INTERVALLE DE FLUCTUATION ET ESTIMATION

I Intervalle de fluctuation

1 Cadre

Une urne contient un très grand nombre de boules blanches et de boules noires dont **la proportion** p de boules blanches **est connue**.
On tire avec remise n boules (qui vont constituer notre échantillon) et on observe la fréquence d'apparition des boules blanches. Cette fréquence observée appartient à un intervalle, qu'on appellera **intervalle de fluctuation** de centre p dont la longueur diminue avec n. On est ici dans le domaine de l'échantillonnage et de l'intervalle de fluctuation.
Dans certain cas il nous est impossible de connaître la proportion p mais on est capable de faire une hypothèse sur sa valeur, on parle de la **prise de décision**. On veut par exemple savoir si une pièce est bien équilibrée. On peut faire l'hypothèse que l'apparition de chaque face est égale à $\frac{1}{2}$ et on va tester cette hypothèse.
Résumé : On utilise un intervalle de fluctuation quand :
- on connait la proportion p de présence du caractère dans la population
- on fait une hypothèse sur la valeur de cette proportion

2 Intervalle de fluctuation asymptotique

Dans cette partie que la proportion p du caractère étudié est connue.

Définition 1

X_n est une variable aléatoire qui suit une loi binomiale $\mathscr{B}(n;p)$.
La variable aléatoire $F_n = \frac{X_n}{n}$ s'appelle **la variable aléatoire fréquence de succès** pour un schéma de Bernoulli de paramètres n et p.

Définition 2

Pour tout α dans $]0;1[$, un **intervalle de fluctuation asymptotique** de la variable aléatoire F_n au seuil $1-\alpha$ est un intervalle déterminé à partir de p et de n et qui contient F_n avec une probabilité d'autant plus proche de $1-\alpha$ que n est grand.

Remarque : il n'existe pas un unique intervalle de fluctuation asymptotique à un seuil donné. La probabilité $P(a \leq F \leq b)$ (a et b étant les bornes de l'intervalle), n'est pas nécessairement égale à $1-\alpha$ mais s'en rapproche quand la taille de l'échantillon devient de plus en plus grande.

Propriété 1

Soit $\alpha \in]0;1[$ et X_n une variable aléatoire qui suit une loi binomiale $\mathscr{B}(n;p)$.
Quand la taille de l'échantillon n devient grande, la probabilité que la fréquence F_n prenne ses valeurs dans l'intervalle $I_n = \left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ se rapproche de $1-\alpha$.
On note $\lim\limits_{n\to+\infty} P(F_n \in I_n) = 1-\alpha$

Remarque : I_n est un intervalle de fluctuation asymptotique de la fréquence F_n au seuil $1-\alpha$.

Preuve

X_n suit la loi binomiale $\mathscr{B}(n;p)$ donc la suite de variables aléatoires $Z_n = \frac{X_n - E(X_n)}{\sigma(X_n)}$ tend, d'après le théorème de Moivre-Laplace, vers la loi normale centrée réduite $\mathscr{N}(0;1)$.
On a donc : $\lim\limits_{n\to+\infty} P(a \leq Z_n \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2} \mathrm{d}x$ pour tous réels a et b tels que $a < b$.

Or $Z_n = \frac{X_n - E(X_n)}{\sigma(X_n)} = \frac{X_n - np}{\sqrt{np(1-p)}} = \frac{n\left(\frac{X_n}{x} - p\right)}{n\frac{\sqrt{p(1-p)}}{\sqrt{n}}} = \frac{F_n - p}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}}$

Donc $\lim\limits_{n\to+\infty} P\left(p + a\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq F_n \leq p + b\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) = \int_a^b \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2} \mathrm{d}x$
Comme, pour tout réel $\alpha \in]0;1[$, il existe un unique réel positif u_α tel que $P(-u_\alpha \leq X \leq u_\alpha) = 1-\alpha$ où X suit une loi normale centrée réduite $\mathscr{N}(0;1)$, on a $\int_{-u_\alpha}^{u_\alpha} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2} \mathrm{d}x$
En prenant $a = -u_\alpha$ et $b = u_\alpha$, on a

$$\lim_{n\to+\infty} P\left(p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq F_n \leq p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) = 1-\alpha$$

