# STATISTIQUES - ESTIMATION

## I Echantillonnage et estimation : introduction

On se situe ici dans 2 domaines des statistiques qui sont ceux de l'« échantillonnage » et de l'« estimation ». Ces 2 domaines ont des contextes d'application différents qu'il faut savoir connaître. (Ces domaines appartiennent au champ des statistiques « inférentielles »)

### 1. Identification de la situation

On considère deux urnes $U_1$ et $U_2$ contenant chacune un très grand nombre de boules rouges et bleues.

| | |
|---|---|
| Dans l'urne $U_1$ , **on connaît la proportion p de boules rouges.**<br><br>On procède à des tirages avec remise de $n$ boules, et on observe la fréquence d'apparition d'une boule rouge. Cette fréquence observée appartient « en général » à un « intervalle de fluctuation » de centre p dont la longueur diminue avec $n$.<br><br>Cet intervalle un « intervalle de fluctuation ».<br><br>On est ici dans le domaine de l'**échantillonnage**, et de l'**intervalle de fluctuation**. | Dans l'urne $U_2$ , **on ignore la proportion de boules rouges.**<br><br>En procédant à des tirages avec remise de $n$ boules, on va essayer d'estimer la proportion p de boules rouges dans l'urne, proportion dont on a aucune idée *a priori*. Cette estimation se fait au moyen d'un « intervalle de confiance ».<br><br>Cet intervalle dépend d'un coefficient, le « niveau de confiance », que l'on attribue à l'estimation.<br><br>On est ici dans le domaine de l'estimation, et de **l'intervalle de confiance**. |

### 2. Quel intervalle utiliser ?

On s'intéresse à une population, dont on étudie un caractère particulier.

| **Echantillonnage** | **Estimation** |
|---|---|
| On utilise un **intervalle de fluctuation** quand :<br><br>- On **connaît la proportion** p de présence du caractère dans la population<br><br>OU<br><br>- On **fait une hypothèse sur la valeur de cette proportion** (on est alors dans le cas de la « prise de décision ») | On utilise un **intervalle de confiance** quand :<br><br>On **ignore la valeur de la proportion p** de présence du caractère dans la population, et on ne formule pas d'hypothèse sur cette valeur. |

***Exemples :***

1) On dispose d'une pièce de monnaie. Comment décider qu'elle est « équilibrée » ou pas ?
On va ici faire l'hypothèse que la fréquence d'apparition de « Pile », par exemple, est égale à 0,5, et on va tester cette hypothèse.
On est dans une **situation d'échantillonnage**.

2) Une usine fabrique des fusées de feux d'artifice. Sur 100 fusées choisies au hasard à l'issue du processus de fabrication et mises à feu, on trouve 12 fusées qui ne fonctionnent pas.
Comment se faire une idée de la proportion des fusées défectueuses dans la production ?
On est dans une **situation d'estimation** : on n'a, au départ, aucune idée de la valeur de la proportion étudiée dans la population.

# II Rappels de 2nde et 1ère

## 1 Déterminer un intervalle de fluctuation (Seconde)

La proportion d'Écossais porteurs du gène celtique à l'origine de la rousseur est d'environ $p = 0{,}40$.
Sur un échantillon de 50 Écossais tous issus du même bassin de population, 17 sont porteurs de ce gène.

**a)** Déterminer, pour un échantillon de taille $n = 50$, l'intervalle de fluctuation au seuil de 95 % de la forme :

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

**b)** Calculer la fréquence $f$ du gène celtique dans l'échantillon. Peut-on considérer, au seuil de 95 %, que la population de porteurs du gène celtique est la même dans ce bassin de population que parmi l'ensemble des Écossais ?

**1** **a)** L'intervalle de fluctuation au seuil de 95% est :

$$\left[0{,}40 - \frac{1}{\sqrt{50}} ; 0{,}40 + \frac{1}{\sqrt{50}}\right] \approx [0{,}25 ; 0{,}55]$$

**b)** $\frac{17}{50} = 0{,}34$ et $0{,}34 \in [0{,}25 ; 0{,}55]$.
On ne peut pas rejeter cette hypothèse.

## 2 Déterminer un intervalle de fluctuation (Première)

On reprend la situation de l'exercice 1 et on note X la variable aléatoire qui compte le nombre de personnes qui ont le gène celtique dans l'échantillon.

**a)** Expliquer pourquoi X suit la loi binomiale $\mathcal{B}(50 ; 0{,}4)$.

**b)** Lire sur la table ci-contre :

- le plus petit nombre entier $a$ tel que $P(X \leqslant a) > 0{,}025$ ;
- le plus petit nombre entier $b$ tel que $P(X \leqslant b) \geqslant 0{,}975$.

