I Intervalle de fluctuation

Contexte : Dans une certaine population, la proportion d'individus présentant le caractère C est p.
Que peut-on dire de la fréquence f de C sur un échantillon aléatoire de taille n ?

Population
Caractère C
de proportion p
connue ou
supposée connue

Échantillon
de taille n
Fréquence f
de C ?

I.1 Mise en place de la notion d'intervalle de fluctuation

I.1.1 Définition

Soit X_n une variable aléatoire suivant une loi binomiale $\mathcal{B}(n;p)$ et α un réel de l'intervalle $]0;1[$. On pose $F_n = \dfrac{X_n}{n}$ la variable aléatoire fréquence du succès.

`Définition` : Tout intervalle $[a;b]$ tel que : $P(F_n \in [a;b]) \geqslant 1-\alpha$ peut-être considéré comme un intervalle de fluctuation de F_n au seuil de $1-\alpha$

Remarque 1 L'intervalle $[0;1]$ est-il un intervalle de fluctuation ? ...

I.1.2 Intervalles de fluctuation vus au lycée

- **En Seconde** : Vous avez vu, que sous les conditions $0,3 < p < 0,7$, un intervalle de fluctuation de la variable $F_n = \dfrac{X_n}{n}$ au seuil de 95% est l'intervalle :
$$\left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$$
- **En première** : On rechercher l'intervalle qui « symétrise » les probabilités que X_n soit à l'extérieur. En pratique, on cherche le plus petit entier a pour lequel $P(X_n \leqslant a)$ est strictement supérieur à 0,025 et le plus petit entier b pour lequel $p(X_n \leqslant b)$ est supérieur ou égal à 0,975. On prend alors comme intervalle de fluctuation de $F_n = \dfrac{X_n}{n}$ l'intervalle :
$$\left[\frac{a}{n}; \frac{b}{n}\right]$$

Remarque 2 D'autres intervalles de fluctuation sont possibles, on peut par exemple prendre celui qui a une amplitude minimale, ou le plus petit intervalle centré sur l'espérance de F_n. $(E(F_n) = p)$

EXERCICE 1 On joue à Pile ou Face avec une pièce truquée telle que $P(Pile) = 0,3$. On la lance 100 fois et on note X_{100} la variable aléatoire qui compte le nombre de Pile. X_{100} suit une .... Donner trois intervalles I_{100} de fluctuation de F_{100} au seuil de 95% et calculer la probabilité des événements $F_{100} \in I_{100}$ pour chacun de ces intervalles.

I.2 Intervalle de fluctuation asymptotique

I.2.1 Retour sur le théorème de Moivre-Laplace

Théorème : Soit X_n une variable aléatoire suivant une $\mathcal{B}(n;p)$. On pose $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$, variable centrée et réduite associée à X_n.
Alors, pour tous réels a et b, $a < b$, on a :

$$\boxed{\lim_{n\to+\infty} P(Z_n \in [a;b]) = \int_a^b \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-t^2/2}\,\mathrm{d}t = P(Z \in [a;b])} \text{ où } Z \text{ suit la loi } \mathcal{N}(0;1)$$

I.2.2 Retour sur le nombre u_α

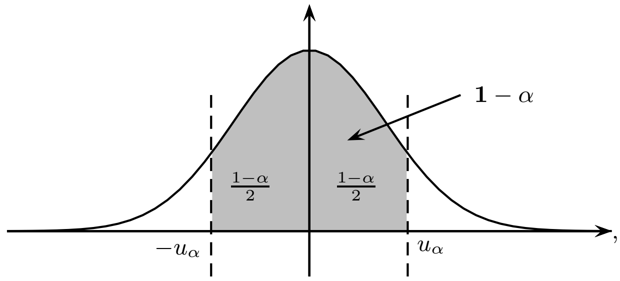
Théorème et Définition : Soit Z une variable aléatoire suivant une $\mathcal{N}(0;1)$, $\forall \alpha \in [0;1]$, il existe un unique nombre u_α tel que

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

Valeurs de u_α à connaître

α	$1-\alpha$	u_α	Interprétation
0,05	0,95	1,96	$P(-1,96 \leqslant Z \leqslant 1,96) \approx 0,95$
0,01	0,99	2,58	$P(-2.58 \leqslant Z \leqslant 2.58) \approx 0,99$

,



,

I.2.3 Intervalle de fluctuation asymptotique

Théorème : Soit $\alpha \in]0;1[$ et X_n une variable aléatoire qui suit une loi binomiale $\mathcal{B}(n;p)$.
La probabilité que F_n prenne ses valeurs dans l'intervalle

$$I_n = \dots$$

se rapproche de .... quand la taille de l'échantillon devient grande. On note ....

Démonstration : X_n une variable aléatoire qui suit une loi binomiale $\mathcal{B}(n;p)$.

