Estimation des grands écarts

Références : [Les01] p.16-20 ([Ouv07] et [Gou08]).

Théorème 0.1 *Soit $p \in]0,1[$. Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires indépendantes de même loi de Bernoulli de paramètre p. Soit la variable aléatoire $S_n = \sum_{i=1}^n X_i \sim \mathcal{B}(n,p)$.*
On pose pour $\epsilon \in]0, 1-p[$,

$$h_+(\epsilon) = (p+\epsilon)\log\left(\frac{p+\epsilon}{p}\right) + (1-p-\epsilon)\log\left(\frac{1-p-\epsilon}{1-p}\right)$$

Alors :

1. *$h_+(\epsilon) > 0$;*
2. *pour tout $n \geq 1$,*
$$\mathbb{P}\left(\frac{S_n}{n} \geq p+\epsilon\right) \leq e^{-nh_+(\epsilon)}$$
3. *et cette inégalité est optimale, ie*
$$\lim_{n\longrightarrow+\infty} \frac{1}{n}\log\mathbb{P}\left(\frac{S_n}{n} \geq p+\epsilon\right) = -h_+(\epsilon)$$

Démonstration

Étape 1 Montrons le (1) et le (2).
Soit $t > 0$ fixé. Déterminons la probabilité suivante :

$$\begin{aligned}\mathbb{P}\left(\frac{S_n}{n} \geq p+\epsilon\right) &= \mathbb{P}(S_n \geq n(p+\epsilon)) \\ &= \mathbb{P}(S_n - n(p+\epsilon) \geq 0) \\ &= \mathbb{P}(t(S_n - np - n\epsilon) \geq 0) \\ &= \mathbb{P}(e^{t(S_n-np-n\epsilon)} \geq 1)\end{aligned}$$

par bijectivité de l'exponentielle. Ainsi d'après l'inégalité de Markov, on a :

$$\begin{aligned}\mathbb{P}\left(\frac{S_n}{n} \geq p+\epsilon\right) &\leq \mathbb{E}(e^{t(S_n-np-n\epsilon}) \\ &= e^{-nt(p+\epsilon)}\mathbb{E}(e^{tS_n}) \\ &= e^{-nt(p+\epsilon)}\sum_{k=1}^n e^{tk}\mathbb{P}(S_n = k) \quad \text{d'après le lemme de transfert ;} \\ &= e^{-nt(p+\epsilon)}\sum_{k=1}^n e^{tk}C_n^k p^k(1-p)^{n-k}\end{aligned}$$

La formule du binôme de Newton nous donne alors :

$$\begin{aligned}\mathbb{P}\left(\frac{S_n}{n} \geq p+\epsilon\right) &\leq e^{-nt(p+\epsilon)}(1-p+pe^t)^n \\ &= e^{-nt(p+\epsilon)-n\log(1-p+pe^t)}\end{aligned}$$

Cela étant vrai $\forall t > 0$, on obtient finalement :

$$\mathbb{P}\left(\frac{S_n}{n} \geq p + \epsilon\right) \leq e^{-nh}$$

où $h = \sup_{t>0}\{t(p+\epsilon) - \log(1 - p + pe^t)\}$.
Étudions alors la fonction :

$$\begin{aligned} g : \mathbb{R}_+^* &\longrightarrow \mathbb{R} \\ t &\longmapsto t(p+\epsilon) - \log(1 - p + pe^t) \end{aligned}$$

On remarque que $g(0) = 0$ et que g est dérivable, de dérivée :

$$g'(t) = p + \epsilon - \frac{pe^t}{1 - p + pe^t}$$

d'où $g'(0) = \epsilon > 0$, ce qui prouve que la borne supérieure de g est strictement positive.
Cherchons le point où la dérivée s'annule :

$$\begin{aligned} g'(t) = 0 &\Leftrightarrow p + \epsilon - \frac{pe^t}{1 - p + pe^t} = 0 \\ &\Leftrightarrow p + \epsilon = \frac{pe^t}{1 - p + pe^t} \\ &\Leftrightarrow (p+\epsilon)(1 - p + pe^t) = pe^t \\ &\Leftrightarrow (p+\epsilon)(1-p) = pe^t(1 - p - \epsilon) \\ &\Leftrightarrow e^t = \frac{(p+\epsilon)(1-p)}{p(1-p-\epsilon)} \\ &\Leftrightarrow t = \log\left(\frac{(p+\epsilon)(1-p)}{p(1-p-\epsilon)}\right) \end{aligned}$$

