# Échantillonnage et estimation

Dans ce chapitre, on s'intéresse à un caractère dans une population donnée dont la proportion est notée $p$. Cette proportion sera dans quelques cas connue (échantillonnage), dans certains cas supposée connue (prise de décision) et dans d'autres cas inconnue (estimation).

Pour des raisons généralement économiques, on étudie le caractère, non pas sur la population entière, mais sur des échantillons de taille $n$ extraits de cette population. Pour ce faire, on prélève au hasard des individus de cette population un par un **avec remise**. On constitue ainsi un échantillon aléatoire de taille $n$.
Dans une situation tel qu'un sondage on pourrait donc être amené à interroger deux fois la même personne. Pour éviter cela on procède successivement et **sans remise**. Si la taille de l'échantillon n'excède pas 10 % de la taille de la population entière, ce prélèvement ne modifie pas sensiblement la proportion du caractère dans la population.

## 1. Échantillonnage et prise de décisions

### ❖ Intervalle de fluctuation d'une fréquence

Considérons une population dans laquelle un caractère qualitatif est répandu avec une proportion connue $p$.
Prélevons un échantillon de taille $n$ de cette population et soit $X$ la variable aléatoire qui compte le nombre d'individu possédant le caractère ; cette variable suit la loi binomiale de paramètres $n$ et $p$. La variable aléatoire $F = \frac{X}{n}$ est la fréquence aléatoire du caractère dans l'échantillon.

**Exemple A**

Dans une urne contenant 4 boules rouges et 6 boules blanches, on effectue 100 tirages au hasard avec remise.
Le nombre $X$ de boules rouges suit la loi binomiale de paramètres $n = 100$ et $p = 0{,}4$. La fréquence de boules rouges est donnée par la variable aléatoire $F = \frac{X}{n} = \frac{X}{100}$.
À l'aide de la calculatrice, on peut effectuer divers calculs. Par exemple :

- La probabilité qu'on observe 42 % de boules rouges est
$$P(F = 0{,}42) = P(X = 42) \approx 0{,}0742.$$
- La probabilité qu'on observe au moins 40 % de boules rouges est
$$P(F \geq 0{,}4) = P(X \geq 40) = 1 - P(X \leq 39) \approx 0{,}538.$$
- La probabilité qu'on observe entre 31 % et 50 % de boules rouges est
$$P(0{,}31 \leq F \leq 0{,}50) = P(31 \leq X \leq 50) = P(X \leq 50) - P(X \leq 30) \approx 0{,}958.$$

**Définition.** Soit $\alpha \in ]0;1[$. *On appelle intervalle de fluctuation au seuil de* $1-\alpha$ *un intervalle* $I$ *tel que* $P(F \in I) \geq 1 - \alpha$.

En seconde, on a admis que l'intervalle $\left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$ est un intervalle de fluctuation au seuil de 95 %. Le résultat était énoncé sous cette forme.

**Théorème 1.** Si $p$ est la proportion d'un caractère dans une population (avec $0{,}2 \leq p \leq 0{,}8$), alors pour un échantillon de taille $n$ avec $n \geq 25$, la fréquence $f$ du caractère dans l'échantillon appartient à l'intervalle $\left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$ avec une probabilité 95 %.

**Exemple A**

L'intervalle de fluctuation au seuil de 95 % de la fréquence de boules rouges observées dans des échantillons de taille 100 est

$$\left[0{,}4 - \frac{1}{\sqrt{100}}\ ; 0{,}4 + \frac{1}{\sqrt{100}}\right] = [0{,}3; 0{,}5].$$

Cela signifie qu'avec une probabilité de 95 %, la fréquence de boules rouges dans un échantillon de taille 100 sera comprise entre 30 % et 50 %

## ❖ Intervalle de fluctuation et loi binomiale (rappel de Première S)

La loi binomiale permet de calculer très exactement des intervalles de fluctuations au seuil de 95 %. On cherche deux entiers $a$ et $b$ tels que $P(a \leq X \leq b) \geq 0{,}95$. Pour cela, on prendra les plus petits entiers $a$ et $b$ tels

