⧫ Chapitre 16. — Échantillonnage

Dans tout ce qui suit, on s'intéresse à un expérience aléatoire et on note A un évènement relatif à cette expérience. Un loi de probabilité sur l'univers ayant été choisie pour modéliser l'expérience, on note p la probabilité de l'évènement A.

1 Fluctuation d'échantillonnage

Définition 1. — Soit n un entier naturel non nul. Un échantillon de taille n de l'expérience aléatoire est une série de résultats obtenus en réalisant n fois l'expérience aléatoire dans les mêmes conditions.

À un tel échantillon, on associe la fréquence f de A définie par $f = \dfrac{n_A}{n}$ où n_A est le nombre de fois où A a été réalisé lors des n expériences effectuées.

Exemple. — On lance un dé équilibré et on considère l'évènement A : « obtenir un nombre supérieur ou égal à 5 ». On modélise l'expérience par l'équiprobabilité sur $\{1, 2, 3, 4, 5, 6\}$.

Si on lance 20 fois ce dé, on obtient un échantillon de taille 20 de l'expérience. Si les résultats sont

$$5 - 6 - 2 - 3 - 4 - 4 - 6 - 5 - 3 - 1 - 5 - 2 - 3 - 2 - 2 - 4 - 3 - 3 - 1 - 2$$

alors $n_A = 5$ et la fréquence de A est $f = \dfrac{5}{20} = \dfrac{1}{4}$.

Si on relance 20 fois ce même dé, on obtient un autre échantillon, par exemple :

$$5 - 4 - 5 - 6 - 2 - 3 - 2 - 6 - 4 - 1 - 5 - 3 - 1 - 6 - 5 - 6 - 4 - 3 - 6 - 5.$$

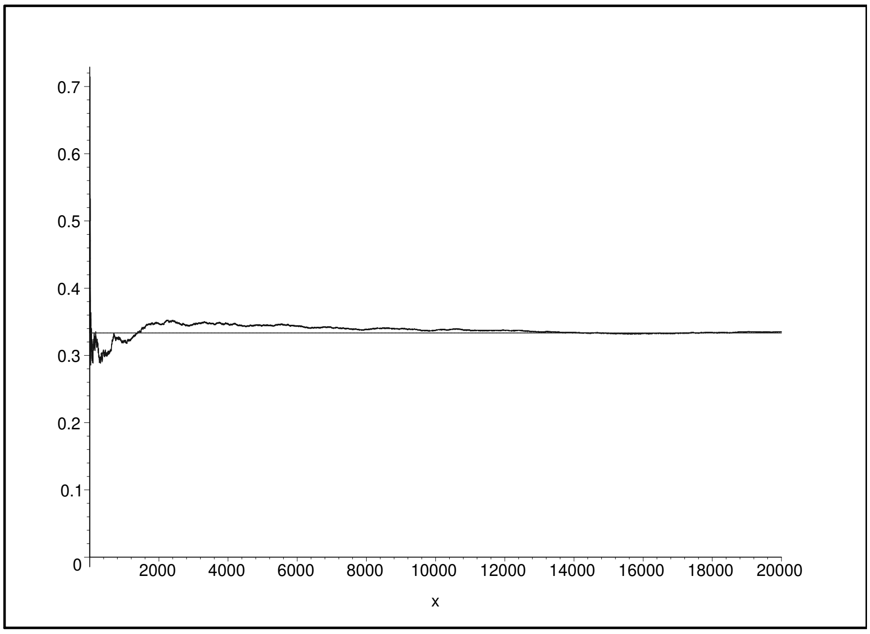
Pour ce second échantillon, $n_A = 10$ et la fréquence de A est $f = \dfrac{10}{20} = \dfrac{1}{2}$.

Lorsqu'on considère plusieurs échantillons et qu'on calcule, pour chaque échantillon, la fréquence du même évènement A, on trouve en général des fréquences différentes et cela même si les échantillons ont la même taille. Ainsi, la fréquence de A varie en fonction de l'échantillon. Ce phénomène est appelé la **fluctuation d'échantillonnage**.

