Échantillonnage

I Fluctuation d'échantillonnage

1) Fluctuation d'échantillonnage

Définition : En statistiques, **un échantillon de taille n** est la liste des n résultats obtenus par n répétitions indépendantes de la même expérience.

Exemple 1 : L'expérience consiste à lancer 100 fois de suite un dé numéroté de 1 à 6, bien équilibré, et à noter après chaque lancer le chiffre qui apparaît sur la face supérieure du dé lorsque celui-ci s'est immobilisé.

On répète deux fois cette expérience et on obtient donc deux échantillons de taille 100.

Dans le tableau ci-dessous sont notées les fréquences d'apparition de chaque chiffre pour chaque échantillon.

Chiffre	1	2	3	4	5	6
Fréquences échantillon n° 1	0,15	0,17	0,18	0,14	0,17	0,19
Fréquences échantillon n° 2	0,15	0,16	0,18	0,16	0,18	0,17

On constate que les distributions de fréquence des deux échantillons sont différentes : c'est ce que l'on appelle **la fluctuation d'échantillonnage.**

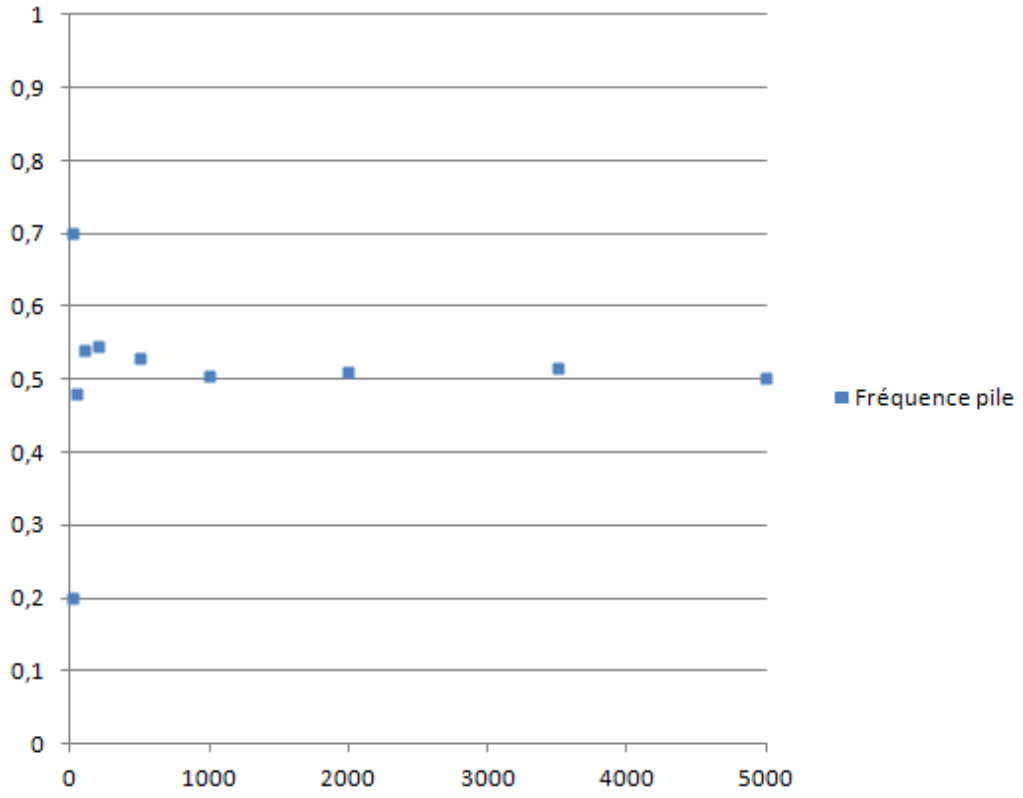
2) Loi des grands nombres

Exemple 2 :

L'expérience consiste à lancer n fois une pièce équilibrée et à calculer la fréquence d'apparition du côté pile.

Voici les résultats obtenus en fonction de la taille de l'échantillon n :

n	10	20	50	100	200	500	1000	2000	3500	5000
fréq pile	0,2	0,7	0,48	0,54	0,55	0,53	0,51	0,51	0,51	0,5



Propriété : Pour une expérience donnée, dans le modèle défini par une loi de probabilité, les distributions de fréquences obtenues dans des séries de taille n se rapprochent de la loi de probabilité quand n devient grand.

Dans l'exemple précédent, lorsque la taille de l'échantillon devient grande la fréquence d'apparition du côté pile se rapproche de 0,5.

II Intervalle de fluctuation

1) Intervalle de fluctuation et prévision

On considère une expérience aléatoire dont le résultat est soit un succès avec la probabilité p, soit un échec avec la probabilité 1 – p (appelée expérience de Bernoulli).

On souhaite réaliser un échantillon de taille n de cette expérience. En raison de la fluctuation d'échantillonnage, la fréquence qui sera observée pour le caractère ne sera pas nécessairement égale à la probabilité attendue. Cependant, sous certaines conditions et en admettant un risque d'erreur on peut donner un intervalle contenant cette fréquence.

Théorème : On répète n fois une expérience aléatoire dont la probabilité de succès est p.

