Chapitre 15

Échantillonnage

L'objectif de ce chapitre est de faire percevoir, sous une forme expérimentale, la loi des grands nombres, la fluctuation d'échantillonnage et le principe de l'estimation d'une probabilité par une fréquence observée sur un échantillon.

Le saviez-vous ?

Le varroa destructor est un acarien parasite de l'abeille qui est en partie responsable de l'importante diminution du nombre d'abeilles depuis les années 2000. Pour détecter sa présence au sein d'un rucher et commencer le traitement, l'apiculteur doit récolter un échantillon de 300 abeilles sur au moins 10 % de ses ruches.

1. Activités de découverte

a) Je préfère les acidulés

Une marque de confiseries produit des bonbons avec deux parfums différents, citron vert et mangue en quantités égales. Les bonbons sont ensuite placés aléatoirement dans des boites de 100. On a relevé le nombre n de bonbons au citron vert dans 40 boîtes.

Boîte	1	2	3	4	5	6	7	8	9	10
n	56	50	57	47	54	46	53	50	53	52

Boîte	11	12	13	14	15	16	17	18	19	20
n	49	46	46	51	50	47	43	59	49	48

Boîte	21	22	23	24	25	26	27	28	29	30
n	54	56	47	40	49	47	42	53	53	58

Boîte	31	32	33	34	35	36	37	38	39	40
n	52	60	43	51	50	58	46	56	49	59

1. Le nombre de bonbons au citron vert, est-il le même dans toutes les boîtes ?
2. Dans combien de boîte a-t-on trouvé le même nombre de bonbons de chaque parfum ?
3. Quelles sont les proportions minimale et maximale, en pourcentage, de bonbons au citron vert pour ces 40 boîtes ?
4. Calculer l'écart maximal entre la proportion de bonbons au citron vert dans une boîte et 50 %.
5. Quelle est la proportion de boîtes respectant $0,45 \leqslant p \leqslant 0,55$?

b) Dés ronds

Mario a acheté un lot de dés sphériques. Ces dés sont tous identiques, hormis la couleur. Un tel dé est conçu de sorte que, quand on le lance, il se stabilise avec un numéro de 1 à 6 sur le dessus. Mario se demande si ces dés sont réellement équilibrés.

1. Il lance 200 fois un dé et obtient à 34 reprises le nombre 1. Quelle est la fréquence d'apparition du nombre 1 ?
2. Le tableau ci-dessous donne l'effectif de chaque nombre pour les 200 lancers. Calculer la fréquence d'apparition de chaque nombre.

Nombres	1	2	3	4	5	6
Effectifs	34	32	30	34	35	35

3. Représenter ces données par un diagramme en bâtons.
4. Que peut-on penser de l'équilibrage du dé ?

2. Échantillon, simulation et fluctuation

Expérience aléatoire

Une expérience aléatoire est une expérience renouvelable dont les résultats possibles sont connus sans qu'on puisse déterminer lequel sera réalisé.

Exemples d'expériences aléatoires :
- *le lancer de dé ;*
- *un sondage d'opinion avant une élection ;*
- *le tirage de jetons dans une urne ou de cartes dans un jeu.*

Échantillon

Un échantillon de taille n est constitué des résultats de n répétitions indépendantes de la même expérience.

Exemples d'échantillons :
- *on lance une pièce 50 fois et on regarde si on obtient pile ;*
- *on tire 20 fois une carte d'un jeu de 32 cartes en la remettant et on regarde si c'est un cœur ;*
- *on interroge 1 000 personnes et on leur demande si elles voteront.*

Fluctuation d'échantillonnage

Deux échantillons de même taille issus de la même expérience aléatoire ne sont généralement pas identiques.
On appelle fluctuation d'échantillonnage les variations des fréquences des valeurs relevées.

Labo 1.