Exemple : On dispose d'une urne contenant un grand nombre de boules blanches et noires. La proportion de boules blanches contenues dans l'urne est $p = 0,3$.
On tire successivement avec remises $n = 50$ boules.
Soit X_{50} la variable aléatoire dénombrant le nombre de boules blanches tirées.
X_{50} suit la loi binomiale $\mathscr{B}(50;0,3)$.
Prenons $\alpha = 0,05$ alors, d'après le cours sur la loi normale, $u_{0,05} = 1,96$.

$$I_{50} = \left[0,3 - 1,96 \times \frac{\sqrt{0,3 \times 0,7}}{\sqrt{50}} ; 0,3 + 1,96 \times \frac{\sqrt{0,3 \times 0,7}}{\sqrt{50}}\right] \approx [0,172;0,428]$$

Attention aux arrondis : on arrondit par défaut pour la borne inférieure et par excès pour la borne supérieure afin de s'assurer d'avoir une probabilité d'au moins $1 - \alpha$.
Avec 50 tirages dans cette urne, la fréquence d'apparition d'une boule blanche appartient donc à $[0,172;0,428]$ avec une probabilité de $0,95$.
Pour 500 tirages, on obtient :

$$I_{500} = \left[0,3 - 1,96 \times \frac{\sqrt{0,3 \times 0,7}}{\sqrt{500}} ; 0,3 + 1,96 \times \frac{\sqrt{0,3 \times 0,7}}{\sqrt{500}}\right] \approx [0,259;0,341]$$

Les intervalles de fluctuation, pour un même seuil, se resserrent lorsqu'on augmente le nombre de tirages.

Propriété 2

Un intervalle de fluctuation asymptotique au seuil de confiance 95% de la fréquence F_n d'un caractère dans un échantillon de taille n est

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

Remarque : En seconde, on avait défini un intervalle de fluctuation $\left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}\right]$. Celui-ci contient l'intervalle de fluctuation asymptotique au seuil de 95%.
L'intervalle de fluctuation asymptotique au seuil de 95% est centré sur la proportion p ce qui n'est pas nécessairement le cas pour l'intervalle de fluctuation déterminé en 1re à l'aide de la loi binomiale.

3 Prise de décision

Dans ce paragraphe, la proportion du caractère étudié n'est pas connue mais on va émettre l'hypothèse qu'elle est égale à p. Le but de la prise de décision est de valider ou invalider l'hypothèse faite sur la proportion p.

Propriété 3 (Règle de décision)

On appelle f la fréquence observée du caractère étudié d'un échantillon de taille n. On fait l'hypothèse suivante : " la proportion de ce caractère dans la population est p".
On appelle I l'intervalle de fluctuation asymptotique au seuil de 95%.

1. Si $f \in I$ alors on accepte l'hypothèse faite sur la proportion p
2. Si $f \notin$alors on rejette l'hypothèse faite sur la proportion p au risque de 5%.

Remarques :

1. Le risque de rejetter à tord l'hypothèse faite sur p sachant qu'elle est vraie est approximativement égale à 5%.
2. Pour pouvoir utiliser un intervalle de fluctuation au seuil de 95% déterminé précédemment il faut que $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$.
   Sinon on utilise l'intervalle vu en 1re avec la loi binomiale : on recherche les plus petits entiers a et b tels que :
   — $P(X \leq a) > 0,025$
   — $P(X \leq b) \geq 0,975$
   L'intervalle cherché est alors $\left[\frac{a}{n}; \frac{b}{n}\right]$.

Méthode : Prendre une décision à l'aide d'un intervalle de fluctuation
Un fabriquant commande auprès de son fournisseur deux types de pièces : A et B. Il demande 900 pièces de chaque sorte.
Au moment de la livraison, on contrôle au hasard 50 pièces et on constate que 19 sont des modèles A.
Peut-on affirmer que la commande est respectée par le fournisseur ?

