L'échantillonnage

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. De plus, pour que l'échantillon soit représentatif, il faut absolument varier la provenance des personnes interrogées.
- 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 ?

À la ligne 9, on utilise la fonction `random.random()` qui retourne un nombre aléatoire entre 0 et 1. Voici par exemple ce que donne le code suivant :

```
import random

print(random.random())
```

```
0.5163334704714878
```

Question 5

Avez vous testé le programme 1 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

2.1 Le phénomène de fluctuation

Définition 3: Fluctuations

Lorsque l'on regarde **plusieurs** échantillons d'une seule expérience aléatoire, on observe une **fluctuation** des résultats. Cette fluctuation 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.

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 précédent donne, par exemple, le résultat suivant :

```
[0.4, 0.6, 1.0, 0.8, 0.6, 0.8, 0.6, 0.6, 0.6, 0.6]
```

```python
import random

p = 0.6
n=5
m=10

nbPortantLunettes = 0
resultatsObtenus = []

for echantillon in range(m) : #on va former 10 échantillon
    for lycee in range(n) : #de cinq lycéens
        if random.random() < p :
            nbPortantLunettes = nbPortantLunettes + 1
    frequenceLunettes = nbPortantLunettes / n
    resultatsObtenus.append(frequenceLunettes) #on ajoute le résultat obtenu
    nbPortantLunettes = 0  #on remet le compteur à 0

print(resultatsObtenus)
```

Listing 2: Illustration du phénomène de fluctuation

Exemple 5

Vous pouvez constater une **forte** fluctuation. Les fréquences observées sont très variées !

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.

2.2 Illustration du phénonène de fluctuation

On a construit des graphiques qui montre la fréquence observée dans plusieurs échantillons avec :

— en abscisse les différents échantillons

— en ordonnées la fréquence observées dans chaque échantillon.

Chaque point dans les graphique qui suivent représentent donc un échantillon. Sa hauteur indique la fréquence observée dans cet échantillon.

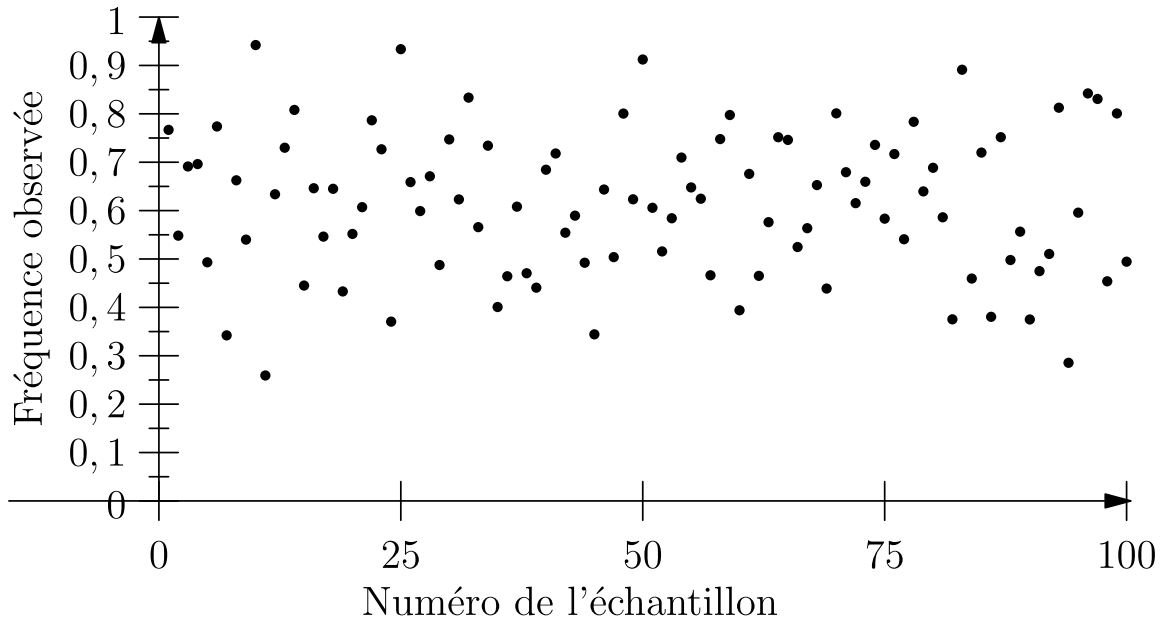

Figure 1 – Voilà ce qui se passe quand la taille de l'échantillon est trop petite..