Notation :
- n est le nombre d'éléments de l'échantillon. C'est l'**effectif** ou la **taille de l'échantillon**. On dit que l'échantillon est de taille n.
- f est la **fréquence** du caractère observé dans l'échantillon.
- p est la **proportion effective** du caractère observé dans la population.

Remarque : plus la taille de l'échantillon augmente, plus les fréquences f observées se rapprochent de p.

Simulation informatique : on demande à l'opérateur de saisir les valeurs de la taille de l'échantillon n puis de la proportion du caractère p. Le programme affiche la fréquence f observée dans l'échantillon.

```
Saisir n
Saisir p
s ← 0
pour i allant jusqu'à n faire
    x ← valeur aléatoire comprise entre 0 et 1
    si x ⩽ p alors
        s ← s + 1
    fin
fin
f ← s/n
Afficher f
```

```
from random import*
n = int(input("n = "))
p = float(input("p = "))
s = 0
for i in range(n):
        if random() <= p :
                s = s+1
print("f = ",s/n)
```

3. Prise de décision : intervalle de fluctuation (p est connue)

Protocole : soit une population pour laquelle on étudie la proportion d'un caractère.
On émet une hypothèse sur la proportion p du caractère étudié dans la population. On considère donc p comme connue car elle a une valeur conjecturée.
Un échantillon de taille n de cette population est prélevé et on détermine une fréquence observée f_o du caractère étudié.
La question : peut-on, à partir de l'observation de f_o, valider la conjecture faite sur p ?
La fréquence observée, f_o, est-elle proche ou éloignée de la probabilité ou proportion théorique, p ?

Intervalle de fluctuation

L'intervalle de fluctuation **au seuil de 95%** , relatif aux échantillons de taille n, est l'intervalle centré autour de p qui contient la fréquence observée f_o dans un échantillon de taille n avec une probabilité égale à 0,95 environ.

Remarques :

- Il n'existe pas d'intervalle dans lequel on trouverait f_o avec certitude (à moins de prendre l'intervalle $[0\ ;\ 1]$) à cause de la fluctuation d'échantillonnage.
- Cet intervalle peut être obtenu de façon approchée à l'aide de simulations.

Propriété

Soit p la proportion effective d'un caractère d'une population comprise entre 0,2 et 0,8 et f_o la fréquence du caractère dans un échantillon de taille n supérieure ou égale à 25.
f_o appartient à l'intervalle $\left[p - \frac{1}{\sqrt{n}}\ ;\ p + \frac{1}{\sqrt{n}}\right]$ avec une probabilité d'environ 0,95.

Remarque : la taille de l'intervalle de fluctuation $\left(\frac{2}{\sqrt{n}}\right)$ diminue si n augmente.

Méthode : pour prendre une décision
Dans les conditions de la définition et de la propriété :

- On émet une hypothèse sur la proportion du caractère de la population p.
- On détermine l'intervalle de fluctuation au seuil de 95% de la proportion p dans des échantillons de taille n.
  - Si f_o n'appartient pas à cet intervalle, on rejette l'hypothèse faite sur p **avec un risque d'erreur de 5%**.
  - Si f_o appartient à cet intervalle, on ne rejette pas l'hypothèse faites sur p.

Application : dans la réserve indienne d'Aamjiwnaag, située au Canada, à proximité d'industries chimiques, il est né entre 1999 et 2003, 132 enfants dont 46 garçons. Est ce normal ?

4. Estimation : intervalle de confiance (p est inconnue)

Lois des grands nombres

Dans une population, la proportion d'individus présentant un certain caractère est p. On prélève dans cette population un échantillon aléatoire de taille n. On note f la fréquence d'apparition du caractère dans cet échantillon.
Lorsque n **est grand**, sauf exception, la **fréquence observée** f **est proche de la proportion** p.

Dans le cas d'un échantillon de n répétitions indépendantes d'une expérience aléatoire à deux issues, succès et échec, lorsque n est grand, la fréquence observée f du succès dans l'échantillon est proche de la probabilité p du succès.

