L’échantillonnage

Table des matières

1 Introduction

1.1 Vocabulaire

Définition 1: Résultat aléatoire

Un résultat est aléatoire s'il est impossible de le déterminer à l'avance avec certitude.

Exemple 1

- Le résultat d'un lancer de dé est aléatoire.
- La taille que vous aurez adulte est aléatoire.
- La quantité de pluie récoltée sur un mètre carré au lycée demain est un résultat aléatoire.

Définition 2: Expérience aléatoire

Une expérience aléatoire est une expérience qui admet un résultat aléatoire.

Exemple 2

Lancer un dé est une expérience aléatoire.

Définition 3: 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 3

- 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 4: 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.

1.2 Détour «culturel»

Question 3

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 ?

2 Exemples en Python

2.1 Simuler le hasard en Python

Pour simuler le hasard en python, il faut commencer par la ligne `import random` dans votre programme.
Puis, vous pouvez utiliser la fonction `random.ramdom()`, qui retourne un nombre aléatoire compris entre 0 et 1. Voici un exemple :

```
import random

print(random.random()) #la fonction print permet d'afficher le nombre
↪  aléatoire
```

```
0.5163334704714878
```

Question 4

Testez ce code, et relancez-le plusieurs fois. Qu'observe-t-on ?

Vous pouvez aussi utiliser la fonction `random.randint` qui permet de tirer un nombre entier entre deux nombres donnés. Par exemple :

```
import random

print(random.randint(1, 36))
```

Donne comme résultat :

```
27
```

Question 5

Testez ce code, et relancez-le plusieurs fois. Qu'observe-t-on ?

2.2 Utiliser Python pour simuler des expériences aléatoires.

Exemple 4

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

Listing 1: Simulation de 100 lancers de dés à 6 faces équilibrées.

Vous reconnaissez la fonction `random.randint` expliquée à la section précédente à la ligne 6.

Question 6

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

Exemple 5

Un élève à $p = 80\%$ de chance d'avoir la moyenne à ses contrôles de mathématique. Il y a $n = 10$ contrôles cette année. On peut simuler son nombre de réussites à l'aide de Python. Regardez le programme 2.

```
1  import random
2
3  p = 0.8 #probabilité de réussite d'un contrôle
4  n = 10 #nombres de controles
5
6  controleReussi = 0 #variable qui va compter le nombre de contrôles réussis
7
8  for i in range(n) :
9      if random.random() < p:
10         controleReussi = controleReussi + 1 #cette variable augmente SI la
           ↪ condition est réalisée
11 print(controleReussi)
```

Listing 2: 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, comme expliqué à la section précédente. Le mot clé `if` permet ici de tester si le nombre tiré au hasard est inférieure à $p = 0{,}8$. Ainsi, dans 80% des cas, cette condition sera réalisée, et on pourra à la ligne 10 augmenter de 1 le nombre de contrôles réussis !

Question 7

Testez le programme 2 sur basthon. Comment modifier ce programme pour simuler non pas 10 mais 100 contrôles ?

Question 8

1. Comment modifier ce programme pour montrer la **proportion** de contrôles réussis, et non le nombre de contrôles réussis ?
2. Vers quoi cette proportion s'approche-t-elle lorsque n devient de plus en plus grand ?

3 Fluctuations

3.1 Le phénomène de fluctuation

Définition 5: Fluctuations

Lorsque l'on regarde **plusieurs** échantillons d'une même expérience aléatoire, on observe une **fluctuation** des résultats. Le nombre de réussite obtenue varie aléatoirement entre chaque échantillon.

Exemple 6

Dans un lycée fictif, la proportion de personnes portant des lunettes est de $p = 0{,}6 = 60\%$. On forme des échantillons de taille $n = 5$ élèves. Combien peut-il y avoir de lycéens qui portent des lunettes dans un échantillon ?
Pour le savoir, je peux effectuer **plusieurs** échantillons de taille $n = 5$. Je peux former $m = 10$ échantillons de taille $n = 5$ (on a ainsi interrogé 50 lycéens). Dans chaque échantillon, on compte le nombre de lycéens qui portent des lunettes. Il peut y en avoir 0, 1, 2, 3, 4 ou 5.
La simulation en python de cette expérience est donné au programme 3.

```python
import random

p = 0.6
n=5
m=10

nbPortantLunettes = 0
resultatsObtenus = []

for echantillon in range(m) : #on va former 10 échantillons
    for lycee in range(n) : #de cinq lycéens
        if random.random() < p : #si un nombre au hasard est inférieur à p
            nbPortantLunettes = nbPortantLunettes + 1 #on ajoute de 1 le nombre de lycéen portant des lunettes
    frequenceLunettes = nbPortantLunettes / n
    resultatsObtenus.append(frequenceLunettes) #on ajoute le résultat obtenu
    nbPortantLunettes = 0  #on remet le compteur à 0

print(resultatsObtenus)
```