Lorsque $\mathbf{n \geqslant 25}$ et lorsque et $\mathbf{0,2 \leqslant p \leqslant 0,8}$ la fréquence observée du nombre de succès appartient à l'intervalle $\mathbf{\left[p - \frac{1}{\sqrt{n}} \ ; \ p + \frac{1}{\sqrt{n}} \right]}$ avec un **probabilité d'au moins 0,95**.

Cet intervalle est appelé **intervalle de fluctuation de taille n au seuil de 95 %.**

Cette propriété signifie que dans 95% des cas, la fréquence d'apparition d'un succès est située dans l'intervalle centré en p d'amplitude $\frac{2}{\sqrt{n}}$.

Exemple 3 : prévision

Un joueur tire une carte dans un jeu de 32 cartes puis il la remet à l'intérieur du jeu.

Il gagne s'il obtient un « cœur ». Le joueur répète 100 fois l'expérience.

La probabilité d'un succès est donc égale à 0,25 à chaque tirage.

On a $p - \frac{1}{\sqrt{n}} = 0,25 - \frac{1}{\sqrt{100}} = 0,15$ et $p + \frac{1}{\sqrt{n}} = 0,25 + \frac{1}{\sqrt{100}} = 0,35$

On peut donc affirmer que dans 95% des cas, la fréquence de succès du joueur observée après les 100 tirages appartient à l'intervalle [0,15 ; 0,35].

On peut prévoir que dans 95 % des cas, sur 100 tirages, le joueur gagnera entre 15 et 35 fois.

Remarque :

Si le joueur répète 1000 fois l'expérience l'intervalle de fluctuation de la fréquence du succès est [0,24 ; 0,26]

Cet intervalle est de moindre amplitude (plus « précis ») mais le risque d'erreur est toujours de 5%.

2) Prise de décision

Exemple 4 : *Dans une commune de 50 000 habitants, la proportion de femmes est 0,5.*

Au conseil municipal, composé de 43 personnes, il y a 17 femmes.

Peut-on affirmer qu'au conseil municipal, la parité homme-femme n'est pas respectée ?

La taille du conseil municipal (43) étant très faible par rapport à la taille de la population (50 000) on peut le considérer comme un échantillon de la population.

On a $p - \frac{1}{\sqrt{n}} = 0,5 - \frac{1}{\sqrt{43}} \simeq 0,347$ (arrondi par défaut) et $p + \frac{1}{\sqrt{n}} = 0,5 + \frac{1}{\sqrt{43}} \simeq 0,653$ (arrondi par excès)

La fréquence des femmes dans l'échantillon est égale à $\frac{17}{43} \simeq 0,4$.

$0,4 \in [0,347 ; 0,653]$ donc au seuil de 95% il n'y a pas lieu d'affirmer que la parité n'est pas respectée.

Exemple 5 : *26% des Français se déclarent allergiques aux pollens de fleurs.*

Dans un échantillon de 400 personnes on trouve 130 personnes allergiques.

Le nombre d'allergiques dans cet échantillon est-il normal ?

$p - \frac{1}{\sqrt{n}} = 0,26 - \frac{1}{\sqrt{400}} = 0,21$ et $p + \frac{1}{\sqrt{n}} = 0,26 + \frac{1}{\sqrt{400}} = 0,31$

Donc l'intervalle de fluctuation au seuil de 95 % est [0,21 ; 0,31]

On a $\frac{130}{400} = 0,325$ et $0,325 \notin [0,21 ; 0,31]$ donc avec un risque d'erreur de 5% on peut affirmer que cet échantillon est anormal. Il semble y avoir une proportion trop importante de personnes allergiques.

III Intervalle de confiance

On va maintenant étudier comment estimer une proportion inconnue dans une population à partir d'un échantillon.

Dans le cas d'une Dans les conditions du théorème précédent on a dans 95% des cas :

$$p - \frac{1}{\sqrt{n}} \leqslant f \leqslant p + \frac{1}{\sqrt{n}}$$

on en déduit : $$-f - \frac{1}{\sqrt{n}} \leqslant -p \leqslant -f + \frac{1}{\sqrt{n}}$$

d'où $$f + \frac{1}{\sqrt{n}} \geqslant p \geqslant f - \frac{1}{\sqrt{n}}$$

On raisonne alors de la manière suivante :

Pour estimer une probabilité inconnue p, on prélève un échantillon de taille n pour lequel on obtient une fréquence f.

La probabilité p appartient à l'intervalle $\left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}} \right]$ dans 95 % des cas.

Cet intervalle est appelé **intervalle de confiance de p au niveau de confiance 95%.**

Il ne faut pas confondre intervalle de fluctuation et intervalle de confiance.

Exemple 6 : *Lors d'une élection, un sondage portant sur 1 000 personnes donne 350 votants pour un candidat.*

Quelles informations peut-on obtenir sur la proportion de votants pour ce candidat dans la population ?

La fréquence des votants pour ce candidat dans cet échantillon est $f = \frac{350}{1000} = 0{,}35$.

On a $f - \frac{1}{\sqrt{n}} \simeq 0{,}318$ et $f + \frac{1}{\sqrt{n}} \simeq 0{,}382$ donc on peut estimer, avec un risque de se tromper de 5 % que ce candidat obtiendra entre 31,8% et 38,2% des voix.