Théorème Central Limite pour les quantiles

Référence : *Probabilités pour les non probabilistes*, Walter Appel.
Leçons : 228 [1], 229[1], 241[1], 261 [2], 262[2].
Seule la démonstration du lemme 1 est dans la référence.

Lemme 1 (*de Dini*)

Soit $f_n : \mathbb{R} \to [0,1]$ une suite de fonctions croissantes convergeant simplement avec $\lim_{-\infty} f_n(x) = 0$ et $\lim_{+\infty} f_n(x) = 1$ vers une fonction f continue vérifiant les mêmes propriétés sur les limites alors la convergence est même uniforme.

Démonstration : La convergence simple implique f est encore croissante à valeur dans $[0,1]$.
Le TVI nous donne que $f(\overline{\mathbb{R}}) = [0,1]$.
Soit $\varepsilon > 0$ et $q \in \mathbb{N}$ tel que $\frac{1}{q} \leqslant \varepsilon$. On pose $\beta_k = \frac{k}{q}$, avec $k \in [\![0, q]\!]$.
Comme $\mathrm{Im} f = [0,1]$, il existe une suite $-\infty = a_0 < a_1 < \ldots < a_q = +\infty$ telle que $f(a_k) = \beta_k$ pour chaque k.
Maintenant comme il y a un nombre fini de k, on peut choisir un $N \in \mathbb{N}$ tel que

$$\forall n \geqslant N, \forall k \in [\![0, q]\!],\ |f_n(a_k) - f(a_k)| \leqslant \varepsilon$$

Soit $x \in \mathbb{R}$, on a $x \in [a_k, a_{k+1}]$ pour un certains k, et on a

$$\begin{cases} f(a_{k+1}) - \varepsilon \leqslant f(a_k) \leqslant f(x) \leqslant f(a_{k+1}) \leqslant f(a_k) + \varepsilon \\ f(a_k) - \varepsilon \leqslant f_n(a_k) \leqslant f_n(x) \leqslant f_n(a_{k+1}) \leqslant f(a_{k+1}) + \varepsilon \end{cases}$$

En soustrayant ces deux lignes, on obtient $-2\varepsilon \leqslant f(x) - f_n(x) \leqslant 2\varepsilon$. Donc $\|f - f_n\|_\infty \leqslant 2\varepsilon$. ■

Théorème 1

Soit $0 < p < 1$ et (X_n) une suite de variables aléatoires indépendantes et identiquement distribuées de fonction de répartition F. On suppose que F est dérivable en α_p le quantile d'ordre p (ie $\alpha_p = \inf\{t \in \mathbb{R} \mid F(t) \geqslant p\}$) avec $F'(\alpha_p) > 0$. Alors on a

$$\sqrt{n}(\widehat{\alpha_{p,n}} - \alpha_p) \Rightarrow \mathcal{N}\left(0, \frac{p(1-p)}{F'(\alpha_p)^2}\right)$$

où $\widehat{\alpha_{p,n}}$ est l'estimateur empirique du quantile d'ordre p (ie $\widehat{\alpha_{p,n}} = \inf\{t \in \mathbb{R} \mid F_n(t) \geqslant p\}$ avec $F_n(t) = \frac{1}{n}\sum \mathbf{1}_{X_i \leqslant t}$).

Démonstration : On note Φ la fonction de répartition de loi normale centrée réduite. Soit $t \in \mathbb{R}$ et $A > 0$ un réel que l'on choisira à la fin de la démonstration.
On pose $G_n(t) := \mathbb{P}\left(\frac{\sqrt{n}(\widehat{\alpha_{p,n}} - \alpha_p)}{A} \leqslant t\right) = \mathbb{P}\left(\widehat{\alpha_{p,n}} \leqslant \alpha_p + \frac{tA}{\sqrt{n}}\right)$. L'objectif est d'avoir $G_n(t) \to \Phi(t)$.
Or $\inf\{u \in \mathbb{R} \mid F(u) \geqslant p\} \leqslant x \Leftrightarrow F(x) \geqslant p$. Pour le voir, on utilise la continuité à droite de F.
Donc $G_n(t) = \mathbb{P}\left(p \leqslant F_n\left(\alpha_p + \frac{tA}{\sqrt{n}}\right)\right) = \mathbb{P}\left(np \leqslant \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leqslant \alpha_p + tA/\sqrt{n}\}}\right)$.
On définit la variable $Z_n := \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leqslant \alpha_p + tA/\sqrt{n}\}} \sim \mathrm{Bin}(n, \Delta_{n,t})$,

1. Démonstration du lemme 1 et théorème 1.
2. Démonstration du théorème 1, lemme 2 et de l'application.

où $\Delta_{n,t} = \mathbb{P}\left(X_1 \leqslant \alpha_p + \frac{tA}{\sqrt{n}}\right) = F\left(\alpha_p + \frac{tA}{\sqrt{n}}\right) \to F(\alpha_p) = p$ car F est continue en α_p.
Maintenant, on va modifier l'intérieur de la probabilité pour essayer de trouver un TCL.