Listing 3: Illustration du phénomène de fluctuation

Plusieurs choses à noter dans ce programme :

1. La ligne 8 permet de créer une liste, appelées `resultatsObtenus`, qui va stocker nos résultats. À la ligne 8, la liste est encore vide.
2. La fonction `append` utilisée à la ligne 15 permet d'ajouter la **fréquence observée** de lycéens portant des lunettes dans la liste `resultatsObtenus`.
3. La ligne 13 permet d'augmenter de 1 la valeur de la variable `nbPortantLunettes`. On retrouve cette technique dans beaucoup de programme.
4. 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.
5. Faites bien attention aux nombres d'espaces des lignes 15 et 16.

L'exécution du programme m'a donné le résultat suivant :

```
[0.4, 0.6, 0.8, 0.8, 1.0, 0.6, 0.6, 0.2, 0.6, 0.4]
```

Exemple 7

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

Question 9

Utilisez basthon pour reproduire ce programme. Prenez le temps de le comprendre.

Question 10

Une fois que vous avez compris ce programme, changez la valeur de n. Que remarquez vous ?

Reponse 2

Vous devriez avoir observé en modifiant le programme 3 la proposition suivante :

Proposition 1

La **fluctuation** est un phénomène d'autant **plus important** que la taille des échantillons est **petite**.

3.2 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.

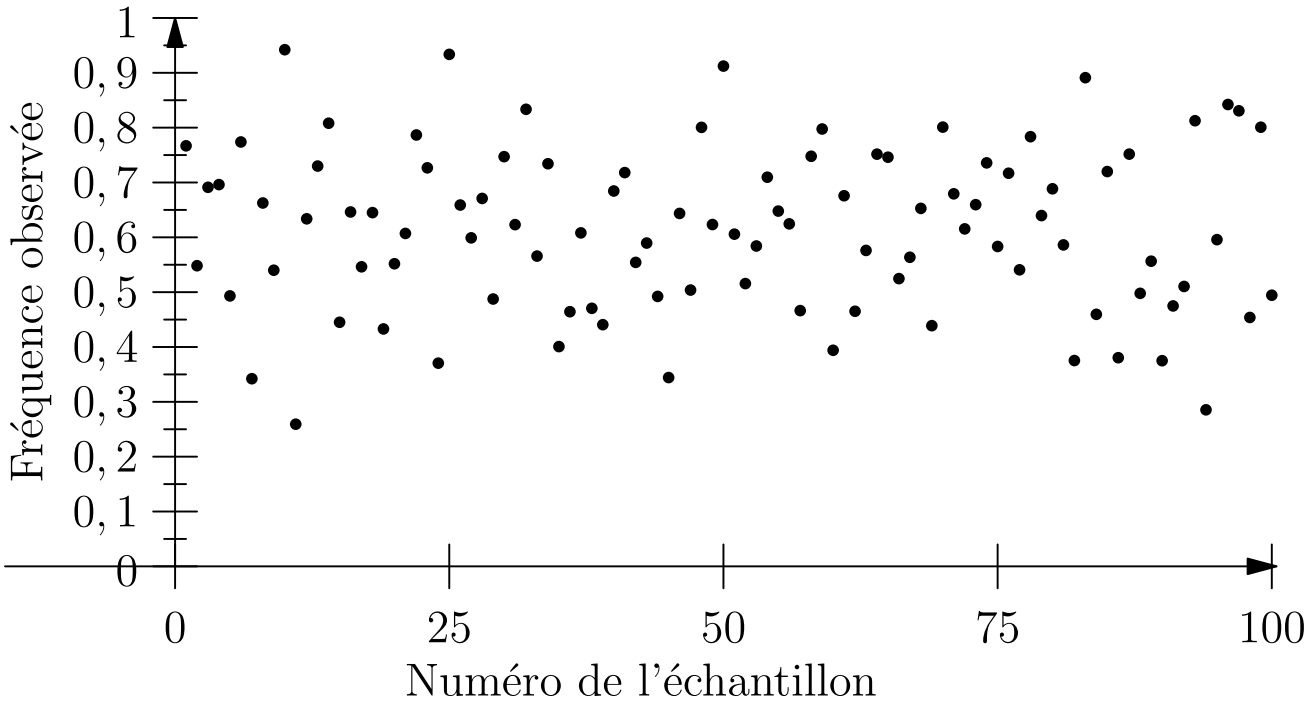

Figure 3.1 – Voilà ce qui se passe quand la taille de l'échantillon est trop petite..

