in \mathbb{R} \quad \psi_Z(t) \leq \frac{(b-a)^2 t^2}{8},$$

où $\psi_Z(t) := \log(\mathbb{E}(\exp(tZ)))$. Dans la suite on se donne une telle v.a. Z.

1. Montrer que ψ_Z est bien définie.

2. Montrer que, si Y est à valeurs dans $[a,b]$, alors

$$\mathrm{Var}(Y) \leq \frac{(b-a)^2}{4}.$$

3. En déduire que, pour tout $t \in \mathbb{R}$, $\psi_Z''(t) \leq (b-a)^2/4$.

4. Conclure.

Dans une deuxième partie on va prouver l'inégalité de Hoeffding. Soient $X_1, \dots, X_n$ des variables aléatoires indépendantes telles que $X_i \in [a_i, b_i]$ p.s. et $\mathbb{E}(X_i) = 0$. On note $S = \sum_{i=1}^n X_i$.

5. Montrer que, pour tout $\varepsilon \geq 0$ et $t \in \mathbb{R}$,

$$\mathbb{P}\left(S \geq \varepsilon\right) \leq \exp\left(-t\varepsilon + \psi_S(t)\right).$$

6. En choisissant soigneusement t, montrer que, pour tout $\varepsilon \geq 0$,

$$\mathbb{P}\left(S \geq \varepsilon\right) \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),$$

ce qui constitue l'inégalité de Hoeffding.

7. Application: si $X_1, \dots, X_n$ sont i.i.d. de loi $\mathcal{B}(p)$, montrer que, pour tout $\varepsilon \geq 0$,

$$\mathbb{P}\left(|\bar{X}_n - p| \geq \varepsilon\right) \leq 2\exp\left(-2n\varepsilon^2\right).$$

Solution 2. — 1. Z étant bornée par $M = |a| \vee |b|$, $\exp(tZ)$ est bornée par $\exp(tM)$, et est donc intégrable. Par ailleurs $\mathbb{E}(\exp(tZ)) > 0$, ψ_Z est donc bien définie.

2. Soit Y à valeurs dans $[a,b]$. On a

$$\begin{aligned}\mathrm{Var}(Y) &\leq \mathbb{E}\left(Y - \frac{a+b}{2}\right)^2 \\ &\leq \left(\frac{b-a}{2}\right)^2 \qquad (a \leq Y \leq b).\end{aligned}$$

3. On commence par remarquer que $f : (t,u) \mapsto \exp(tu)$ est dérivable en t, de dérivée $u\exp(tu)$ qui est majorée par $M\exp(t_0 M) \in L_1(P_Z)$ P_Z presque sûrement, pour tout $t \in [-t_0, t_0]$. ψ_Z est donc dérivable sur $\mathbb{R}$, de dérivée

$$\psi_Z'(t) = \frac{\mathbb{E}(Z\exp(tZ))}{\mathbb{E}(\exp(tZ))}.$$

En utilisant des arguments similaires, on peut montrer que ψ_Z' est dérivable sur $\mathbb{R}$, avec

$$\begin{aligned}\psi_Z''(t) &= \frac{\mathbb{E}(Z^2\exp(tZ))}{\mathbb{E}(\exp(tZ))} - \left(\frac{\mathbb{E}(Z\exp(tZ))}{\mathbb{E}(\exp(tZ))}\right)^2 \\ &= \mathrm{Var}(Y),\end{aligned}$$

où

$$\frac{dP_Y}{dP_Z}(x) = \frac{\exp(tx)}{\mathbb{E}(\exp(tZ))}.$$

Comme $P_Y \ll P_Z$, on en déduit $Y \in [a,b]$ p.s., et la question 2 donne alors

$$\psi_Z''(t) \leq \frac{(b-a)^2}{4}.$$

4. Comme $\psi'_Z(0) = \mathbb{E}(Z) = 0$, on déduit que, pour $t \geq 0$,

$$\psi'_Z(t) \leq t\frac{(b-a)^2}{4}.$$

Comme $\psi_Z(0) = 0$, on déduit, pour $t \geq 0$,

$$\psi_Z(t) \leq \frac{t^2(b-a)^2}{8}.$$

Pour traiter le cas $t \leq 0$, il suffit de remarquer que $\psi_Z(t) = \psi_{-Z}(-t)$ et que $-Z$ satisfait les mêmes hypothèses que Z.

5. Soient $\varepsilon \geq 0$ et $t \in \mathbb{R}$. On a

$$\begin{aligned}\mathbb{P}(S \geq \varepsilon) &= \mathbb{P}(\exp(tS) \geq \exp(t\varepsilon)) \\ &\leq \exp(-t\varepsilon)\mathbb{E}(\exp(tS)) \qquad \text{(Markov)} \\ &= \exp(\psi_S(t) - t\varepsilon).\end{aligned}$$

6. Les $X_1, \dots, X_n$ étant indépendants, on a

$$\begin{aligned}\psi_S(t) &= \log\left(\mathbb{E}\exp\left(\sum_{i=1}^n tX_i\right)\right) \\ &= \sum_{i=1}^n \psi_{X_i}(t) \\ &\leq \frac{t^2 v}{2},\end{aligned}$$

avec $v = \frac{1}{4}\sum_{i=1}^n (b_i - a_i)^2$. On en déduit

$$\log\left(\mathbb{P}(S \geq \varepsilon)\right) \leq \frac{t^2 v}{2} - t\varepsilon.$$

Le terme de droite est minimal en $t_* = \frac{\varepsilon}{v}$, ce qui mène à

$$\begin{aligned}\mathbb{P}(S \geq \varepsilon) &\leq \exp\left(\frac{t_*^2 v}{2} - t_*\varepsilon\right) \\ &\leq \exp\left(-\frac{\varepsilon^2}{2v}\right).\end{aligned}$$

7. En notant $Y_i = X_i - \mathbb{E}(X_i)$ et $S = \sum_{i=1}^n Y_i$, on a

$$\begin{aligned}\mathbb{P}\left(|\bar{X}_n - p| \geq \varepsilon\right) &= \mathbb{P}\left(S \geq n\varepsilon\right) + \mathbb{P}\left(S \leq -n\varepsilon\right) \\ &\leq 2\exp\left(-\frac{n^2\varepsilon^2}{2v}\right),\end{aligned}$$

avec $v = n/4$, ce qui donne le résultat.