Estimation d'une proportion

Dans une population, la proportion p d'individus présentant un certain caractère est inconnue. On prélève dans cette population un échantillon aléatoire de taille n. On note f_0 la fréquence d'apparition du caractère dans l'échantillon.
La fréquence observée f_0 est appelée **estimation** de la proportion p.

Remarque : l'estimation obtenue dépend de l'échantillon considéré, donc il y a plusieurs estimations possibles d'une même proportion p.

L'intervalle de fluctuation permet d'avoir un intervalle où se situe la proportion inconnue p avec une probabilité de 0,95%.

Propriété

On considère un échantillon de taille $n(\ n \geqslant 25)$ tel que $f_o \in [0,2;0,8]$.
Alors p appartient à l'intervalle $\left[f_o - \dfrac{1}{\sqrt{n}}\ ;\ f_o + \dfrac{1}{\sqrt{n}}\right]$ avec une probabilité de 0,95.

Intervalle de confiance

Un intervalle de confiance **au seuil de 95%** , relatif aux échantillons de taille n, est
un intervalle centré autour de f_0 où se situe la proportion p du caractère dans la population avec une probabilité égale à 95% .
L'intervalle $\left[f_o - \dfrac{1}{\sqrt{n}}\ ;\ f_o + \dfrac{1}{\sqrt{n}}\right]$ est donc appelé intervalle de confiance au seuil de 95%.

Méthode : pour estimer la proportion d'un caractère
- On réalise un échantillon de taille n et on y obtient une fréquence observée f_o.
- On construit l'intervalle de confiance à partir de n et f_o.

La proportion réelle dans la population se situe dans cet intervalle **avec une probabilité d'environ 0,95**.

Application : le 4 mai 2007 soit deux jours avant le second tour des élections présidentielles, on publie le sondage suivant réalisé auprès de 992 personnes :

S. Royal	*: 45%*
N. Sarkozy	*: 55%*

Interpréter ce sondage.

Remarque : les sondages sont souvent réalisés auprès d'environ 1000 personnes car cela permet de connaître la proportion d'un candidat à 3% près.

5. Caisse à outils

Labo 2.

Comprendre une fonction écrite en Python $\Longrightarrow$ l'instruction **random()** renvoie une valeur décimale aléatoire comprise entre 0 et 1. Pour être opérationnelle, le module random doit être importé. On compare la valeur aléatoire obtenue à la probabilité du succès de l'expérience simulée ; si elle est inférieure ou égale on comptabilise un succès. Ensuite on divise le nombre de succès par la taille de l'échantillon (le nombre de répétitions de l'expérience) pour déterminer la fréquence observée f_0 dans l'échantillon. La saisie de la commande **nb_freq(n,p)** en remplaçant n et p par leur valeur numérique permet d'obtenir le nombre de succès et la fréquence observée.

Définir fonction nb_freq
$s \leftarrow 0$
pour *i allant jusqu'à* n **faire**
  $x \leftarrow$ valeur aléatoire comprise entre 0 et 1
  si $x \leqslant p$ **alors**
    $s \leftarrow s+1$
  fin
fin
Renvoyer s et s/n

```
from random import*
def nb_freq(n,p):
  s = 0
  for i in range(n):
    if random() <= p :
      s = s+1
  return(s,s/n)
```

Application : on prend un jeu de 32 cartes et l'on gagne si l'on tire un des quatre as.

1. *Quelle est la probabilité de gagner ?*
2. *Pour simuler cette expérience écrire une fonction en langage Python nommée Tirage qui affichera gagné ou perdu.*
3. *Pour simuler n répétitions de cette expérience écrire une fonction en langage Python nommée RepTir qui affichera la fréquence observée puis le nombre de parties gagnées.*

Estimer une proportion $\Longrightarrow$ on divise le nombre de succès constatés dans l'échantillon par sa taille pour obtenir la fréquence observée.