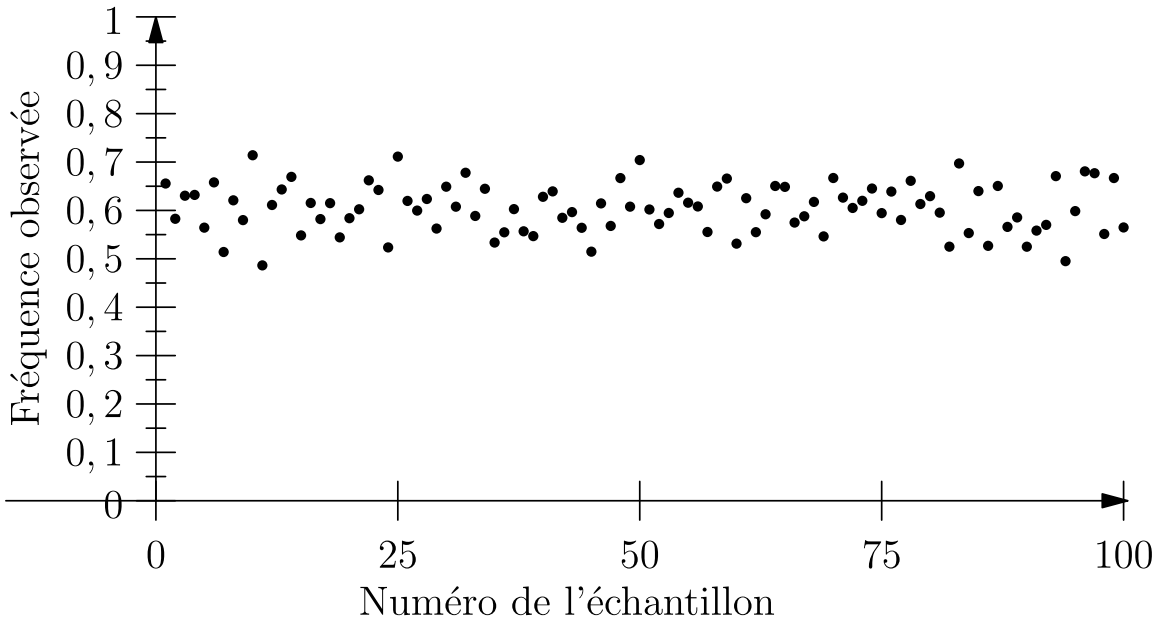

Figure 2 – Voilà ce qui se passe quand la taille de l'échantillon plus grande.

