CHAPITRE 13 : FLUCTUATION ET ESTIMATION

I Introduction

I.1 Identification de la situation

On se situe, dans ce chapitre, dans deux domaines des statistiques, qui sont ceux de « l'échantillonnage » et de « l'estimation » .
Apprenons à reconnaître les différents contextes d'applications.

Exemples :
On considère deux urnes U_1 et U_2 contenant chacune un très grand nombre de boules, rouges ou bleues.

1. Dans l'urne U_1, **on connaît la proportion p de boules rouges**.

   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 lorsque n augmente.

   Cet intervalle est un « intervalle de fluctuation » .

   On est ici dans le domaine de **l'échantillonnage** et de l'**intervalle de fluctuation**.

2. Dans l'urne U_2, **on ignore la proportion de boules rouges**.

   En procédant à des tirages avec rmise de n boules, on va essayer d'estimer la proportion p de boules rouges dans l'urne.

   Cette estimation se fait au moyen d'un « intervalle de confiance » .

   Cet intervalle dépend d'un coefficient, le « niveau de confiance » , que l'on attribue à l'estimation.

   On est ici dans le domaine de l'**estimation**, et de l'**intervalle de confiance**.

Questions	**Réponses**
1. Question	☐ Proposition 1 ☐ Proposition 2 ☐ Proposition 3

I.2 Intervalle de fluctuation ou intervalle de confiance ?

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

Deux cas

- **Échantillonnage**

  On utilise un **intervalle de fluctuation** quand :

  – on **connaît la proportion** p de présence du caractère étudié dans la population,

  – OU, on **fait une hypothèse sur la valeur de cette proportion** (on est alors dans le cas de la « prise de décision » .

- **Estimation**

  On utilise un **intervalle de confiance** quand :

  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 de 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 production.

Exercice :

Un enfant a réalisé 4 040 lancers d'une pièce de monnaie, et il a obtenu 2 048 fois le résultat « Pile » .

Nous allons nous intéresser au problème suivant : « Peut-on considérer que la pièce utilisée est équilibrée ? »

1. Prendre une décision en utilisant l'intervalle de fluctuation au seuil de 95% sous la forme étudiée en classe de 2^{nde}.
2. Reprendre le même travail, en utilisant un intervalle de fluctuation au seuil 95% obtenu à partir de la loi binomiale vue en classe de $1^{\text{ère}}$.

Solution :

1. En classe de 2^{nde}, on a vu que pour $n > 25$ et $0,2 < p < 0,8$, l'intervalle $\left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$ constitue un intervalle de fluctuation au seuil 95%.

   Ici, si on appelle p la « proportion théorique » de résultats « Pile » obtenus sur un très grand nombre de lancers, on veut tester l'hypothèse selon laquelle la pièce est équilibrée, c'est-à-dire l'hypothèse $p = 0,5$.

   Pour l'échantillon considéré, on a $n = 4\,040$; les conditions d'application sont donc réalisées, et on obtient l'intervalle de fluctuation

   $$\left[0,5 - \frac{1}{\sqrt{4\,040}}; 0,5 + \frac{1}{\sqrt{4\,040}}\right]$$

   En arrondissant à 10^{-4} près, on obtient l'intervalle de fluctuation $I_1 = [0,4843; 0,5157]$.

   La fréquence observée dans l'échantillon est donnée par :

   $$f = \frac{2\,048}{4\,040} \approx 0,5069 \text{ à } 10^{-4} \text{ près}$$

   Ainsi, on a $f \in I_1$, donc on accepte l'hypothèse de pièce équilibrée au seuil de confiance de 95%.

2. En supposant que la pièce est équilibrée, la variable aléatoire X qui dénombre les résultat « Pile » obtenus parmi les $4\,040$ lancers réalisés suit une loi binomiale de paramètres $n = 4\,040$ et $p = 0,5$.

   On a vu en classe de $1^{\text{ère}}$ que l'ntervalle de fluctuation à 95% associé à la variable aléatoire X est l'intervalle $I_2 = \left[\frac{k_1}{n}; \frac{k_2}{n}\right]$, où k_1 est le plus petit entier k vérifiant $P(X \leq k) > 0,025$ et k_2 le plus petit entier k vérifiant $P(X \leq k) \geq 0,975$.

