Intervalles de confiance

Définition 1. *Soient Y une v.a. réelle et $\alpha \in]0, 1[$. On appelle quantile d'ordre α de Y le nombre q_α tel que*

$$q_\alpha = \inf \{y \in \mathbb{R}: F_Y(y) \geq \alpha\}.$$

Propriétés:

1. On a $\mathbb{P}(Y \leq q_\alpha) = F_Y(q_\alpha) \geqslant \alpha$. Si Y est une v.a. continue $F_Y(q_\alpha) = \alpha$.
2. Si Y est une v.a. continue la fonction $q_\alpha:]0, 1[\rightarrow \{x: f_Y(x) > 0\}$ est bijective et continue.
3. Si Y est une v.a. continue alors pour tout $0 \leq \beta \leq \gamma \leq 1$:
$$\mathbb{P}(q_\beta < Y \leq q_\gamma) = \mathbb{P}(Y \leq q_\gamma) - \mathbb{P}(Y \leq q_\beta) = \gamma - \beta.$$
4. Si f_Y est une fonction paire (= la loi de Y est symétrique autour de zéro, $-Y$ a la même loi de Y) alors $q_{1-\alpha} = -q_\alpha$.
5. $q_{1/2}$ est la médiane. $q_{1/4}$ le premier quartile.

Problème

Une entreprise reçoit d'un de ses fournisseurs un lot de pièces qui doit "normalement" contenir une proportion $\theta \leq 10\%$ de pièces défectueuses. L'entreprise voudrait, par examen d'un échantillon de taille n, décider entre $\theta \leq 10\%$ et $\theta > 10\%$, sachant qu'elle acceptera le lot dans le premier cas et le rejettera dans le deuxième cas.

On définit

$$X_i = \begin{cases} 1 & \text{si la pièce prélevée est défectueuse ;} \\ 0 & \text{sinon.} \end{cases}$$

$X_1, \ldots, X_n$ sont n variables iid de loi de Bernoulli de paramètre θ qui composent l'échantillon $\mathbf{X}$. L'EMV est $\hat{\theta}_n = \overline{X}_n$ (égale à l'estimateur par méthode des moments).

Supposons $n = 100$ et que on observe $\overline{X}_n = 0.195$.

Question: Quelle décision l'entreprise doit prendre? Accepter ou rejeter le lot? Et, sur quel critère l'entreprise doit se baser pour prendre sa décision?

Définition 2. *Soit $\mathcal{P} = \{\mathbb{P}_\theta: \theta \in \Theta\}$ un modèle paramétrique. On dispose d'un échantillon $\mathbb{X} = (X_1, \ldots, X_n)$ de n v.a. iid $\sim \mathbb{P}_\theta$. Soient A_n et B_n deux statistiques. On dira que $[A_n, B_n]$ est un intervalle de confiance de niveau $1 - \alpha$ pour θ si*

$$\mathbb{P}_\theta(A_n \leq \theta \leq B_n) = 1 - \alpha$$

pour tout $\theta \in \Theta$.

On dira que $[A_n, B_n]$ est un intervalle de confiance de niveau asymptotiquement égal à $1-\alpha$ pour θ si

$$\lim_{n\to\infty} \mathbb{P}_\theta(A_n \leq \theta \leq B_n) = 1-\alpha$$

pour tout $\theta \in \Theta$.

Remarque: Dans les applications on utilise souvent les valeurs $\alpha = 0.05, 0.01$.

Exemple 3. Soit $X \sim \mathcal{N}(\mu, 1)$, $\mu \in \mathbb{R}$ notre modèle paramétrique. Soient ζ_α les quantiles de la v.a. Gaussienne standard (centrée et réduite). On pose $A_n = \overline{X}_n - \zeta_\gamma/\sqrt{n}$ et $B_n = \overline{X}_n - \zeta_\beta/\sqrt{n}$.

On veut déterminer β et γ dans $[0, 1]$ tels que $[A_n, B_n]$ soit un intervalle de confiance de niveau $1-\alpha$ pour μ.

