L'échantillonnange

Table des matières

1 Vocabulaire et expérimentations en Python

Définition 1: Échantillonage

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

```python
import random

p = 0.8
n = 10

controleReussi = 0

for i in range(n) :
    if random.random() < p:
        controleReussi = controleReussi + 1
        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

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 obtenu
16     nbPortantLunettes = 0  #on remet le compteur à 0
17
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 :

```
[0.6, 0.6, 0.6, 0.4, 0.6, 0.6, 0.8, 0.8, 0.8, 0.0]
```

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 ?

Proposition 1

Mathématiquement, on peut quantifier la fluctuations. C'est-a-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 à une grande chance d'appartenir à l'intervalle $[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}]$

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 $]0; +\infty[$:

$$\begin{aligned}\sqrt{n} &> \frac{1}{0.005}\\ \sqrt{n} &> 200\\ n &> 200^2\\ n &> 40000\end{aligned}$$

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

Question 8

Pourquoi 0.05 et pas 0.1 dans la proposition précédente ?

Question 9

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

Exemple 6

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 10

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 11

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.