Exercice: un résultat utile de convergence

Exercice 3 (Convergence des quantiles). — Soit X v.a. sur $\mathbb{R}$ de fonction de répartition F. Pour $p \in [0,1]$, le **quantile** d'ordre p de X est défini par

$$q(p) = \inf \{t \in \mathbb{R} \mid F(t) \geq p\}.$$

Soient $X_1, \ldots, X_n$ i.i.d. de même loi que X. On se donne $p \in [0,1]$. On note q_n le quantile empirique $q_n(p)$ (associé à la fonction de répartition empirique $F_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i \leq t}$). On suppose que F est strictement croissante en $q(p)$, que l'on abrègera en q.

1. Soit $u \in \mathbb{R}$. Montrer que
$$q \leq u \Leftrightarrow F(u) \geq p.$$
2. Soit $\varepsilon > 0$. Montrer que
$$\mathbb{P}(q_n \leq q - \varepsilon) \leq \mathbb{P}\left(\sum_{i=1}^n \mathbb{1}_{X_i \leq q-\varepsilon} \geq np\right).$$
3. En déduire l'existence de $\delta_-(\varepsilon) > 0$ tel que
$$\mathbb{P}(q_n \leq q - \varepsilon) \leq \exp(-2n\delta_-(\varepsilon)^2).$$
4. Soit $\varepsilon > 0$. Montrer que
$$\mathbb{P}(q_n > q + \varepsilon) \leq \mathbb{P}\left(\sum_{i=1}^n \mathbb{1}_{X_i \leq q+\varepsilon} < np\right).$$
5. En déduire l'existence de $\delta_+(\varepsilon) > 0$ tel que
$$\mathbb{P}(q_n > q + \varepsilon) \leq \exp(-2n\delta_+(\varepsilon)^2).$$
6. En déduire que $q_n \xrightarrow[n\to+\infty]{\text{p.s.}} q$.

Solution 3. — 1. Par définition de q on a $F(u) \geq p \Rightarrow u \geq q$. Ensuite, F étant cadlàg, $F(q) \geq p$. On en déduit que si $u \geq q$, $F(u) \geq F(q) \geq p$, par croissance de F.

2. D'après la question 1,
$$q_n \leq u \Leftrightarrow F_n(u) \geq p.$$
On en déduit
$$\begin{aligned}\mathbb{P}(q_n \leq q - \varepsilon) &= \mathbb{P}(F_n(q-\varepsilon) \geq p)\\ &\leq \mathbb{P}\left(|\{i \mid X_i \leq q-\varepsilon\}| \geq np\right)\\ &= \mathbb{P}\left(\sum_{i=1}^n \mathbb{1}_{X_i \leq q-\varepsilon} \geq np\right).\end{aligned}$$

3. On a

$$\mathbb{P}\left(\sum_{i=1}^{n} \mathbb{1}_{X_i \le q-\varepsilon} \ge np\right) = \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbb{1}_{X_i \le q-\varepsilon} - F(q-\varepsilon)) \ge p - F(q-\varepsilon)\right)$$
$$\le \exp(-2n\delta_-(\varepsilon)^2),$$

en utilisant l'inégalité de Hoeffding, avec $\delta_-(\varepsilon) = p - F(q-\varepsilon)$. Par définition de q, on a $F(q-\varepsilon) < p$, et donc $\delta_-(\varepsilon) > 0$.

4. Pour $\varepsilon > 0$, on a

$$\mathbb{P}(q_n > q+\varepsilon) = \mathbb{P}(F_n(q+\varepsilon) < p)$$
$$= \mathbb{P}\left(\sum_{i=1}^{n} \mathbb{1}_{X_i < q+\varepsilon} < np\right).$$

5. On a

$$\mathbb{P}\left(\sum_{i=1}^{n} \mathbb{1}_{X_i < q+\varepsilon} < np\right) \le \mathbb{P}\left(\sum_{i=1}^{n} \mathbb{1}_{X_i \le q+\varepsilon/2} < np\right)$$
$$= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbb{1}_{X_i \le q+\varepsilon/2} - F(q+\varepsilon/2)) < p - F(q+\varepsilon/2)\right)$$
$$\le \exp(-2n\delta_+(\varepsilon)^2),$$

en utilisant l'inégalité de Hoeffding, avec $\delta_+(\varepsilon) = F(q+\varepsilon/2) - p$. Comme F est strictement croissante en q, $\delta_+(\varepsilon) > 0$.

6. Pour $\varepsilon > 0$, le théorème de Borel-Cantelli donne

$$\mathbb{P}\left(\limsup_n \{|q_n - q| > \varepsilon\}\right) = 0,$$

ce dont on déduit

$$\limsup_n |q_n - q| \le \varepsilon,$$

p.s., pour tout $\varepsilon > 0$, et donc $q_n \to q$ p.s. (prendre $\varepsilon = 1/k$).