Échantillonnage

1 Simuler des échantillons

On considère une **population** dans laquelle on connaît la proportion des individus présentant un certain **caractère**. On extrait de cette population étudiée un petit groupe d'individus et on observe dans ce groupe la **fréquence** du caractère considéré.

Définition

Un **échantillon de taille n** est la liste de n résultats obtenus par n répétitions indépendantes d'une même expérience aléatoire.

Exemples

- On a compté le nombre d'élèves inscrits en seconde générale et technologique lors d'une année scolaire. Il y a 214 767 garçons et 209 542 filles.
  De l'ensemble de ces élèves, on sélectionne, au hasard, un groupe de 10 000 adolescents.
  On a ici prélevé un échantillon de taille 10 000 de l'ensemble des élèves de seconde (la population étudiée).
- On tire au sort le nom de 400 électeurs français ayant exprimé un choix lors de l'élection présidentielle de 2012. On constitue ainsi un échantillon de taille 400 de la population. Est-il certain, ou probable, que les électeurs de François Hollande sont majoritaires, comme ce fût le cas pour l'ensemble de la population ?
  C'est ce type de question qu'étudie l'échantillonnage.

Remarque

Pour réaliser une simulation d'un certain nombre d'échantillons de tailles n, on utiliser souvent le tableur et Python.

Exemple : à l'aide d'un tableur (début)

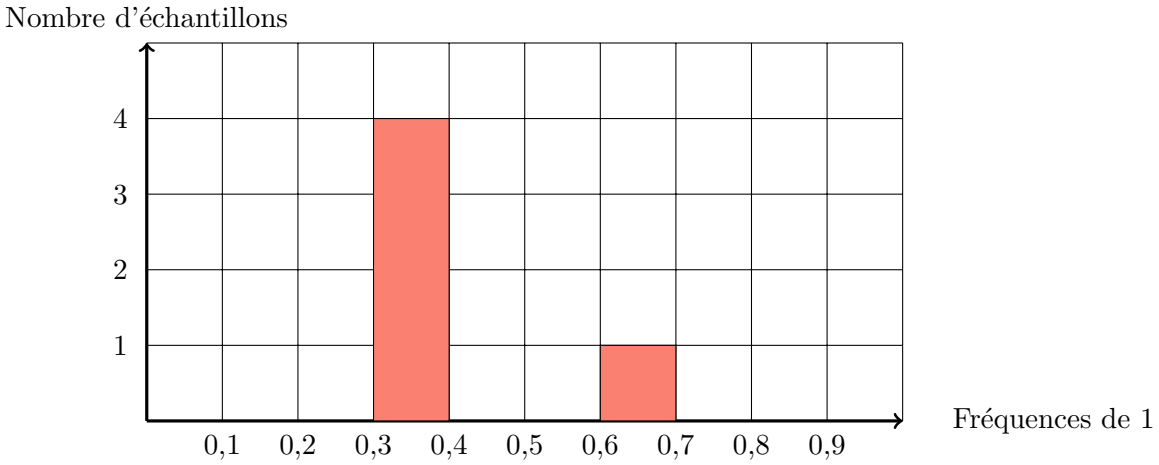
La fonction **ALEA()** du tableur permet de renvoyer un nombre aléatoire réel de l'intervalle [0 ; 1[. En ajoutant 0,3 on obtient un nombre de l'intervalle [0,3 ; 1,3[. La fonction **ENT** permet ensuite d'arrondir à l'entier inférieur. Ainsi, **ENT(ALEA()+0,3)** permet d'obtenir le 1 dans 30% des cas et 0 dans 70% des cas.

On va réaliser 5 échantillons de taille 10 de cette expérience aléatoire sur un tableau.
Voici l'extrait obtenu :

Exemple : à l'aide d'un tableur (fin)

	A	B	C	D	E
1	**Échantillon n°1**	**Échantillon n°2**	**Échantillon n°3**	**Échantillon n°4**	**Échantillon n°5**
2	1	0	1	0	0
3	0	0	0	0	0
4	1	1	0	0	0
5	1	0	0	0	0
6	0	1	1	1	1
7	0	0	0	0	0
8	0	0	1	0	0
9	0	0	1	0	0
10	0	0	1	1	1
11	0	1	1	1	1

Pour l'échantillon n°1, la fréquence de 1 est de 0,3. Pour l'échantillon n°2, elle est de 0,3. Pour l'échantillon n°3, elle est de 0,6. Pour l'échantillon n°4, elle est de 0,3. Enfin, pour l'échantillon n°5, elle est de 0,3. On obtient l'histogramme suivant :



Cet histogramme n'est pas très utile puisqu'il ne permet pas de mettre en avant certains résultats. En effet, il faudrait que le nombre d'échantillons réalisés soit beaucoup plus important et/ou que la taille de l'échantillon soit plus élevé.

