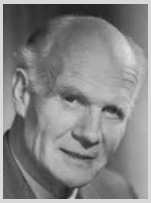
Chapitre XVIII : Fluctuation et estimation



Histoire des maths : Ce chapitre fait l'étude d'outils permettant d'estimer des paramètres à l'aide d'intervalles Ces outils peuvent ensuite être utilisés pour valider ou invalider des hypothèses relatives à un choix de modèle probabiliste. On parlera alors de tests statistiques.

Ces derniers ont été inventés par **Karl Pearson** au début du XXe siècle dans le cadre de ses travaux de biométrie. Ils ont été systématisés par son fils **Egon Pearson** (à gauche) aux côtés de **Jerzy Neyman** (à droite). Aujourd'hui les tests statistiques servent en biostatistiques, économie, sociologie, physique statistique, …

C'est grâce à ces tests que l'on peut faire un choix de modèle probabiliste adéquat à l'étude d'un problème donné. *Source Wikipedia*

Echantillonnage – Prise de décision	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 **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 à l'aide d'une expérience. Le résultat de l'expérience va nous permettre d'accepter ou rejeter l'hypothèse de départ.	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. On obtient ainsi une fréquence d'apparition qui va nous permettre d'estimer la proportion p à l'aide d'un **intervalle de confiance**. 

I. Echantillonnage

On étudie un caractère et on suppose que **la proportion $p$de ce caractère est connue**.

Définition : fréquence de succès

Soit $X_n$une variable aléatoire qui suit une loi binomiale $B(n,p)$, la variable aléatoire $F_n = \frac{X_n}{n}$représente la fréquence de succès pour un schéma de Bernoulli de paramètres n et p.

Propriété

Si la variable aléatoire $X_n$suit la loi binomiale $B(n,p)$, alors pour tout $\alpha \in]0;1[$, on a :
la probabilité que la fréquence $F_n$prenne ses valeurs dans l'intervalle :

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

Autrement dit : $\lim\limits_{n \to +\infty} P(F_n \in I_n) = 1 - \alpha$

Cet intervalle contient $F_n$avec une probabilité d'autant plus proche de $1 - \alpha$que $n$est grand.

Cette approximation est valable dès que $n \geq 30$, $np \geq 5$, $n(1-p) \geq 5$

Définition : intervalle de fluctuation asymptotique

I_n est appelé un intervalle de fluctuation asymptotique de la variable aléatoire F_n au seuil de confiance $1 - \alpha$.

Remarque : La probabilité définie dans la propriété se rapproche de 1−α sans être nécessairement égale d'où l'emploi du terme "asymptotique".

ROC 10 : Démonstration exigible

Soit X_n une variable aléatoire qui suit la loi binomiale $B(n,p)$.

D'après le théorème de Moivre-Laplace, la suite de variables aléatoires $Z_n = \frac{X_n - np}{\sqrt{np(1-p)}}$ suit une loi normale centrée réduite.

A savoir : $\lim\limits_{n \to +\infty} P(a \leq Z_n \leq b) = P(a \leq N \leq b)$ où N suit une loi normale centrée réduite $N(0,1)$.

Or $a \leq Z_n \leq b \Leftrightarrow a \leq \frac{X_n - np}{\sqrt{np(1-p)}} \leq b$

On factorise par n au numérateur et au dénominateur :

$$a \leq \frac{X_n - np}{\sqrt{np(1-p)}} \leq b \Leftrightarrow a \leq \frac{n(\frac{X_n}{n} - p)}{n\frac{\sqrt{p(1-p)}}{\sqrt{n}}} \leq b \Leftrightarrow a\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq F_n - p \leq b\frac{\sqrt{p(1-p)}}{\sqrt{n}} \Leftrightarrow$$

$$p + a\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq F_n \leq p + b\frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

Or lorsque X est une va qui suit une loi normale centrée réduite, alors pour tout $\alpha \in]0\,;\,1[$, il existe un unique u_α tel que $P(-u_\alpha \leq X \leq u_\alpha) = 1 - \alpha$

Prenons $a = -u_\alpha$ et $b = u_\alpha$, ainsi :

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

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

Intervalle de fluctuation au seuil de confiance $0,95$

Nous savons, d'après le chapitre précédent que $P(-1,96 \leq X \leq 1,96) = 0,95$

Ici $\alpha = 0,05$ et $u_\alpha = 1,96$

Ainsi

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

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} est une va qui suit une loi binomiale $B(50\,;\,0,3)$

Pour 50 tirages :

On calcule l'intervalle de fluctuation au seuil 0,95 :

$$I_{50} = [0,3 - 1,96\frac{\sqrt{0,3(1-0,3)}}{\sqrt{50}}\,;\, 0,3 + 1,96\frac{\sqrt{0,3(1-0,3)}}{\sqrt{50}}] = [0,173\,;\,0,427]$$

La fréquence d'apparition d'une boule blanche est comprise dans l'intervalle [0,173 ; 0,427] avec une probabilité de 0,95.

Pour 500 tirages :

On calcule l'intervalle de fluctuation au seuil 0,95 :

$$I_{500} = [0,3 - 1,96\frac{\sqrt{0,3(1-0,3)}}{\sqrt{500}} ; 0,3 + 1,96\frac{\sqrt{0,3(1-0,3)}}{\sqrt{500}}] = [0,26 ; 0,34]$$

La fréquence d'apparition d'une boule blanche est comprise dans l'intervalle [0,26 ; 0,34] avec une probabilité de 0,95.
On constate que l'intervalle, pour un même seuil, se resserre fortement lorsqu'on augmente le nombre de tirages.

