Théorème central limite

RIFFAUT Antonin, ROSTAM Salim

Avril 2014

Proposition 1. *Soit X une variable aléatoire de loi $\mathcal{N}(0,1)$. Sa fonction caractéristique est donnée par*

$$\varphi_X(t) = e^{-\frac{t^2}{2}}, \qquad \forall t \in \mathbb{R}.$$

Démonstration. La fonction φ_X est définie par

$$\varphi_X(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{itx-\frac{x^2}{2}}\,\mathrm{d}x, \qquad \forall t \in \mathbb{R}.$$

Elle est de classe $\mathcal{C}^1$ sur $\mathbb{R}$, par le théorème de dérivation sous le signe intégrale, appliqué à la fonction $g : (x,t) \in \mathbb{R}^2 \longmapsto e^{itx-\frac{x^2}{2}}$:

- pour tout $t \in \mathbb{R}$, $x \longmapsto g(x,t)$ est mesurable et intégrable ;
- pour tout $x \in \mathbb{R}$, $t \longmapsto g(x,t)$ est de classe $\mathcal{C}^1$, de dérivée
$$\frac{\partial g}{\partial t}(x,t) = ixe^{itx-\frac{x^2}{2}};$$
- pour tout $(x,t) \in \mathbb{R}^2$,
$$\left|\frac{\partial g}{\partial t}(x,t)\right| = |x|e^{-\frac{x^2}{2}} =: \psi(x),$$
avec ψ intégrable.

On peut donc écrire

$$\begin{aligned}
\varphi_X'(t) &= \frac{i}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \left(xe^{-\frac{x^2}{2}}\right) e^{itx}\,\mathrm{d}x \\
&= \frac{i}{\sqrt{2\pi}} \left\{ \left[-e^{-\frac{x^2}{2}} e^{itx}\right]_{-\infty}^{+\infty} + it\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}e^{itx}}\,\mathrm{d}x \right\} \\
&= -t\varphi_X(t),
\end{aligned}$$

d'où la formule annoncée (sachant que $\varphi_X(0) = 1$). ∎

Théorème 2 (TCL)**.** *Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires réelles i.i.d. de carré intégrable. Alors, en notant $S_n = X_1 + \cdots + X_n$, on a*

$$\frac{1}{\sqrt{n}} \frac{S_n - n\mathbb{E}[X_1]}{\sqrt{\mathrm{Var}(X_1)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Démonstration. Quitte à remplacer X_i par $\frac{X_i - \mathbb{E}[X_1]}{\sqrt{\mathrm{Var}(X_1)}}$, on peut supposer que $\mathbb{E}[X_1] = 0$ et que $\mathrm{Var}(X_1) = 1$. En vertu du théorème de Lévy, il s'agit alors de montrer que $\varphi_{\frac{S_n}{\sqrt{n}}}$ converge simplement vers la fonction $t \longmapsto e^{-\frac{t^2}{2}}$ sur $\mathbb{R}$.

Observons que

$$\varphi_{\frac{S_n}{\sqrt{n}}}(t) = \varphi_{X_1}\left(\frac{t}{\sqrt{n}}\right)^n, \qquad \forall t \in \mathbb{R};$$

d'autre part, comme X_1 admet des moments jusqu'à l'ordre 2, alors φ_{X_1} est deux fois dérivable en 0, et $\varphi'_{X_1}(0) = i\mathbb{E}[X_1] = 0$, $\varphi''_{X_1}(0) = -\mathbb{E}[X_1^2] = -1$; le développement limité de φ_{X_1} au voisinage de 0 s'écrit donc

$$\varphi_{X_1}(t) = 1 - \frac{t^2}{2} + o(t^2),$$

ce qui permet d'en déduire, à $t \in \mathbb{R}$ fixé, un développement asymptotique de $\varphi_{\frac{S_n}{\sqrt{n}}}(t)$:

$$\varphi_{\frac{S_n}{\sqrt{n}}}(t) = \left(1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)^n.$$

Pour n suffisamment grand, $1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right) \in B\left(1, \frac{1}{2}\right) \subset \mathbb{C}$, ce qui permet d'écrire, avec la détermination principale du logarithme sur $\mathbb{C} \setminus \mathbb{R}^{-*}$:

$$\begin{aligned}\varphi_{\frac{S_n}{\sqrt{n}}}(t) &= e^{n\log\left(1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)} \\ &= e^{-\frac{t^2}{2} + o(1)} \underset{n \to +\infty}{\longrightarrow} e^{-\frac{t^2}{2}}.\end{aligned}$$

Le théorème central limite est ainsi démontré. ■

Utilisation du TCL en statistique. Un institut de sondage interroge n personnes avant une élection : n_{A} répondent qu'elles voteront pour le candidat A et n_{B} répondent qu'elles voteront pour le candidat B. On désire savoir à quel point l'on peut se fier à ce sondage.