On effectue la tableau de variation de g et on se rend compte que g est maximale en $t = \log\left(\frac{(p+\epsilon)(1-p)}{p(1-p-\epsilon)}\right)$, donc :

$$\begin{aligned} h &= \sup_{t>0}\{t(p+\epsilon) - \log(1 - p + pe^t)\} \\ &= \log\left(\frac{(p+\epsilon)(1-p)}{p(1-p-\epsilon)}\right)(p+\epsilon) - \log(1 - p + p\frac{(p+\epsilon)(1-p)}{p(1-p-\epsilon)}) \\ &= (p+\epsilon)\log\left(\frac{p+\epsilon}{p}\right) + (p+\epsilon)\log\left(\frac{1-p}{1-p-\epsilon}\right) - \log\left(1 - p + \frac{(p+\epsilon)(1-p)}{1-p-\epsilon}\right) \\ &= (p+\epsilon)\log\left(\frac{p+\epsilon}{p}\right) + (p+\epsilon)\log\left(\frac{1-p}{1-p-\epsilon}\right) - \log\left(\frac{(1-p)(1-p-\epsilon) + (p+\epsilon)(1-p)}{1-p-\epsilon}\right) \\ &= (p+\epsilon)\log\left(\frac{p+\epsilon}{p}\right) + (p+\epsilon)\log\left(\frac{1-p}{1-p-\epsilon}\right) - \log\left(\frac{1-p}{1-p-\epsilon}\right) \\ &= (p+\epsilon)\log\left(\frac{p+\epsilon}{p}\right) + (1-p-\epsilon)\log\left(\frac{1-p-\epsilon}{1-}\right) \\ &= h_+(\epsilon) \end{aligned}$$

Étape 2 Montrons l'assertion (3). L'estimation des grands écarts obtenue dans l'étape 1 s'écrit :

$$\mathbb{P}\left(\frac{S_n}{n} \geq p + \epsilon\right) \leq e^{-nh_+(\epsilon)}$$

Nous cherchons alors une minoration. On pose $k_n = \lceil n(p+\epsilon) \rceil$ le nombre entier immédiatement supérieur à $n(p+\epsilon)$. On a alors :

$$\mathbb{P}(S_n \geq n(p+\epsilon)) \geq \mathbb{P}(S_n = k_n)$$

et nous allons maintenant montrer que :

$$\lim_{n\longrightarrow+\infty}\frac{1}{n}\log\mathbb{P}(S_n=k_n)=-h_+(\epsilon)$$

car cela assurera que :

$$\liminf_{n\longrightarrow+\infty}\frac{1}{n}\mathbb{P}(S_n\geq n(p+\epsilon))\geq -h_+(\epsilon)$$

On a :

$$\mathbb{P}(S_n=k_n)=\frac{n!}{k_n!(n-k_n)!}p^{k_n}(1-p)^{n-k_n}$$

Ainsi en appliquant la formule de Stirling aux trois suites de factorielles, on obtient :

$$\begin{aligned}
\mathbb{P}(S_n=k_n) &\sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n\frac{1}{\sqrt{2\pi k_n}}\left(\frac{e}{k_n}\right)^{k_n}\frac{1}{\sqrt{2\pi(n-k_n)}}\left(\frac{e}{n-k_n}\right)^{n-k_n}p^{k_n}(1-p)^{n-k_n}\\
&\sim \sqrt{\frac{2\pi n}{(2\pi k_n)(2\pi(n-k_n))}}\left(\frac{n}{e}\right)^n\left(\frac{ep}{k_n}\right)^{k_n}\left(\frac{e(1-p)}{n-k_n}\right)^{n-k_n}\\
&\sim \sqrt{\frac{2\pi n}{4\pi^2k_n(n-k_n)}}\left(\frac{n}{e}\right)^{k_n}\left(\frac{n}{e}\right)^{n-k_n}\left(\frac{ep}{k_n}\right)^{k_n}\left(\frac{e(1-p)}{n-k_n}\right)^{n-k_n}\\
&\sim \sqrt{\frac{n}{2\pi k_n(n-k_n)}}\left(\frac{n}{e}\frac{ep}{k_n}\right)^{k_n}\left(\frac{n}{e}\frac{e(1-p)}{n-k_n}\right)^{n-k_n}\\
&\sim \frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{k_n(n-k_n)}}\left(\frac{np}{k_n}\right)^{k_n}\left(\frac{n(1-p)}{n-k_n}\right)^{n-k_n}
\end{aligned}$$