$$P(X \leq a) > 0{,}025 \text{ et } P(X \leq b) \geq 0{,}975.$$

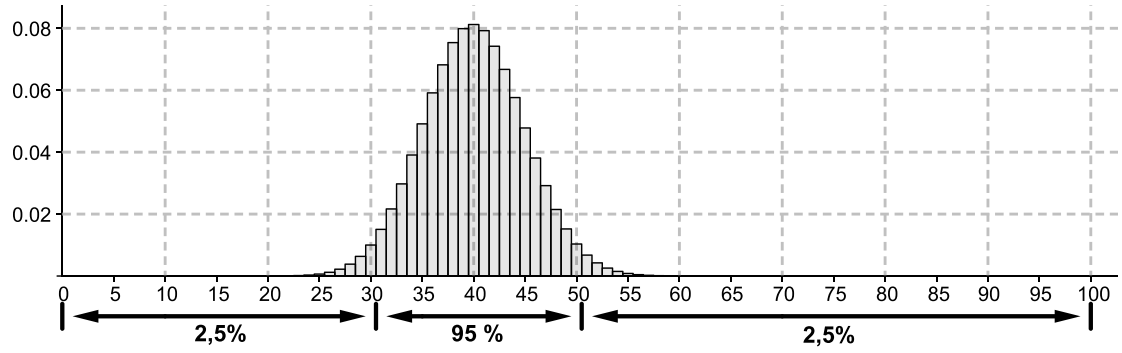
Puisque $a$ est le plus petit entier vérifiant $P(X \leq a) > 0{,}025$, on a $P(X \leq a - 1) \leq 0{,}025$ et donc

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a - 1) \geq 0{,}975 - 0{,}025 = 0{,}95.$$

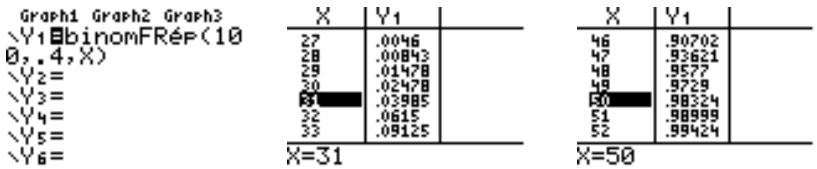
Il en résulte en divisant par $n$ que $P\left(\frac{a}{n} \leq F \leq \frac{b}{n}\right) \geq 0{,}95$, et donc l'intervalle $\left[\frac{a}{n}; \frac{b}{n}\right]$ est un intervalle de fluctuation au seuil de 95 %.

**Exemple A**

On cherche deux entiers $a$ et $b$ tels que $P(a \leq X \leq b) \geq 0{,}95$.




À l'aide d'un tableur ou d'une calculatrice, on constate que $a = 31$ est le plus petit entier tel que $P(X \leq a) > 0{,}025$ et $b = 50$ est le plus petit entier tel que $P(X \leq b) \geq 0{,}975$. Par conséquent on a $P(31 \leq X \leq 50) \geq 0{,}95$ ou encore $P(0{,}31 \leq F \leq 0{,}50) \geq 0{,}95$.




Donc $[0{,}31; 0{,}50]$ est un intervalle de fluctuation au seuil de 95 % de la fréquence de boules rouges observées dans les échantillons de taille 100.

Avec la méthode de seconde on avait trouvé $[0{,}3; 0{,}5]$, ce qui à peine moins précis, mais avec des calculs bien plus simples.

## ❖ Intervalle de fluctuation asymptotique et loi normale

Soit $X_n$ une variable aléatoire suivant la loi binomiale de paramètres $n$ et $p$. D'après le théorème de Moivre-Laplace, sa loi « converge » vers celle de la loi normale d'espérance $\mu = np$ et d'écart-type $\sigma = \sqrt{np(1-p)}$ lorsque $n$ tend vers $+\infty$.
En pratique dès que $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$ on utilise cette approximation.

**Exemple A**

Le nombre $X$ de boules rouges suit la binomiale de paramètres $n = 100$ et $p = 0{,}4$, sa loi est donc « proche » d'une loi normale $\mathcal{N}(40; 24)$.
Cherchons un intervalle de fluctuation asymptotique au seuil de 95 % pour $F$ grâce à cette approximation. D'après le théorème de Moivre-Laplace, la loi de la variable aléatoire $Z = \frac{X-40}{\sqrt{24}}$ peut être approximée par celle d'une loi $\mathcal{N}(0; 1)$ et on sait qu'alors pour une telle loi, il existe un unique réel positif $u$ tel que $P(-u \leq Z \leq u) = 0{,}95$ ($u \approx 1{,}96$), donc

$$P\left(-u \leq \frac{X-40}{\sqrt{24}} \leq u\right) = 0{,}95 \Leftrightarrow P\left(40 - \sqrt{24}u \leq X \leq 40 + \sqrt{24}u\right) = 0{,}95$$

d'où $P(30{,}4 \leq X \leq 49{,}6) = 0{,}95$, puis $P(0{,}304 \leq F \leq 0{,}496) = 0{,}95$. Cela montre que $[0{,}304; 0{,}496]$ est un intervalle de fluctuation au seuil de 95 % de $F$.