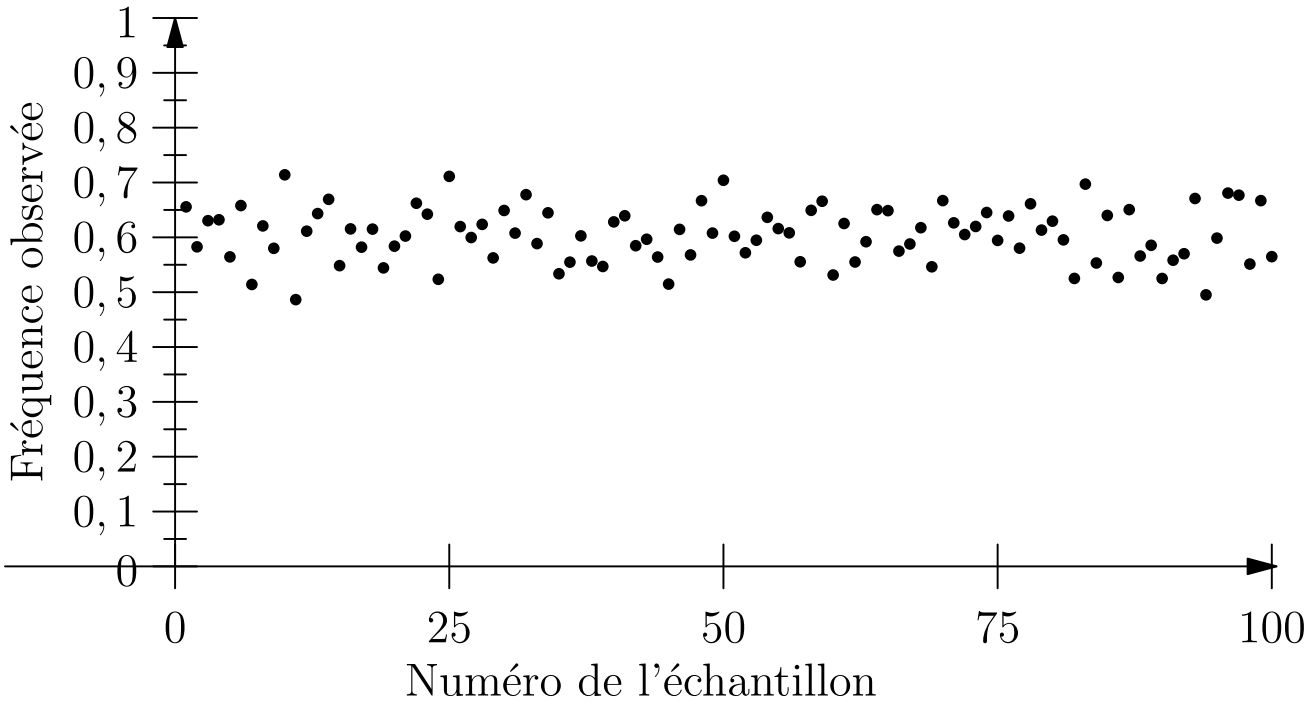

Figure 3.2 – Voilà ce qui se passe quand la taille de l'échantillon plus grande.