   On détermine k_1 et k_2 à l'aide d'un tableur (ou d'une calculatrice...mais le tableau est très long, ici).

   On obtient $k_1 = 1\,958$ et $k_2 = 2\,082$, qui donnent comme intervalle de fluctuation :

   $$I_2 = \left[\frac{1958}{4\,040}; \frac{2082}{4\,040}\right]$$

   soit en arrondissant à 10^{-4} près $I_2 = [0,4847; 0,5153]$.

   La fréquence observée $f \approx 0,5069$ vérifie $f \in I_2$.

   On est ici aussi conduit à accepter l'hypothèse de pièce équilibrée au seuil de confiance de 95%.

II Intervalles de fluctuation

II.1 Intervalle de fluctuation asymptotique

On dispose d'un urne contenant un très grand nombre de boules rouges et bleues.
On sait que la proportion de boules rouges dans l'urne est égale à $p = 0,4$.
Si on tire, successivement avec remise, n boules dans l'urne (n entier naturel non nul), et si on appelle X_n la variable aléatoire dénombrant les boules rouges tirées, alors X_n suit une loi binomiale $\mathscr{B}(n; p)$.

Théorème et définition

Soit X_n une variable aléatoire suivant une loi binomiale $\mathscr{B}(n; p)$, et α un réel tel que $0 < \alpha < 1$.
Si X est une variable aléatoire suivant la loi normale centrée réduite $\mathscr{N}(0; 1)$, on appelle u_α l'unique réel tel que :

$$P(-u_\alpha \leq X \leq u_\alpha) = 1 - \alpha$$

- On appelle I_n l'intervalle :

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

  Alors,

$$\lim_{n \to +\infty} P\left(\frac{X_n}{n} \in I_n = 1 - \alpha\right)$$

- L'intervalle I_n contient la fréquence $F_n = \dfrac{X_n}{n}$ avec une probabilité qui se rapproche de $1 - \alpha$ lorsque n augmente :
  on dit que c'est l'**intervalle de fluctuation asymptotique de F_n au seuil** $1 - \alpha$.

Démonstration (exigible) :

On pose $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$ et on applique le théorème de Moivre-Laplace.
Si X suit la loi normale $\mathscr{N}(0; 1)$: $\lim\limits_{n \to +\infty} P(Z_n \in [-u_\alpha; u_\alpha]) = P(X \in [-u_\alpha; u_\alpha] = 1 - \alpha$.
Or, $Z_n \in [-u_\alpha; u_\alpha] \Leftrightarrow -u_\alpha \leq \dfrac{X_n - np}{\sqrt{np(1-p)}} \leq u_\alpha \Leftrightarrow np - u_\alpha\sqrt{np(1-p} \leq X_n \leq np + u_\alpha\sqrt{np(1-p)}$.
$Z_n \in [-u_\alpha; u_\alpha] \Leftrightarrow p - u_\alpha \dfrac{\sqrt{p(1-p}}{\sqrt{n}} \leq \dfrac{X_n}{n} \leq p + u_\alpha \dfrac{\sqrt{p(1-p)}}{\sqrt{n}}$.
Donc, $Z_n \in [-u_\alpha; u_\alpha] \Leftrightarrow \dfrac{X_n}{n} \in I_n$.

Exemple :
On tire 50 boules de l'urne décrite ci-dessus, et on souhaite déterminer un intervalle de fluctuation au seuil 0,9 (c'est-à-dire $\alpha = 0,1$).
À l'aide de la calculatrice, on trouve 1,645 comme valeur approchée à 10^{-3} près de $u_{0,1}$.

On obtient donc comme intervalle de fluctuation :

$$I_{50} = \left[0,4 - u_{0,1}\frac{\sqrt{0,4\times 0,6}}{\sqrt{50}} ; 0,4 + u_{0,1}\frac{\sqrt{0,4\times 0,6}}{\sqrt{50}}\right]$$

soit $I_{50} = [0,286;0,514]$.
Ainsi, en effectuant 50 tirages dans cette urne, la fréquence d'apparition d'une boule rouge est comprise entre 0,28 et 0,514 avec une probabilité d'environ 0,9.
Pour 500 tirages, au même seuil de 0,9, on obtient $I_{500} = [0,364;0,436]$.