Plus généralement, on a le résultat suivant.

**Théorème 2.** Si $X_n$ suit la loi binomiale de paramètres $n$ et $p$ ($p$ fixé), alors pour tout réel $\alpha \in ]0; 1[$, on a $\lim\limits_{n\to+\infty} P(F_n \in I_n) = 1 - \alpha$ où $F_n = \frac{X_n}{n}$, $I_n = \left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ et $u_\alpha$ est le réel tel que $P(-u_\alpha \leq Z \leq u_\alpha) = 1 - \alpha$ lorsque $Z$ suit la loi $\mathcal{N}(0; 1)$.

**Démonstration (exigible).** Soit $Z$ une variable aléatoire suivant une loi $\mathcal{N}(0; 1)$ et posons $Z_n = \frac{X_n - np}{\sqrt{np(1-p)}}$. D'après le théorème de Moivre-Laplace,

$$\lim_{n\to+\infty} P(-u_\alpha \leq Z_n \leq u_\alpha) = P(-u_\alpha \leq Z \leq u_\alpha) = 1 - \alpha.$$

Or l'inégalité $-u_\alpha \leq Z_n \leq u_\alpha$ équivaut à

$$-u_\alpha \leq \frac{X_n - np}{\sqrt{np(1-p)}} \leq u_\alpha \Leftrightarrow np - u_\alpha\sqrt{np(1-p)} \leq X_n \leq np + u_\alpha\sqrt{np(1-p)}$$

$$\Leftrightarrow p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq F_n \leq p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \Leftrightarrow F_n \in I_n. \quad \blacksquare$$

Si les conditions d'approximation du théorème de Moivre-Laplace sont réunies, on peut alors énoncer le résultat suivant.

**Corollaire 3.** Si $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$, l'intervalle

$$\left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$$

est un intervalle de fluctuation au seuil de $1 - \alpha$ de la variable aléatoire fréquence $F_n$.

Cet intervalle de fluctuation est dit asymptotique car il n'est valide que si les conditions sur $n$ et $p$ sont satisfaites.

On rappelle qu'on a notamment $u_{0,05} \approx 1{,}96$, donc un intervalle de fluctuation asymptotique au seuil de 95 % de $F_n$ est $\left[p - 1{,}96\ \frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1{,}96\ \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$.

La borne inférieure de l'intervalle de fluctuation IF sera arrondie par défaut, et la borne supérieure par excès pour ne pas réduire l'intervalle et prendre le risque qu'il ne vérifie plus $P(F \in \text{IF}) \geq 1 - \alpha$.

**Exemple A**

Cherchons un intervalle de fluctuation au seuil de confiance de 97 % de la fréquence de boules rouges dans les échantillons de taille 100.
Avec la commande `FracNormale(1-0.97/2)`, on trouve que $u_{0,03} \approx 2{,}17$, donc

$$\text{IF} = \left[0{,}4 - 2{,}17\frac{\sqrt{0{,}4\times(1-0{,}4)}}{\sqrt{100}};\ 0{,}4 + 2{,}17\frac{\sqrt{0{,}4\times(1-0{,}4)}}{\sqrt{100}}\right] \approx [0{,}293; 0{,}507].$$

On peut justifier le résultat vu en Seconde.

**Théorème 4.** Soit $X_n$ une variable aléatoire suivant la loi binomiale de paramètres $n$ et $p$. Pour $n$ assez grand, $P\left(p - \frac{1}{\sqrt{n}} \leq \frac{X_n}{n} \leq p + \frac{1}{\sqrt{n}}\right) \geq 0{,}95$.

