

L'échantillonnage

Table des matières

1 Vocabulaire et expérimentations en Python	1
2 Fluctuations et estimation, et la loi des grands nombres	4
2.1 Fluctuations	4
2.2 La loi des grands nombres	7
2.2.1 Énoncé de la loi des grands nombres	7
2.2.2 Illustration de la loi des grands nombres avec une expérience aléatoire simple.	7
2.2.3 Programme Python	9
2.2.4 Les résultats sur un graphique	11
2.3 L'estimation	13

1 Vocabulaire et expérimentations en Python

Définition 1: Échantillonnage

Lorsqu'on réalise une expérience aléatoire plusieurs fois de manière indépendante, la compilation des résultats représente ce que l'on appelle des échantillons.

Exemple 1

- Les sondages sont des formes d'échantillon, si on choisit d'interroger des personnes totalement au hasard, et non issues, par exemple du même quartier.
- Simuler une expérience aléatoire sur ordinateur permet d'effectuer des échantillons sur un grand nombre de données.
- Lancer cent fois un dé de 6 faces et noter les résultats obtenus est un échantillon. On s'attend à avoir autant de 1 que de 6, mais rien ne le garanti.

Question 1

Inventez, à partir d'une expérience aléatoire, un échantillon que l'on pourrait étudier.

Question 2

En réfléchissant le moins possible, donner cinq nombres entre 1 et 6. Pensez-vous avoir des échantillons semblables à ceux obtenus à partir d'un dé à 6 faces ?

Définition 2: Taille d'un échantillon

La taille d'un échantillon représente le nombre de fois où l'on a répété l'expérience aléatoire. On notera souvent n la taille de l'échantillon.

Exemple 2

On considère l'expérience aléatoire de lancer un dé à 6 faces. On lance ce dé exactement cent fois. On obtient un échantillon de l'expérience aléatoire «lancer un dé à 6 faces» de taille $n = 100$. Souvent, on notera n la taille de l'échantillon.

On peut programmer l'expérience précédente en Python avec quelques lignes :

```
1 from random import randint
2
3 n = 100
4
5 for i in range(n) :
6     print(randint(1, 6))
```

La fonction `randint` utilisée à la ligne 6 permet de générer un nombre aléatoire entier, ici entre 1 et 6.

Question 3

Modifier le programme précédent pour simuler cent lancers d'une pièce de monnaie.

Question 4

De quoi parle la page <https://arxiv.org/abs/2310.04153?>!

Reponse 1

La page en question mentionne une expérience (réelle !) d'un échantillon de 350 757 lancers de pièces de monnaie (oui, oui) pour vérifier combien de fois ces pièces retombaient sur «Pile». Un lancer de pièce est-il équilibré ?

Exemple 3

On peut aussi simuler une expérience aléatoire qui a une probabilité p de succès, avec $p \in [0; 1]$. Par exemple, si un élève à $p = 80\%$ de chance d'avoir la moyenne à ses contrôles de mathématique, et qu'il y a $n = 10$ contrôle cette année, on peut simuler son nombre de réussite à l'aide du programme 1

```
1 import random
2
3 p = 0.8
4 n = 10
5
6 controleReussi = 0
7
8 for i in range(n) :
9     if random.random() < p:
10         controleReussi = controleReussi + 1
11     print(controleReussi)
```

Listing 1: Combien de contrôle vais-je réussir ?

Question 5

Avez vous testé ce programme sur basthon ? Comment modifier ce programme pour simuler non pas dix mais cent contrôles ? Comment modifier ce programme pour montrer la **proportion** de contrôles réussis, et non le nombre de contrôle réussi (il s'agit de faire une division, mais de quoi par quoi?). Vers quoi cette proportion s'approche-t-elle lorsque n devient de plus en plus grand ? En un seul programme Python vous avez accès à l'intégralité des questions que l'on peut poser dans ce chapitre.