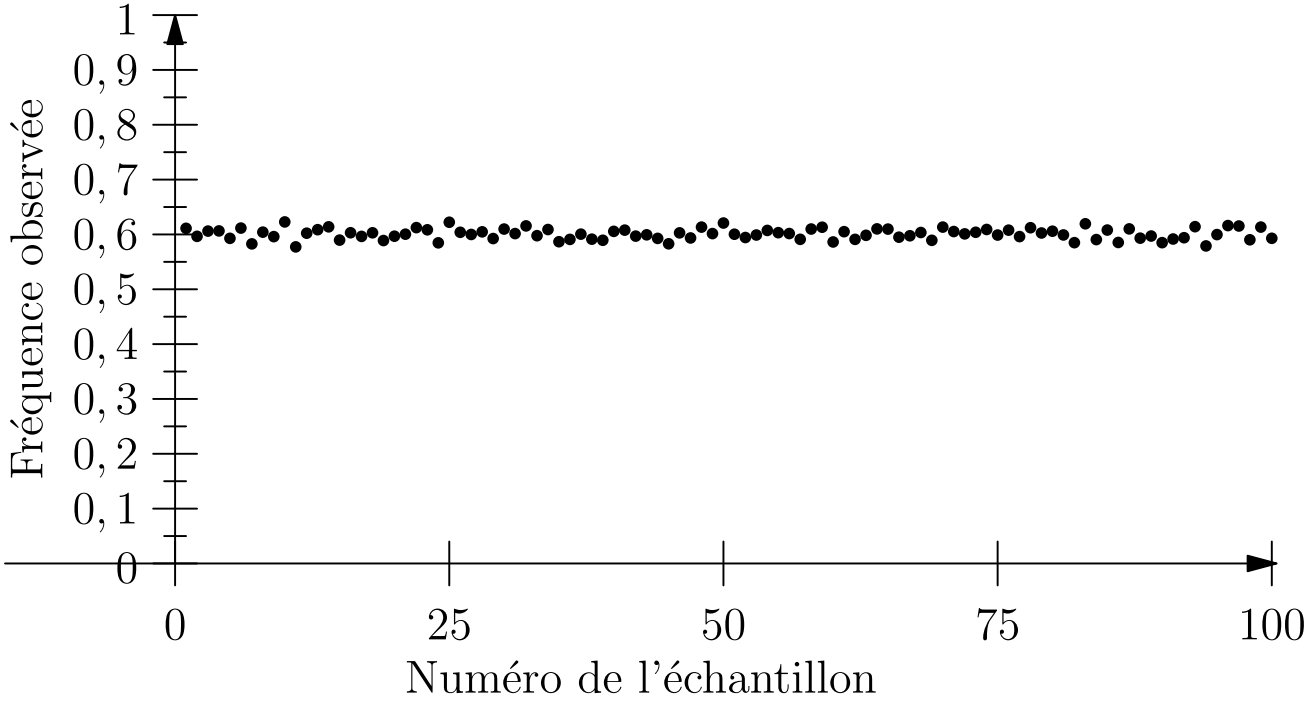

Figure 3.3 – Et voilà ce qui se passe pour une très grande taille d'échantillons.

Question 11

Dans la figure 3.1, 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.

4 La loi des grands nombres et le théorème centrale limite : énoncés et illustration

4.1 Énoncé de la loi des grands nombres et du théorème centrale limite

La loi des grands nombres admet un énoncé mathématique (très) compliqué. Voici ici une version simplifiée qu'il faut garder en tête et comprendre :

Proposition 2: La loi des grands nombres – version simplifiée

La fréquence observée dans un échantillon s'approche de la probabilité de l'événement observée lors la taille de l'échantillon augmente.

De plus, il existe un théorème qui nous permettra de quantifier les fluctuations que nous avons observées dans la section précédente. C'est le théorème centrale limite.

Proposition 3: Théorème centrale limite – version simplifiée

Pour un échantillon de taille n, l'écart entre la fréquence observée f et la probabilité de l'événement aléatoire observée p est *très souvent* de $\frac{1}{\sqrt{n}}$

Nous allons expliquer et appliquer, et illustrer ces deux théorèmes dans la suite de ce cours.

4.2 Application de la loi des grands nombres et du théorème centrale limite.

On peut utiliser ces théorèmes pour :

1. **Anticiper les fluctuations** d'une fréquence observée f à partir de la probabilité de l'événement associé p.
2. **Estimer** la probabilité p à partir de f

Proposition 4: Anticiper les fluctuations de f à partir 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) :

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

La fréquence observée f a de *très grande chance* de ne pas être très loin de p.

Exemple 8

Nous anticipons les fluctuations de f à partir de p dans la section suivante.

Proposition 5: Estimer la probabilité p à partir de la fréquence observée f

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) :

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

La probabilité p recherchée a de *très grande chance* d'être «emprisonnée» par deux bornes déterminées par f et n.

Proposition 6

(pour les plus curieux) Peut-on résumer ces inégalités en un seule ? Et bien oui, et voici comment : [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,

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*.

4.3 Anticiper les fluctuations d'une expérience aléatoire : exemple concret.

4.3.1 Expérience aléatoire étudiée

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.
Ici, on connait la probabilité p de l'événement «obtenir un six». On va donc chercher à anticiper les fluctuations de cet événement dans des échantillons.

Question 12

Quelle est la probabilité d'avoir un succès ?