$$\text{On écrit } G_n(t) = \mathbb{P}\left(\frac{Z_n - n\Delta_{n,t}}{\sqrt{n\Delta_{n,t}(1-\Delta_{n,t})}} \geqslant \frac{np - n\Delta_{n,t}}{\sqrt{n\Delta_{n,t}(1-\Delta_{n,t})}}\right)$$
$$= \mathbb{P}\left(\frac{Z_n - n\Delta_{n,t}}{\sqrt{n\Delta_{n,t}(1-\Delta_{n,t})}} \geqslant -\frac{\sqrt{n}(\Delta_{n,t} - p)}{\sqrt{\Delta_{n,t}(1-\Delta_{n,t})}}\right)$$

On pose $Z_n^* = \dfrac{Z_n - n\Delta_{n,t}}{\sqrt{n\Delta_{n,t}(1-\Delta_{n,t})}}$ et $c_{n,t} = \dfrac{\sqrt{n}(\Delta_{n,t}) - p}{\sqrt{\Delta_{n,t}(1-\Delta_{n,t})}}$.
On ne peut pas appliquer le TCL sur Z_n^* car c'est une somme de n variables iid **qui dépendent de** n. C'est pour cela qu'il faut travailler un peu.

Lemme 2

On a quand même $Z_n^* \Rightarrow \mathcal{N}(0,1)$.

Démonstration : On remarque que $\mathbf{1}_{\{X_i \leqslant \alpha_p + tA/\sqrt{n}\}} = \underbrace{\mathbf{1}_{\{X_i \leqslant \alpha_p\}}}_{:=Y_i} + \underbrace{\mathbf{1}_{\{\alpha_p < X_i \leqslant \alpha_p + tA/\sqrt{n}\}}}_{:=R_i}$. Les $(Y_i)_i$ sont iid et $Y_1 \sim \mathcal{B}(F(\alpha_p) = p)$. On rappelle que $\Delta_{n,t} = F\left(\alpha_p + \frac{tA}{\sqrt{n}}\right)$.

$$\text{On a } Z_n^* = \frac{\sqrt{p(1-p)}}{\sqrt{\Delta_{n,t}(1-\Delta_{n,t})}} \left[\underbrace{\frac{\sum Y_i - np}{\sqrt{np(1-p)}}}_{(1)} + \underbrace{\frac{np - n\Delta_{n,t}}{\sqrt{np(1-p)}}}_{(2)} + \underbrace{\frac{\sum R_i}{\sqrt{np(1-p)}}}_{(3)}\right].$$

Par un TCL, (1) tend vers $\mathcal{N}(0,1)$.

$$\text{Pour (2), on écrit } \frac{np - n\Delta_{n,t}}{\sqrt{np(1-p)}} = -\sqrt{n}\frac{F(\alpha_p + tA/\sqrt{n}) - F(\alpha_p)}{\sqrt{p(1-p)}}$$
$$= -\frac{tA}{\sqrt{p(1-p)}} \frac{F(\alpha_p + tA/\sqrt{n}) - F(\alpha_p)}{\frac{tA}{\sqrt{n}}} \to -\frac{tA}{\sqrt{p(1-p)}} F'(\alpha_p)$$

Maintenant, montrons que (3) **converge en proba** vers $\frac{tA}{\sqrt{p(1-p)}} F'(\alpha_p)$. On pose $T_n := \frac{1}{\sqrt{n}} \sum R_i$.
On a $\mathrm{Var}[T_n] = (F(\alpha_p + tA/\sqrt{n}) - F(\alpha_p))[1 - (F(\alpha_p + tA/\sqrt{n}) - F(\alpha_p))] \to 0$. Donc par l'inégalité de Chebychev, $\mathbb{P}(|T_n - \mathbb{E}[T_n]| > \varepsilon) \to 0$.
Or $\mathbb{E}[T_n] = \sqrt{n}(F(\alpha_p + tA/\sqrt{n}) - F(\alpha_p)) \to tAF'(\alpha_p)$. Donc (3) converge bien vers la valeur voulue en probabilité.
On applique le théorème de Slutsky et obtenir la convergence souhaitée du lemme. ■

Comme Z_n^* converge en loi vers $\mathcal{N}(0,1)$, les fonctions de répartitions de ces variables (qui sont croissantes) convergent simplement vers Φ en ces points de non discontinuité (théorème Portemanteau). Or $\mathcal{N}(0,1)$ est à densité, donc Φ est continue et donc la convergence se fait sur $\mathbb{R}$ et on vérifie les hypothèses du lemme 1, on a alors

$$\sup_{x \in \mathbb{R}} |\mathbb{P}(Z_n^* \leqslant x) - \Phi(x)| = o_n(1)$$

Comme Φ est sans atome, on peut mettre une inégalité stricte dans la probabilité précédente.