— Le fabriquant a commandé autant de pièces de chaque sorte. On peut donc supposer que la proportion de pièce A est égale à $0,5$.
  La taille de notre l'échantillon est $n = 50$.
— Avant de déterminer un intervalle de fluctuation il faut faut vérifier si les paramètres n et p répondent aux conditions imposées :
  $n = 50 \geq 30$, $np = 50 \times 0,5 = 25 \geq 5$ et $n(1-p) = 50 \times 0,5 = 25 \geq 5$.
— Un intervalle de fluctuation asymptotique au seuil de $0,95$ est :

$$I_{50} = \left[0,5 - 1,96 \times \frac{\sqrt{0,5 \times 0,5}}{\sqrt{50}}; 0,5 + 1,96 \times \frac{\sqrt{0,5 \times 0,5}}{\sqrt{50}}\right] \approx [0,361; 0,639]$$

La fréquence observée est $f = \frac{19}{50} = 0,38 \in I_{50}$. On peut donc accepter l'hypothèse faite.

II Estimation

1 Cadre

Une urne contient un très grand nombre de boules blanches et de boules noires dont on ignore la proportion p de boules blanches. Cette proportion **n'est pas connue**.
On tire avec remise n boules afin d'estimer la population de boules blanches. Elle est obtenue à l'aide d'un **intervalle de confiance** construit selon un niveau de confiance fixé au début de l'estimation.

Remarque :

L'intervalle de fluctuation $I_n = \left[p - 1,96 \times \frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 1,96 \times \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ peut être simplifié en $\left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}\right]$.

En effet la fonction f définie sur $[0;1]$ par $f(x) = x(1-x) = x - x^2$ admet un maximum en $0,5$ et $f(0,5) = 0,25$. Par conséquent $0 \leq \sqrt{p(1-p)}\sqrt{0,25} = 0,5$.

Ainsi $0 \leq 1,96\frac{\sqrt{p(1-p}}{\sqrt{n}} \leq \frac{1}{\sqrt{n}}$

2 Estimation

Propriété 4

X_n est une variable aléatoire qui suit une loi binomiale $\mathscr{B}(n;p)$ et $F_n = \frac{X_n}{n}$ est la variable aléatoire fréquence associée à X_n.

Pour n suffisamment grand, p appartient à l'intervalle $J_n = \left[F_n - \frac{1}{\sqrt{n}} ; F_n + \frac{1}{\sqrt{n}}\right]$ avec une probabilité supérieure ou égale à $0,95$.

Définition 3

Soit f une fréquence observée du caractère étudié sur un échantillon de taille n. On appelle **intervalle de confiance** de la proportion p **au niveau de confiance** $0,95$, l'intervalle $\left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}}\right]$.

Remarques :

1. Cet intervalle a une amplitude de $\frac{2}{\sqrt{n}}$. Donc plus la taille de l'échantillon est grande, plus les intervalles de confiance sont précis.
2. — Un niveau de confiance $0,95$ signifie que dans 95 cas sur 100, on peut affirmer que p appartient à l'intervalle de confiance.
   — p ne correspond pas nécessairement au centre de l'intervalle de confiance. La position de p dans l'intervalle de confiance ne peut pas être déterminée.
   — p étant inconnu, on ne peut pas vérifier si les conditions énoncées sur n et p vues précédemment sont vraies.
   On va cependant, vérifier que : $n \geq 30$, $nf \geq 5$ et $n(1-f) \geq 5$

Exemple : Entre 2 tours d'une élection on interroge 1250 personnes sur leurs intentions de vote. 748 déclarent avoir l'intention de voter pour le candidat A.
On suppose que les votes seront conformes aux intentions de votes, le candidat A peut-il penser qu'il sera élu ?

— Nous avons un échantillon de taille $n = 1250$ et la fréquence observée est $f = \dfrac{748}{1250} = 0,5984$.
— Avant de déterminer un intervalle de confiance, on va contrôler que les paramètres n et f vérifient les conditions imposées :
$n = 1250 \geq 30$, $nf = 1250 \times 0,5984 = 748 \geq 5$ et $n(1-f) = 502 \geq 5$.
— Un intervalle de confiance au seuil de 95% est :

$$I_{1250} = \left[0,5984 - \frac{1}{\sqrt{1250}} ; 0,5984 + \frac{1}{\sqrt{1250}}\right] \approx [0,570 ; 0,627]$$

Les intentions de vote pour le candidat A sont donc supérieures à 50%. Ce sondage permet donc de dire que le candidat A devrait être élu.