ECHANTILLONNAGE ET FLUCTUATION

A partir d'une population, on appelle « échantillon » un sous-ensemble de cette population obtenu par prélèvement aléatoire.

Le nombre d'individus de l'échantillon est appelé la taille n de l'échantillon.

Si on effectue une même série statistique auprès de plusieurs échantillons d'une même population, alors on observe une fluctuation des fréquences d'échantillonnage.

Plus la taille de l'échantillon est grande, et plus cette fluctuation est resserrée autour de fréquence « idéale ».

Généralisation :

Supposons une population dont une partie A est en proportion connue $\boldsymbol{p}$ (on peut aussi dire que, si l'on tire au hasard un individu dans la population, la probabilité d'obtenir un individu de A est $\boldsymbol{p}$). Considérons les échantillons de taille $\boldsymbol{n}$.

On peut établir que dans 95% de ces échantillons de taille n, la fréquence des individus de A dans l'échantillon est dans l'intervalle $\left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}\right]$.

Cet intervalle est appelé intervalle de fluctuation au seuil de 95%, ou intervalle de confiance au seuil de 95%.

Remarques :

1. Ce résultat n'est en fait valable que si $0{,}2 \leq p \leq 0{,}8$ (autrement dit, la partie A est en proportion « moyenne » dans la population) et si $n \geq 25$ (autrement dit, si l'échantillon est de taille « suffisante »).
2. Ce résultat signifie qu'il est « rare » (5% des cas) de trouver un échantillon dans lequel la fréquence de la partie A n'est pas dans l'intervalle indiqué.
   On pourra alors considérer qu'un tel échantillon n'est pas « normal », avec un risque de se tromper assez faible (5% des cas).
3. On peut remarquer que plus la taille des échantillons est grande, plus l'intervalle de fluctuation est petit.
   Pour $n = 100$, on a $\frac{1}{\sqrt{n}} = 0{,}1$; pour $n = 1\,000$, on a $\frac{1}{\sqrt{n}} \approx 0{,}013$ et pour $n = 2\,500$, on a $\frac{1}{\sqrt{n}} = 0{,}02$.
4. Il existe bien sûr des résultats concernant des intervalles de fluctuation à d'autres seuils que 95% (entre autres à 90% ou à 99%). Ce sera pour l'année prochaine.

Exemple 1 : sachant que dans la population française il y a autant d'hommes que de femmes (à peu près), on voudrait savoir si deux entreprises respectent le principe de parité dans leur recrutement. Le principe de parité est-il respecté dans chacune des deux entreprises ?

Exemple 2 : environ 26% de la population française se déclare allergique aux pollens de fleurs. Lors d'une enquête auprès des 400 salariés d'une entreprise, 123 personnes se déclarent allergiques aux pollens. Est-ce inquiétant ?

On utilise aussi l'échantillonnage pour la prise de décision à partir d'un échantillon d'une population statistiquement connue.

Dans certains cas, nous avons une population dont une partie A est en proportion $\boldsymbol{p}$ inconnue. On peut aussi dire que, si l'on tire au hasard un individu dans la population, la probabilité d'obtenir un individu de A est p).

Considérons les échantillons de taille n.

On peut établir que pour 95% de ces échantillons de taille n, la fréquence f des individus de A dans l'échantillon est telle que :

$$p \in \left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}}\right]$$

Cet intervalle est souvent appelé fourchette de sondage au seuil de 95%.

On a les mêmes remarques que pour l'intervalle de confiance.

Exemple : lors d'une campagne électorale, un sondage portant sur 1 024 personnes prises au hasard indique que 528 d'entre elles ont l'intention de voter pour le candidat A. Au journal de 20h, le présentateur de journal affirme : « si l'élection avait eu lieu aujourd'hui, le candidat A serait élu avec plus de 51,5% des voix ». A-t-il raison de dire cela ?
- à première vue :
- calcul de la fourchette de sondage au seuil de 95% :
On peut dire (et encore, avec un risque de 5% de se tromper) que si l'élection avait eu lieu aujourd'hui, le pourcentage de voix pour le candidat A serait compris entre 48,4% et 54,7%. On est donc loin de pouvoir affirmer qu'il serait élu.

Source : Académie en ligne.