Ce que l'on cherche est idéalement le résultat de l'élection, c'est-à-dire la proportion $p \in]0, 1[$ de voix que va récolter A (si p vaut 0 ou 1 alors normalement on ne dépense pas d'argent pour faire un tel sondage). Chaque personne sondée est représentée par une variable aléatoire X_i : en supposant que la personne ne peut répondre que A ou B, la variable X_i suit donc une loi binomiale de paramètre p. De plus, on peut supposer que les variables $(X_i)_{1 \leqslant i \leqslant n}$ sont indépendantes.

On désire trouver un *intervalle de confiance* pour p, *i.e.* un intervalle $I \subseteq \mathbb{R}$ tel que $\mathbb{P}(p \in I) = 1 - \alpha$ pour un certain α fixé à l'avance (typiquement $\alpha = 5\%$). Pour déterminer I, on va utiliser le TCL : avec $\overline{X_n} := \frac{S_n}{n}$ et $G \sim \mathcal{N}(0, 1)$ on a, d'après le TCL :

$$Y_n := \sqrt{n}\frac{\overline{X_n} - p}{\sqrt{p(1-p)}} \xrightarrow[n \to \infty]{\mathcal{L}} G$$

donc d'après la caractérisation en termes de fonctions de répartition :

$$\forall a \in \mathbb{R}_+,\ \mathbb{P}(-a \leqslant Y_n \leqslant a) \underset{n \to \infty}{\longrightarrow} \Phi(a) - \Phi(-a) =: C_a$$

où Φ désigne la fonction de répartition de G. Ainsi, si a est choisi tel que $C_a = 1-\alpha$ on voit apparaître $\mathbb{P}(p \in E_n) \xrightarrow[n\to\infty]{} 1-\alpha$ pour $E_n \subseteq \mathbb{R}$ un certain ensemble ; pour obtenir un encadrement de p (*i.e.* transformer E_n en un intervalle), on va utiliser le lemme de Slutsky.

On sait que :

- (Y_n) converge en loi vers G ;
- $(\overline{X_n})$ converge en probabilité vers la *constante* p (d'après la loi faible des grands nombres) ;

donc par le lemme de Slutsky la suite des couples $((Y_n, \overline{X_n}))$ converge en loi vers (G, p). Comme l'on peut composer la convergence en loi avec des fonctions continues, on en déduit que la suite $\left(Y_n \frac{\sqrt{p(1-p)}}{\sqrt{\overline{X_n}(1-\overline{X_n})}}\right)_n$ converge en loi vers $G\frac{\sqrt{p(1-p)}}{\sqrt{p(1-p)}} = G$. Cela signifie en particulier que :

$$\forall a \in \mathbb{R}_+,\ \mathbb{P}(-a \leqslant \sqrt{n}\frac{\overline{X_n}-p}{\sqrt{\overline{X_n}(1-\overline{X_n})}} \leqslant a) \xrightarrow[n\to\infty]{} C_a$$

d'où, avec $I_{a,n} := \left[-a\sqrt{\frac{\overline{X_n}(1-\overline{X_n})}{n}} + \overline{X_n}, a\sqrt{\frac{\overline{X_n}(1-\overline{X_n})}{n}} + \overline{X_n}\right]$:

$$\forall a \in \mathbb{R}_+,\ \mathbb{P}(p \in I_{a,n}) \xrightarrow[n\to\infty]{} C_a$$

Reste à remarquer les choses suivantes :

- $\overline{X_n} = \frac{n_A}{n}$;
- on veut $C_a = 1-\alpha$ donc il faut choisir a tel que $\Phi(a) - \Phi(-a) = 1-\alpha$: comme la densité de G est paire, on a $\Phi(-a) = 1-\Phi(a)$ d'où $a = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$.

Remarque. L'intervalle $I_{a,n}$ trouvé est en fait un intervalle de confiance *asymptotique*, puisque $\mathbb{P}(p \in I_{a,n})$ tend (quand $n\to\infty$) vers $C_a = 1-\alpha$. Pour n assez grand, cette probabilité sera donc proche de $1-\alpha$.

Références

[BL] Ph. Barbe, M. Ledoux, *Probabilité*, Belin.