**c)** Donner l'intervalle de fluctuation au seuil de 95 % de la forme $\left[\frac{a}{n} ; \frac{b}{n}\right]$.

**d)** Peut-on considérer, au seuil de 95 %, que la population de porteurs du gène celtique est la même dans ce bassin de population que parmi l'ensemble des Écossais ?

| | A | B |
|---|---|---|
| 1 | $k$ | $P(X \leq k)$ |
| 2 | 0 | 8,08281E-012 |
| 3 | 1 | 2,77510E-010 |
| 13 | ... | ... |
| 14 | 11 | 0,0056876858 |
| 15 | 12 | 0,0132505247 |
| 16 | 13 | 0,0279883646 |
| 17 | 14 | 0,0539550349 |
| 28 | ... | ... |
| 29 | 25 | 0,9426562395 |
| 30 | 26 | 0,9685944469 |
| 31 | 27 | 0,9839652365 |
| 32 | 28 | 0,9923825737 |

**2** **a)** On peut modéliser le choix d'une personne comme une épreuve de Bernoulli. On appelle succès une personne est rousse, sa probabilité est 0,40.
Il y a une répétition de 50 épreuves identiques et indépendantes. L'expérience décrite est bien un schéma de Bernoulli.
La variable aléatoire qui prend pour valeur le nombre de succès, c'est-à-dire le nombre de personnes rousses suit la loi binomiale de paramètres $n = 50$ et $p = 0{,}4$.

**b)** $a = 13$ et $b = 27$

**c)** L'intervalle de fluctuation au seuil de 95 % est :
$[0{,}26 ; 0{,}54]$.

**d)** $\frac{17}{50} = 0{,}34$ et $0{,}34 \in [0{,}26 ; 0{,}54]$.
On ne peut pas rejeter cette hypothèse.

# III Intervalle de fluctuation asymptotique

## 1. Intervalle

$X_n$ est une variable aléatoire qui suit la loi binomiale $\mathcal{B}(n, p)$.
Elle indique le nombre de succès dans un schéma de Bernouilli d'ordre $n$ et de paramètre $p$.
La variable aléatoire $F_n = \dfrac{X_n}{n}$ indique la fréquence de succès lors des $n$ épreuves.

**Théorème :**

Soit $X_n$ une variable aléatoire suivant la loi $\mathcal{B}(n, p)$ et $F_n = \dfrac{X_n}{n}$.

Pour tout réel $\alpha \in ]0;1[$, on a :

$$\lim_{n\to+\infty} P\left(F_n = \frac{X_n}{n} \in I_n\right) = 1-\alpha \quad \text{où} \quad I_n = \left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$$

$u_\alpha$ étant le réel tel que $P(-u_\alpha \le Z \le u_\alpha) = 1-\alpha$ où $Z$ suit la loi $\mathcal{N}(0\ ;1)$.

L'intervalle $I_n$ est un **intervalle de fluctuation asymptotique au seuil 1 – α.**

**ROC**

**Démonstration :**

D'après le théorème de Moivre-Laplace, sous certaines conditions ($n \ge 30$, $np \ge 5$ et $n(1-p) \ge 5$), on peut approcher la loi de la variable aléatoire $\dfrac{X_n - np}{\sqrt{np(1-p)}}$ par la loi $\mathcal{N}(0\ ;1)$.

Ainsi, la loi de cette variable aléatoire associée à la fréquence de succès $F_n$ peut être approchée par la loi normale $\mathcal{N}(0\ ;1)$.

On pose $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$.

D'après le théorème de Moivre-Laplace, $\lim\limits_{n\to+\infty} P(-u_\alpha \le Z_n \le u_\alpha) = P(-u_\alpha \le Z \le u_\alpha)$ où $Z \sim N(0;1)$.

Or

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

$$\Leftrightarrow -u_\alpha \sqrt{np(1-p)} \le X_n - np \le u_\alpha \sqrt{np(1-p)}$$

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

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

Donc $\lim\limits_{n\to+\infty} P\left(p - u_\alpha \dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \le \dfrac{X_n}{n} \le p + u_\alpha \dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right) = 1-\alpha$

**Propriété**

Comme $u_{0,05} \simeq 1,96$, un intervalle de fluctuation asymptotique de $F_n$ au seuil de 95 % est :

