Chapitre 7 : Fluctuation, estimation

I – Rappel sur la notion d'échantillonnage

Définition 1: Un **échantillon de taille *n*** est constitué des résultats de n répétitions indépendantes de la même expérience.

Remarque : L'**échantillonnage** est le prélèvement d'un échantillon dans une population.

Définition 2: On appelle **fluctuation d'échantillonnage** la variation de la distribution des fréquences d'un échantillon à l'autre.

Exemple de fluctuation d'échantillonnage

On simule sur ordinateur le jeu suivant :

On lance un dé à 6 faces. Si on fait 5 ou 6, on gagne, sinon on perd.

Face	gagné	perdu
Échantillon 1 : fréquence en %	36	64
Échantillon 2 : fréquence en %	32	68

On constate que la distribution des fréquences varie d'un échantillon à l'autre.

II – Intervalle de fluctuation asymptotique au seuil de 95 %

Propriété 1: Soit $n \in \mathbb{N}$ et $p \in [0;1]$ tels que $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$.

L'**intervalle de fluctuation asymptotique au seuil de 95 %** de la fréquence d'apparition f d'un succès sur un échantillon de taille n est $I_n = \left[p - 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$.

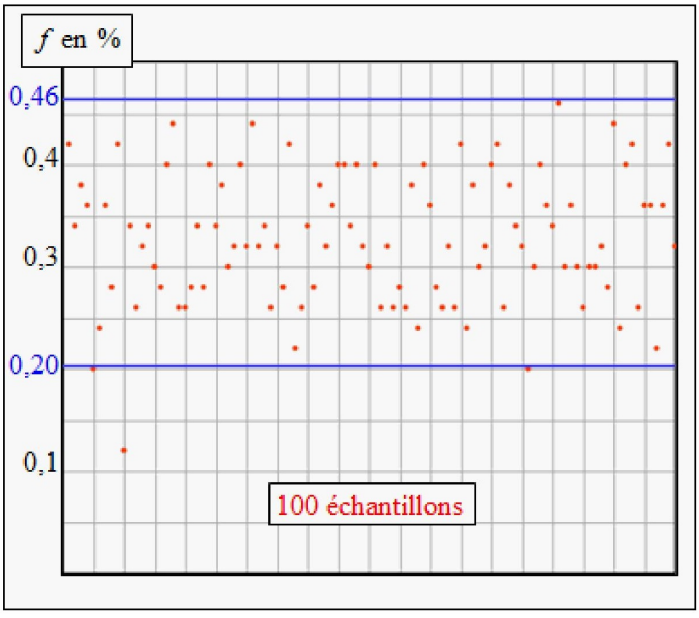
Exemple d'intervalle de fluctuation asymptotique au seuil de 95%

On reprend le jeu précédent et on constitue 100 échantillons de Bernoulli de taille $n = 50$.

La probabilité du succès est $p = \frac{2}{6} = \frac{1}{3}$.

$$p - 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}} = \frac{1}{3} - 1{,}96\frac{\sqrt{\frac{1}{3}\left(1-\frac{1}{3}\right)}}{\sqrt{50}} \approx 0{,}20$$

$$p + 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}} = \frac{1}{3} + 1{,}96\frac{\sqrt{\frac{1}{3}\left(1-\frac{1}{3}\right)}}{\sqrt{50}} \approx 0{,}46$$



L'intervalle de fluctuation asymptotique au seuil de 95% est donc : $I_{50} = [0{,}20;0{,}46]$

Il y a 3 points à l'extérieur des 2 lignes donc 97% des points du nuage sont dans la zone représentant l'intervalle de fluctuation asymptotique au seuil des 95%.

Exemple : Prendre une décision

Une fameuse marque de bonbons chocolatés vend des paquets constitués de bonbons de cinq couleurs différentes. Les bonbons marron sont indiqués sur l'emballage comme représentant 20% de l'ensemble des bonbons. Pour vérifier cette affirmation, les élèves d'une classe de terminale se sont amusé à observer un échantillon de 690 bonbons et ont dénombré 140 bonbons marron.

a) Que peut-on penser de la proportion annoncée ?

b) Dans un échantillon similaire, il y avait 152 bonbons jaunes pour une proportion annoncée de 20% et 125 bonbons rouges pour une proportion annoncée de 10%. Que peut-on conclure de ces résultats ?

