Inégalité de Hoeffding

↳ Réf : Bernis & Bernis p216-224.

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé.

Lemme : Soit X une variable aléatoire réelle, centrée, bornée presque sûrement par 1, alors $\forall t \in \mathbb{R}$, $L_X(t) = \mathbb{E}(e^{tX}) \leq e^{t^2/2}$.

Théorème : (Inégalité de Hoeffding) : Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires réelles, indépendantes, centrées. On suppose de plus que les $(X_n)_{n\in\mathbb{N}^*}$ sont bornées ps : $\forall n \in \mathbb{N}^*, \exists C_n > 0$, $|X_n| \leq C_n$ ps. Alors en notant pour $n \in \mathbb{N}^*$, $S_n = \sum_{k=1}^{n} X_k$, on a pour tout $\varepsilon > 0$, pour tout $n \in \mathbb{N}^*$, $\mathbb{P}(|S_n| > \varepsilon) \leq 2\exp\left(\frac{-\varepsilon^2}{2\sum_{k=1}^{n} C_k^2}\right)$

261-264

Application 1 : Soit $n \in \mathbb{N}^*$. On étudie le modèle statistique $(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), \mathcal{B}(p)_{p\in]0,1[})$. Soit $X_1, \dots X_n$ un n-échantillon de loi $\mathcal{B}(p)$ pour $p \in]0,1[$. Soit $\alpha \in]0,1[$, alors un intervalle de confiance de niveau $1-\alpha$ pour p est :

$$I_{1-\alpha} = \left[\frac{1}{n}S_n - \sqrt{\frac{2}{n}\ln\left(\frac{2}{\alpha}\right)}, \frac{1}{n}S_n + \sqrt{\frac{2}{n}\ln\left(\frac{2}{\alpha}\right)}\right].$$

262-266

Application 2 : Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de VA réelles, indépendantes, centrées. On suppose toujours que $\forall n \in \mathbb{N}^*, \exists C_n > 0$, $|X_n| \leq C_n$ ps. Soit $\alpha > 0$. On suppose de plus qu'il existe $\beta > 0$ tel que pour tout $n \in \mathbb{N}^*$, $\sum_{k=1}^{n} C_k^2 \leq n^{2\alpha-\beta}$. Alors $\frac{S_n}{n^\alpha} \xrightarrow[n\to+\infty]{ps} 0$.

Preuve du lemme : Soit $x \in [-1,1]$, alors pour tout $t \in \mathbb{R}$, on a $\frac{1-x}{2} + \frac{1+x}{2} = 1$, $\frac{1+x}{2} \in [0,1]$ et $\frac{1-x}{2} \times (-t) + \frac{1+x}{2} t = tx$, donc par convexité de l'exponentielle, on en déduit que pour tout $t \in \mathbb{R}$,

$$\exp(tx) \leq \frac{1-x}{2} e^{-t} + \frac{1+x}{2} e^{t}. \quad (*)$$

Puisque X est bornée p.s, pour tout $t \in \mathbb{R}$, tX est bornée p.s, donc $\exp(tX)$ l'est également et elle est donc intégrable. Cela assure l'existence de la transformée de Laplace sur $\mathbb{R}$ tout entier. Soit $t \in \mathbb{R}$, alors

$$L_X(t) = \mathbb{E}(\exp(tX)) \leq \frac{1}{2}e^{-t}\mathbb{E}(1-X) + \frac{1}{2}e^{t}\mathbb{E}(1+X)$$ par $(*)$ et linéarité de l'espérance

$$\leq \frac{1}{2}e^{-t} + \frac{1}{2}e^{t} \quad \text{car } \mathbb{E}(X) = 0$$

$$\leq \cosh(t).$$

Or pour tout $k \in \mathbb{N}^*$, on a,

$2^k k! = 2 \times 4 \times \ldots \times 2(k-1) \times 2k \leq (2k)!$, donc pour tout $t \in \mathbb{R}$

$\cosh(t) = \sum_{k=0}^{+\infty} \frac{t^{2k}}{(2k)!} \leq \sum_{k=0}^{+\infty} \frac{t^{2k}}{2^k k!} = e^{t^2/2}$, et ainsi

$\mathbb{E}(e^{tX}) \leq e^{t^2/2}$.