Comme k_n et $n-k_n$ divergent vers l'infini et $\mathbb{P}(S_n=k_n)\longrightarrow 0$ alors on peut passer au logarithme dans l'équivalence précédente, ie :

$$\begin{aligned}
\log\mathbb{P}(S_n=k_n) &\sim \log\left(\frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{k_n(n-k_n)}}\left(\frac{np}{k_n}\right)^{k_n}\left(\frac{n(1-p)}{n-k_n}\right)^{n-k_n}\right)\\
&\sim \log\left(\frac{1}{\sqrt{2\pi}}\right)+\log\left(\sqrt{\frac{n}{k_n(n-k_n)}}\right)+k_n\log\left(\frac{np}{k_n}\right)+(n-k_n)\log\left(\frac{n(1-p)}{n-k_n}\right)\\
&\sim -\frac{1}{2}\log(2\pi)+\frac{1}{2}\log\left(\frac{n}{k_n(n-k_n)}\right)+k_n\log\left(\frac{np}{k_n}\right)+(n-k_n)\log\left(\frac{n(1-p)}{n-k_n}\right)
\end{aligned}$$

Or par définition $k_n=\lceil n(p+\epsilon)\rceil$, donc $k_n=n(p+\epsilon)+o(n)$ et $n-k_n=n(1-p-\epsilon)+o(n)$, donc :

$$\frac{n}{k_n(n-k_n)}=\frac{n}{n(p+\epsilon)n(1-p-\epsilon)+o(n)}=\frac{1}{(p+\epsilon)n(1-p-\epsilon)+o(n)}\longrightarrow 0$$

Donc

$$\lim_{n\longrightarrow+\infty}\frac{1}{2n}\log\left(\frac{n}{k_n(n-k_n)}\right)\longrightarrow 0$$

Maintenant :

$$\begin{aligned}
k_n\log\left(\frac{np}{k_n}\right) &= n(p+\epsilon)\log\left(\frac{p}{p+\epsilon}\frac{n(p+\epsilon)}{k_n}\right)+(k_n-n(p+\epsilon))\log\left(\frac{np}{k_n}\right)\\
&= n(p+\epsilon)\log\left(\frac{p}{p+\epsilon}\right)+n(p+\epsilon)\log\left(\frac{n(p+\epsilon)}{k_n}\right)+(k_n-n(p+\epsilon))\log\left(\frac{np}{k_n}\right)
\end{aligned}$$

Ainsi en utilisant le fait que $k_n-n(p+\epsilon)$ reste borné (par 1), on en déduit que :

$$\lim_{n\longrightarrow+\infty}\frac{1}{n}k_n\log\left(\frac{np}{k_n}\right)=(p+\epsilon)\log\frac{p}{p+\epsilon}$$

De la même façon, on montre que :

$$\lim_{n\longrightarrow+\infty}\frac{1}{n}(n-k_n)\log\left(\frac{n(1-p)}{n-k_n}\right)=(1-p-\epsilon)\log\frac{1-p}{1-p-\epsilon}$$

Donc on en déduit que :

$$\frac{1}{n}\log(\mathbb{P}(S_n = k_n)) \longrightarrow -h_+(\epsilon)$$

Donc puisque $\mathbb{P}(S_n \geq n(p+\epsilon)) \geq \mathbb{P}(S_n = k_n)$, alors

$$\liminf_{n \longrightarrow +\infty} \frac{1}{n}\mathbb{P}(S_n \geq n(p+\epsilon)) \geq -h_+(\epsilon)$$

Or d'après le point (2) démontré dans l'étape 1, on a :

$$\limsup_{n \longrightarrow +\infty} \frac{1}{n}\mathbb{P}(S_n \geq n(p+\epsilon)) \leq -h_+(\epsilon)$$

D'où l'égalité voulue.