Reponse 3

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 9

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 10

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

4.3.2 Simulation à l'aide d'un programme Python

On utilise Python pour réaliser plusieurs échantillons de l'expérience aléatoire décrite plus haut.

```python
import random

def lancerDes() :
    return random.randint(1, 6)

n = 5

nombreEchantillons = 15

listeFrequencesObservees = [] #liste qui contiendra les fréquences observées

for echantillon in range(nombreEchantillons) :
    succes = 0

    for experience in range(n) :
      d = lancerDes()
      if d == 6 :
          succes = succes + 1

    f = succes / n
    listeFrequencesObservees.append(f) #on ajoute la fréquence f à la liste

print(listeFrequencesObservees)
```

Listing 4: Simulation de plusieurs échantillons de l'expérience aléatoire et calcul de la fréquence observée au sein de chaque échantillon

```
[0.0, 0.0, 0.0, 0.0, 0.2, 0.0, 0.2, 0.0, 0.0, 0.0, 0.6, 0.2, 0.2, 0.0, 0.2]
```

Question 13

À quoi sert la fonction `lancerDes()` ?

Question 14

Combien d'échantillons ont été simulés ici ?

Question 15

Quelle est la taille de chaque échantillon ?

Question 16

Pourquoi le programme précédent a-t-il deux boucles `for` **imbriquées** (aux ligne 12 et 15) ?

Question 17

Combien de lancers de dé ont-ils été simulés par le programme ?

Question 18

Expliquer le rôle de la ligne 16

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

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

Question 19

En regardant le résultat obtenu :

1. pourquoi obtient-on une liste de nombres compris entre 0 et 1 ?
2. Combien d'échantillons n'ont **aucun six** ?

Question 20

1. Testez vous aussi le programme **plusieurs fois** !
2. Changer les valeurs `n` et `nombreEchantillon`.

4.3.3 Vérifier sur un graphique que la plupart des fluctuations sont bien comprises entre $p - \frac{1}{\sqrt{n}}$ et $p + \frac{1}{\sqrt{n}}$

Nous regardons $m = 1000$ échantillons de lancers de l'expérience aléatoire décrite plus haut. Nous vérifierons que la fréquence f observée est comprise entre $p - \frac{1}{\sqrt{n}}$ et $p + \frac{1}{\sqrt{n}}$ *la plupart du temps*.
Pour chaque graphique donnée à la page suivante :

— nous avons représenté la «bande» issues de la loi des grands nombres, qui nous dit que la fréquence observée à des grandes chances d'être 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 $y = p$. Ici, $p = \frac{1}{6}$ puisque la fréquence attendue de 6 est $\frac{1}{6}$.

— On regarde ce qui se passe pour une taille d'échantillon $n = 50$, puis $n = 100$ et enfin $n = 1000$, sur $m = 1000$ échantillons.

Les trois graphiques obtenus sont donnés à la page suivante.