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

**Remarque** : Cette intervalle peut être simplifié par l'intervalle

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

En effet la fonction $x \mapsto x(1-x) = x - x^2$ est une fonction du second degré qui s'annule en 0 et 1, elle admet donc un maximum (coefficient négatif devant $x^2$) en 0,5. On a alors $f(0,5) = 0,25$. Elle est positive entre 0 et 1. On a alors :

$$0 \leqslant p(1-p) \leqslant 0,25 \quad \Leftrightarrow \quad 0 \leqslant \sqrt{p(1-p)} \leqslant \sqrt{0,25} = 0,5$$

On en déduit alors que : $0 \leqslant 1,96\sqrt{p(1-p)} \leqslant 1$

On a alors $0 \leqslant 1,96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant \dfrac{1}{\sqrt{n}}$

On a ainsi $I_n \subset J_n$. On a alors dans la plupart des cas $P(F_n \in J_n) \geqslant 0,95$

## 2. Prise de décision

On considère une population dans laquelle on **suppose** que la proportion d'un certain caractère est $p$.
On **observe** $f_{obs}$ comme fréquence de ce caractère dans un échantillon de taille $n$.
Soit l'hypothèse : « la proportion de ce caractère dans la population est $p$ ».
Si les conditions d'approximation $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$ sont remplies, alors un intervalle de fluctuation asymptotique au seuil de 95 % est $I = \left[ p - 1,96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 1,96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$ et

**la règle de décision** est la suivante :

- Si $f_{obs} \in I$, alors on **accepte** l'hypothèse que la proportion est p.
- Si $f_{obs} \notin I$, alors on **rejette** cette hypothèse au seuil considéré.

***Exercice :***

Pour créer ses propres colliers, on peut acheter un kit contenant des perles de cinq couleurs différentes (marrons, jaunes, rouges, vertes et bleues), dans des proportions affichées sur le paquet.

Ainsi les perles marron et les perles jaunes sont annoncées comme représentant chacunes 20% de l'ensemble des perles tandis que les perles rouges sont annoncées à 10%.

On veut vérifier cette information. Pour cela, on choisit d'observer un échantillon aléatoire de perles et de construire un intervalle de fluctuation asymptotique au seuil de 95% pour la proportion de perles marron.

On constitue donc un échantillon, que l'on considère aléatoire, de 690 perles. On a dénombré 140 perles marron.

La **prise de décision** est la suivante : si la proportion de perles marron dans l'échantillon n'appartient pas à intervalle de fluctuation, on rejette l'hypothèse selon laquelle les perles marron représentent 20% des perles

a) Déterminer l'intervalle de fluctuation asymptotique I au seuil de 95% pour la proportion de perles marron.

b) Calculer la proportion de perles marron dans l'échantillon. Que peut-on en conclure ?

c) Dans le même échantillon, il y avait 152 perles jaunes et 125 perles rouges. Que peut-on conclure de ces résultats ?

a) En ce qui concerne les perles marron, on a : $n = 690$ et $p = 0,2$, donc :

$$n \geqslant 30 \quad np = 138 \geqslant 5 \quad \text{et} \quad n(1-p) = 552 \geqslant 5$$

Nous sommes bien dans les hypothèses du théorème de Moivre-Laplace.

On calcule ensuite

$$p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{2}} = 0,2 - 1,96\frac{\sqrt{0,2 \times 0,8}}{\sqrt{690}} \simeq 0,1702$$

$$p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{2}} = 0,2 + 1,96\frac{\sqrt{0,2 \times 0,8}}{\sqrt{690}} \simeq 0,2298$$

On a donc : $I = [0,17; 0,23]$

b) On calcule la fréquence $f_m = \frac{140}{690} \simeq 0,203$

Comme $f_m \in I$, **on ne peut pas rejeter** l'hypothèse selon laquelle les perles marron représentent 20% des perles.

c) On calcule la fréquence des perles jaunes : $f_j = \frac{152}{690} \simeq 0,220$

Comme $f_j \in I$, **on ne peut pas rejeter** l'hypothèse selon laquelle les perles jaunes représentent 20% des perles.

Pour les perles rouges, il faut calculer un nouvel intervalle de fluctuation :

$$p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{2}} = 0,1 - 1,96\frac{\sqrt{0,1 \times 0,9}}{\sqrt{690}} \simeq 0,0886$$

$$p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{2}} = 0,1 + 1,96\frac{\sqrt{0,1 \times 0,9}}{\sqrt{690}} \simeq 0,1224$$

On a donc : $I' = [0,08; 0,13]$ (on prend l'intervalle par excès)

On calcule la fréquence des perles rouges : $f_r = \frac{125}{690} \simeq 0,18$

Comme $f_r \notin I'$, **on doit rejeter** l'hypothèse selon laquelle les perles rouges représentent 10% des perles.

