ESTIMATION

Probabilités - Chapitre 4

Table des matières

I Intervalle de fluctuation

I 1 Définition

Définition

Soit n un entier naturel non nul et p un réel strictement compris entre 0 et 1.
Soit X_n une variable aléatoire qui suit la loi binomiale $\mathcal{B}(n,p)$.
Soit α un réel tel que $0<\alpha<1$ et soient a et b deux réels tels que $a<b$.

On dit que l'intervalle $[a\,;b]$ est un **intervalle de fluctuation de** X_n au seuil de $1-\alpha$ si et seulement si :

$$P(a \leqslant X_n \leqslant b) \geqslant 1-\alpha$$

Remarque :
Cette définition est une généralisation de l'intervalle de fluctuation vue en Seconde et en Première.

I 2 Intervalle de fluctuation asymptotique

Théorème

Soit n un entier naturel non nul et p un réel strictement compris entre 0 et 1.
Soit X_n une variable aléatoire qui suit la loi binomiale $\mathcal{B}(n,p)$.

Alors pour tout réel α de $]0\,;1[$, on a :

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

avec u_α le nombre réel tel que

$$P(-u_\alpha \leqslant Z \leqslant u_\alpha) = 1-\alpha$$

lorsque Z est une variable aléatoire suivant la loi normale centrée réduite.

On appelle alors **variable fréquence**, la variable aléatoire

$$F_n = \frac{X_n}{n}$$

qui à tout échantillon de taille n associe la fréquence f obtenue.

Remarque :
I_n est l'intervalle de fluctuation asymptotique au seuil de $1-\alpha$ de la variable aléatoire fréquence F_n.
Le mot « asymptotique » vient du passage à la limite de l'intervalle I_n.

La loi binomiale $\mathcal{B}(n,p)$ peut alors être assimilée à la loi normale $\mathcal{N}(np, np(1-p))$.

Démonstration (BAC) :

Posons $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$. D'après le théorème de Moivre-Laplace :

$$\lim_{n\to+\infty} P(-u_\alpha \leqslant Z_n \leqslant u_\alpha) = P(-u_\alpha \leqslant Z \leqslant u_\alpha) \text{ avec } Z \hookrightarrow \mathcal{N}(0\,;1)$$

Ainsi, d'après les propriétés de la loi normale centrée réduite, pour tout α de $]0\,;1[$, il existe un unique réel strictement positif u_α tel que :

$$P(-u_\alpha \leqslant Z \leqslant u_\alpha) = 1-\alpha$$

Et ainsi,

$$\lim_{n\to+\infty} P(-u_\alpha \leqslant Z_n \leqslant u_\alpha) = 1-\alpha$$

De plus,

$$\begin{aligned}
&-u_\alpha \leqslant Z_n \leqslant u_\alpha \\
\Leftrightarrow\ &-u_\alpha\sqrt{np(1-p)} \leqslant X_n - np \leqslant u_\alpha\sqrt{np(1-p)} \\
\Leftrightarrow\ &np - u_\alpha\sqrt{np(1-p)} \leqslant X_n \leqslant np + u_\alpha\sqrt{np(1-p)} \\
\Leftrightarrow\ &p - u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant \frac{X_n}{n} \leqslant p + u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}
\end{aligned}$$

Donc $\lim\limits_{n\to+\infty} P\left(\dfrac{X_n}{n} \in I_n\right) = 1-\alpha$.

Remarque :
L'intervalle de fluctuation asymptotique au seuil de 95% correspond donc à l'intervalle I_n dans le cas où $\alpha = 0,05$ (et ainsi $u_\alpha \approx 1,96$) et est alors (à connaître) :

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

A noter que cet intervalle peut être simplifié par l'intervalle

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

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

$$\begin{aligned}
&0 \leqslant p(1-p) \leqslant 0,25 \\
\text{donc } &0 \leqslant \sqrt{p(1-p)} \leqslant \sqrt{0,25} \\
\text{donc } &0 \leqslant \sqrt{p(1-p)} \leqslant 0,5
\end{aligned}$$

On en déduit alors que :

$$0 \leqslant 1,96\sqrt{p(1-p)} \leqslant 1$$

On a alors $0 \leqslant 1,96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant \dfrac{1}{\sqrt{n}}$, et ainsi $I_n \subset J_n$.