Remarque :
La longueur de l'intervalle a, pour un même seuil, été divisée par plus de 3 en passant de 50 à 500 tirages !!

Propriété

L'intervalle de fluctuation asymptotique au seuil de 95% pour une variable aléatoire X_n suivant une loi binomiale $\mathscr{B}(n;p)$ est l'intervalle :

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

Démonstration :
Cette propriété découle de celle vue au chapitre 11 concernant la loi normale.
On a vu en effet que $u_{0,05} \approx 1,96$.

Remarque :
En majorant $1,96\sqrt{p(1-p)}$ par 1, on retrouve l'intervalle de fluctuation présenté en classe de 2^{nde}.
La majoration précédente résulta de ce qui suit :
$p(1-p) = -p^2 + p$ est un trinôme du second degré.
Il admet un maximum en $p = 0,5$, qui est égal à 0,25.
Ainsi, pour tout $p \in [0;1], 0 \leq p(1-p) \leq 0,25$.
La fonction racine carrée étant croissante sur $\mathbb{R}_+$, on en déduit : $0 \leq \sqrt{p(1-p)} \leq 0,5$,
puis $0 \leq 1,96\sqrt{p(1-p)} \leq 1,95 \times 0,5 \leq 2 \times 0,5 = 1$.

II.2 Exercice

Énoncé

Dans un casino, il a été décidé que les « machines à sous » doivent être réglées sur une fréquence de gain du joueur de $g = 0,06$.

Une fréquence inférieure est supposée faire « fuir » le client, et une fréquence supérieure est susceptible de ruiner le casino.

Trois contrôleurs différents vérifient une même machine.

Le premier a joué 50 fois et gagné 2 fois,

le second a joué 120 fois et gagné 14 fois,

le troisième a joué 400 fois et gagné 30 fois.

En utilisant des intervalles de fluctuation asymptotiques au seuil 95%, examiner dans chaque cas la décision à prendre par le contrôleur, à savoir accepter ou rejeter l'hypothèse $g = 0,06$.

Solution :

- **Premier contrôleur** Le contrôle a porté sur 50 parties, donc $n = 50$.
  On a bien $n \geq 30$, mais $np = 50 \times 0,06 = 3$ et donc $np < 5$.
  On n'est pas dans les conditions d'application d'un intervalle de fluctuation asymptotique.
  Ce contrôle ne peut rien donner de probant en termes de prise de décision.
- **Deuxième contrôleur**
  Le contrôle a porté sur 120 parties, donc $n = 120$, $np = 7,2$ et $n(1-p) = 112,8$.
  Les conditions d'utilisation d'un intervalle de fluctuation asymptotique sont donc réunies.
  L'intervalle de fluctuation asymptotique au seuil 95% pour un échantillon de taille $n = 120$ est égal à :
  $$I_{120} = \left[0,06 - 1,96\frac{\sqrt{0,06 \times 0,94}}{\sqrt{120}} ; 0,06 + 1,96\frac{\sqrt{0,06 \times 0,94}}{\sqrt{120}}\right]$$
  soit en arrondissant les bornes à 10^{-4} près : $I_{120} = [0,0175 ; 0,1025]$.
  La fréquence observée par le deuxième contrôleur est $f_2 = \dfrac{14}{120} \approx 0,1167$.
  Cette fréquence est à l'extérieur de l'intervalle de fluctuation I_{120}, et le second contrôleur est conduit à rejeter l'hypothèse $g = 0,06$.
- **Troisième contrôleur**
  La contrôle a porté sur 400 parties, donc $n = 400$, $np = 24$ et $n(1-p) = 376$.
  Les conditions d'utilisation d'un intervalle de fluctuation asymptotique sont donc réunies.
  L'intervalle de fluctuation asymptotique au seuil 95% pour un échantillon de taille $n = 400$ est égal à :
  $$I_{400} = \left[0,06 - 1,96\frac{\sqrt{0,06 \times 0,94}}{\sqrt{400}} ; 0,06 + 1,96\frac{\sqrt{0,06 \times 0,94}}{\sqrt{400}}\right]$$
  soit en arrondissant les bornes à 10^{-4} près : $I_{400} = [0,0367 ; 0,0833]$.
  La fréquence observée par le deuxième contrôleur est $f_3 = \dfrac{30}{400} = 0,075 \in I_{400}$.
  Ce dernier est donc amené à accepter l'hypothèse $g = 0,06$.