2 Loi faible des grands nombres

Cependant, malgré la fluctuation d'échantillonnage, lorsqu'on calcule la fréquence de l'évènement A sur des échantillons dont les tailles sont de plus en plus grandes, les résultats obtenus tendent à se stabiliser et à se rapprocher de la probabilité théorie p de l'évènement A.

Exemple. — Pour n compris entre 1 et 20 000, on a simulé des échantillons de taille n d'un lancer de dé et calculer la fréquence de l'évènement « Obtenir un nombre supérieur ou égal à 5 ». Le graphique ci-dessous donne cette fréquence en fonction de n. La droite horizontale est la droite d'équation $y = p$ c'est-à-dire ici $y = \frac{1}{3}$.



On voit sur ce graphique que, pour cet échantillon, effectivement la fréquence observée se rapproche de la probabilité théorique lorsque n augmente.

C'est un résultat général appelé loi faible des grands nombres qu'on peut résumer ainsi : « Lorsque n est grand, sauf exception, la fréquence f est proche de la probabilité p ». Il faut comprendre le « sauf exception » de la manière suivante : si on considère différents échantillons de taille n avec n grand, il est rare de tomber sur un échantillon pour lequel f diffère sensiblement de p et cela est d'autant plus rare que n grand.

On peut estimer cette « rareté » à l'aide de fonctions Python.

```
from random import *
from math import *

def fréquence(n):
  c=0
  for i in range(n):
    a=randint(1,6)
    if (a>=5):
      c+=1
  return(c/n)

def nb_écart(n,m,e):
  nb=0
  for i in range(m):
    f=fréquence(n)
    if (abs(1/3-f)>e):
      nb+=1
  return(nb)
```

La fonction `fréquence(n)` simule un échantillon de taille `n` d'un lancer de dé équilibré et renvoie la fréquence de l'évènement A : « obtenir un nombre supérieur ou égal à 5 ». La fonction `nb_écart(n,m,e)` simule `m` échantillons de taille `n` d'un lancer de dé et renvoie le nombre d'échantillons pour lesquels la fréquence f de A diffère de la probabilité théorique $\frac{1}{3}$ de plus de `e`. On appellera ces échantillons des échantillons *exceptionnels*.

Voici les résultats obtenus pour `m=1000` et différentes valeurs de `n` et `e`

`e` \ `n`	50	100	500	1000	5000	10000
0,5	0	0	0	0	0	0
0,1	147	39	0	0	0	0
0,05	450	279	16	1	0	0
0,01	882	820	659	504	121	34

Ainsi, par exemple, pour `n=500` et `e=0,05`, on a trouvé 16 ce qui signifie que sur les 1000 échantillons de taille 500, 16 donnaient une fréquence différentes de $\frac{1}{3}$ de plus de 0,5 (en plus ou en moins).

On voit que, pour une précision donnée, plus on augmente la taille de l'échantillon, plus le nombre d'échantillons *exceptionnels* diminue et plus on augmente la précision (c'est-à-dire plus on diminue `e`), plus le nombre d'échantillons *exceptionnels* augmente.

La loi faible des grands nombres dit précisément que, pour une précision donnée, la probabilité de tomber sur un échantillon *exceptionnels* tend vers 0 lors n augmente indéfiniment.

Pour finir, on peut montrer qu'en général, pour des échantillons de taille n, f appartient à l'intervalle $\left[p - \frac{1}{\sqrt{n}}\,;p + \frac{1}{\sqrt{n}}\right]$ dans au moins 95% des cas.

Ceci peut se vérifier expérimentalement. Voici les résultats obtenus grâce à la fonction `nb_écart`, toujours avec `m=1000` et pour différentes valeurs de `n` en prenant `e=`$\frac{1}{\sqrt{\mathtt{n}}}$:

`n`	100	500	1000	5000	10000
nombre d'échantillons exceptionnels	42	26	36	28	41

Ainsi, on obtient entre $\frac{28}{1000} = 2{,}8\%$ et $\frac{42}{1000} = 4{,}2\%$ d'échantillons *exceptionnels*. Ainsi, dans tous les cas, les échantillons *exceptionnels* représentent moins de 5% des échantillons.