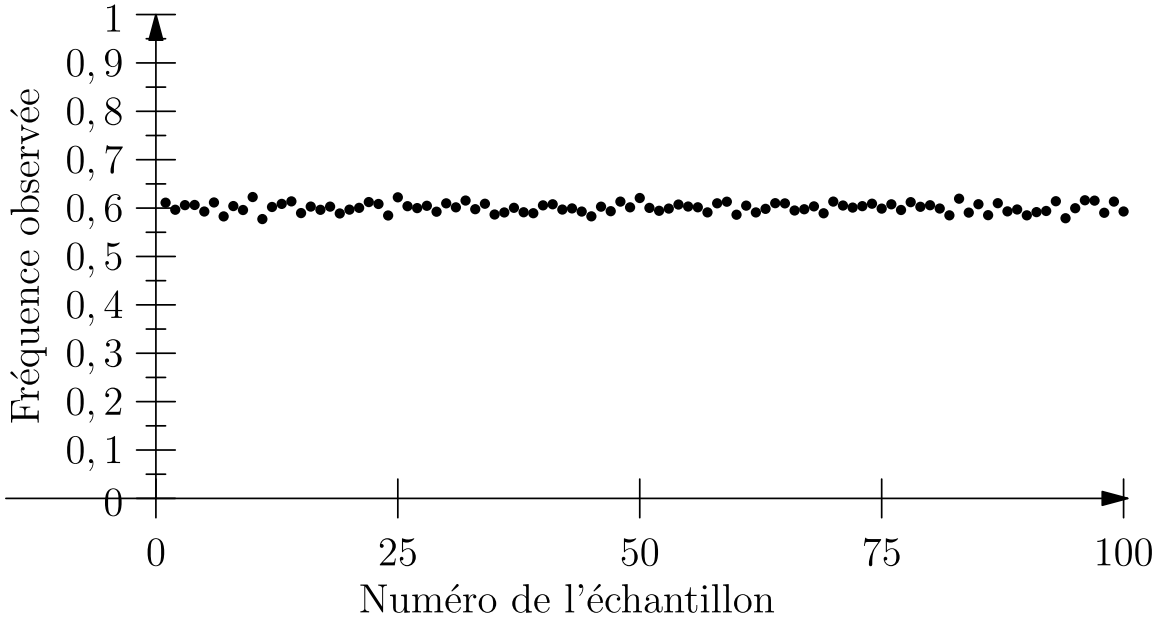

Figure 3 – Et voilà ce qui se passe pour une très grande taille d'échantillons.

Question 8

Dans la figure 1, et par lecture graphique :

1. quelle est la fréquence observée du premier échantillon ?
2. la fréquence observée du dernier échantillon ?
3. Quelle est la fréquence minimale observée ?
4. Quelle est la fréquence maximale observée ?

Qu'est-ce qui a été modifié entre chaque graphique ? Que remarque-t-on ? Prenez le temps de **rédiger** votre réponse, pour vérifier que vous maîtrisé le vocabulaire de chapitre.

3 La loi des grands nombres : énoncés et illustration

Proposition 1: La loi des grands nombres, version simplifiée

Mathématiquement, on peut quantifier les fluctuations montrées à la section précédente. 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. J'ai découpé l'énoncé en plusieurs «points de vue» qui seront tous utilisés par la suite. C'est bien le même théorème mathématique, mais qui est exploité selon deux angles différents.

Proposition 2: Premier point de vue : comment f s'approche de p ?

f s'approche de p au fur et à mesure que la taille de l'échantillon augmente. 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

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

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.

Proposition 3: Deuxième point de vue : comment f est coincé par la valeur de p et n ?

f ne peut pas être très loin de p.
Si on note f la fréquence observée au sein d'un échantillon, et p la probabilité de l'expérience aléatoire répétée, alors f à une grande chance d'appartenir à l'intervalle

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

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

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

C'est comme si f ne pouvait pas aller très loin de p.

Question 9

Qu'est-ce qui a changé selon les deux points de vue ?

Proposition 4

(pour les plus curieux) Pourquoi deux point de vue avec le même théorème ? Et bien c'est à cause de cette inégalité [a] :

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

Où $|\cdot|$ désigne la valeur absolue, autrement dit, ici $|f - p|$ mesure la distance qu'il y a entre f et p, et ce que dit la loi des grands nombres c'est qu'il y a des grandes chances pour que cette distance soit plus petite que $\frac{1}{\sqrt{n}}$. Donc, soit on part de f pour dire que p n'est pas très loin, soit à l'opposé on part de p pour dire que f ne sera pas très loin. Dans le premier cas, vous essayer **d'estimer** la valeur de p à partir de la valeur des fréquences observées f (comme pour un sondage). Alors que dans le second cas vous contrôlé les **fluctuations** de f autour de p (comme pour l'expérience sur les dés que j'illustre plus loin).

a. Cette inégalité est fausse en générale (voir les points rouges de l'illustration sur le dé aux sections suivantes). Elle a juste *des grandes chances d'être vraie*.