Preuve du théorème : Soit $n \in \mathbb{N}^*$. Puisque pour tout $k \in \mathbb{N}^*$, la variable $\frac{X_k}{c_k}$ est bornée par 1 ps et centrée, on peut utiliser le lemme et donc $\forall t \in \mathbb{R}$, $\mathbb{E}(\exp(t\frac{X_k}{c_k})) \leq e^{t^2/2}$.

Ainsi, on obtient pour $t \in \mathbb{R}$:

$L_{S_n}(t) = \mathbb{E}(\exp(t\sum_{k=1}^{n} X_k)) = \mathbb{E}(\prod_{k=1}^{n} \exp(tX_k))$

$= \prod_{k=1}^{n} \mathbb{E}(\exp(t c_k \frac{X_k}{c_k}))$ par indépendance des $(\exp(tX_k))_{k \in \mathbb{N}^*}$

$\leq \prod_{i=1}^{n} \exp(\frac{t^2 c_k^2}{2}) = \exp(\frac{t^2}{2}\sum_{k=1}^{n} c_k^2)$.

Maintenant que nous avons obtenu une majoration sur $\mathbb{E}(e^{tS_n})$, nous allons utiliser l'inégalité de Markov pour obtenir une inégalité de la forme souhaitée.

Soit $\varepsilon > 0$. Puisque l'exponentielle est strictement croissante, on a $\{S_n > \varepsilon\} \subset \{\exp(tS_n) > \exp(t\varepsilon)\}$ (et même égalité mais ce n'est pas nécessaire ici) pour tout $t \in \mathbb{R}_+^*$.

Pour tout $t \in \mathbb{R}_+^*$, la variable $\exp(tS_n)$ est positive, on peut donc lui appliquer l'inégalité de Markov et ainsi :

$\mathbb{P}(S_n > \varepsilon) \leq \mathbb{P}(\exp(tS_n) > \exp(t\varepsilon)) \leq \frac{\mathbb{E}(\exp(tS_n))}{\exp(t\varepsilon)}$

$\leq \exp(\frac{t^2}{2}\sum_{k=1}^{n} c_k^2 - t\varepsilon)$

Nous allons maintenant optimiser l'inégalité pour obtenir la meilleure borne possible. On note $a = \frac{1}{2}\sum_{k=1}^{n} c_k^2$, on cherche donc à optimiser la fonction $t \mapsto at^2 - \varepsilon t$. Cette fonction admet son minimum en $t = \frac{\varepsilon}{2a} > 0$, qui vaut $a\frac{\varepsilon^2}{4a^2} - \frac{\varepsilon^2}{2a} = \frac{-\varepsilon^2}{4a} = -\frac{\varepsilon^2}{2\sum_{k=1}^{n} c_k^2}$. Ainsi,

$\mathbb{P}(S_n > \varepsilon) \leq \exp\left(\frac{-\varepsilon^2}{2\sum^{n} c_k^2}\right)$.

Enfin, la suite $(-X_n)_{n\in\mathbb{N}}$ vérifie les mêmes hypothèses que $(X_n)_{n\in\mathbb{N}}$ donc on a également $\mathbb{P}(-S_n>\varepsilon) \leq \exp\left(\frac{-\varepsilon^2}{2\sum_{k=1}^{n} c_k^2}\right)$. Or

$$\mathbb{P}(|S_n|>\varepsilon) = \mathbb{P}(\{S_n>\varepsilon\} \sqcup \{S_n<-\varepsilon\})$$
$$= \mathbb{P}(S_n>\varepsilon) + \mathbb{P}(S_n<-\varepsilon)$$
$$= \mathbb{P}(S_n>\varepsilon) + \mathbb{P}(-S_n>\varepsilon).$$