II. 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.

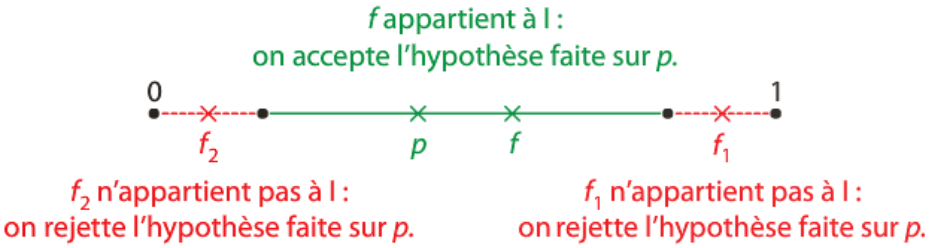
L'intervalle de fluctuation au seuil de 95% signifie qu'il y a 95% de chances que la fréquence soit dans l'intervalle I_n.

Propriété

On considère une population dans laquelle on suppose que la proportion du caractère étudié soit p.
On observe f comme fréquence de ce caractère dans un échantillon de taille n.
On fait l'hypothèse : "la proportion de ce caractère dans la population est p".
Si I est l'intervalle de fluctuation de la fréquence à 95% dans les échantillons de taille n, alors la règle de décision est la suivante :

- si $f \in I$, on considère que l'hypothèse selon laquelle la proportion du caractère est pn'est pas remise en question et on l'accepte.
- si $f \notin I$, on rejette l'hypothèse selon laquelle cette proportion du caractère est p

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



Autre seuil possible :

• Au lieu du coefficient 95 %, on peut choisir d'autres coefficients.
Le plus fréquemment utilisé après 95 % est 99 %, soit un seuil de risque de 1 %.
Dans ce cas $u_{0,01} \approx 2,58$, ce qui donne l'intervalle de fluctuation asymptotique suivant :

$$\left[p - 2,58\frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 2,58\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right].$$

III. Estimation d'une proportion

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

Si la variable aléatoire X_n suit une loi binomiale $B(n,p)$, $F_n = \frac{X_n}{n}$ est la fréquence associée à X_n alors pour n suffisamment grand, l'intervalle $J_n = [F_n - \frac{1}{\sqrt{n}} ; F_n + \frac{1}{\sqrt{n}}]$ contient la proportion $p$avec une probabilité supérieure à 0,95.
On suppose que $n \geq 30, np \geq 5, n(1-p) \geq 5$

Dém : Soit X_n une variable aléatoire qui suit une loi binomiale $B(n,p)$.

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

En effet, la fonction $x \to x(1-x) = x - x^2$ est une fonction polynôme du seconde degré qui s'annule en 0 et 1.

Elle admet un maximum en $-\frac{b}{2a} = \frac{1}{2}$comme $a < 0$. Et on a $f(0,5) = 0,25$. Elle est positive sur $[0\,;\,1]$.

Ainsi sur $[0\,;\,1]$, $0 \leq p(1-p) \leq 0,25 \Leftrightarrow 0 \leq \sqrt{p(1-p)} \leq \sqrt{0,25} = 0,5$

On en déduit que : $0 \leq 1,96\sqrt{p(1-p)} \leq 1$et alors $0 \leq 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq \frac{1}{\sqrt{n}}$

On en conclut que $I_n \subset J_n$ et donc $P(F_n \in J_n) \geq 0,95$

Avec cet intervalle simplifié : $p - \frac{1}{\sqrt{n}} \leq F_n \leq p + \frac{1}{\sqrt{n}}$ soit $F_n - \frac{1}{\sqrt{n}} \leq p \leq F_n + \frac{1}{\sqrt{n}}$

Comme $P(p - \frac{1}{\sqrt{n}} \leq F_n \leq p + \frac{1}{\sqrt{n}}) \geq 0,95$alors l'intervalle $J_n = [F_n - \frac{1}{\sqrt{n}} ; F_n + \frac{1}{\sqrt{n}}]$contient p avec une probabilité supérieure 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 :

$$J_n = [f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}}]$$

Remarques :
- Il n'est pas vrai d'affirmer que p est égal au centre de l'intervalle de confiance. Il n'est pas possible d'évaluer la position de p dans l'intervalle de confiance.
- p étant inconnu, il n'est pas possible de vérifier si les conditions énoncées sur n et p en introduction de chapitre sont vérifiées.

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 $\left[0,54 - \frac{1}{\sqrt{100}} ; 0,54 + \frac{1}{\sqrt{100}}\right] = [0,44 ; 0,64]$.

Exemple :

Voici les résultats d'un sondage IPSOS réalisé avant l'élection présidentielle de 2002 pour Le Figaro et Europe 1, les 17 et 18 avril 2002 auprès de 989 personnes, constituant un échantillon national représentatif de la population française âgée

a) Nous sommes dans les trois hypothèses d'approximation :
$989 \geqslant 30$, $138 \leqslant np \leqslant 198$ et $791 \leqslant n(1-p) \leqslant 851$

- Pour J.M. Le Pen $I_2 = [0,108\ ;\ 0,172]$

b) Les résultats sont bien dans les intervalles de confiance.

c) Les trois intervalles de confiance ont une intersection non vide :
$I_1 \cap I_2 \cap I_3 = [0,168\ ;\ 0,172]$
Il n'était donc pas possible de donner le classement final des trois candidats.
Tous les classements étaient possibles.