2 Fluctuations et estimation, et la loi des grands nombres

2.1 Fluctuations

Définition 3: Fluctuations

Lorsque l'on regarde **plusieurs** échantillons d'une seule expérience aléatoire, on observe une fluctuations des résultats. Cette fluctuations est d'autant **plus grande** que la taille de l'échantillon est **petite**.

Exemple 4

Prenons une simulation pour comprendre. Disons que la proportion de personnes portant des lunettes est de $p = 0,6$ au lycée. Si je choisis $n = 5$ élèves du lycée, combien je peux avoir de lycéens à lunettes ?

Pour le savoir, je peux effectuer **plusieurs** échantillon de taille $n = 5$. Par exemple, je peux former $m = 10$ échantillons de taille $n = 5$ et compter pour chacun combien je trouve de lycéen portant des lunettes.

```
1 import random
2
3 p = 0.6
4 n=5
5 m=10
6
7 nbPortantLunettes = 0
8 resultatsObtenus = []
9
10 for echantillon in range(m) : #on va former 10 échantillon
11     for lycee in range(n) : #de cinq lycéens
12         if random.random() < p :
13             nbPortantLunettes = nbPortantLunettes + 1
14     frequenceLunettes = nbPortantLunettes / n
15     resultatsObtenus.append(frequenceLunettes) #on ajoute le résultat
16     ↪ obtenu
17     nbPortantLunettes = 0 #on remet le compteur à 0
18 print(resultatsObtenus)
```

Listing 2: Illustration du phénomène de fluctuation

Ici, la ligne 8 permet de créer une liste qui va stocker nos résultats. La fonction `append` utilisée à la ligne 15 permet d'ajouter la **fréquence** de lycée portant des lunettes **au sein de l'échantillon** dans la liste `resultatsObtenus`.

Vous noterez que la variable `frequenceLunettes` définies à la ligne 14 n'est pas obligatoire, mais sert à comprendre le programme. On compte le nombre de lycéen qui porte des lunettes, et on divise par n le nombre de lycée de chaque échantillon.

Faites bien attention aux nombres d'espaces des lignes 15 et 16.

Le programme donné dans le listing ?? donne, par exemple, le résultat suivant :

```
[0.2, 0.4, 0.6, 0.8, 0.4, 0.6, 0.8, 0.6, 0.4, 0.8]
```

Exemple 5

Vous pouvez constater une **forte** fluctuation.

Question 6

Aller, on code tout ça en python, et on regarde ce que cela donne en vrai ! Que se passe-t-il si on **augmente la taille de l'échantillon** ? Par exemple avec $n = 100$?

Question 7

Pourquoi une «grande» taille d'échantillon diminue les fluctuations ?

Exemple 6

Les exemples qui suivent illustrent ce qui peut se passer selon la taille choisie des échantillons. Vous verrez la fluctuation baisser au fur et à mesure que la taille des échantillons augmentent. Nous verrons un exemple de ce phénomène juste après.

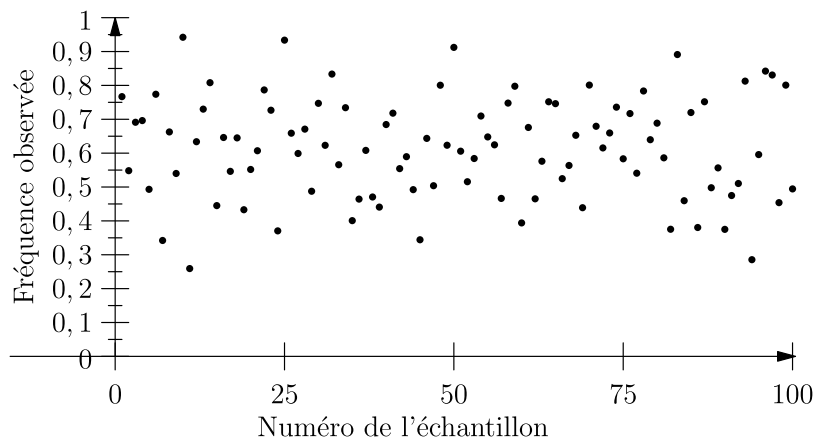


Figure 1 – Voilà ce qui se passe quand la taille de l'échantillon est trop petite..