Méthode :

La règle de décision est la suivante : si la fréquence observée de bonbons marron dans l'échantillon n'appartient pas à l'intervalle de fluctuation asymptotique, alors on rejette l'hypothèse selon laquelle les bonbons marron représentent 20% des bonbons.

a) • *On détermine l'intervalle de fluctuation asymptotique au seuil de 95% avec n = 690 et p = 0,2*

$$p-1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}=0{,}2-1{,}96\frac{\sqrt{0{,}2x0{,}8}}{\sqrt{690}}\approx 0{,}17015$$

$$p+1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}=0{,}2+1{,}96\frac{\sqrt{0{,}2x0{,}8}}{\sqrt{690}}\approx 0{,}22984$$

Avec des valeurs approchées à 10^{-2} près, l'intervalle de fluctuation asymptotique au seuil de 95% est :

$J=[0{,}17;0{,}23]$

• *On calcule la fréquence observée f :*

Il y a 140 bonbons marron sur 690 bonbons, par conséquent :

$$f=\frac{140}{690}\approx 0{,}2029\approx 0{,}20$$

• *Prise de décision :*

On constate que $f\in[0{,}17;0{,}23]$ donc **on ne peut pas rejeter** l'hypothèse selon laquelle les bonbons marron représentent 20% des bonbons.

Pour rejeter ou non l'hypothèse sur la proportion de bonbons marron, les étapes sont les suivantes :

On constate que les conditions sont bien remplies : n =690 (donc $n\geq 30$), np = 138 (donc $np\geq 5$), $n(1-p)$ = 552 (donc $n(1-p)\geq 5$).

On choisit une valeur approchée par défaut (resp. par excès) de la borne inférieure (resp. supérieure).

On calcule la fréquence observée de bonbons marron dans l'échantillon.

On applique la règle de décision : si la fréquence observée n'appartient pas à l'intervalle, on rejette l'hypothèse ; si la fréquence appartient à l'intervalle, on ne rejette pas l'hypothèse.

b) • *Pour les bonbons jaunes :*

n = 690 et p = 0,2 comme pour les bonbons marrons, donc l'intervalle de fluctuation asymptotique au seuil de 95% est le même qu'à la question a).

$f' = \frac{152}{690} \approx 0,22$ or $f' \in [0,17;0,23]$ donc **on ne peut pas rejeter** l'hypothèse selon laquelle les bonbons jaunes représentent 20% des bonbons.

• *Pour les bonbons rouges :*

$n = 690$ et $p = 0,1$

$p - 1,96\frac{\sqrt{p(1-p}}{\sqrt{n}} = 0,1 - 1,96\frac{\sqrt{0,2x0,8}}{\sqrt{690}} \approx 0,07$ valeurs arrondie à 10^{-2} près par défaut.

$p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}} = 0,1 + 1,96\frac{\sqrt{0,1x0,9}}{\sqrt{690}} \approx 0,13$ valeurs arrondie à 10^{-2} près par défaut.

L'intervalle de fluctuation asymptotique au seuil de 95% est donc : $I = [0,07;0,13]$

Or $f'' = \frac{125}{690} \approx 0,18$ et $f'' \notin [0,07;0,13]$ donc **on rejette** l'hypothèse selon laquelle les bonbons rouges représentent 10% des bonbons.

III – Estimation d'une proportion à partir d'un échantillon

Propriété 4: Soit f la fréquence observée dans un échantillon de taille n et p la proportion que l'on veut estimer (avec $n \in \mathbb{N}$ et $p \in [0;1]$ tels que $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$).

L'intervalle $\left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}} \right]$ contient p avec une probabilité d'au moins 95 %.

Définition 3: L'intervalle $\left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}} \right]$ est appelé **intervalle de confiance de p au niveau de confiance de 0,95** (On dit aussi avec un risque de 5 %).

Exemple d'estimation :

Dans un jeu, on constate que sur 100 joueurs prélevés au hasard, il y a 45 gagnants.

On souhaite estimer la proportion théorique p de gagnants, sur l'ensemble des joueurs.

On suppose que p est compris entre 0,2 et 0,8.