# III Estimation par intervalles de confiance

## 1. Présentation

Pour des raisons de coût et de faisabilité, on ne peut étudier un certain caractère sur l'ensemble d'une population. La proportion $p$ de ce caractère est donc inconnue.

On cherche alors à estimer $p$ à partir d'un échantillon de taille $n$. On calcule alors la fréquence $f_{obs}$ des individus de cet échantillon ayant ce caractère.

***Estimation :*** On estime la proportion $p$ par un intervalle de confiance déterminé à partir de $f_{obs}$ et de $n$ selon un niveau de confiance $1 - \alpha$.

***Remarque*** : La fréquence $f_{obs}$ 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.

## 2. Intervalle de confiance

On suppose que les 3 conditions sont remplies : $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$.

La variable $F_n$ prend ses valeurs dans l'intervalle $\left[ p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}} \right]$ d'où le résultat : la proportion inconnue $p$ est telle que $P\left( F_n - \frac{1}{\sqrt{n}} \leq p \leq F_n + \frac{1}{\sqrt{n}} \right) \approx 0,95$.

***Démonstration*** :

Or : $p - \frac{1}{\sqrt{n}} \leqslant F_n \leqslant p + \frac{1}{\sqrt{n}} \Leftrightarrow -\frac{1}{\sqrt{n}} \leqslant F_n - p \leqslant \frac{1}{\sqrt{n}} \Leftrightarrow -F_n - \frac{1}{\sqrt{n}} \leqslant -p \leqslant -F_n + \frac{1}{\sqrt{n}}$ ; donc :
$p - \frac{1}{\sqrt{n}} \leqslant F_n \leqslant p + \frac{1}{\sqrt{n}} \Leftrightarrow F_n - \frac{1}{\sqrt{n}} \leqslant p \leqslant F_n + \frac{1}{\sqrt{n}}$. Ainsi, $P\left( F_n - \frac{1}{\sqrt{n}} \leqslant p \leqslant F_n + \frac{1}{\sqrt{n}} \right) \approx 0,95$.

***Définition :***

On observe une fréquence $f_{obs}$ sur un échantillon de taille $n$. On appelle **intervalle de confiance de $p$ au niveau de confiance de 95 %** l'intervalle $\left[ f_{obs} - \frac{1}{\sqrt{n}} ; f_{obs} + \frac{1}{\sqrt{n}} \right]$.

***Remarques*** :

- On utilise cet intervalle dès que : $n \geq 30$, $nf \geq 5$ et $n(1-f) \geq 5$
- Cet intervalle de confiance a pour amplitude $\frac{2}{\sqrt{n}}$. Ainsi si l'on souhaite encadrer $p$ dans un intervalle de longueur $a$, on doit avoir : $\frac{2}{\sqrt{n}} \leq a \Leftrightarrow n \geq \frac{4}{a^2}$
- On admet que l'intervalle $\left[ f - 1,96\frac{\sqrt{f(1-f)}}{\sqrt{n}} ; f + 1,96\frac{\sqrt{f(1-f)}}{\sqrt{n}} \right]$ est aussi un intervalle de confiance $p$ au niveau de confiance de 95 %.

***Exercice :***

Deux candidats se présentent à une élection. Un sondage portant sur un échantillon de 1 200 personnes donne 53 % des suffrages au candidat A.

1. Déterminer, au niveau de confiance 95 %, un intervalle de confiance de la proportion $p$ des votants pour le candidat A.
2. Au seuil de confiance 95 %, le candidat A peut-il croire en sa victoire ?
3. Quelle devrait être la taille minimale de l'échantillon afin que l'amplitude de l'intervalle de confiance de cette proportion soit inférieure à 4 % .

1. On a $n$ = 1 200 et $f$ = 0,53 donc $nf = 636$ et $n(1-f) = 564$ et les conditions sont remplies.
   Un intervalle de confiance de $p$ au niveau de confiance de 95 % est
   $\left[ 0,53 - \frac{1}{\sqrt{1200}} ; 0,53 + \frac{1}{\sqrt{1200}} \right] \approx [0,501 ; 0,559]$
2. 0,501 > 0,5 donc au seuil de 95 %, le candidat A peut croire en sa victoire.
3. On doit avoir $\frac{2}{\sqrt{n}} \leq 0,04 \Leftrightarrow \frac{\sqrt{n}}{2} \geq \frac{1}{0,04} \Leftrightarrow \sqrt{n} \geq 50 \Leftrightarrow n \geq 2500$ . La taille de l'échantillon doit être au minimum de 2 500.