# Chapitre 8

# ÉCHANTILLONNAGE ET ESTIMATION

# I FLUCTUATION D'ÉCHANTILLONNAGE

## 1 INTERVALLE DE FLUCTUATION AU SEUIL DE 95%

On s'intéresse à un caractère de proportion $p$ connue au sein d'une population.
On considère la variable aléatoire $F_n$ qui à chaque échantillon aléatoire de taille $n$ associe la fréquence du caractère étudié.

### DÉFINITION

On appelle intervalle de fluctuation de $F_n$ au seuil de 95%, tout intervalle $[\alpha;\beta]$ tel que la probabilité $P\left(F_n \in [\alpha;\beta]\right) \geqslant 0{,}95$

EXEMPLE

En première partie de soirée une série a attiré près de 6,2 millions de téléspectateurs soit 34 % de part d'audience.
Déterminons un intervalle de fluctuation de la part d'audience de cette série pour un échantillon de taille 100.
Soit $X$ la variable aléatoire qui correspond au nombre de téléspectateurs qui ont regardé cette série dans un échantillon de 100 personnes ayant regardé la télévision en première partie de soirée.
Le nombre de téléspectateurs en première partie de soirée est suffisamment important pour considérer que la variable $X$ suit la loi binomiale de paramètres $n = 100$ et $p = 0{,}34$.
Le plus petit entier $a$ tel que $P(X \leqslant a) > 0{,}025$ est 25 et, le plus petit entier $b$ tel que $P(X \leqslant b) \geqslant 0{,}975$ est 43.
Un intervalle de fluctuation à 95% de la fréquence des téléspectateurs qui ont regardé cette série dans un échantillon de taille 100 est :

$$I = \left[\frac{25}{100};\frac{43}{100}\right] \text{ soit } I = [0{,}25;0{,}43]$$

## 2 INTERVALLE DE FLUCTUATION ASYMPTOTIQUE AU SEUIL DE 95%

On appelle intervalle de fluctuation asymptotique au seuil de 95% de la variable aléatoire $F_n$, l'intervalle :

$$I_n = \left[p - 1{,}96 \times \sqrt{\frac{p(1-p)}{n}}; p + 1{,}96 \times \sqrt{\frac{p(1-p)}{n}}\right].$$

### INTERPRÉTATION

L'intervalle $I_n$ contient la fréquence $F_n$ avec une probabilité proche de 0,95 pourvu que $n$ soit suffisamment grand.
En pratique, on utilise l'intervalle de fluctuation asymptotique au seuil 0,95 dès que :

$$n \geqslant 30,\ np \geqslant 5 \text{ et } n(1-p) \geqslant 5.$$

EXEMPLE

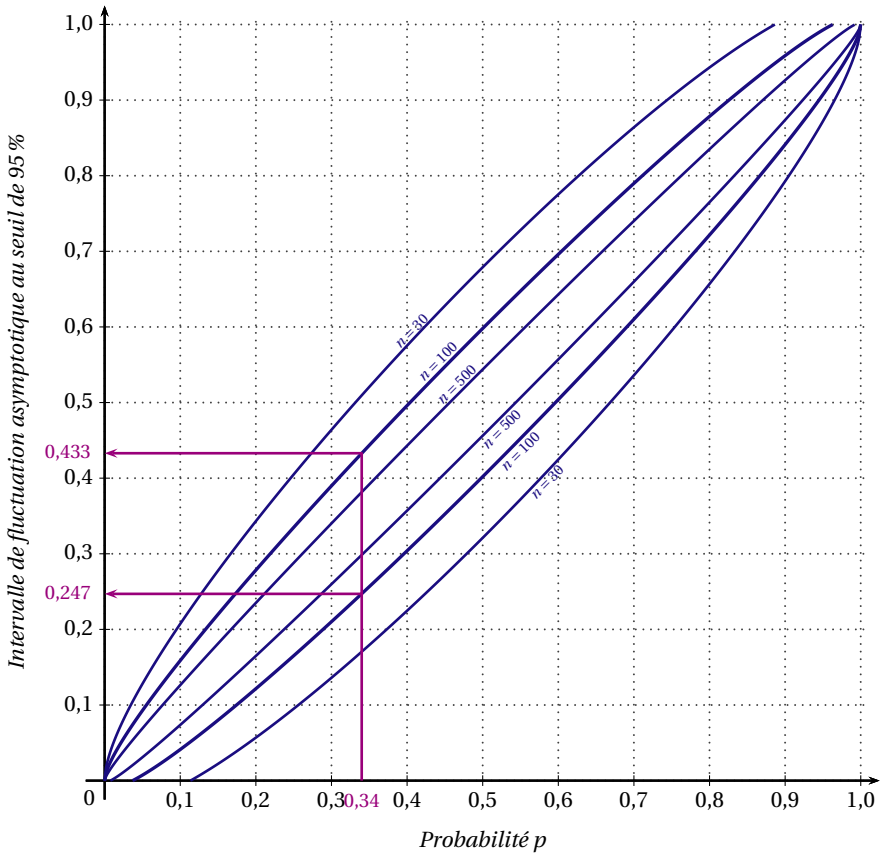
Avec $p = 0{,}34$ et $n = 100$ on a $np = 34$ et $n(1-p) = 66$, les critères d'approximation sont vérifiés.
L'intervalle de fluctuation asymptotique à 0,95 sur un échantillon de taille 100 est :