La fréquence observée dans l'échantillon de taille n = 100 est $f_{obs} = \frac{45}{100} = 0,45$

Vérifions les conditions d'application de l'intervalle de confiance au seuil de 95%

$nf_{obs} = 45 \geq 5$ et $n(1-f_{obs}) = 55 \geq 5$. On peut donc appliquer la propriété 4.

L'**intervalle de confiance** de p au niveau de confiance 0,95 est :

$$I_c = \left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}} \right] = \left[0,45 - \frac{1}{\sqrt{100}} ; 0,45 + \frac{1}{\sqrt{100}} \right] = [0,35;0,55]$$

Cet intervalle contient la proportion p avec une probabilité d'au moins 95%.

Méthode : déterminer la taille de l'échantillon

Le gérant d'une salle de concert veut estimer la proportion de places vendues non occupées afin d'organiser la surréservation.
Quelle taille d'échantillon de ses clients doit-il étudier pour obtenir un intervalle de confiance au niveau de confiance 0,95 de longueur au plus 0,04 ?

Mise en équation du problème :

n est la taille de l'échantillon et f est la fréquence qui serait observée dans l'échantillon.

L'intervalle de confiance au niveau de confiance 0,95 est : $I_c = \left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}} \right]$

C'est un intervalle centré sur f avec un écart de $\frac{1}{\sqrt{n}}$ de part et d'autre de f.

La longueur de l'intervalle est donc $\frac{2}{\sqrt{n}}$.

On cherche donc la plus petite valeur de n telle que $\frac{2}{\sqrt{n}} \leq 0,04$.

La fonction inverse est décroissante sur $]0;+\infty[$.

On passe à l'inverse dans chaque membre de l'inégalité, et on obtient une inégalité de sens contraire.

Résolution de l'inéquation :

$$\frac{2}{\sqrt{n}} \leq 0,04 \quad \Leftrightarrow \quad \frac{\sqrt{n}}{2} \geq \frac{1}{0,04} \quad \Leftrightarrow \quad \sqrt{n} \geq \frac{2}{0,04} \quad \Leftrightarrow \quad \sqrt{n} \geq 50 \quad \Leftrightarrow \quad n \geq 2500$$

Conclusion : le gérant doit étudier un échantillon d'au moins 2500 clients afin d'obtenir un intervalle de confiance au niveau 0,95 de longueur au plus 4%.

Méthode : Comprendre les risques de l'estimation

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 que cet échantillon est 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 suivants : 20% pour J. Chirac, 18% pour L. Jospin et 14% pour J-M. Le Pen.
Les médias se préparent pour un second tour entre J. Chirac et L. Jospin.

a) 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.

b) Le 21 avril, les résultats du premier tour sont les suivants : 19,88% pour J. Chirac, 16,18% pour L. Jospin et 16,86% pour J-M. Le Pen.
Que peut-on penser des pourcentages de voix recueillies par chaque candidat ?

c) Pouvait-on, au vu de ce sondage, écarter avec un niveau de confiance de 95% l'un des 3 candidats pour le second tour ?

a) L'échantillon étudié comprend 989 personnes. Pour chaque candidat, l'intervalle de confiance au niveau de confiance de 0,95 est de la forme :

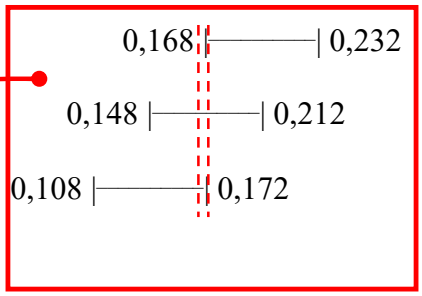
$$\left[f - \frac{1}{\sqrt{989}} ; f + \frac{1}{\sqrt{989}} \right].$$

La valeur unique en pourcentage donnée par l'institut de sondage est entachée d'une imprécision de ± 3 points environ.

Cela donne :

Candidat	J. Chirac	L. Jospin	J-M. Le Pen
Fréquence observée	0,2	0,18	0,14
Intervalle de confiance	[0,168 ;0,232]	[0,148 ;0,212]	[0,108 ;0,172]
Résultat du scrutin	0,1988	0,1618	0,1686

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



c) Ces trois intervalles de confiance ont une intersection non vide : [0,168 ;0,172]. Il n'est donc pas possible, avec un niveau de confiance de 0,95 de désigner le classement final des trois candidats.