La v.a. $\bar{X}_n$ est une Gaussienne de moyenne μ et variance $1/n$ donc

$$\mathbb{P}(A_n \leqslant \mu \leqslant B_n) = \mathbb{P}(A_n \leqslant \mu, \mu \leqslant B_n) = \mathbb{P}(\overline{X}_n - \zeta_\gamma/\sqrt{n} \leqslant \mu, \mu \leqslant \overline{X}_n - \zeta_\beta/\sqrt{n})$$

$$= \mathbb{P}(\sqrt{n}(\overline{X}_n - \mu) \leqslant \zeta_\gamma, \zeta_\beta \leqslant \sqrt{n}(\overline{X}_n - \mu)) = \mathbb{P}(\zeta_\beta \leqslant \sqrt{n}(\overline{X}_n - \mu) \leqslant \zeta_\gamma)$$

$$= \mathbb{P}(\zeta_\beta \leqslant Z \leqslant \zeta_\gamma) = \mathbb{P}(Z \leqslant \zeta_\gamma) - \mathbb{P}(Z \leqslant \zeta_\beta)$$

où $Z \sim \mathcal{N}(0, 1)$. Par la définition des quantiles Gaussiens on a que $\mathbb{P}(\mathrm{Z} \leqslant \zeta_r) = r$ pour tout $r \in]0, 1[$ et donc

$$1-\alpha = \mathbb{P}(A_n \leqslant \mu \leqslant B_n) = \gamma - \beta$$

est la condition à imposer sur γ, β pour avoir un intervalle de confiance à niveau $1-\alpha$.

Remarque 4.

- Il existe un nombre infini des intervalles de confiance de niveau $1-\alpha$.
- Si $\beta \neq 0$ et $\gamma \neq 1$ on parlera d'un intervalle de confiance bilatérale.
- Si $\beta = 0$ $(\gamma = 1-\alpha)$ ou si $\gamma = 1$ $(\beta = \alpha)$ on parlera d'un intervalle de confiance unilatéral.
- Si $\beta = \alpha/2$ et $\gamma = 1-\alpha/2$ on parlera d'un intervalle de confiance bilatéral symétrique.
- Valeurs utiles de ζ_α : $\zeta_{1/2} = 0$, $\zeta_{0.9} = 1.28$, $\zeta_{0.9} = 1.645$, $\zeta_{0.975} = 1.96$, $\zeta_{0.995} = 2.58$.

Remarque 5. Dans le cas Gaussien où l'échantillon est tiré de la loi $\mathcal{N}(\mu, \sigma_0^2)$ avec variance σ_0^2 connue, les intervalles plus utilisé sont

- Les intervalles unilatéraux

$$\mathbb{P}(\mu \geqslant \bar{X}_n - \frac{\sigma_0}{\sqrt{n}}\zeta_{1-\alpha}) = \mathbb{P}(\mu \leqslant \bar{X}_n + \frac{\sigma_0}{\sqrt{n}}\zeta_{1-\alpha}) = 1-\alpha;$$

- L'intervalle bilatérale symétrique:

$$\mathbb{P}(\bar{X}_n - \frac{\sigma_0}{\sqrt{n}}\zeta_{1-\alpha/2} \leqslant \mu \leqslant \bar{X}_n + \frac{\sigma_0}{\sqrt{n}}\zeta_{1-\alpha/2}) = 1-\alpha.$$

Exemple 6. Reprenons le problème introductif. $X \sim \text{Ber}(\theta)$. L'EMV pour θ est $\overline{X}_n$. Par le TCL:

$$\frac{\sqrt{n}(X_n - \theta)}{\sqrt{\theta(1-\theta)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Par la loi des grandes nombres $\overline{X}_n \xrightarrow{p.s.} \theta$ et donc par le lemme de continuité appliqué à la fonction $g(x) = \sqrt{x(1-x)}$ on a aussi

$$g(\overline{X}_n) = \sqrt{X_n(1-X_n)} \xrightarrow{p.s.} \sqrt{\theta(1-\theta)}.$$

On peut conclure par le lemme de Slutsky que

$$\frac{\sqrt{n}(X_n - \theta)}{\sqrt{X_n(1-X_n)}} = \frac{\sqrt{n}(X_n - \theta)}{\sqrt{\theta(1-\theta)}} \frac{\sqrt{\theta(1-\theta)}.}{X_n(1-X_n)} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Donc asymptotiquement l'intervalle de confiance symétrique bilatérale pour θ est donné par

$$\bar{X}_n - n^{-1/2}(\overline{X}_n(1-\overline{X}_n))^{1/2}\zeta_{1-\alpha/2} \leqslant \theta \leqslant \bar{X}_n + n^{-1/2}(\overline{X}_n(1-\overline{X}_n))^{1/2}\zeta_{1-\alpha/2}.$$

Application: Si on fixe $\alpha = 0.05$. Pour la valeur observé de $\overline{X}_n = 0.195$ ($n = 100$) on a que l'intervalle de confiance trouvé dans l'exemple précèdent est

$$\theta \in [0.117, 0.273]$$

(vérifier). Ce qui permet de rejeter le lot avec niveau de confiance 95%.