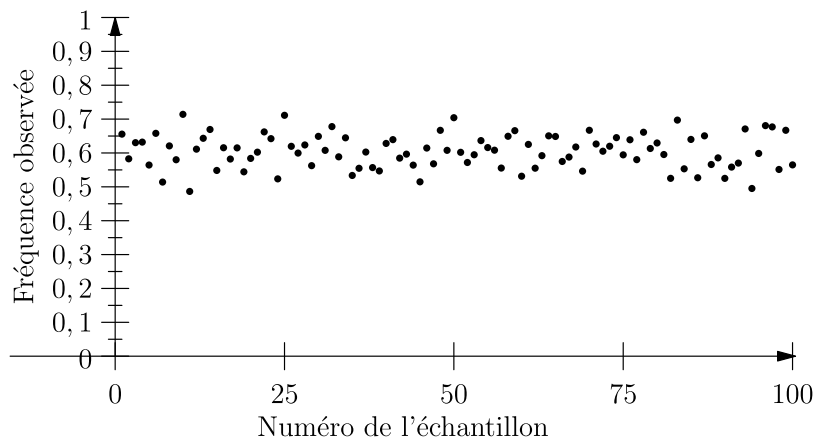


Figure 2 – Voilà ce qui se passe quand la taille de l'échantillon plus grande.

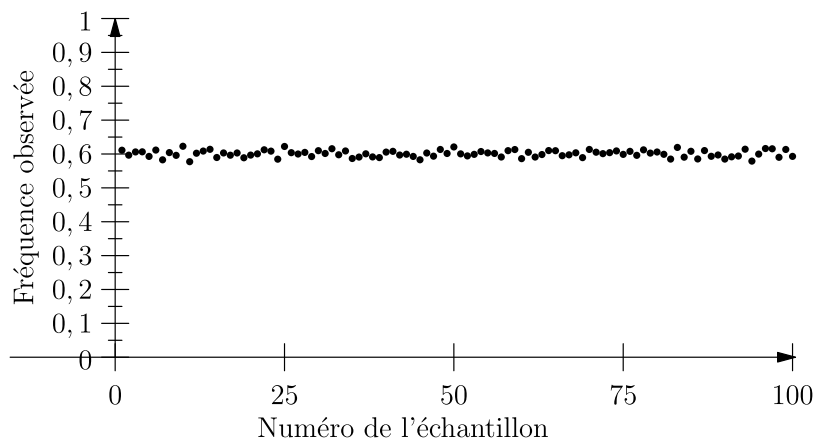


Figure 3 – Et voilà ce qui se passe pour une très grande taille d'échantillons.

2.2 La loi des grands nombres

2.2.1 Énoncé de la loi des grands nombres

Proposition 1: La loi des grands nombres

Mathématiquement, on peut quantifier les fluctuations montrées à la section précédente. C'est-à-dire que l'ordre de grandeur de la fluctuation est donné par la formule $\frac{1}{\sqrt{n}}$. Le véritable énoncé mathématique de la loi des grands nombres est compliqué. Voici une version plus simple d'accès.

Si on note f la fréquence observée au sein d'un échantillon, et p la probabilité de l'expérience aléatoire répétée, alors p a une grande chance d'appartenir à l'intervalle

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

On peut le formuler avec une double inégalité (merci le chapitre sur les intervalles) :

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

C'est comme si on «emprisonnait» la probabilité que l'on cherchait entre deux bornes qui sont de plus en plus proche à mesure que n est grand.

2.2.2 Illustration de la loi des grands nombres avec une expérience aléatoire simple.

On va considérer l'expérience aléatoire d'un lancer de dé équilibré à 6 faces, et on va s'intéresser à l'obtention d'un six.

