Chapitre 11

Intervalles de confiance

Nous allons conclure ce cours d'initiation au calcul de probabilités sur une application à un sujet d'actualité : celle des sondages. Je prendrai l'exemple du sondage IFOP–Wanadoo-actu effectué les 12 et 13 mai 2005 (OUI : 46%, NON : 54%). Rappelons tout-de-suite qu'il s'agit d'un cliché de l'opinion à ces dates, et pas un pronostic du résultat, ne serais-ce que parce que seuls 67% des partisans du OUI se déclarent sûrs de leur choix et 33% pouvoir encore changer d'avis (et 76%-24% pour les "déclarants" du NON). Nous allons aborder la question de ce que l'on peut conclure de ce sondage avec un peu de certitude sur l'opinion à cette date.

11.1 Modélisation

Parmi les N personnes de la population qui sont inscrites sur les listes electorales et déclarant aller voter le 29 mai, il y a une *proportion* p voulant voter OUI et $1-p$ voulant noter NON.

On procède alors à un *sondage*, un protocole (subtile) cherchant à tirer "au hasard" et de façon "indépendantes" n personnes a qui on demandera leur intention de vote x_1, x_2, ..., x_n, avaec $x_i = 1$ si la i-ème personne interrogée déclare vouloir voter OUI.

On modélise cette interrogation de n personnes par un n-échantillon X_1, ..., X_n de v.a. i.i.d., avec $X_i = \mathcal{B}(1,p)$. Il est naturel de choisir pour modèle des $X_i \rightsquigarrow \mathcal{B}(1,q)$, avec $q \in]0,1[$, puisque seule deux réponses sont possibles (on a écarté les réponses différentes de OUI et NON), codée par 1 et 0, respectivement. Le choix de $q = p$, la proportion d'electeurs voulant voter OUI peut se motiver par le choix suivant : on définit $X_i : \Omega_i \longrightarrow \{0,1\}$, avec $\Omega_i = \{\omega_1, \ldots, \omega_N\}$, chacun des $\omega_j \in \Omega_i$ représantant un des N électeurs inscrits ; on postule que chacun de électeurs a la même chance d'être interrogé lors de la i-ème interrogation, et on choisit donc une loi uniforme sur Ω_i, d'où $\mathbb{P}(\{\omega\}) = \frac{1}{N}$ pour tout $\omega \in \Omega_i$. Soient $A := \{X_i = 1\}$ les électeurs voulant voter OUI et $a := \mathrm{Card}\,(A)$ ne nombre de ces électeurs voulant voter OUI. ; on a alors $\mathbb{P}(\{X_i = 1\}) = \mathbb{P}(A) = \frac{a}{N} = p$, par définition de p. Ceci permet de choisir la *loi* de $X_i \rightsquigarrow \mathcal{B}(1,p)$. A noter que pour avec un modèle avec n v.a. $X_1, \ldots, X_n$ indépendantes, il faut alors choisir un autre Ω ; par exemple $\Omega = \Omega_1 \times \ldots \times \Omega_n$, ou plus simplement $\Omega = \{0,1\}^n$, avec $\mathbb{P}(\{(\omega_i, \ldots, \omega_n)\}) = p^{\omega_1+\ldots+\omega_n}(1-p)^{(1-\omega_1)+\ldots+(1-\omega_n)}$ et $X_i(\omega_i, \ldots, \omega_n) = \omega_i$. Notons que toutefois toutes ces précisions de sont pas indispensables : tout ce dont nous avons besoin, c'est que les X_i soient i.i.d. de loi $\mathcal{B}(1,p)$, où p est la proportion d'électeurs voulant voter OUI.

11.2 Domaine de confiance pour l'estimation de p

La question maintenant est d'estimer la valeur de la proportion p et de juger de la qualité de cette estimation. Dans notre modèle p est l'espérance commune des X_i et nous avons vu que la moyenne $\boxed{P_n := \frac{1}{n}(X_1 + \ldots + X_n)}$ est un estimateur qui converge (en probabilité) vers cette espérance commune : c'est essentiellement la loi des grands nombres (LGN).

Si les n personnes interrogées on répondu $x_1, \ldots, x_n$, en pratique on donne l'estimation

$$\boxed{\hat{p} = p_n = \frac{1}{n}(x_1, \ldots, x_n)}.$$

La question est alors de savoir quelle confiance attacher à cette estimation $\hat{p}$. C'est l'objet de la définition d'un domaine de confiance au seuil α, par exemple $\alpha = 5\%$.

Définition : On dit que l'intervalle (aléatoire) $\mathcal{D}_n = [P_n - \Delta^-, P_n + \Delta^+]$ est un domaine de confiance pour l'estimation de p au seuil α si et seulement si $\boxed{\mathbb{P}(\{p \notin \mathcal{D}_n\}) \leq \alpha}$.