$$\text{Maintenant, on a } \Phi(t) - G_n(t) = \Phi(t) - \mathbb{P}(Z_n^* \geqslant -c_{n,t}) = \Phi(t) - 1 + \mathbb{P}(Z_n^* < -c_{n,t})$$
$$= \Phi(t) - [\Phi(c_{n,t}) + \Phi(-c_{n,t})] + \mathbb{P}(Z_n^* < -c_{n,t}) \quad \text{(par la symétrie de } \mathcal{N}(0,1))$$

Donc $|\Phi(t) - G_n(t)| \leqslant o_n(1) + |\Phi(t) - \Phi(c_{n,t})|$.
On va maintenant choisir A pour avoir $c_{n,t} \to t$.
Ce sont des calculs similaires que dans la démonstration du lemme 2, il faut reconnaître un taux d'accroissement. Il faut alors prendre $A = \dfrac{\sqrt{p(1-p)}}{F'(\alpha_p)}$.

Comme la convergence simple des fonctions de répartition implique la convergence en loi, nous avons

$$\frac{\sqrt{n}(\widehat{\alpha_{p,n}} - \alpha_p)}{\frac{\sqrt{p(1-p)}}{F'(\alpha_p)}} \Rightarrow \mathcal{N}(0,1) \qquad \left(\text{ie } \sqrt{n}(\widehat{\alpha_{p,n}} - \alpha_p) \Rightarrow \mathcal{N}\left(0, \frac{p(1-p)}{F'(\alpha_p)^2}\right)\right)$$

■

Par rapport au TCL, on n'a pas besoin de considérer des v.a. avec un moment d'ordre 2. Ce théorème peut nous aider dans les rares cas où la méthode des moments en statistique ne peut pas être utilisée, par exemple avec des lois de Cauchy. L'hypothèse de dérivation est satisfaite lorsque l'on considère des v.a. à densité f avec f continue en α_p, ce qui est encore le cas pour les lois de Cauchy. Vous m'avez compris, nous allons parler de loi de Cauchy pour notre application.

Application 1

Pour chaque $\theta \in \mathbb{R}$, on note P_θ la loi sur $\mathbb{R}$ de densité

$$f_\theta(x) = \frac{1}{\pi(1 + (x-\theta)^2)}$$

L'objectif est de définir et d'étudier un estimateur du paramètre du modèle $\left(\mathbb{R}^n, \{P_\theta^{\otimes n}\}\right)$.

Démonstration : Les f_θ sont continues donc par le théorème fondamental de l'analyse les F_θ sont dérivables.

On remarque que $\frac{1}{2} = \displaystyle\int_{-\infty}^{\theta} f_\theta(x)\,\mathrm{d}x$, donc $\theta = \alpha_{1/2}$ car F_θ est strictement croissante.

Donc par le théorème 1, on a $\sqrt{n}(\widehat{\alpha_{1/2,n}} - \theta) \Rightarrow \mathcal{N}\left(0, \frac{\pi^2}{4}\right)$. ■

Application 2

Comparer les intervalles de confiance obtenus avec le TCL et le TCLQ pour le modèle $(\mathbb{R}^n, \{\mathcal{E}(\alpha)^{\otimes n}\}_\alpha)$. [à voir si c'est possible]

Cette méthode possède un gros défaut. On ne connait pas forcément la valeur de α_p ni F. Dans notre application 1, nous avons eu de la chance car $f_\theta(\theta)$ valait toujours la même chose. Ce qui n'est pas tout le temps le cas. Avec une δ-méthode, on peut réussir à enlever le $F'(\alpha_p)$ de la loi limite mais il faut donc connaître F, ce qui n'est jamais le cas :(. Il faut donc reconnaitre une faiblesse assez importante de cette méthode.

La preuve du théorème 1 est une adaptation de celle-ci, où l'argument utilisant le théorème de Berry-Esseen a été remplacé par le lemme de Dini et le lemme 2.

Théorème 2 (*de Berry-Esseen*)

Soit $(X_i)_i$ une suite iid de variables aléatoires ayant un moment d'ordre 3. On note m sa moyenne, σ^2 sa variance et $\rho = \mathbb{E}[|X_1 - m|^3]$, alors

(i) $\sqrt{n}(S_n - m) \Rightarrow \mathcal{N}(0, \sigma^2)$ *(TCL)*

(ii) $\left|\mathbb{P}\left(\frac{\sqrt{n}[S_n - m]}{\sigma} \leqslant x\right) - \Phi(x)\right| \leqslant \frac{33}{4}\frac{\rho}{\sqrt{n}\sigma^3}$, *pour tout $x \in \mathbb{R}$,*

où Φ est la fonction de répartition $\mathcal{N}(0,1)$ et la constante $\frac{33}{4}$ s'améliorant au fil du temps.