On appellera «succès» le fait d'obtenir un six.

Question 8

Quelle est la probabilité d'avoir un succès ?

Reponse 2

Il y a 6 faces, et le dé est supposé équilibré. Donc, nous avons une chance sur 6 d'obtenir un six. Ainsi, la probabilité p recherchée est :

$$p = \frac{1}{6}$$

On considère des échantillons de taille n . On va faire évoluer n pour voir les fluctuations. Dans cet échantillon de taille n , on observe une certaine fréquence de succès f .

Exemple 7

Si je fixe la taille de l'échantillon à $n = 5$. On peut obtenir par exemple la série suivante :

$$6, 1, 4, 2, 6$$

La fréquence observée est alors $f = \frac{2}{5} = 40\%$, car on observe 2 succès parmi l'échantillon de taille 5.

Maintenant, on va multiplier le nombre d'échantillon, puis calculer f_m la fréquence moyenne des fréquences observée.

Exemple 8

En continuant sur des échantillons de taille $n = 5$, j'obtiens une deuxième série qui pourrait être la suivante :

$$3, 3, 6, 5, 2$$

On obtient une fréquence $f = \frac{1}{5} = 20\%$. Pour l'instant, notre fréquence moyenne observée f_m vaut :

$$f_m = \frac{\frac{2}{5} + \frac{1}{5}}{2} = \frac{\frac{3}{5}}{2} = \frac{3}{10} = 30\%$$

Il faudrait accumuler d'avantage d'échantillon pour espérer retrouver la valeur de $p = \frac{1}{6} \approx 16,67\%$.

On récapitule :

- n est la taille des échantillons
- p est la probabilité de succès de l'expérience aléatoire, que l'on connaît ici, mais qui parfois est à estimer (c'est l'objectif des statistiques)
- f est la fréquence de succès observée dans un échantillon.
- f_m est la fréquence moyenne des fréquence de chaque échantillon.

Maintenant que l'on a tout ça en tête, on peut simuler cette expérience en Python.

2.2.3 Programme Python

```

1  from random import randint
2  def lancerDes() :
3      return randint(1, 6)
4
5  n = 5
6
7  nombreEchantillon = 10
8
9  succes = 0
10
11 listeFrequenceObserve = []
12
13 for echantillon in range(nombreEchantillon) :
14     for experience in range(n) :
15         if lancerDes() == 6 :
16             succes = succes + 1
17         f = succes / n
18         listeFrequenceObserve.append(f)
19         succes = 0
20
21 print(listeFrequenceObserve)

```

Question 9

Comment a-t-on calculer la fréquence observée à la ligne 17 ?
Expliquer le rôle de la ligne 16 (tester le programme sans pour voir ce qui se passe).

J'obtiens les fréquences observées suivantes après exécution du programme :

[0.2, 0.0, 0.0, 0.0, 0.0, 0.2, 0.4, 0.0, 0.4, 0.2]

Vous remarquez que parfois je n'ai obtenu aucun 6 (cela m'est arrivé 5 fois, pas de chance !).

Faites tourner vous aussi le programme **plusieurs fois**, en comprenant ce qui est affiché.

Question 10

Ajouter une fonction qui permet de calculer la fréquence moyenne f_m à partir de la liste `listeFrequenceObserve`

Je vous laisse le temps de réfléchir, le résultat est à la page d'après !

Reponse 3

On peut coder la fonction moyenne comme le montre le code suivant. Vous pouvez placer cette définition en dessus de la première ligne de votre code.

```
def moyenne(liste) :  
    somme = 0  
    tailleListe = len(liste)  
  
    for element in liste :  
        somme = somme + element  
    return somme/tailleListe
```

Ensuite, il faut appliquer cette fonction sur notre liste, ce que l'on peut faire avec :

```
print(moyenne(listeFrequenceObserve))
```

Il existe aussi des méthodes très classe en python qui permettent de faire la même chose en une seule ligne de code. Voici un exemple :

```
def moyenne(liste):  
    return sum(liste) / len(liste)
```

Question 11

Qu'est-ce qu'on peut dire de la fréquence observée f_m sur les échantillons de taille $n = 5$? S'approche-t-on de $p = \frac{1}{6}$?