Question 10

Les mots en gras de la proposition précédente devraient vous rappeler le titre de la section. Non ?

3.1 Étude d'une expérience aléatoire simple – obtenir un six sur un dé

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 11

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 $f = \frac{1}{5} = 20\%$.

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 succés observée dans un échantillon.

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

3.2 Simulation à l'aide d'un programme Python

```
from random import randint
def lancerDes() :
    return randint(1, 6)

n = 5

nombreEchantillon = 10

succes = 0

listeFrequenceObserve = []

for echantillon in range(nombreEchantillon) :
    for experience in range(n) :
      if lancerDes() == 6 :
          succes = succes + 1
    f = succes / n
    listeFrequenceObserve.append(f)
    succes = 0

print(listeFrequenceObserve)
```

Question 12

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 13

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 14

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

3.3 Vérification du deuxième point de vue sur des graphique

Voici ce que l'on obtient lorsque l'on représente la même chose sur un graphique. Nous allons appeler le nombre d'échantillons regardé m, et on va prendre $m = 1000$, parce que pourquoi pas. Ce qui nous intéresse, c'est de vérifier que la loi des grands nombres fonctionne, c'est-à-dire que dans la plupart des cas, la fréquence f observée est compris entre $p - \frac{1}{\sqrt{n}}$ et $p + \frac{1}{\sqrt{n}}$.
Pour cela :

— On a représenté la «bande» issues de la loi des grands nombres, qui nous dit que la fréquence observée à des grandes chances d'etre comprise entre $p - \frac{1}{\sqrt{n}}$ et $p + \frac{1}{\sqrt{n}}$.

— Les points en **vert** représentent les échantillons pour lesquels la fréquence observée est effectivement dans la bande prévue par la loi des grands nombres.

— Les points en **rouge** au contraire représente les points qui sont hors de la bande.

— La droite pointillée en rouge représente l'axe $y = p$ autour de laquelle les fréquences observées fluctuent.

— On a fait varier la taille de l'échantillon (c'est-à-dire n ici), et à chaque fois on a regardé $m = 1000$ échantillons différents. Oui, certains calculs ont pris une poignée de secondes pour être affiché dans cette page !