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

Exemple :

On lance 120 fois un dé à 6 faces bien équilibré et on note X la variable aléatoire qui associe le nombre de fois que le dé affiche la face 6 sur les 120 lancers.
Déterminer l'intervalle de fluctuation asymptotique au seuil de 95% de la fréquence d'apparition du 6 dans un échantillon de 120 lancers :

$$p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}} = \frac{1}{6} - 1,96\frac{\sqrt{\frac{1}{6}\times\frac{5}{6}}}{\sqrt{120}} \approx 0,099 \text{ (valeur approchée par défaut*)}$$

$$p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}} = \frac{1}{6} + 1,96\frac{\sqrt{\frac{1}{6}\times\frac{5}{6}}}{\sqrt{120}} \approx 0,234 \text{ (valeur approchée par excès*)}$$

* : afin que la probabilité soit d'au moins 95%, il faut choisir une valeur approchée par défaut pour la borne inférieure, et par excès pour la borne supérieure.

Donc $I_n = [0,099\,;0,234]$ pour la variable aléatoire fréquence $\dfrac{X}{120}$.

I 3 Prise de décision

Propriété

Soit f_{obs} la fréquence observée d'un caractère sur un échantillon de taille n issu d'une population donnée. On suppose que $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$.
Test d'hypothèse : on fait une conjecture sur la valeur de la proportion p du caractère étudié dans la population toute entière.
Soit I_n l'intervalle de fluctuation asymptotique au seuil de 95%.
- Si $f_{obs} \in I_n$, on ne peut pas rejeter l'hypothèse faite sur p.
- Si $f_{obs} \notin I_n$, on rejete l'hypothèse faite sur p.

Exemple :

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 marrons et les perles jaunes sont annoncées comme représentant chacune 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 marrons.

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

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

1. Déterminer l'intervalle de fluctuation asymptotique I au seuil de 95% pour la proportion de perles marrons dans un échantillon de taille 690 (on donnera les résultats arrondis à 10^{-3} près).
2. Calculer la proportion de perles marrons dans l'échantillon. Que peut-on en conclure ?
3. Dans le même échantillon, il y avait 152 perles jaunes et 125 perles rouges.
   Que peut-on conclure de ces résultats ?

Correction :

1. En ce qui concerne les perles marrons, on a : $n = 690$ et $p = 0,2$ donc :

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

Les hypothèses du théorème de Moivre-Laplace sont vérifiées, on calcule ensuite :
$p - 1,96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}} = 0,2 - 1,96\dfrac{\sqrt{0,2 \times 0,8}}{\sqrt{690}} \approx 0,170$
$p + 1,96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}} = 0,2 + 1,96\dfrac{\sqrt{0,2 \times 0,8}}{\sqrt{690}} \approx 0,230$
On a donc $I \approx [0,170\,;0,230]$

2. On calcule la fréquence observée : $f_{obs} = \dfrac{140}{690} \approx 0,203$.
Comme $f_{obs} \in I$, **on ne peut pas rejeter** l'hypothèse selon laquelle les perles marrons représentent 20% des perles.

3. On calcule la fréquence observée des perles jaunes : $f_j = \dfrac{152}{690} \approx 0,220$.
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 :
$p - 1,96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}} = 0,1 - 1,96\dfrac{\sqrt{0,1 \times 0,9}}{\sqrt{690}} \approx 0,077$
$p + 1,96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}} = 0,1 + 1,96\dfrac{\sqrt{0,1 \times 0,9}}{\sqrt{690}} \approx 0,123$
On a donc $I' \approx [0,077\,;0,123]$.
On calcule la fréquence observée des perles rouges : $f_r = \dfrac{125}{690} \approx 0,18$.
Comme $f_r \notin I'$, **on doit rejeter** l'hypothèse selon laquelle les perles rouges représentent 10% des perles.

II Estimation

II 1 Intérêt d'une estimation

Pour des raisons de coût et de faisabilité, on ne peut pas 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 observée f_{obs} des individus de cet échantillon ayant ce caractère.
On **estime** alors la proportion p par un intervalle de confiance déterminée à partir de la fréquence f_{obs} et de la taille n de l'échantillon.

