EXERCICE CORRIGÉ TYPE : INTERVALLE DE CONFIANCE

On commence par rappeler le théorème de la limite centrale

Théoreme 1 (limite centrale)**.** *Soit (X_n) une suite de v.a. i.i.d. d'espérance commune μ et de variance commune $\sigma^2 < \infty$. Alors*

$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \to \mathcal{N}(0,1) \; .$$

Exercice : La société Audimat, désirant évaluer la proportion p de téléspectateurs regardant une certaine émission de télévision, interroge n téléspectateurs par téléphone. On note X la v.a. qui, à tout individu, fait correspondre le nombre 1 s'il a regardé l'émission, et 0 sinon.

On note $(X_1, \ldots, X_n)$, le n-échantillon de X ainsi obtenu ; X_i étant la réponse de l'individu numéroté i.

Le but est d'estimer la proportion p, en donnant un intervalle de confiance au risque α, avec $\alpha \in]0;1[$ «petit» (typiquement $\alpha = 5\%$ ou $\alpha = 1\%$).

X suit une loi de Bernoulli de paramètre p. Donc, pour tout i, on a $X_i \sim \mathcal{B}(p)$. On utilise l'estimateur suivant pour p :

$$\hat{p}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \; .$$

(à (re)faire : calculer son biais et son risque quadratique.)

On cherche un intervalle de confiance I pour p de sorte que :

$$\mathbb{P}\{p \in I\} \geq 1 - \alpha \; .$$

avec $\alpha \in [0;1]$ fixé.

Plus précisément, on a de bonnes raisons de penser que $\hat{p}_n$ soit une bonne estimation du paramètre p. On va donc chercher cet intervalle sous la forme $I = [\hat{p} - t; \hat{p} + t]$ où $t \in \mathbb{R}_+$. On cherche alors t tel que

$$\begin{aligned} \mathbb{P}\{p \in [\hat{p}_n - t; \hat{p}_n + t]\} \geq 1 - \alpha \; &\Leftrightarrow \mathbb{P}\{|\hat{p}_n - p| < t\} \geq 1 - \alpha \\ &\Leftrightarrow \mathbb{P}\left\{ \frac{|\hat{p}_n - p|}{\sqrt{p(1-p)/n}} < \frac{t}{\sqrt{p(1-p)/n}} \right\} \geq 1 - \alpha \; . \end{aligned}$$

Or, si n est suffisament grand, le TLC nous dit que $\frac{|\hat{p}_n - p|}{\sqrt{p(1-p)/n}}$ est approximativement de loi $\mathcal{N}(0,1)$. Soit alors $Z \sim \mathcal{N}(0,1)$, et soit z l'unique valeur telle que

$$\mathbb{P}\{|Z| < z\} = 1 - \alpha \; .$$

La valeur de z nous est donnée par les tables de la loi normale. Par exemple, si $\alpha = 0.95$, alors $z = 1.96$. Si $\alpha = 0.99$, alors $z = 2.57$. z est donc maintenant connu, et on peut poursuivre le calcul. On a donc :

$$\mathbb{P}\left\{\frac{|\hat{p}_n - p|}{\sqrt{p(1-p)/n}} < \frac{t}{\sqrt{p(1-p)/n}}\right\} \geq 1-\alpha\ ,$$

avec

$$\frac{t}{\sqrt{p(1-p)/n}} = z\ .$$

Malheureusement, cette dernière équation ne nous permet pas de déterminer t, puisqu'elle contient deux inconnues : t et p. Mais si n est grand, p peut être approché par $\hat{p}_n$ d'après la loi des grands nombres. On arrive alors à

$$t \simeq z\sqrt{\hat{p}_n(1-\hat{p}_n)/n}\ .$$

On peut donc parier (avec un risque α d'avoir tort) que

$$p \in [\hat{p}_n - z\sqrt{\hat{p}_n(1-\hat{p}_n)/n}; \hat{p}_n + z\sqrt{\hat{p}_n(1-\hat{p}_n)/n}]\ .$$

Application numérique : On fixe $\alpha = 1\%$. On suppose que la société Audimat a interrogé 200 personnes. Sur ces 200 personnes, 40 ont répondu avoir regardé l'émission de télévision. On peut donc dire (avec 1% de risque de se tromper) que la proportion p de téléspectateurs ayant regardé cette émission est telle que

$$\begin{aligned} p &\in [0.2 - 2.57\sqrt{0.2(0.8)/200}; 0.2 + 2.57\sqrt{0.2(0.8)/200}] \\ &\in [0.127; 0.272]\ . \end{aligned}$$

Si on avait fixé $\alpha = 5\%$, on aurait trouvé

$$p \in [0.144; 0.255]\ .$$

Il faut ici comprendre que si on refait la même expérience (compter le nombre de personnes ayant regardé l'émission sur 200 personnes interrogées au hasard) un grand nombre de fois, 99% (ou 95%, selon le choix de α) des intervalles de confiance obtenus contiennent la vraie valeur de p. Il s'agit donc bien d'un pari que l'on fait sur notre échantillon. Il est tout à fait possible que nous n'ayons pas de chance, et que notre échantillon de données soit parmi les 1% (ou 5% selon α) de «mauvais» échantillons.