Exemple : à l'aide de Python

On va créer un programme qui permet de lancer un dé à 6 faces. On considère le succès « Obtenir le 4 ». Cette expérience suit une loi de Bernoulli avec $p = \frac{1}{6} \approx 0,167$.

```
from random import*
def echantillon(n):
    L=[]
    for i in range (n):
        x=randint(1,6)
        L.append(x)
    print(L)
```

Si on souhaite un échantillon de taille 10 (donc 10 lancers de dé), voici ce que l'on obtient :

```
echantillon(10)
```

>>> [5, 6, 4, 1, 4, 3, 1, 4, 6, 5]

La fréquence de 4 est ainsi de 0,3. On est loin des 0,167 théoriques. Nous allons donc augmenter le nombre d'échantillons simulés ainsi que la taille de chacun des échantillons. Le programme va également compter le nombre de « 4 » obtenus et nous afficher la fréquence pour chaque échantillon.

```
from random import*
def echantillon(n):
    c=0
    for i in range (n):
        x=randint(1,6)
        if x==4:
            c=c+1
    return(c/n)

def PlusieursEchantillons(N,n):
    L=[]
    for i in range(N):
        L.append(echantillon(n))
    print(L)
```

Voici ce que l'on obtient pour 10 échantillons de taille 100 :

```
PlusieursEchantillons(10,100)
```

>>> [0.21, 0.17, 0.21, 0.15, 0.15, 0.12, 0.19, 0.17, 0.14, 0.17]

Là encore, il est possible de réaliser un histogramme de la situation.

Exemple : un nuage de points

On souhaite réaliser 400 échantillons de taille 50 du précédent exemple. Cela signifie que l'on va lancer 50 fois un dé et observer la fréquence de 4 obtenu. Et cela, 400 fois.
Python va ensuite afficher le nuage de points correspondant :

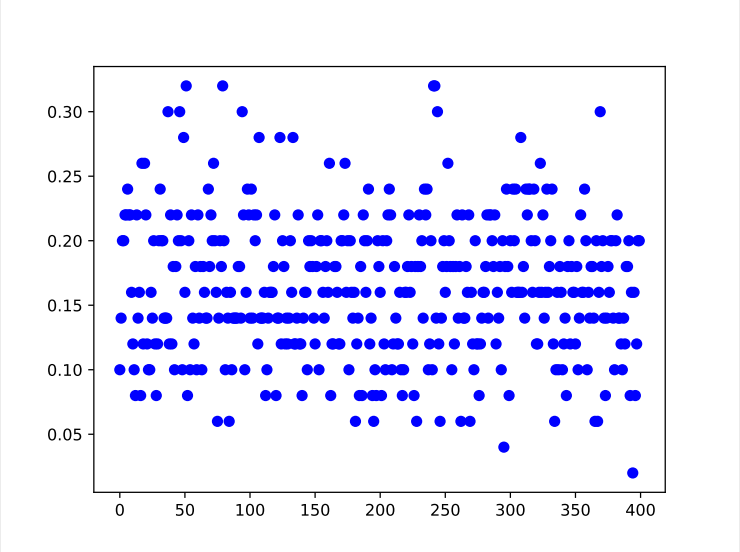
```
from random import*
import matplotlib.pyplot as plt

def echantillon(n):
    c=0
    for i in range (n):
        x=randint(1,6)
        if x==4:
            c=c+1
    return(c/n)

def PlusieursEchantillons(N,n):
    L=[]
    for i in range(N):
        L.append(echantillon(n))
    return(L)
A=PlusieursEchantillons(400,50)

def abscisse(N):
    B=[]
    for i in range(400):
        B.append(i)
    return(B)

plt.plot(abscisse(400),PlusieursEchantillons(400,50),"ob")
plt.show()
```



2 Étudier des échantillons

Définition : fluctuation d'échantillonnage

Soient N et n deux entiers naturels.
On considère N échantillons de taille n dont on étudie la fréquence d'apparition d'un caractère. Cette fréquence varie d'un échantillon à un autre : c'est ce que l'on appelle la **fluctuation d'échantillonnage**.

Exemple

D'après le nuage de points de l'exemple précédent, la fréquence de 4 varie. Elle fluctue entre 0,02 et 0,32 mais elle semble tout de même se « concentrer » dans l'intervalle [0,1 ; 0,2].

Remarque

Plus la taille des échantillons est grande, plus les fréquences observées vont se rapprocher de la fréquence théorique : la probabilité. Les fluctuations sont alors moins importantes.

Propriété

Soient N et n deux entiers naturels. On considère N échantillons de taille n dont la probabilité théorique est notée p. Soit $s = \dfrac{1}{2\sqrt{n}}$.

- En moyenne, environ 68% des fréquences sont dans l'intervalle $[p - s\,;\, p + s]$;
- En moyenne, environ 95% des fréquences sont dans l'intervalle $[p - 2s\,;\, p + 2s]$;
- En moyenne, environ 99% des fréquences sont dans l'intervalle $[p - 3s\,;\, p + 3s]$;

Remarques

- Le nombre s coïncide avec l'écart-type observé dans les fréquences obtenues.
- On peut aussi retenir $P(p - s \leqslant X \leqslant p + s) \approx 0,68$
  $P(p - 2s \leqslant X \leqslant p + 2s) \approx 0,95$
  $P(p - 3s \leqslant X \leqslant p + 3s) \approx 0,99$

Exemple

On reprend nos 400 échantillons de taille 50. La probabilité p théorique est de $\dfrac{1}{6} \approx 0,167$.

De plus, $s = \dfrac{1}{2\sqrt{n}} = \dfrac{1}{2\sqrt{50}} \approx 0,071$.

Ainsi, environ 68% des fréquences sont dans l'intervalle [0,096 ; 0,238] et environ 95% des fréquences sont dans l'intervalle [0,025 ; 0,309].