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 donnant une estimation par un intervalle.

II 2 Intervalle de confiance

On suppose que les trois conditions d'approximations sont remplies :

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

Théorème

Soit F_n la variable aléatoire fréquence qui à chacun des échantillons de taille n associe la fréquence du caractère dans cet échantillon.
La proportion inconnue p est telle que :

$$P\left(F_n - \frac{1}{\sqrt{n}} \leqslant p \leqslant F_n + \frac{1}{\sqrt{n}}\right) \geqslant 0,95$$

Démonstration :

On a vu plus haut que l'intervalle de fluctuation au seuil de 95% pouvait être simplifié par $\left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}\right]$.
On a donc : $p - \frac{1}{\sqrt{n}} \leqslant F_n \leqslant p + \frac{1}{\sqrt{n}}$
$\Leftrightarrow -\frac{1}{\sqrt{n}} \leqslant F_n - p \leqslant \frac{1}{\sqrt{n}}$
$\Leftrightarrow -F_n - \frac{1}{\sqrt{n}} \leqslant -p \leqslant -F_n + \frac{1}{\sqrt{n}}$
$\Leftrightarrow F_n - \frac{1}{\sqrt{n}} \leqslant p \leqslant F_n + \frac{1}{\sqrt{n}}$
Ainsi : $P\left(F_n - \frac{1}{\sqrt{n}} \leqslant p \leqslant F_n + \frac{1}{\sqrt{n}}\right) \geqslant 0,95$.

Définition

On observe la fréquence f_{obs} sur un échantillon de taille n et p désigne la proportion inconnue d'apparition du caractère dans la population entière. On appelle **intervalle de confiance** de p au niveau asymptotique de 95% l'intervalle :

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

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 :

$$\frac{2}{\sqrt{n}} \leqslant a \Leftrightarrow n \geqslant \frac{4}{a^2}$$

Remarque importante :

Pour un intervalle de confiance, puisque l'on veut que $P\left(F_n - \frac{1}{\sqrt{n}} \leqslant p \leqslant F_n + \frac{1}{\sqrt{n}}\right)$ soit **supérieur ou égal** à $0,95$, il faut arrondir les bornes de l'intervalle de confiance ainsi :

- On arrondit la borne inférieure de l'intervalle de confiance par une valeur approchée **par défaut**.
- On arrondit la borne supérieure de l'intervalle de confiance par une valeur approchée **par excès**.

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 de 18 ans et plus et inscrite sur les listes électorales.

On suppose cet échantillon constitué de manière aléatoire (même si en pratique ce n'est pas le cas). Les intentions de vote au premier tour pour les principaux candidats sont les suivantes :

Jacques Chirac : 20%
Lionel Jospin : 18%
Jean-Marie Le Pen : 14%

Les médias se préparent pour un second tour entre Jacques Chirac et Lionel Jospin.

1. Déterminer, pour chaque candidat, l'intervalle de confiance au niveau de confiance de 0,95 de la proportion inconnue d'électeurs ayant l'intention de voter pour lui.
2. Le 21 avril, les résultats du premier tour des élections sont les suivantes :
   Jacques Chirac : 19,88%
   Lionel Jospin : 16,18%
   Jean-Marie Le Pen : 16,86%
   Les pourcentages de voix recueillies par chaque candidat sont-ils bien dans les intervalles de confiance précédents ?
3. Pouvait-on, au vu de ce sondage, écarter avec un niveau de confiance de 0,95, l'un de ces trois candidats du second tour ?

Correction :

1. Les trois hypothèses d'approximation sont bien vérifiées :
   $989 \geqslant 30$, $138 \leqslant np \leqslant 198$ et $797 \leqslant n(1-p) \leqslant 851$.
   On calcule $\dfrac{1}{\sqrt{n}} = \dfrac{1}{\sqrt{989}} \approx 0,032$.
   On obtient alors les intervalles de confiance à 0,95 suivants :
   - Pour Jacques Chirac : $I_1 = [0,168\,;0,232]$
   - Pour Lionel Jospin : $I_2 = [0,148\,;0,212]$
   - Pour Le Pen : $[0,108\,;0,172]$
2. Les résultats sont bien dans les intervalles de confiance.
3. 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 !