Et ainsi on obtient $\boxed{\mathbb{P}(|S_n|>\varepsilon) \leq 2\exp\left(\frac{-\varepsilon^2}{2\sum_{k=1}^{n} c_k^2}\right).}$

<u>Preuve de l'application 1</u> : Soit $\alpha\in]0,1[$. On applique l'inégalité de Hoeffding à la suite $(X_n-p)_{n\in\mathbb{N}^*}$ qui est bornée p.s par 1, alors pour $\varepsilon>0$, on a $\mathbb{P}\left(\left|\frac{S_n}{n}-p\right|>\varepsilon\right) = \mathbb{P}\left(\left|\sum_{k=1}^{n}(X_k-p)\right|>n\varepsilon\right)$

$$\leq 2\exp\left(-\frac{n\varepsilon^2}{2}\right).$$

On pose $\varepsilon_\alpha = \sqrt{\frac{2}{n}\ln\left(\frac{2}{\alpha}\right)}$, de sorte à avoir $\alpha = 2\exp\left(-\frac{n\varepsilon_\alpha^2}{2}\right)$.

Alors $\mathbb{P}\left(\left|\frac{S_n}{n}-p\right|\leq\varepsilon_\alpha\right) = 1-\mathbb{P}\left(\left|\frac{S_n}{n}-p\right|>\varepsilon_\alpha\right) \geq 1-\alpha$.

Donc $I_{1-\alpha} = \left[\frac{S_n}{n}-\varepsilon_\alpha, \frac{S_n}{n}+\varepsilon_\alpha\right]$ est un intervalle de confiance de niveau $1-\alpha$ pour p.

<u>Preuve de l'application 2</u> : Soit $\varepsilon>0$ et $n\in\mathbb{N}^*$. L'inégalité de Hoeffding nous donne $\mathbb{P}(|S_n|>n^\alpha\varepsilon) \leq 2\exp\left(-\frac{n^{2\alpha}\varepsilon^2}{2\sum_{k=1}^{n}c_k^2}\right) \leq 2\exp\left(-\frac{n^\beta\varepsilon^2}{2}\right)$

Or $\exp\left(-\frac{n^\beta\varepsilon^2}{2}\right) = \underset{n\to+\infty}{o}\left(\frac{1}{n^2}\right)$ donc par critère de comparaison des séries à termes positifs, on en déduit que la série $\sum_{n\geq1}\mathbb{P}(|S_n|>n^\alpha\varepsilon)$ converge. (Si manque de temps on peut conclure dès maintenant, c'est presque du cours*) Par le lemme de Borel Cantelli, $\mathbb{P}\left(\limsup_{n\to+\infty}\left\{\left|\frac{S_n}{n^\alpha}\right|>\varepsilon\right\}\right)=0$ (et ce pour tout $\varepsilon>0$ donc). (*en invoquant tout de même B.C. et il faut savoir faire !)

En particulier, pour tout $\varepsilon\in\mathbb{Q}_+^*$,

$$1 = \mathbb{P}\left(\liminf_{n\to+\infty}\left\{\left|\frac{S_n}{n^\alpha}\right|\leq\varepsilon\right\}\right) = \mathbb{P}\Big(\underbrace{\bigcup_{n\in\mathbb{N}^*}\bigcap_{m\geq n}\left\{\left|\frac{S_m}{m^\alpha}\right|\leq\varepsilon\right\}}_{F_\varepsilon \text{ (de mesure 1)}}\Big)$$

On pose $F = \bigcap_{\varepsilon\in\mathbb{Q}_+^*} F_\varepsilon$, alors F est de mesure 1 car $\mathbb{Q}$ est dénombrable, et $\forall\omega\in F, \forall\varepsilon\in\mathbb{Q}_+^*, \exists n\in\mathbb{N}^*, \forall m\geq n, \left|\frac{S_m(\omega)}{m^\alpha}\right|\leq\varepsilon$, d'où $\frac{S_n}{n^\alpha}\xrightarrow[n\to+\infty]{p.s.}0$.

