Les fourchettes de sondage

Une population de taille N est constituée d'individus ayant une certaine propriété (par exemple des personnes désirant voter pour le candidat A) et d'individus possédant la propriété contraire (par exemple des personnes ne désirant pas voter pour le candidat A).
Le problème consiste à déterminer la proportion p d'individus ayant cette propriété. S'il est possible d'interroger tout le monde, il n'y a aucun problème pour déterminer p. C'est ce qui se passe habituellement à chaque élection.

Mais si on ne peut interroger qu'une partie de cette population, on dit qu'on procède par **sondage** et dans ce cas on ne pourra qu'estimer la valeur de p.

Si on est très prudent, on cherchera un intervalle contenant p avec une probabilité de 95%.
Si on est encore plus prudent, on peut être amené à augmenter l'amplitude de l'intervalle de façon que la probabilité de se tromper soit alors de 1%. A la limite si l'intervalle est [0 ; 1] on est sûr de ne pas faire d'erreur mais est-ce dans ce cas bien intéressant ?

La méthode consistant à dire que p appartient à un certain intervalle en se basant sur les données d'un échantillon est appelée **estimation par intervalle de confiance**.

Un sondage peut être modélisé par exemple par un tirage de n boules d'une population de N boules dont une proportion p est de couleur rouge et une proportion $1-p$ est de couleur verte.

Si n est petit devant N ou si le tirage se fait boule par boule avec remise alors on démontre que la loi de la variable aléatoire X qui donne le nombre de boules rouges obtenues suit approximativement (ou réellement en cas de remise) une loi binomiale de paramètres n et p.

Si de plus p n'est pas trop petit et si n est assez grand (supérieur à 30) alors on démontre que :

$$\frac{X-np}{\sqrt{np(1-p)}}$$

suit approximativement une loi normale centrée réduite, la probabilité pour que $\frac{X-np}{\sqrt{np(1-p)}}$ appartienne à l'intervalle $[a\,;\,b]$ étant égale à : $\frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{t^2}{2}}\,dt$.

Comme $\frac{X-np}{\sqrt{np(1-p)}}$ suit à peu près une loi normale centrée réduite on est sûr à 95% que :

$$-1{,}96 \le \frac{X-np}{\sqrt{np(1-p)}} \le 1{,}96.$$

Ce résultat vient du fait que : $\frac{1}{\sqrt{2\pi}}\int_{-1{,}96}^{1{,}96} e^{-\frac{t^2}{2}}\,dt \approx 0{,}95$.

Cette double inégalité s'écrit encore :

$$\frac{X}{n} - \frac{1{,}96\sqrt{np(1-p)}}{n} \le p \le \frac{X}{n} + \frac{1{,}96\sqrt{np(1-p)}}{n}.$$

On obtient ainsi un encadrement de p. Mais cet encadrement est en fait inutilisable car le majorant et le minorant dépendent eux-mêmes du nombre p inconnu.
Cependant comme $\sqrt{p(1-p)} \le 0{,}5$ pour tout p, la double inégalité peut se réduire à : $\frac{X}{n} - \frac{0{,}98}{\sqrt{n}} \le p \le \frac{X}{n} + \frac{0{,}98}{\sqrt{n}}$ ou encore à $\frac{X}{n} - \frac{1}{\sqrt{n}} \le p \le \frac{X}{n} + \frac{1}{\sqrt{n}}$.

Ceci revient à dire qu'on est sûr à 95% que p appartient à l'intervalle $\left[\hat{p} - \frac{1}{\sqrt{n}}\,;\,\hat{p} + \frac{1}{\sqrt{n}}\right]$ avec $\hat{p} = \frac{X}{n}$ qui n'est autre que la proportion de boules rouges obtenues lors du prélèvement de l'échantillon.

Cet intervalle «aléatoire» est fixé dès que le tirage des n boules a eu lieu ; il suffit alors de prendre pour valeur de $\hat{p}$ le nombre $\frac{k}{n}$ où k est le nombre de boules rouges effectivement obtenues.

De manière analogue on trouve les intervalles au niveau 90%, intervalles d'amplitude plus réduite, et les intervalles au niveau 99% au contraire plus larges.

La seule possibilité pour diminuer l'amplitude d'un intervalle est alors d'augmenter n. Mais cela entraîne un coût de sondage plus élevé.