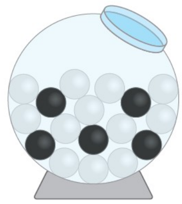
FLUCTUATION ET ESTIMATION

Dans ce chapitre, on va étudier deux domaines des statistiques qu'il faut savoir distinguer :

Echantillonnage	Estimation
Une urne contient un très grand nombre de boules blanches et de boules noires dont **on connaît la proportion *p*** de boules blanches. On tire avec remise *n* boules (échantillon) et on observe la fréquence d'apparition des boules blanches. Cette fréquence observée appartient à un intervalle, appelé **intervalle de fluctuation** de centre *p*. Dans le cas où on ne connaît pas 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 un dé est bien équilibré. On peut faire l'hypothèse que l'apparition de chaque face est égale à 1/6 et on va tester cette hypothèse.	Une urne contient un très grand nombre de boules blanches et de boules noires dont **on ignore la proportion *p*** de boules blanches. On tire avec remise *n* boules dans le but d'estimer la proportion *p* de boules blanches. Cette estimation s'obtient à l'aide d'un **intervalle de confiance** construit selon un niveau de confiance que l'on attribue à l'estimation. 

Conditions sur les paramètres : Dans tout le chapitre, sauf mention contraire, la taille de l'échantillon n et la proportion p du caractère étudié dans la population vérifient :

$$n \geq 30 \quad , \quad n \times p \geq 5 \quad \text{et} \quad n \times (1-p) \geq 5 \quad .$$

I. Echantillonnage

1) Intervalle de fluctuation asymptotique

Dans ce paragraphe, **on suppose que la proportion *p* du caractère étudié est connue**.

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 remise n = 50 boules.

Soit X_{50} la variable aléatoire dénombrant le nombre de boules blanches tirées.

X_{50} suit la loi binomiale $B(50\,;0,3)$. En effectuant 50 tirages dans cette urne, on va prouver dans ce chapitre que la fréquence d'apparition d'une boule blanche est comprise dans l'intervalle [0,173 ; 0,427] avec une probabilité de 0,95.

Cet intervalle s'appelle l'intervalle de fluctuation asymptotique au seuil 0,95 (ou 95%).

Définition : X_n est une variable aléatoire qui suit une loi binomiale $B(n;p)$.

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*.

Propriété 1: Soit $\alpha \in \left]0;1\right[$ et X_n une variable aléatoire qui suit une loi binomiale $B(n;p)$.

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 $1-\alpha$ quand la taille de l'échantillon n devient grande. On note : $\lim\limits_{n \to +\infty} P(F_n \in I_n) = 1-\alpha$.

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

Exemple :

Démontrons le résultat donné dans l'exemple en début de paragraphe :

$I_{50}=\left[0,3-1,96\times\frac{\sqrt{0,3\times0,7}}{\sqrt{50}}\,;0,3+1,96\times\frac{\sqrt{0,3\times0,7}}{\sqrt{50}}\right]$ car $u_{0,05}=1,96$. Soit $I_{50}=[0,173\,;0,427]$.

Pour 500 tirages, on obtient : $I_{500}=\left[0,3-1,96\times\frac{\sqrt{0,3\times0,7}}{\sqrt{500}};0,3+1,96\times\frac{\sqrt{0,3\times0,7}}{\sqrt{500}}\right]=[0,26\,;0,34]$

On constate que l'intervalle, pour un même seuil, se resserre fortement lorsqu'on augmente le nombre de tirages.

Définition : On appelle intervalle de fluctuation au seuil 0,95 de la variable aléatoire fréquence l'intervalle : $\left[p-1,96\times\frac{\sqrt{p(1-p)}}{\sqrt{n}}\,;p+1,96\times\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$.

2) Prise de décision

Dans ce paragraphe, la proportion du caractère étudié n'est pas connue mais est supposée être égale à *p*. La prise de décision consiste à valider ou invalider l'hypothèse faite sur la proportion *p*.

Propriété (Règle de décision) : Soit f la fréquence du caractère étudié d'un échantillon de taille n.
Soit l'hypothèse : "La proportion de ce caractère dans la population est *p*."
Soit *I* l'intervalle de fluctuation asymptotique au seuil 0,95.

- Si $f\in I$, alors on accepte l'hypothèse faite sur la proportion *p*.
- Si $f\notin I$, alors on rejette l'hypothèse faite sur la proportion *p*.

Remarque :

On peut interpréter cette propriété par le fait que la probabilité qu'on rejette à tord l'hypothèse sur *p* sachant qu'elle est vraie est approximativement égale à 5%.

II. Estimation

Dans ce paragraphe, **on suppose que la proportion *p* du caractère étudié est inconnue**.
C'est le problème inverse de celui de l'échantillonnage. A partir de la fréquence observée sur un échantillon, on va estimer la proportion *p* d'un caractère dans la population tout entière.

Propriété : (admise) X_n est une variable aléatoire qui suit une loi binomiale $B(n;p)$.

$F_n=\frac{X_n}{n}$ est la 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 : 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 $\left[f-\frac{1}{\sqrt{n}};f+\frac{1}{\sqrt{n}}\right]$.

Exemple : On dispose d'une urne contenant un grand nombre de boules blanches et noires. La proportion de boules blanches contenues dans l'urne n'est pas connue. On réalise un tirage de 100 boules et on obtient 54 boules blanches. La fréquence observée est donc $f=0,54$.
L'intervalle de confiance de la proportion de boule blanche dans l'urne au niveau de confiance 95% est $\left[0,54-\frac{1}{\sqrt{100}}\,;0,54+\frac{1}{\sqrt{100}}\right]=[0,44\,;0,64]$.

Démonstration propriété 1 :

X_n suit la loi binomiale $B(n;p)$ donc la suite de variables aléatoires $Z_n = \frac{X_n - E(X_n)}{\sigma(X_n)}$ tend vers la loi normale centrée réduite $N(0;1)$ et d'après le théorème de Moivre-Laplace, on a :

$\lim_{n \to +\infty} P\left(a \le Z_n \le b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$, pout tous réels a et b avec $a < b$.

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

Donc $\lim_{n \to +\infty} P\left(p + a\frac{\sqrt{p(1-p)}}{\sqrt{n}} \le F_n \le p + b\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$

Comme, pour tout réel $\alpha \in \left]0;1\right[$, il existe un unique réel positif u_α tel que $P(-u_\alpha \le X \le u_\alpha) = 1 - \alpha$ où X suit une loi normale centrée réduite $N(0;1)$, on a :

$\int_{-u_\alpha}^{u_\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 1 - \alpha$.

En prenant $a = -u_\alpha$ et $b = u_\alpha$, on a : $\lim_{n \to +\infty} P\left(p - u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}} \le F_n \le p + u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) = 1 - \alpha$.