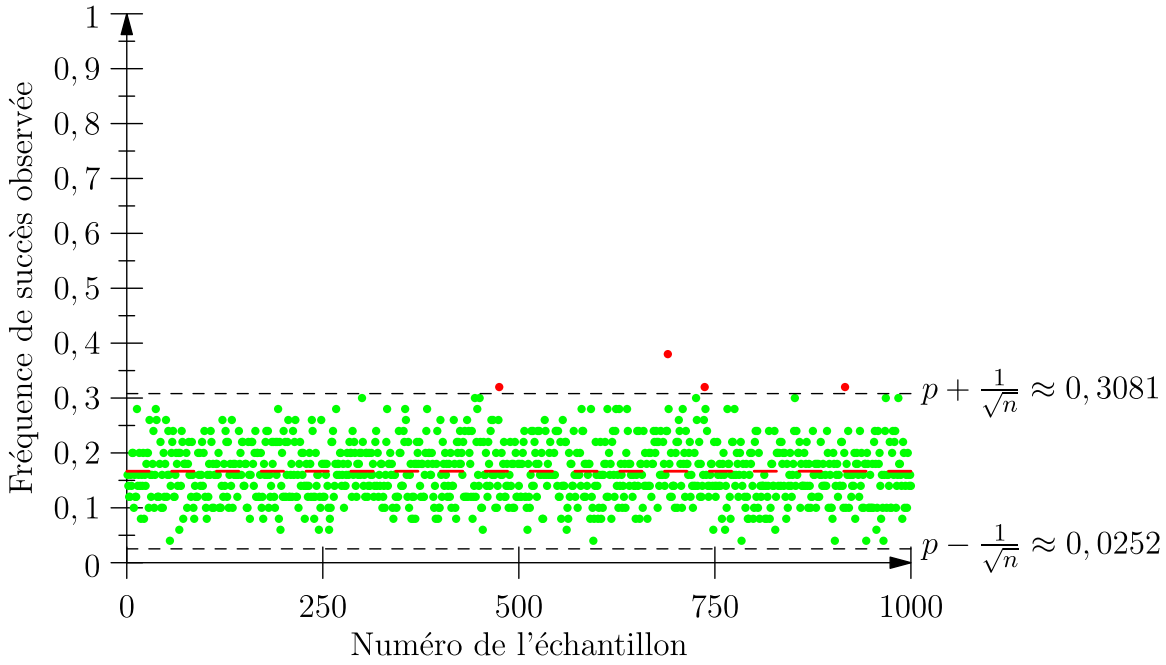

Figure 4 – Une taille d'échantillon $n = 50$

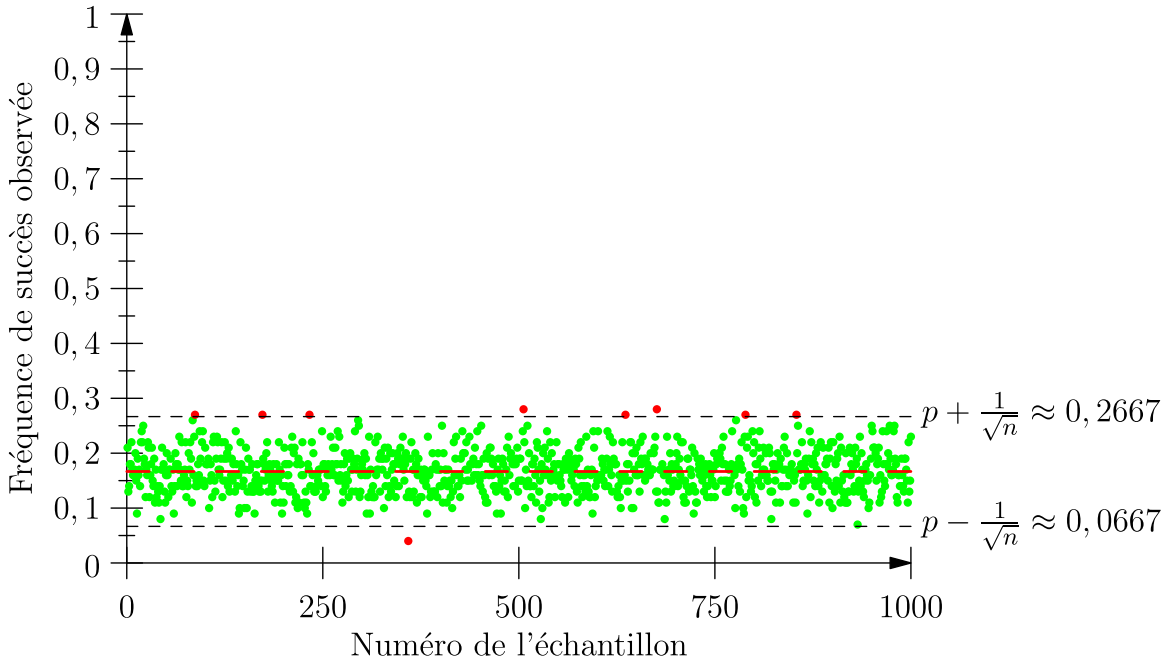

Figure 5 – Une taille d'échantillon $n = 100$

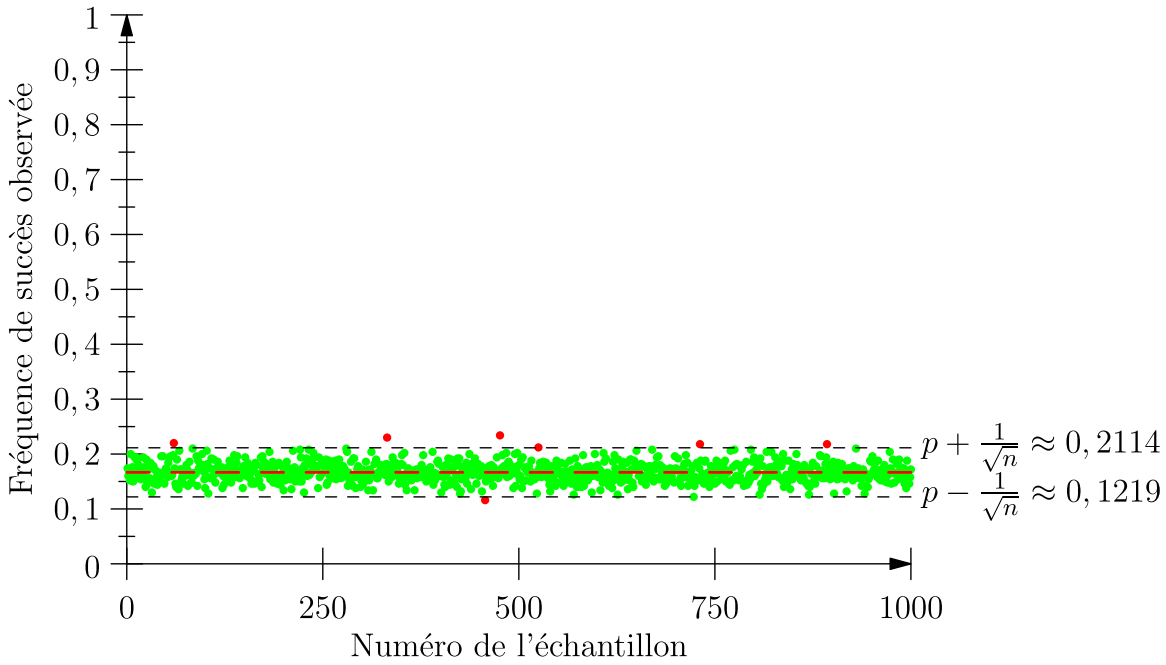

Figure 6 – Une taille d'échantillon $n = 1000$

Question 15

Regardez attentivement ces graphes, et répondez pour chacun d'eux aux questions suivantes :

1. Que représente l'axe des abscisses ?
2. Que représente l'axe des ordonnées ?
3. Combien y'a-t-il eu d'échantillons simulés ?
4. Combien de lancer de dés ont été réalisés par l'ordinateur pour afficher ce graphique ?
5. En proportion, combien il y a de points rouges ? Même question pour les points verts.
6. Quelle est la valeur de p ?
7. Pourquoi il y a deux «bandes» autour de p ?
8. Retrouver les valeurs de $p + \frac{1}{\sqrt{n}}$ et $p - \frac{1}{\sqrt{n}}$ sur votre calculatrice.

Et voici des questions qui portent sur l'ensemble des graphiques.

1. Quel est le seul paramètre parmi n, m, p, f que l'on a changé entre chaque graphique ?
2. Que peut-on dire de la «bande» d'emprisonnement entre chaque graphe ?
3. Que représente les points rouges ? Et les points verts ? Essayer **d'écrire** une phrase complète pour que vous maîtrisiez le plus possible les concepts du chapitre.
4. Vers quelle valeur les expressions $p + \frac{1}{\sqrt{n}}$ et $p - \frac{1}{\sqrt{n}}$ vont s'approcher lors que l'on va augmenter la taille de l'échantillon ?

4 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 5

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 16

Pourquoi 0.05 et pas 0.1 dans la proposition précédente ?

Question 17

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 18

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 19

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.