**Démonstration.** Posons $Z_n = \frac{X_n - np}{\sqrt{np(1-p)}}$ et $a_n = P(-2 \leq Z_n \leq 2)$. D'après le théorème de Moivre-Laplace, $\lim\limits_{n\to+\infty} a_n = P(-2 \leq Z \leq 2)$ où $Z$ suit la loi $\mathcal{N}(0; 1)$. Soit $L$ cette limite.
Puisque $P(-2 \leq Z \leq 2) = 2P(Z \leq 2) - 1 \approx 0{,}954$, on a $L > 0{,}95$, et il existe donc un rang $n_0$ tel que pour $n \geq n_0$, on ait $a_n \geq 0{,}95$.
Par ailleurs $a_n = P\left(p - \frac{2}{\sqrt{n}}\sqrt{p(1-p)} \leq \frac{X_n}{n} \leq p + \frac{2}{\sqrt{n}}\sqrt{p(1-p)}\right)$.
Pour tout $p \in [0; 1]$, une étude de la fonction $f: p \mapsto p(1-p)$ montre que $0 \leq p(1-p) \leq \frac{1}{4}$.
En effet, sa dérivée est $f'(p) = 1 - 2p$, elle atteint donc un maximum égal à $f\left(\frac{1}{2}\right) = \frac{1}{4}$.
Il en résulte que $\frac{2}{\sqrt{n}}\sqrt{p(1-p)} \leq \frac{1}{\sqrt{n}}$. Ainsi

$$\left[p - \frac{2}{\sqrt{n}}\sqrt{p(1-p)}; p + \frac{2}{\sqrt{n}}\sqrt{p(1-p)}\right] \subset \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right],$$

donc si $n \geq n_0$, il vient $P\left(p - \frac{1}{\sqrt{n}} \leq \frac{X_n}{n} \leq p + \frac{1}{\sqrt{n}}\right) \geq a_n \geq 0{,}95$. ■

## ❖ Prise de décision

Dans ce paragraphe, la proportion du caractère étudié est supposée être égale à $p$.
La prise de décision consiste, à partir d'un échantillon de taille $n$, à valider ou non cette hypothèse faite sur la proportion $p$. Pour ce faire,

- on calcule la fréquence observée $f$ du caractère étudié dans cet échantillon ;
- on détermine un intervalle de fluctuation au seuil de 95 % de la fréquence ;
- enfin on applique la règle de décision suivante
  - si la fréquence observée $f$ appartient à l'intervalle de fluctuation, on ne rejette pas l'hypothèse faite sur $p$ ;
  - si la fréquence observée $f$ n'appartient pas à l'intervalle de fluctuation, on rejette l'hypothèse faite sur la proportion $p$ avec un risque de 5 % de se tromper.

**Remarque.** Le risque de 5 % signifie que la probabilité que l'on rejette à tort l'hypothèse faite sur la proportion $p$ alors qu'elle est vraie est approximativement égale à 5 %. C'est une probabilité conditionnelle.
Dans le cas où l'on ne rejette pas l'hypothèse faite sur la proportion $p$, le risque d'erreur n'est pas quantifié.
On peut bien entendu changer le seuil de 95 % si nécessaire.

**Exemple A**

Une machine fabrique des composants électroniques qui présentent un défaut de fabrication avec une probabilité 0,4.
On constitue un échantillon de taille 100 de composants issus de la fabrication. On en détecte 49 présentant un défaut de fabrication. La machine nécessite-t-elle d'être réparée ?

On fait l'hypothèse que la machine est bien réglée, donc que $p = 0{,}4$.
Un intervalle de fluctuation au seuil de 95 % de la fréquence de pièces présentant un défaut sur les échantillons de taille 100 est $[0{,}304; 0; 496]$.
Ici $f = \frac{49}{100} = 0{,}49$ appartient à l'intervalle de fluctuation, donc on ne rejette pas l'hypothèse $p = 0{,}4$.
On peut donc dire que la machine ne nécessite pas d'être réparée, au seuil de confiance de 95 %.

## 2. Estimation et intervalle de confiance

Le problème posé ici est « l'inverse » de celui de l'échantillonnage : à partir de la fréquence $f$ observée sur un échantillon, comment peut-on estimer la proportion $p$ correspondante dans la population entière ?

**Théorème 5.** Soit $F_n$ la variable aléatoire fréquence qui, à tout échantillon de taille $n$ extrait d'une population dans laquelle la proportion d'un caractère est $p$, associe la fréquence obtenue.
Alors l'intervalle $\left[F_n - \frac{1}{\sqrt{n}} ; F_n + \frac{1}{\sqrt{n}}\right]$ contient, pour $n$ assez grand, la proportion $p$ avec une probabilité d'au moins 0,95.