1. Écrire M.L avec $a = -u_\alpha$ et $b = u_\alpha$.

2. « Revenir » à la variable X_n dans $P(Z_n \in [-u_\alpha; u_\alpha])$.

3. Conclure

Remarque 3 ▷ Lorsqu'une suite $(u_n)_{n\in\mathbb{N}}$ converge vers ℓ, pour n « assez grand » le terme u_n de rang n constitue une approximation de ℓ.
Ici, la situation est inverse, la suite $(P(F_n \in I_n))_{n\geqslant 1}$ admet $1-\alpha$ comme limite. On doit se poser la question du choix de conditions pour lesquelles $P(F_n \in I_n)$ est « proche » de $1-\alpha$.
Les conditions communément admises sont :

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

On peut donc sous ces conditions, dire que $P(F_n \in I_n)$ est quasiment égal à $1-\alpha$.

De sorte que, l'intervalle $I_n = \left[p - u_\alpha \dfrac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + u_\alpha \dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ est un intervalle de fluctuation « approchée » de la variable fréquence $F_n = \dfrac{X_n}{n}$ au seuil $1-\alpha$. (au sens vu au I.1.1)

Approchée, en effet, car la suite de terme général $P\left(\dfrac{X_n}{n} \in I_n\right)$ n'est pas **monotone**, on ne peut pas savoir si la probabilité de l'intervalle est supérieure ou inférieure à la limite $1-\alpha$.

▷ Pour $n = 44$ et $p = 0,5$, l'intervalle I_n continent celui de $1S$; il contient donc plus de 95% des fréquences d'échantillons de taille 44.
Pour $n = 60$ et $p = 0,5$, l'intervalle I_n est inclus dans celui de 1S : peut-être contient-il moins de 95% des fréquences observées sur des échantillons de taille 60 !

Définition : Un intervalle de fluctuation asymptotique de la variable aléatoire $F_n = \dfrac{X_n}{n}$ 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.

Ainsi d'après le théorème vu en I.2.3, l'intervalle

$$\mathbf{I_n} = \left[\mathbf{p} - \mathbf{u}_\alpha \frac{\sqrt{\mathbf{p(1-p)}}}{\sqrt{\mathbf{n}}} ; \mathbf{p} + \mathbf{u}_\alpha \frac{\sqrt{\mathbf{p(1-p)}}}{\sqrt{\mathbf{n}}}\right]$$

est un intervalle de fluctuation asymptotique de F_n au seuil $1-\alpha$.

I.2.4 Exemple d'utilisation

On admet que dans la population d'enfants de 11 à 14 ans d'un département français le pourcentage d'enfants ayant eu une crise d'asthme dans leur vie est de **13%**.
Un médecin d'une ville de ce département est surpris du nombre important d'enfants le consultant ayant eu une crise d'asthme et en informe les services sanitaires. Ceux-ci décident d'entreprendre une étude et d'évaluer la proportion d'enfants de 11 à 14 ans ayant déjà eu des crises d'asthme.
Il sélectionnent de manière aléatoire **100** jeunes de 11 à 14 ans de la ville.

La règle de décision prise est la suivante : si la proportion observée est supérieure à la borne supérieure de l'intervalle de fluctuation asymptotique au seuil de 95% alors une investigation plus complète sera mise en place afin de rechercher les facteurs de risques pouvant expliquer cette proportion élevée.

1. Déterminer l'intervalle de fluctuation asymptotique au seuil de 95% de la proportion de jeunes de 11 à 14 ans ayant eu une crise d'asthme dans un échantillon de taille 100. (vérifier au préalable que les conditions d'utilisation de l'expression de l'intervalle sont réalisées : schéma de Bernoulli et conditions avec n et p)
2. L'étude réalisée auprès de 100 personnes a dénombré **19** jeunes ayant déjà eu des crises d'asthme. Que pouvez-vous en conclure ?
3. Le médecin n'est pas convaincu par cette conclusion et déclare que le nombre de jeunes interrogées était insuffisant pour mettre en évidence qu'il y avait plus de jeunes ayant eu des crises d'asthme dans sa ville que dans le reste du département. Combien faudrait-il prendre de sujets pour qu'une proportion observée de 19% soit en dehors de l'intervalle de fluctuation asymptotique ?

II Estimation

II.1 Principe de l'étude d'un caractère C dans une population P

Deux méthodes sont possibles :

- La méthode exhaustive qui consiste à recenser tous les individus de la population. Cette méthode, en raison de son coût et de sa durée, est fort peu employée.

- La méthode des sondages qui consiste à n'étudier qu'un échantillon E, extrait de la population, et à induire, à partir des résultats observés sur cet échantillon des résultats concernant la population entière. La difficulté est d'obtenir un échantillon représentatif de la population. Nous admettrons qu'en réalisant un tirage au sort, cette condition est réalisée.