$$I_{100} = \left[0{,}34 - 1{,}96 \times \sqrt{\frac{0{,}34 \times 0{,}66}{100}}; 0{,}34 + 1{,}96 \times \sqrt{\frac{0{,}34 \times 0{,}66}{100}}\right]$$

Soit avec des valeurs approchées à $10^{-3}$ près des bornes de l'intervalle, $I_{100} \approx [0{,}247;0{,}433]$.

REMARQUE

On a tracé ci-dessous, sur l'intervalle $]0;1[$, les courbes représentatives des fonctions $f_{\text{inf}} = p - 1{,}96 \times \sqrt{\frac{p(1-p)}{n}}$ et $f_{\text{sup}} = p + 1{,}96 \times \sqrt{\frac{p(1-p)}{n}}$ associées aux bornes des intervalles de fluctuation asymptotique au seuil de 95 % pour les valeurs de $n = 30$, $n = 100$ et $n = 500$.




### 3 DÉCISION À PARTIR DE LA FRÉQUENCE D'UN ÉCHANTILLON

Quand les critères d'approximation sont vérifiés, l'intervalle de fluctuation asymptotique $I_n$ permet de déterminer des seuils de décision :

— pour accepter ou rejeter l'hypothèse selon laquelle $p$ est la proportion d'un caractère dans la population ;

— pour déterminer si un échantillon issu de la population est représentatif.

On formule l'hypothèse que la proportion d'un caractère dans la population est $p$.
On prélève dans la population un échantillon de taille $n$ et on note $f$ la fréquence observée du caractère étudié.
Lorsque $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$ on pose :

$$I_n = \left[ p - 1{,}96 \times \sqrt{\frac{p(1-p)}{n}} ; p + 1{,}96 \times \sqrt{\frac{p(1-p)}{n}} \right].$$

— Si la fréquence observée $f$ n'appartient pas à l'intervalle $I_n$, alors on rejette l'hypothèse selon laquelle $p$ est la proportion du caractère étudié dans la population avec un risque d'erreur de 5 %.

— Si la fréquence observée $f$ appartient à l'intervalle $I_n$, alors l'hypothèse selon laquelle $p$ est la proportion du caractère étudié dans la population est acceptée.

EXEMPLE

Dans un forum on a constaté que 28 personnes sur 100 ont regardé la série dont la part d'audience a été estimée à 34 %. Ce résultat remet-il en question l'estimation de la part d'audience de la série ?

La fréquence observée de la part d'audience dans l'échantillon de taille 100 est : $f = \dfrac{28}{100} = 0{,}28$.
L'intervalle de fluctuation asymptotique au seuil de 95 % de la part d'audience de la série dans les échantillons de taille 100 est $I_{100} = [0{,}247 ; 0{,}433]$.
Comme $0{,}28 \in [0{,}247 ; 0{,}433]$, l'estimation d'une part d'audience de 34 % pour la série n'est pas remise en cause.

## II INTERVALLE DE CONFIANCE

On cherche à estimer avec un certain niveau de confiance, la proportion $p$ **inconnue** d'un caractère au sein d'une population à partir d'un échantillon de taille $n$.

### 1 DÉFINITION

Soit $f$ la fréquence observée d'un caractère dans un échantillon de taille $n$.
Sous les conditions usuelles d'approximation $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$, l'intervalle $\left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}\right]$ est un intervalle de confiance au niveau de confiance 0,95 de la proportion inconnue $p$ dans la population.

#### REMARQUES

— En pratique, les conditions de validité de la formule peuvent être vérifiées à posteriori.