En d'autres termes, il y a une probabilité au plus égale à α de se tromper si l'on affirme que $p \in \mathcal{D}_n$.
Pour déterminer un tel intervalle de confiance, il convient de réécrire l'évènement

$$E_n := \{p \in \mathcal{D}_n\} = \{p \in [P_n - \Delta^-, P_n + \Delta^+]\}.$$

Tout d'abord nous avons

$$\begin{aligned} E_n &= \{p \in [P_n - \Delta^-, P_n + \Delta^+]\} = \{p \geq P_n - \Delta^-, p \leq P_n + \Delta^+\} = \{P_n \leq p + \Delta^-, P_n \geq p - \Delta^+\} \\ &= \{P_n \in [p - \Delta^+, p + \Delta^-]\}. \end{aligned}$$

Par ailleur, nous allons réécrire $P_n = \mu + \sigma_n Z_n$ où $\mu = \mathbb{E}(P_n)$ et $\sigma_n^2 = \mathrm{Var}\,(P_n)$; ainsi Z_n sera une v.a. centrée ($\mathbb{E}(Z_n) = 0$) et réduite ($\mathrm{Var}\,(Z_n) = 1$. Pour cela, il suffit de poser

$$\mu = \mathbb{E}(P_n) = \mathbb{E}\left(\frac{1}{n}(X_1 + \ldots + X_n)\right) = \frac{1}{n}(\mathbb{E}(X_1) + \ldots + \mathbb{E}(X_n)) = \frac{n}{n}p = p.$$

$$\sigma_n^2 = \mathrm{Var}\,(P_n) = \mathrm{Var}\,\left(\frac{1}{n}(X_1 + \ldots + X_n)\right) = \frac{1}{n^2}(\mathrm{Var}\,(X_1) + \ldots + \mathrm{Var}\,(X_n)) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n},$$

où $\sigma^2 = \mathrm{Var}\,(X_i) = p(1-p)$, d'où finalement $\boxed{\sigma_n = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}}$.

A présent, il est facile de voir que l'évènement E_n peut encore s'écrire

$$E_n = \{Z_n \in [z_-, z_+]\} \text{ avec } z_- = -\frac{\sqrt{n}}{\sigma}\Delta^+ \text{ et } z_+ = +\frac{\sqrt{n}}{\sigma}\Delta^-.$$

11.3 Approximation normale

La préparation que nous avons faite de P_n en l'écrivant $P_n = p + \frac{\sigma}{\sqrt{n}}Z_n$ est dictée par l'usage que nous allons faire du théorème limite central (TLC) ; en effet, ce théorème assure précisément que la v.a. Z_n définie par cette relation tends en loi vers $Z \rightsquigarrow \mathcal{N}(0,1)$. Ainsi, $\mathbb{P}(E_n) = \mathbb{P}(\{Z_n \in [z_-, z_+]\}) = F_{Z_n}(z_+) - F_{Z_n}(z_-) \approx F_Z(z_+) - F_Z(z_-)$ dès que n est assez grand.

Choisissons à présent $\Delta^+ = \Delta^- =: \Delta$, ce qui implique que $z_- = -z_+$. Nous cherchons $\Delta = \Delta_\alpha$ tel que $\mathbb{P}(E_n^c) = \alpha$, c'est à dire $\mathbb{P}(E_n) = 1 - \alpha$. Comme la densité de Z est paire, ceci revient à choisir z_+ tel que $F_Z(z_+) = 1 - \frac{\alpha}{2}$. Comme $z_+ = \frac{\sqrt{n}}{\sigma}\Delta$, nous obtenons

$$\boxed{\Delta_\alpha = \frac{\sigma}{\sqrt{n}}F_Z^{-1}(1-\tfrac{\alpha}{2}) = \frac{\sqrt{p(1-p)}}{\sqrt{n}}F_Z^{-1}(1-\tfrac{\alpha}{2}) \leq \frac{1}{2\sqrt{n}}F_Z^{-1}(1-\tfrac{\alpha}{2}) =: \Delta'_\alpha,}$$

l'inégalité résultant du fait que pour $p \in [0,1]$ on a $p(1-p) \leq \frac{1}{4}$.

En utilisant le tableau de la fonction de répartition F_Z de la loi normale centrée réduite, nous obtenons les valeurs suivantes pour Δ'_α, et donc que $\mathbb{P}(p \notin [P_n - \Delta'_\alpha, P_n + \Delta'_\alpha]) \leq \alpha$.

α	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%
$A = F_Z^{-1}(1-\frac{\alpha}{2})$	2,576	2,326	2,170	2,054	1,960	1,881	1,812	1,751	1,695	1,645
$n = 400$	0,064	0,058	0,054	0,051	0,049	0,047	0,045	0,044	0,042	0,041
$n = 850$	0,044	0,040	0,037	0,035	0,034	0,032	0,031	0,030	0,029	0,028
$n = 1000$	0,041	0,037	0,034	0,032	0,031	0,030	0,029	0,028	0,027	0,026
$n = 2000$	0,029	0,026	0,024	0,023	0,022	0,021	0,020	0,020	0,019	0,018

Valeurs de $\Delta'_\alpha := \frac{1}{2\sqrt{n}}F_Z^{-1}\left(1-\frac{1}{\alpha}\right) \geq \frac{\sigma}{\sqrt{n}}F_Z^{-1}\left(1-\frac{1}{\alpha}\right)$

Exercice : Soient X_1, ..., X_n i.i.d. avec $X_i \rightsquigarrow \mathcal{B}(1,p)$ et $P_n := \frac{1}{n}(X_1 + \ldots + X_n)$ l'estimateur usuel de p. Montrer que si n est suffisamment grand pour justifier l'approximation normale donnée par le TLC ($n \geq 30$), on a que $\alpha := \mathbb{P}(p \notin [P_n - \frac{1}{\sqrt{n}}, P_n + \frac{1}{\sqrt{n}}]$ est inférieur à (et proche de) 5%.