2.2.4 Les résultats sur un graphique

Voici ce que l'on obtient lorsque l'on représente la même chose sur un graphique.

- La droite en rouge représente la valeur de $p = \frac{1}{6}$
- La droite en bleue représente la fréquence observée moyenne f_m

Question 12

Combien y'a-t-il eu d'échantillon dans le graphique précédent ?

Question 13

Qu'est-ce que l'on constate sur ces graphes ?

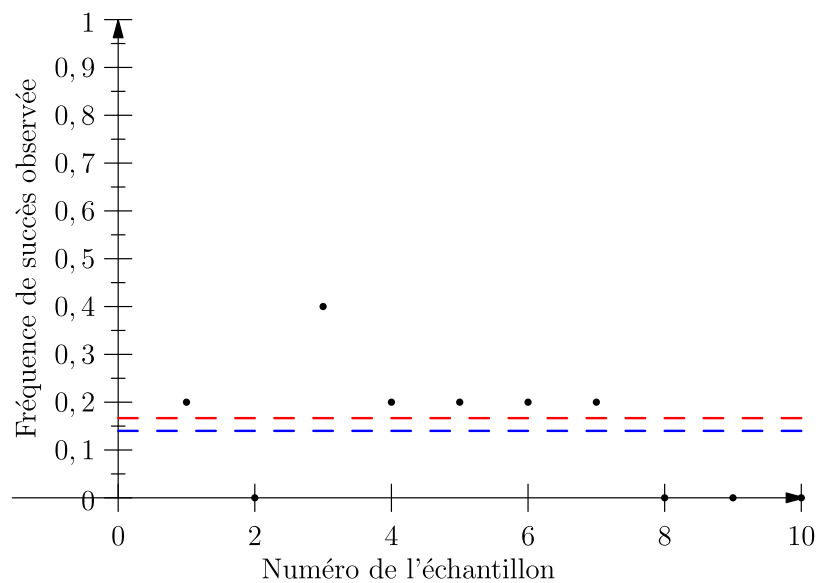


Figure 4 – Une taille d'échantillon $n = 5$

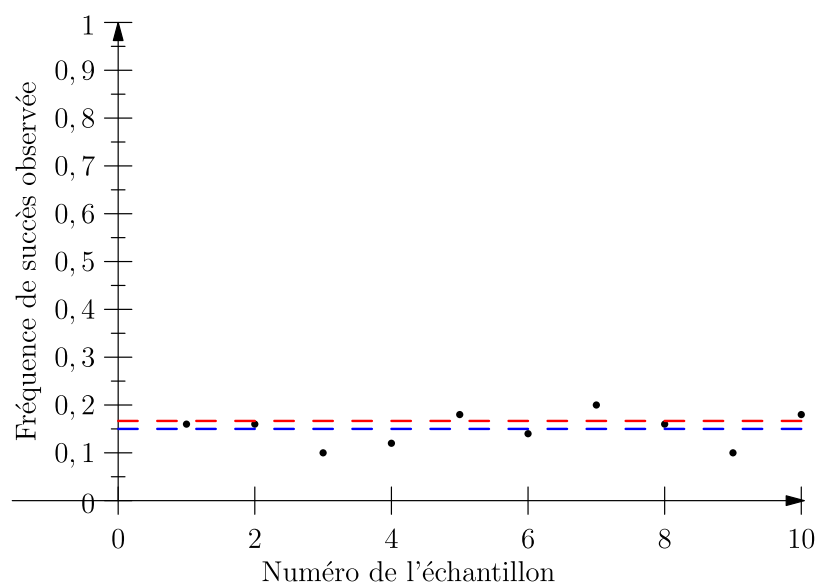


Figure 5 – Une taille d'échantillon $n = 50$

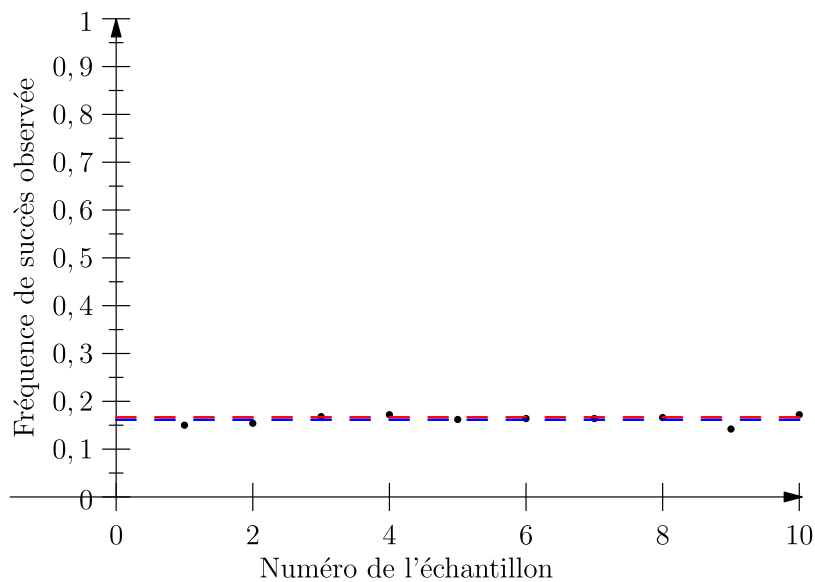


Figure 6 – Une taille d'échantillon $n = 500$

Reponse 4

On constate que la fréquence observée se stabilise autour de la valeur p .

2.3 L'estimation

Définition 4: L'estimation

On se sert de l'échantillonnage pour **estimer** la probabilité p d'une expérience aléatoire. Le phénomène de fluctuation vu précédemment nous incite à avoir une taille d'échantillon **grande** pour s'approcher le plus fidèlement possible de la probabilité p .

Proposition 2

D'après la loi des grands nombres citées plus haut, si on cherche à obtenir p au centième près (c'est-à-dire au pourcent près), il faut une taille d'échantillon n telle que :

$$\frac{1}{\sqrt{n}} < 0.005$$

C'est-à-dire (puisque la fonction inverse est décroissante sur $]0; +\infty[$:

$$\sqrt{n} > \frac{1}{0.005}$$

$$\sqrt{n} > 200$$

$$n > 200^2$$

$$n > 40000$$

Et ce n'est pas une bonne nouvelle ! Cela veut dire que pour obtenir une estimation à pourcent près il faudrait une taille d'échantillon de $n = 40\,000$!

Question 14

Pourquoi 0.05 et pas 0.1 dans la proposition précédente ?

Question 15

Justifier chaque étape de la résolution de l'inéquation de la proposition précédente.

Exemple 9

En pratique le vrai théorème de la loi des grands nombres est un peu plus précis, et on estime qu'une taille «suffisante» d'un échantillon se situe autour de $n = 1000$.

Question 16

Vous êtes en charge d'une étude scientifique de médecine, et vous étudier la propagation d'une maladie rare. Vous surveillez la propagation de la maladie, en essayant de déterminer la probabilité qu'une personne contamine une autre. Expliquez, à l'aide de la loi des grands nombres, pourquoi cela va prendre **beaucoup** de temps.

Question 17

Deux candidats s'affrontent pour la présidentielle. La veille de l'élection, un sondage sur $n = 70$ personnes est organisé. 48% des personnes sondées déclarent voter pour le candidat A, et toutes les personnes interrogées ont voté pour un des deux candidats.

Peut-on assurer la victoire du candidat B? Justifier en utilisant la loi des grands nombres.