Contexte : La fréquence d'individus présentant le caractère C sur un échantillon de taille n est f.
Que peut-on dire de la proportion p de C dans la population ?

Population
Proportion p
du caractère C ?

Échantillon
de taille n
Fréquence f
de C connue

II.2 Intervalle de confiance

II.2.1 Principe général

Étant donné une proportion p inconnue, la procédure d'estimation consiste à utiliser les informations recueillies dans un échantillon sélectionné de manière aléatoire pour obtenir la valeur de la variable F_n destinée à fournir une estimation de p. Cette estimation va varier d'un échantillon à l'autre, il est donc nécessaire d'apprécier l'incertitude en fournissant une estimation par intervalle, appelé **intervalle de confiance de** p.

Propriété : X_n est une variable qui suit une loi binomiale $\mathcal{B}(n;p)$ et on note $F_n = \dfrac{X_n}{n}$.
Ob suppose que $\mathbf{n \geqslant 30}$, $\quad \mathbf{np \geqslant 5}$, $\quad \mathbf{n(1-p) \geqslant 5}$.
p est une proportion inconnue ($\in]0;1[$), l'intervalle $\left[F_n - \dfrac{1}{\sqrt{n}}; F_n + \dfrac{1}{\sqrt{n}}\right]$ contient pour n assez grand, la proportion p avec une probabilité environ égale à 0,95.

`Démonstration :`

1. $\mathbf{n \geqslant 30}$, $\mathbf{np \geqslant 5}$, $\mathbf{n(1-p) \geqslant 5}$. En utilisant l'intervalle de fluctuation asymptotique du paragraphe précédent, prouver que

$$P\left(p - \frac{1}{\sqrt{n}} \leqslant \frac{X_n}{n} \leqslant p - \frac{1}{\sqrt{n}}\right) \approx 0,95.$$

2. En déduire que $P\left(F_n - \dfrac{1}{\sqrt{n}} \leqslant p \leqslant F_n - \dfrac{1}{\sqrt{n}}\right) \approx 0,95$ ce qui se traduit en disant que l'intervalle aléatoire $\left[F_n - \dfrac{1}{\sqrt{n}} : F_n + \dfrac{1}{\sqrt{n}}\right]$ a une probabilité approximativement égale à 0,95 de contenir p, ou encore environ 95% des intervalles $\left[F_n - \dfrac{1}{\sqrt{n}} : F_n + \dfrac{1}{\sqrt{n}}\right]$ contiennent p.

II.2.2 Application

On calcule une fréquence f à partir d'un échantillon de taille n, on détermine l'intervalle $\left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}\right]$.

DÉFINITION : Cet intervalle est appelé intervalle de confiance de la proportion p inconnue au niveau de confiance 0,95.

Remarque 4 En réalisant le tirage d'un échantillon, on obtient un intervalle de confiance de la proportion inconnue p au niveau de confiance de 0,95.
Ainsi, à chaque **choix d'échantillon**, on obtient un **intervalle de confiance différent**.

Exemple 1 : Estimation à partir d'un échantillon.
A et B sont candidats à une élection. La population semble partagée entre les deux candidats. Un journal décide de réaliser un sondage sur un échantillon de 900 personnes et constate que 468 sont favorables à A. Que devrait-il dire à ses lecteurs ?

Exemple 2 : Dans un urne contenant des boules rouges et bleues en proportions inconnues, on effectue des tirages au hasard avec remise.

1. Après avoir effectué 100 tirages, on compte 52 boules rouges et 48 boules bleues. Donner un intervalle de confiance à 95% de la proprotion p de boules rouges dans l'urne.
2. Combien faudrait-il, au minimum, effectuer de tirages pour obtenir un intervalle de confiance à 95% de longueur inférieure ou égale à 0,02 ?

II.3 Intervalle de fluctuation ou Intervalle de confiance

Règle générale :

▷ On utilise un intervalle de fluctuation lorsque la proportion p dans la population est **connue ou si l'on fait une hypothèse sur sa valeur** : on prend alors une décision sur cette hypothèse.
(exercice corrigé page 409)

▷ On utilise un intervalle de confiance lorsque l'on veut estimer une proportion **inconnue** dans une population.
(exercice 8 résolu page 414)

intervalle de fluctuation p connue	intervalle de confiance p inconnue
$\mathbf{n \geqslant 30,\ np \geqslant 5,\ n(1-p) \geqslant 5}$ Asymptotique au seuil $1-\alpha$ $I_F = \left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$	$\mathbf{n \geqslant 30,\ np \geqslant 5,\ n(1-p) \geqslant 5}$ Au niveau de confiance 95% $I_C = \left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}\right]$