

Intervalles de confiance

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

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

Propriétés :

1. On a $P(Y \leq q) = F_Y(q)$. Si Y est une v.a. continue $F_Y(q) = q$.
2. Si Y est une v.a. continue la fonction $q :]0, 1[\rightarrow \{x : f_Y(x) > 0\}$ est bijective et continue.
3. Si Y est une v.a. continue alors pour tout $0 < q < 1$:

$$P(q < Y < q) = P(Y < q) - P(Y \leq q) = 0.$$

4. Si f_Y est une fonction paire (= la loi de Y est symétrique autour de 0) alors $q_{1/2} = 0$.
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 10% de pièces défectueuses. L'entreprise prendrait, par examen d'un échantillon de taille n , décider entre 10% et > 10%, sachant qu'elle acceptera le lot dans le premier cas et le rejettera dans le deuxième cas.

On définit

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

X_1, \dots, X_n sont n variables iid de loi de Bernoulli de paramètre p qui composent l'échantillon. L'EMV est $\hat{p} = \bar{X}_n$ (il est l'estimateur par méthode des moments).

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

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 $P = \{P : \dots\}$ un modèle paramétrique. On dispose d'un échantillon (X_1, \dots, X_n) de n v.a. iid P . 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 si

$$P(A_n \leq \theta \leq B_n) = 1 - \alpha$$

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 si

$$\lim_n P(A_n \leq B_n) = 1 - \alpha$$

pour tout $\epsilon > 0$.

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

Exemple 3. Soit $X \sim N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$ notre modèle paramétrique. Soient les quantiles de la v.a. Gaussienne standard (centrée et réduite). On pose $A_n = \bar{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ et $B_n = \bar{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$.

On veut déterminer et 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 σ^2/n donc

$$\begin{aligned} P(A_n \leq \mu \leq B_n) &= P\left(\bar{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) \\ &= P\left(-z_{\alpha/2} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = P\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right) \\ &= P(Z \leq z_{\alpha/2}) - P(Z \leq -z_{\alpha/2}) \end{aligned}$$

où $Z \sim N(0, 1)$. Par la définition des quantiles Gaussiens on a que $P(Z \leq r) = r$ pour tout $r \in [0, 1]$ et donc

$$1 - \alpha = P(A_n \leq \mu \leq B_n) = 1 - \alpha$$

est la condition à imposer sur $z_{\alpha/2}$ 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 $\alpha \in (0, 1)$ on parlera d'un intervalle de confiance bilatéral.

Si $\alpha = 0$ ($= 1 - \alpha$) ou si $\alpha = 1$ ($= 0$) on parlera d'un intervalle de confiance unilatéral.

Si $\alpha = 1/2$ et $\alpha = 1 - \alpha/2$ on parlera d'un intervalle de confiance bilatéral symétrique.

Valeurs utiles de $z_{\alpha/2}$: $z_{1/2} = 0$, $z_{0.9} = 1.28$, $z_{0.95} = 1.645$, $z_{0.975} = 1.96$, $z_{0.995} = 2.58$.

Remarque 5. Dans le cas Gaussien où le échantillon est tiré de $N(\mu, \sigma^2)$ avec variance σ^2 connue, les intervalles plus utilisés sont

Les intervalles unilatéraux

$$P\left(\bar{X}_n - \frac{\sigma}{\sqrt{n}} \leq \mu\right) = P\left(\bar{X}_n + \frac{\sigma}{\sqrt{n}} \geq \mu\right) = 1 - \alpha;$$

L'intervalle bilatéral symétrique:

$$P\left(\bar{X}_n - \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

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

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

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

$$g(\bar{X}_n) = \frac{1}{\bar{X}_n(1-\bar{X}_n)} \xrightarrow{p.s.} \frac{1}{p(1-p)}.$$

On peut conclure par le lemme de Slutsky que

$$\frac{\bar{n}(X_n - p)}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} = \frac{\bar{n}(X_n - p)}{\sqrt{p(1-p)}} \cdot \frac{\sqrt{p(1-p)}}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} \xrightarrow{L} N(0, 1).$$

Donc asymptotiquement l'intervalle de confiance symétrique bilatéral est donné par

$$\bar{X}_n \pm z_{\alpha/2} \sqrt{\bar{X}_n(1-\bar{X}_n)} \quad \text{avec } z_{\alpha/2} = z_{0.05/2} = 1.96.$$

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

$$[0.117, 0.273]$$

(voir l'exercice). Ce qui permet de rejeter H_0 avec niveau de confiance 95.