Remarques

1. Pour tester une hypothèse à l'aide d'un intervalle de fluctuation asymptotique, on doit vérifier ses conditions d'utilisation :

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

   Ces conditions vérifiées, on détermine l'intervalle de fluctuation I_n au seuil demandé (souvent 95%).

   Si la fréquence observée f n'appartient pas à I_n, on rejette l'hypothèse.

2. L'intervalle de fluctuation étudié est « bilatéral » : on est conduit à rejeter l'hypothèse dans le cas où la fréquence observée est « trop grande » , ou « trop petite » .

III Estimation

III.1 Notion d'intervalle de confiance

On dispose d'une urne contenant un très grand nombre de boules rouges et bleues.
On ignore quelle est la proportion p de boules rouges dans l'urne, et rien ne permet de faire une hypothèse sur la valeur de p.
L'« estimation » consiste à chercher à « deviner (ou estimer) » , avec un certain niveau de confiance, quelle valeur peut prendre p, en s'appuyant sur les informations recueillies en procédant à des tirages au sort aléatoires.
On supposera les trois conditions d'approximation remplies :

$$n \geq 30; \quad np \geq 5; \quad n(1-p) \geq 5$$

Théorème

Soit X_n une variable aléatoire suivant une loi binomiale $\mathscr{B}(n;p)$ où p est la proportion inconnue d'apparition d'un caractère, et $F_n = \dfrac{X_n}{n}$ la fréquence associée à X_n.
Alors, pour n assez grand, p appartient à l'intervalle $\left[F_n - \dfrac{1}{\sqrt{n}}; F_n + \dfrac{1}{\sqrt{n}}\right]$ avec une probabilité supérieure ou égale à 0,95.

Démonstration (exigible)

On sait que $P\left(p-1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq F_n \leq p+1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) \approx 0,95$.
Suite à la dernière remarque du II - 1),
$\left[p-1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}};p+1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right] \subset \left[p-\frac{1}{\sqrt{n}};p+\frac{1}{\sqrt{n}}\right]$.
On a donc : $P\left(p-\frac{1}{\sqrt{n}} \leq F_n \leq p+\frac{1}{\sqrt{n}}\right) \geq 0,95$.
Or, $p-\frac{1}{\sqrt{n}} \leq F_n \leq p+\frac{1}{\sqrt{n}} \Leftrightarrow -\frac{1}{\sqrt{n}} \leq F_n - p \leq \frac{1}{\sqrt{n}} \Leftrightarrow -F_n-\frac{1}{\sqrt{n}} \leq -p \leq -F_n+\frac{1}{\sqrt{n}}$
$\Leftrightarrow F_n-\frac{1}{\sqrt{n}} \leq p \leq F_n+\frac{1}{\sqrt{n}}$.
Ainsi, $P\left(F_n-\frac{1}{\sqrt{n}} \leq p \leq F_n+\frac{1}{\sqrt{n}}\right) \geq 0,95$.

Exemple :
Dans l'urne ci-dessus, on réalise un trage de 100 boules.
On obtient 59 rouges et 41 bleues ; la fréquence observée de sortie du rouge est donc 0,59.
L'intervalle $\left[0,59-\frac{1}{\sqrt{100}};0,59+\frac{1}{\sqrt{100}}\right] = [0,49;0,69]$ est un intervalle de confiance de la proportion de boules rouges dans l'urne au niveau de confiance 95%.

III.2 Précision d'une estimation et taille de l'échantillon

On a vu précédemment qu'en tirant 100 boules de l'urne, l'intervalle de confiance obtenu est de longueur 0,2.
On peut trouver cet intervalle « trop grand » .
En procédant à un tirage de 400 boules, si f est la fréquence observée de sortie du rouge, on obtient un intervalle de confiance au niveau 95% égal à :

$$\left[f-\frac{1}{\sqrt{400}};f+\frac{1}{\sqrt{400}}\right] = [f-0,05;f+0,05]$$

Son amplitude, deux fois moins grande que la précédente, est de 0,1.
Plus généralement, on retiendra :

À retenir

Un intervalle de confiance au niveau 95% est d'amplitude $\frac{2}{\sqrt{n}}$.
Plus la taille de l'échantillon est grande, plus les intervalles de confiance obtenus sont précis.