Lemmes utilisés

Lemme 0.1 (Inégalité de Markov) *Soit $X \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathbb{P})$ positive, alors pour tout $\epsilon > 0$:*

$$\mathbb{P}(X \geq \epsilon) \leq \frac{\mathbb{E}(X)}{\epsilon}$$

A fortiori, on a aussi :

$$\mathbb{P}(X > \epsilon) \leq \frac{\mathbb{E}(X)}{\epsilon}$$

Démonstration Si $X \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathbb{P})$ est positive, soit $D = \{X \geq \epsilon\}$. On a alors les minorations suivantes :

$$\begin{aligned}\mathbb{E}(X) &= \int_\Omega X d\mathbb{P} \\ &\geq \int_D X d\mathbb{P} \\ &\geq \epsilon \mathbb{P}(D)\end{aligned}$$

D'où le résultat. La deuxième inégalité résulte de l'inclusion $(X > \epsilon) \subset (X \geq \epsilon)$.

Lemme 0.2 (Formule de Stirling) *On considère la suite $(u_n)_{n\in\mathbb{N}^*}$ définie par $\forall n \in \mathbb{N}^*$:*

$$u_n = \frac{n^n e^{-n}\sqrt{n}}{n!}$$

Déterminons la nature de la série de terme général $v_n = \log\left(\frac{u_{n+1}}{u_n}\right)$ et déduisons-en l'existence d'un entier $k > 0$ tel que, quand $n \longrightarrow +\infty$:

$$n! \sim k\sqrt{n}n^n e^{-1}$$

et déterminons cette constante à l'aide de la formule de Wallis.

Démonstration Estimons v_n lorsque $n \longrightarrow +\infty$:

$$\begin{aligned}v_n &= \log\left(\left(\frac{n+1}{n}\right)^{n+\frac{1}{2}} e^{-1}\right) \\ &= -1 + \left(n + \frac{1}{2}\right)\log\left(1 + \frac{1}{n}\right) \\ &= -1 + \left(n + \frac{1}{2}\right)\left(\frac{1}{n} - \frac{1}{2n^2} + \mathcal{O}\left(\frac{1}{n^3}\right)\right) \\ &= \mathcal{O}\left(\frac{1}{n^2}\right)\end{aligned}$$

Cette expression nous montre la convergence de la série $\sum v_n$. Comme, on a $v_1+\ldots+v_n = \log u_{n+1} - \log u_1$ pour $n \geq 1$, alors la suite $(\log(u_n))_{n\in\mathbb{N}^*}$ converge. En notant λ sa limite, on voit que (u_n) converge vers $k = e^\lambda > 0$, d'où l'équivalent.
Il nous reste à déterminer la constante k. Nous allons pour ce faire utiliser la formule de Wallis, qui est :

$$\lim_{p\longrightarrow+\infty} \frac{1}{p}\left(\frac{(2p)(2p-2)\ldots 2}{(2p-1)(2p-3)\ldots 1}\right)^2 = \pi$$

Par ailleurs, en utilisant l'équivalent, on a lorsque $p \longrightarrow +\infty$:

$$\frac{1}{p}\left(\frac{(2p)(2p-2)\ldots 2}{(2p-1)(2p-3)\ldots 1}\right)^2 = \frac{1}{p}\left(\frac{2^{2p}(p!)^2}{(2p)!}\right)^2 \sim \frac{2^{4p}}{p}\frac{k^4 p^{4p+2}e^{-4p}}{k^2(2p)^{4p+1}e^{-4p}} = \frac{k^2}{2}$$

donc $\pi = \dfrac{k^2}{2}$, ie $k = \sqrt{2\pi}$. D'où le résultat.

Références

[Gou08] Xavier Gourdon. *Analyse.* Ellipses, 2008.

[Les01] Emmanuel Lesigne. *Pile ou face.* Ellipses, 2001.

[Ouv07] Jean-Yves Ouvrard. *Probabilités 1.* Cassini, 2007.