

Échantillonnage

Prélude - Historique et contexte

En 1713, dans *Ars Conjectandi*, Jacques Bernoulli démontre ce qu'on appelle aujourd'hui la **loi des grands nombres** : en répétant une expérience un grand nombre de fois, la fréquence d'apparition d'un événement se rapproche de sa probabilité. C'est ce principe qui permet aux instituts de sondage de prédire un résultat à partir de quelques milliers de personnes, ou aux industriels de contrôler la qualité d'une production sans tester chaque pièce.

I Échantillon et fluctuation

Définition

Un **échantillon de taille n** est la liste des résultats obtenus en répétant n fois, de manière indépendante, la même expérience aléatoire.

On s'intéresse à une issue de probabilité p et on note f sa **fréquence observée** dans l'échantillon.

Remarque

D'un échantillon à l'autre, la fréquence f varie : c'est la **fluctuation d'échantillonnage**.

Exemple

On lance 20 fois une pièce équilibrée et on compte les « Pile ». Trois élèves obtiennent 8, 11 et 13 « Pile », soit des fréquences 0,40 ; 0,55 et 0,65 alors que $p = 0,5$.

II Loi des grands nombres

Propriété (version vulgarisée)

Lorsque n est grand, sauf exception, la fréquence observée f est proche de la probabilité p .

Propriété - Intervalle de fluctuation

Pour des échantillons de taille n assez grande, dans environ **95 %** des cas, l'écart entre la fréquence observée et la probabilité est inférieur ou égal à $\frac{1}{\sqrt{n}}$:

$$\boxed{|f - p| \leq \frac{1}{\sqrt{n}}} \quad \text{c'est-à-dire} \quad f \in \left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}} \right].$$

Exemple

On lance 400 fois un dé équilibré. La probabilité d'obtenir un 6 est $p = \frac{1}{6} \approx 0,167$. Comme $\frac{1}{\sqrt{400}} = 0,05$, dans environ 95 % des cas la fréquence de 6 sera entre 0,117 et 0,217, c'est-à-dire entre environ 47 et 87 « six ».

III Estimer une probabilité

Méthode

Lorsqu'on ne connaît pas la probabilité p (dé truqué, sondage...), on l'estime par la fréquence observée sur un grand échantillon. D'après la propriété précédente, dans environ 95 % des cas :

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

Exemple

Un sondage auprès de 1 000 personnes donne 52 % d'intentions de vote pour un candidat. $\frac{1}{\sqrt{1000}} \approx 0,032$, donc la proportion réelle est, avec une confiance de 95 %, entre 48,8 % et 55,2 %. On ne peut pas affirmer que le candidat est majoritaire!

IV Simulation

Remarque - Python

On simule N échantillons de taille n d'une expérience à deux issues de probabilité p , et on calcule la proportion des échantillons pour lesquels l'écart entre f et p est inférieur ou égal à $\frac{1}{\sqrt{n}}$.

```
from random import random
from math import sqrt

def frequence(n, p):
    succes = 0
    for i in range(n):
        if random() < p:
            succes = succes + 1
    return succes / n

def proportion_ecart(N, n, p):
    compteur = 0
    for k in range(N):
        f = frequence(n, p)
        if abs(f - p) <= 1 / sqrt(n):
            compteur = compteur + 1
    return compteur / N
```

En pratique, `proportion_ecart(1000, 100, 0.3)` renvoie une valeur proche de 0,95 ou plus.

Remarque

L'intervalle n'est pas toujours symétrique en pratique : pour une probabilité très proche de 0 ou de 1, les fréquences ne peuvent pas sortir de $[0 ; 1]$.