<u>Remarques</u> : * On peut aussi utiliser l'inégalité de Hoeffding à la méthode de Monte-Carlo : sur $I = [0,1]^d$, avec $f \in L^1(I)$ et bornée sur I, on considère $X_n = f(Y_n) - \int_I f$ où $(Y_n)_{n\in\mathbb{N}}$ iid de loi $\mathcal{U}(I)$, alors $\mathbb{E}(X_n) = 0$ et $|X_n| \leq 2\|f\|_\infty$, donc par Hoeffding, on obtient

$$\forall \varepsilon > 0, \quad \mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n} f(Y_k) - \int_I f\right| > \varepsilon\right) \leq 2\exp\left(\frac{-n\varepsilon^2}{8\|f\|_\infty^2}\right)$$

* Concernant l'application 1 : on peut comparer le résultat obtenu avec celui obtenu grâce à Bienaymé-Tchebychev : $\left[\frac{S_n}{n} - \varepsilon'_\alpha, \frac{S_n}{n} + \varepsilon'_\alpha\right]$ où $\varepsilon'_\alpha = \frac{1}{2\sqrt{n\alpha}}$, si on étudie $\varepsilon_\alpha - \varepsilon'_\alpha$, on voit que cette quantité est négative pour $\alpha \in]0;0{,}0297[$* donc sur cet intervalle, l'inégalité de Hoeffding fournit un meilleur intervalle que celle de BT. Cependant pour des valeurs plus grandes (mais classiques comme $\alpha = 0{,}05$) c'est Tchebychev qui est plus performant. *pour $n = 1000$

* Concernant l'application 2 : l'inégalité ne nécessite pas que les variables soient identiquement distribuées, ce qui constitue un avantage face à la loi des grands nombres classiques ! Par exemple, dès que l'on prend une suite de VA indépendantes centrées bornées par 1, on obtient le résultat pour $\alpha > \frac{1}{2}$ (si les variables sont iid, le TCL nous dit qu'il n'y a pas de convergence PS pour $\alpha = \frac{1}{2}$).

* L'inégalité de Hoeffding fait partie de la grande famille des inégalités de concentration. Citons-en quelques autres : Gamer-Chernoff*, Bennett, Bernstein ⇒ cf cours de stat. Ces inégalités sont par exemple très utiles pour obtenir des intervalles de confiance <u>non asymptotiques</u>, contrairement au TCL qui fournit généralement des <u>intervalles asymptotiques</u>

*c'est globalement ce qu'on applique dans la preuve du thm

* Il existe une généralisation de cette inégalité : l'inégalité d'Azuma (aussi appelée Azuma-Hoeffding...) : Soit $(M_n)_{n\in\mathbb{N}}$ une martingale issue de 0 dont les accroissements sont bornés ps : $\forall n \in \mathbb{N}^*, \exists c_n > 0, |M_n - M_{n-1}| \leq c_n$ ps

Alors pour tout $\varepsilon > 0$, $\mathbb{P}(|M_n| \geq \varepsilon) \leq 2\exp\left(\frac{-\varepsilon^2}{2\sum_{k=1}^{n} c_k^2}\right)$.

⇒ cf Cadre & Vial ([illegible])

* L'inégalité de Hoeffding peut être raffinée, en supposant $\forall n \in \mathbb{N}^*$, $\exists a_n, b_n$ tq $a_n \leq X_n \leq b_n$ $(a_n < b_n)$, alors $\mathbb{P}(|S_n| > \varepsilon) \leq 2\exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$ et ça permet par exemple dans le cadre de l'applicat° 1 d'obtenir des bornes encore meilleures $\left(\varepsilon''_\alpha = \sqrt{\frac{1}{2n}\ln\left(\frac{2}{\alpha}\right)}\right)$, et meilleures encore que BT pour $\alpha \in]0;0{,}232[$! (toujours pour $n = 1000$)