Intervalle de fluctuation - estimation

Notion d'échantillon

Définition 1 : Un échantillon de taille *n* est constitué des résultats de *n* répétitions indépendantes de la même expérience

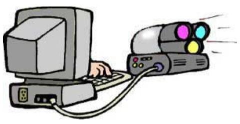
Cas d'une urne

- Avec une urne, un échantillon de taille *n* est constitué de *n* tirages avec remise.
- Si le tirage ne s'effectue pas avec remise, les tirages successifs ne sont pas indépendants car la composition de l'urne varie à chaque fois.

Cas du sondage

- **Convention :** On convient que si le nombre *n* de personnes interrogées est nettement inférieur au nombre total N de personnes, un sondage peut être considéré comme un tirage avec remise dans une urne.

1.1. *Fluctuation d'échantillonnage*



Projection Tableur

Exemple : Simulons 20 échantillons de taille 100 tirages avec remise dont la proportion de boules blanches est 0,6.
On constate que les fréquences observées fluctuent. Ce phénomène est appelé fluctuation d'échantillonnage.

1. Intervalle de fluctuation asymptotique au seuil de 95%

Théorème 1 :
Notons F_n la variable aléatoire qui à tout échantillon de taille *n*, associe la fréquence *f* d'un caractère.

Posons $I_n = \left[p - 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$ où *p* désigne la proportion de ce caractère dans la population.

Alors F_n **prend ses valeurs dans** I_n **avec une probabilité qui s'approche de 0,95 lorsque *n* devient grand** (Ce résultat est admis ici).

Définition 2

L'intervalle $I_n = \left[p - 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$ est appelé **intervalle de fluctuation asymptotique au seuil de 95%** de la variable aléatoire fréquence F_n

Remarques

- On admet que l'on utilise cet intervalle dès que les conditions $n \geq 30$ et $np \geq 5$ et $n(1-p) \geq 5$ sont remplies.
- Le nombre 1,96 qui intervient dans la définition de I_n est celui qui vérifie $P(-1{,}96 \leq Z \leq 1{,}96) \approx 0{,}95$ lorsque Z suit $\mathcal{N}(0\,;1)$.

Exemple :

Avec X qui suit une loi binomiale $\mathcal{B}(100\,;0{,}4)$, n=100 et p=0,4.

L'intervalle de fluctuation asymptotique au seuil de 95% de la fréquence est :

$$I_n = \left[0{,}4 - 1{,}96\frac{\sqrt{0{,}4(1-0{,}4)}}{\sqrt{100}} ; 0{,}4 + 1{,}96\frac{\sqrt{0{,}4(1-0{,}4)}}{\sqrt{100}}\right] \approx [0{,}303;0497]$$

On peut utiliser cette approximation car les conditions sont réunies :

$n \geq 30$, np=40 donc $np \geq 5$ et n(1-p)=60, donc $n(1-p) \geq 5$

TOUJOURS VERIFIER LES CONDITIONS D'UTILISATION

2. Lien avec l'intervalle de fluctuation donné en Seconde

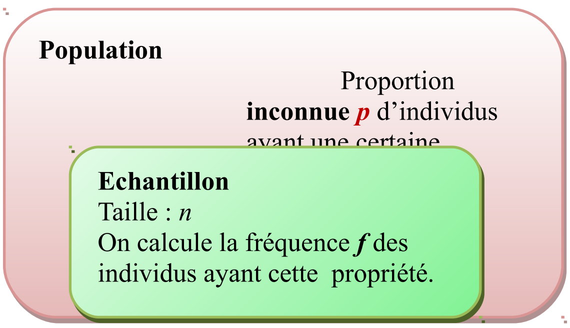
En **seconde** l'intervalle de fluctuation d'une proportion p sur un échantillon de taille n était donné par : $\left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}\right]$.

Le **théorème 1** que nous venons de voir donne en général un résultat un peu plus précis que le théorème analogue vu en seconde.

On peut démontrer en fait que $I_n \subset \left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}\right]$.

1.2. Estimation d'une proportion inconnue à partir d'un échantillon

On suppose maintenant que la proportion p est inconnue



Dans une situation où la population est très importante, pour des raisons de coût, on ne peut pas étudier toute la population

On sélectionne aléatoirement un échantillon de taille n de cette population.

Estimation : on estime la proportion ***p*** par un intervalle de confiance déterminé à partir de f et de n selon un niveau de confiance

Remarque : la fréquence f calculée varie d'un échantillon à l'autre du fait de la fluctuation d'échantillonnage. Il est donc nécessaire d'apprécier l'incertitude en fournissant une estimation par un intervalle.

1. Intervalle de confiance

Théorème 2 :

Notons f la fréquence observée dans un échantillon de taille n et p le pourcentage que l'on veut estimer.

L'intervalle $\left[f-\frac{1}{\sqrt{n}};f+\frac{1}{\sqrt{n}}\right]$ contient la proportion ***p*** avec **une probabilité d'au moins 0,95**

Conditions d'application. Ce sont les mêmes que celles du théorème 1 :

$n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$

Définition :

On dit que l'intervalle $\left[f-\frac{1}{\sqrt{n}};f+\frac{1}{\sqrt{n}}\right]$ est **l'intervalle de confiance de *p* au niveau de confiance 0,95** (on dit aussi avec un risque de 5%).

Remarque : Par rapport au paragraphe précédent, l'intervalle ne peut être centré sur la proportion p que l'on ne connait pas. Il est centré sur la fréquence obtenue à l'aide de l'échantillon.

Exemple

Un magasin en ligne souhaite estimer la proportion p de ses clients satisfaits de ses services, au niveau de confiance de 0,95.

Pour cela, il interroge un échantillon de 772 clients. 82% des clients de cet échantillon se déclarent satisfait.

$\left[0{,}82-\frac{1}{\sqrt{772}};0{,}82+\frac{1}{\sqrt{772}}\right] \approx [0{,}78;0{,}86]$, donc ce magasin en ligne peut affirmer qu'avec un niveau de confiance de 0,95 entre 78% et 86% de ses clients sont satisfaits.

Remarques :

- On admet que l'on utilise cet intervalle dès que les conditions $n \geq 30$ et $np \geq 5$ et $n(1-p) \geq 5$ sont remplies.

Dans notre exemple, les conditions d'application du théorème 2 sont bien remplies :

- n=772, donc $n \geq 30$
- $np \geq 5$ en supposant que p est compris entre 0,7 et 0,9
- $n(1-p) \geq 5$

2. Taille minimale de l'échantillon pour avoir une précision donnée

Au niveau de confiance de 0,95, l'amplitude de l'intervalle de confiance est $\frac{2}{\sqrt{n}}$.

Si on souhaite situer ***p*** dans un intervalle de longueur donnée a, alors $\frac{2}{\sqrt{n}} \leq a$ donc $n \geq \frac{4}{a^2}$