Exemple :
Pour obtenir un intervalle de confiance d'amplitude inférieure à 0,01 de a proportion de boules rouges dans l'urne, il faut procéder à des tirages de n boules, avec $\frac{2}{\sqrt{n}} \leq 0,01$, soit $\frac{4}{n} \leq 10^{-4}$, ou encore $n \geq 4\times 10^4$.

Pour obtenir des intervalles de confiance au niveau 0,95 d'amplitude inférieure à 0,01, il fut procéder à au moins 40 000 tirages...

III.3 Exercice

Énoncé

Une usine fabrique des pièces métalliques, qui son censées résister à certaines contraintes mécaniques.
Le responsable de fabrication souhaite estimer le taux de pièces défectueuses concernant la résistance mécanique dans la production.
Pour cela, il utilise la méthode par intervalle e confiance au niveau 95%, en extrayant au hasard n pièces en fin de production, qui sont soumises à contrainte mécanique jusqu'à la rupture.
En fonction du niveau de confiance à la rupture, on décide de la nature défectueuse ou pas de la pièce.

1. Chaque pièce testée étant détruite, le responsable souhaite limiter la taille de l'échantillon testé, tout en ayant un intervalle de confiance de longueur inférieure à 0,1.
   Quelle taille d'échantillon peut-on lui conseiller ?
2. Il est finalement décidé de mener l'étude sur 500 pièces ; on en trouve 40 défectueuses.
   Quel intervalle de confiance, au niveau de confiance 95 %, obtient-on ?
3. L'année précédente, à l'issue d'un problème grave de rupture d'une pièce, une large étude avait débouché sur 130 pièces défectueuses dans un échantillon de 1 000.
   Peut-on supposer que la mise en place de nouvelles procédures de fabrication a vraiment diminué la proportion de pièces défectueuses ?

Solution :

1. La longueur d'un intervalle de confiance à 95 % est égale à $\dfrac{2}{\sqrt{n}}$.
   On doit donc résoudre l'inéquation : $\dfrac{2}{\sqrt{n}} \leq 0,1$, soit $\dfrac{4}{n} \leq 10^{-2}$, ou encore $n \geq 400$.
   Pour obtenir un intervalle de confiance de longueur inférieure ou égale à 0,1, il faut prendre n au moins égal à 400.
   On peut donc conseiller, pour limiter les coûts de l'étude, de procéder à un tirage de 400 pièces.
2. La fréquence observée est : $f = \dfrac{40}{500} = 0,08$.
   L'intervalle de confiance au niveau 95 % obtenu ici est :
   $$\left[0,08 - \frac{1}{\sqrt{500}} ; 0,08 + \frac{1}{\sqrt{500}}\right]$$
   En arrondissant les bornes à 10^{-4} près, on obtient $I_1 = [0{,}0353 ; 0{,}1247]$.

3. Si on déterminer l'intervalle de confiance à 95 % correspondant à l'échantillon décrit, on obtient une fréquence observée de $f' = \dfrac{130}{1\,000} = 0,13$, d'où un intervalle de confiance 95 % égal à :

$$\left[0,13 - \frac{1}{\sqrt{1\,000}} ; 0,13 + \frac{1}{\sqrt{1\,000}}\right]$$

soit en arrondissant les bornes à 10^{-4} : $I_2 = [0{,}098\,4 ; 0{,}161\,6]$.

Les deux intervalles de confiance obtenus, I_1 et I_2, ne sont pas disjoints ; par exemple, une proportion de 10 % de pièces défectueuses dans la production est compatible avec les deux intervalles de confiance obtenus.

Malgré les résultats assez différents obtenus par les deux études (fréquences observées de 8 % et de 13 % pour les pièces défectueuses), il est possible que la proportion de pièces défectueuses dans la production n'ait pas changé.