— La précision de l'intervalle de confiance est donnée par son amplitude $\frac{2}{\sqrt{n}}$. Plus la taille de l'échantillon est grande, plus les intervalles de confiance obtenus sont précis.

— La différence entre deux fréquences $f_1$ et $f_2$ observées sur deux échantillons est considérée comme significative quand les intervalles de confiance correspondants sont disjoints.
Dans ce cas, on considère que les deux proportions $p_1$ et $p_2$ sont différentes. Dans le cas contraire, on ne peut pas conclure.

#### EXEMPLE

On interroge au hasard 100 clients ayant effectué des achats à la sortie d'une grande surface. Le temps d'attente aux caisses a été jugé raisonnable par 52 personnes interrogées.

1. Peut-on considérer que plus de 50% des clients de cette grande surface estiment que le temps d'attente aux caisses est raisonnable ?
Soit $f = \frac{52}{100} = 0{,}52$ la fréquence des clients qui estiment que le temps d'attente aux caisses est raisonnable.
Les bornes de l'intervalle de confiance au niveau de confiance 95% de la proportion des clients qui estiment que le temps d'attente aux caisses est raisonnable sont :

$$f - \frac{1}{\sqrt{n}} = 0{,}52 - \frac{1}{\sqrt{100}} = 0{,}42 \text{ et } f + \frac{1}{\sqrt{n}} = 0{,}52 + \frac{1}{\sqrt{100}} = 0{,}62$$

On a : $n = 100$, $0{,}42 \leqslant p \leqslant 0{,}62$, $100 \times 0{,}42 \leqslant np \leqslant 100 \times 0{,}62$ et $100 \times (1-0{,}62) \leqslant n\left(1-p\right) \leqslant 100 \times (1-0{,}42)$.
Soit $n \geqslant 30$, $42 \leqslant np \leqslant 62$ et $38 \leqslant n\left(1-p\right) \leqslant 58$. Les conditions d'approximation d'un intervalle de confiance au niveau de confiance 95 % sont vérifiées.
Un intervalle de confiance au niveau de confiance 95 % de la proportion de clients qui estiment que le temps d'attente aux caisses est raisonnable est $[0{,}42; 0{,}62]$.
La borne inférieure de l'intervalle de confiance est 0,42, il est donc possible que moins de 50% des clients trouvent que le temps d'attente aux caisses est raisonnable.

2. Déterminer le nombre minimal de clients qu'il faut interroger pour estimer la proportion $p$ de clients qui trouvent le temps d'attente aux caisses raisonnable avec une précision inférieure à 0,1.
La précision de l'estimation de $p$ est $\frac{2}{\sqrt{n}}$. Pour tout entier naturel $n$ :

$$\begin{aligned}
\frac{2}{\sqrt{n}} < 0{,}1 &\iff \frac{1}{\sqrt{n}} < 0{,}05 \\
&\iff \sqrt{n} > \frac{1}{0{,}05} \\
&\iff \sqrt{n} > 20 \\
&\iff n > 400
\end{aligned}$$

Il faut interroger plus de 400 clients pour obtenir une estimation de la proportion $p$ de clients qui trouvent le temps d'attente aux caisses raisonnable avec une précision inférieure à 0,1.

3. À fréquence observée égale à 0,52, quel nombre de clients aurait-il fallu interroger pour estimer que plus de 50% des clients trouvent que le temps d'attente aux caisses est raisonnable ?

   La borne inférieure de l'intervalle de confiance au niveau de confiance 0,95 sur un échantillon de taille $n$ est $0{,}52 - \dfrac{1}{\sqrt{n}}$ d'où $n$ est solution de l'inéquation :

$$\begin{aligned} 0{,}52 - \frac{1}{\sqrt{n}} \geqslant 0{,}5 &\iff -\frac{1}{\sqrt{n}} \geqslant -0{,}02 \\ &\iff \frac{1}{\sqrt{n}} \leqslant 0{,}02 \\ &\iff \sqrt{n} \geqslant \frac{1}{0{,}02} \\ &\iff \sqrt{n} \geqslant 50 \\ &\iff n \geqslant 2500 \end{aligned}$$

Avec une fréquence observée égale à 0,52, il faudrait un échantillon de taille supérieure à 2500 pour que la proportion $p$ de clients qui trouvent le temps d'attente aux caisses raisonnable appartienne à un intervalle de confiance dont la borne inférieure est supérieure à 0,5.