**Démonstration.** D'après le théorème 4, si $n$ est assez grand,

$$P\left(F_n \in \left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}\right]\right) \geq 0{,}95.$$

Or

$$F_n \in \left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}\right] \Leftrightarrow p - \frac{1}{\sqrt{n}} \leq F_n \leq p + \frac{1}{\sqrt{n}} \Leftrightarrow \begin{cases} p - \frac{1}{\sqrt{n}} \leq F_n \\ F_n \leq p + \frac{1}{\sqrt{n}} \end{cases} \Leftrightarrow \begin{cases} p \leq F_n + \frac{1}{\sqrt{n}} \\ F_n - \frac{1}{\sqrt{n}} \leq p \end{cases}$$

$$\Leftrightarrow F_n - \frac{1}{\sqrt{n}} \leq p \leq F_n + \frac{1}{\sqrt{n}},$$

d'où

$$P\left(p \in \left[F_n - \frac{1}{\sqrt{n}} ; F_n + \frac{1}{\sqrt{n}}\right]\right) \geq 0{,}95. \quad \blacksquare$$

**Définition.** *Soit $f$ la fréquence observée d'un caractère dans un échantillon de taille $n$ extrait d'une population dans laquelle la proportion de ce caractère est $p$.*
*L'intervalle $\left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}}\right]$ est un intervalle de confiance de la proportion $p$ au niveau de confiance 95 %.*

À chaque échantillon de taille $n$ correspond un intervalle de confiance de $p$.

**Exemple**

On cherche à estimer la proportion $p$ d'élèves du lycée pratiquant une activité physique régulière.
Sur 100 élèves interrogés, on en a observé 63 pratiquant un sport. Ainsi l'intervalle de confiance au seuil de 95 % de la proportion d'élèves pratiquant du sport est

$$\text{IC} = \left[\frac{63}{100} - \frac{1}{\sqrt{100}} ; 63 + \frac{1}{\sqrt{100}}\right] = [0{,}53 ; 0{,}73].$$

Un deuxième sondage a révélé 68 élèves sportifs. Cela conduit à un nouvel intervalle de confiance :

$$\text{IC}' = [0{,}58 ; 0{,}78].$$

Le niveau de confiance 95 % indique que si l'on réalise 100 sondages, $p$ se situera dans au moins 95 des intervalles de confiance ainsi construits.

**Remarques.**

- Il est incorrect de dire « la proportion se situe avec une probabilité égale à 95 % dans l'intervalle de confiance » car la proportion n'est pas aléatoire.
- Dans d'autres disciplines on utilise l'intervalle de confiance suivant :
$$\left[f - 1{,}96\ \frac{\sqrt{f(1-f)}}{\sqrt{n}} ; f + 1{,}96\ \frac{\sqrt{f(1-f)}}{\sqrt{n}}\right].$$
- La proportion $p$ étant inconnue, on ne peut pas vérifier si les paramètres $n$ et $p$ satisfont les conditions $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$. Pour remédier à ce problème, on approche la proportion inconnue $p$ par la fréquence observée $f$ sur l'échantillon, puis on vérifie si les conditions $n \geq 30$, $nf \geq 5$ et $n(1-f) \geq 5$ sont satisfaites.

**Exemple**

Un sondage réalisé sur 989 personnes par l'institut IPSOS le 19 avril 2002 avant le premier tour des élections présidentielles créditait Chirac de 20 % des votes, Jospin de 18 % et Le Pen de 14 %.
Les intervalles de confiance des proportions de votes pour ces trois candidats étaient donc respectivement

$$\text{IC}_C = \left[0{,}2 - \frac{1}{\sqrt{989}} ; 0{,}2 + \frac{1}{\sqrt{989}}\right] \approx [0{,}168 ; 0{,}232],$$

$$\text{IC}_J \approx [0{,}148 ; 0{,}212] \text{ et } \text{IC}_{LP} \approx [0{,}108 ; 0 ; 172]$$

Au seuil de confiance 0,95, on constate qu'il était impossible de prédire le classement de ces trois candidats à l'issue du premier tour puisque ces trois intervalles ne sont pas disjoints.
Leurs scores ont été respectivement 19,88 %, 16,18 % et 16,86 %, propulsant Le Pen au second tour contre toute attente (de la part des journalistes, mais pas des statisticiens).