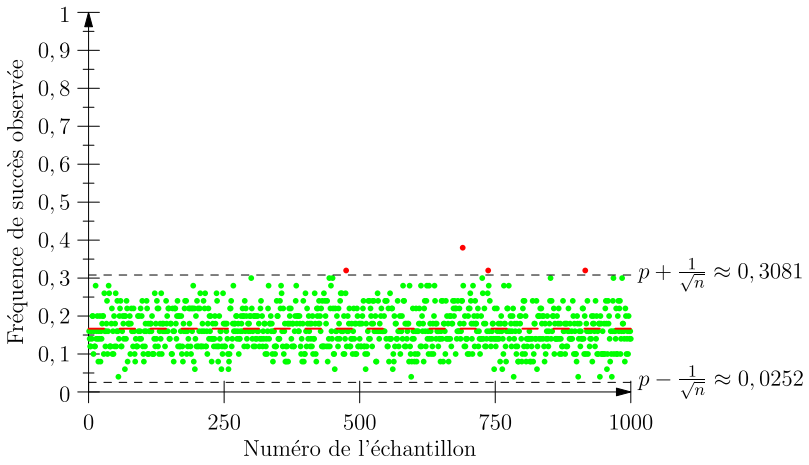

Figure 4.1 – Une taille d'échantillon $n = 50$

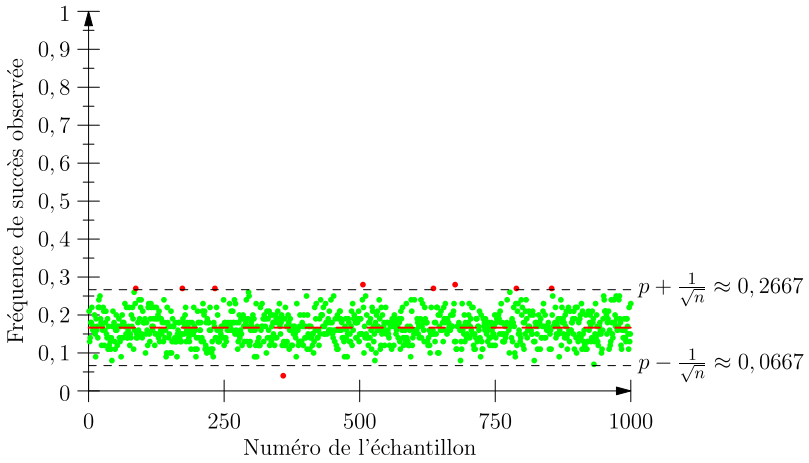

Figure 4.2 – Une taille d'échantillon $n = 100$

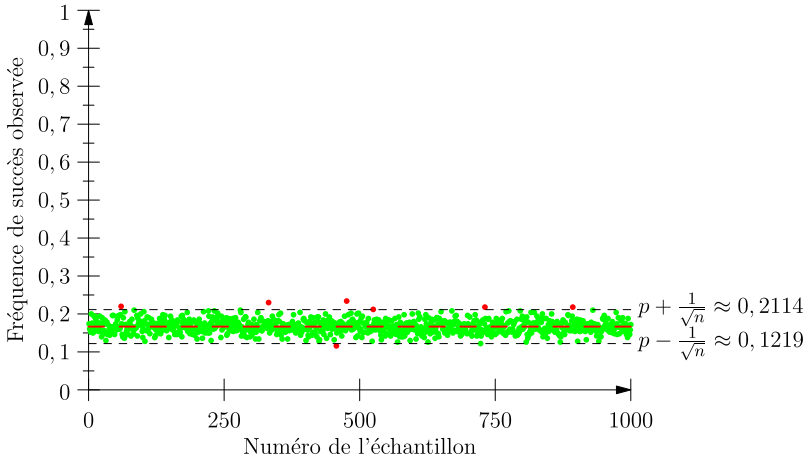

Figure 4.3 – Une taille d'échantillon $n = 1000$

Question 21

Pour chaque graphique :

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. 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.
6. Quelle est la proportion de points verts ?
7. Quelle est la valeur de p ?
8. Pourquoi il y a deux «bandes» autour de p ?
9. Retrouver les valeurs de $p + \frac{1}{\sqrt{n}}$ et $p - \frac{1}{\sqrt{n}}$ sur votre calculatrice.

Question 22

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. 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 ?

5 L'estimation

Définition 6: L'estimation

D'après les théorèmes de la loi des grands nombres et du théorème centrale limite, la fréquence observée f permet d'**estimer** la probabilité p d'une expérience aléatoire. Procédez ainsi, c'est effectuer une « estimation ».

Exemple 11

— Les sondages sont des exemples d'estimation

— On utilise des techniques statistiques d'estimation en médecine, en physique, en chimie, dans l'étude du climat... Et même en informatique !

Question 23

La question qui est au centre du domaine de l'estimation est la suivante : « Quelle est la taille de l'échantillon qui permet d'estimer p de façon suffisament précise ?». Nous allons répondre à cette question dans la proposition suivante.

Proposition 7

D'après les théorèmes citées plus haut, si on cherche à obtenir p au centième près (c'est-à-dire au pourcent près), alors on essaye d'avoir une taille d'échantillon n assez grande pour que la fréquence observée f soit à distance 0.005 de p :
Or la distance entre f et p a *de grande chance* d'être inférieure ou égale à $\frac{1}{\sqrt{n}}$.
Ainsi, on cherche une taille d'échantillon n telle que :

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

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

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

Cela signifie que pour obtenir une estimation au pourcent près il faudrait une taille d'échantillon de $n \geq 40\,000$!

Question 24

Pourquoi 0.005 et pas 0.01 dans la proposition précédente ?

Question 25

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

Exemple 12

En pratique le vrai théorème centrale limite, ou les modèles utilisés en statistique permettent d'être plus précis. On estime généralement qu'une taille d'échantillon autour de $n = 1000$ est suffisante.

Question 26

Deux candidates 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 la candidate A. Toutes les personnes interrogées ont voté pour l'une des deux candidates. Peut-on assurer la victoire de la candidate B ? Justifier.