Application : dans une population, on prélève un échantillon de 400 individus parmi lesquels 92 sont porteurs du marqueur d'une pathologie. Quelle estimation, en pourcentage, de la proportion d'individus potentiellement malade au sein de cette population obtient-on ?

6. Algorithmes

Calcul de l'intervalle de fluctuation d'une proportion p au seuil de confiance de 95%.
L'utilisateur saisit la valeur de la proportion p puis celle de la taille de l'échantillon n. Puis on utilise les formules permettant de calculer les borne de l'intervalle : $\left[p - \frac{1}{\sqrt{n}} \ ; \ p + \frac{1}{\sqrt{n}}\right]$.

Labo 3.

Saisir n
Saisir p
Afficher $p - 1/\sqrt{n}$
Afficher $p + 1/\sqrt{n}$

```
from math import*
n = float(input("n = "))
p = float(input("p = "))
print("I = [",p-1/sqrt(n)," ; ",p+1/sqrt(n),"]")
```

Simulation du tirage d'un échantillon.
Dans un laboratoire, on étudie les capacités de mémorisation d'une souris. L'animal se déplace dans un labyrinthe présentant deux sorties possibles. De la nourriture est placée à seulement une de ces sorties, toujours la même. On a observé que la souris trouve la bonne sortie dans 74 % des cas.
La variable s compte le nombre de succès de la souris. Elle est initialisée à 0. La boucle **Pour** permet de répéter les 120 expériences de l'échantillon, la variable i est le compteur de boucles. La condition $x < 0,74$ est réalisée dans 74 % des cas et correspond au succès de la souris. f est la fréquence de réussite de la souris.

$s \leftarrow 0$
pour *i allant de 1 à 120* **faire**
  $x \leftarrow$ valeur aléatoire comprise entre 0 et 1
  si *$x < 0,74$* **alors**
    $s \leftarrow s + 1$
  fin
fin
$f \leftarrow s/120$
Afficher f

```
from random import*
s = 0
for i in range(1,121):
        x = random()
        if x < 0.74 :
                s = s+1
print("f = ",s/120)
```

Simulation de N échantillons de taille n.
On fait appel à des fonctions qui permettent dans l'ordre de calculer le nombre de succès, tirage inférieur à la proportion p, puis de calculer la fréquence observée dans l'échantillon. Enfin on comptabilise les échantillons pour lesquels l'écart entre proportion et fréquence est suffisamment faible pour en calculer la proportion.
Pour exécuter le programme, saisir par exemple la commande repet_echan(30,100,0.7) ; donc $N = 30$, $n = 100$ et $p = 0,7$.

Définir fonction nombre_succes
nb_succes $\leftarrow 0$
pour *compteur allant jusqu'à* n **faire**
  si *valeur aléatoire* $< p$ **alors**
    nb_succes $\leftarrow$ *nb_succes* $+ 1$
  fin
fin
Renvoyer *nb_succes*

Définir fonction frequence_succes
Renvoyer nombre_succes(n,p)/n

Définir fonction repet_echan
$s \leftarrow 0$
pour *i allant jusqu'à* N **faire**
  $f \leftarrow$ frequence_succes(n,p)
  si $abs(p - f) \leqslant 1/\sqrt{n}$ **alors**
    $s \leftarrow s + 1$
  fin
fin
Renvoyer s/N

```
from math import*
from random import*

def nombre_succes(n,p):
        nb_succes = 0
        for compteur in range(n):
        if random()<p:
            nb_succes = nb_succes + 1
    return nb_succes

def frequence_succes(n,p):
    return nombre_succes(n,p)/n

def repet_echan(N,n,p):
    s = 0
    for i in range(N):
        f = frequence_succes(n,p)
        if abs(p-f) <= 1/sqrt(n):
            s = s+1
    return s/N
```

7. Évaluations

Devoir en temps libre n° 15 : Échantillonnage

Il est rappelé que la qualité de la rédaction, la clarté et la précision des raisonnements entreront pour une part importante dans l'apprécitaion des copies. Le barème est donné à titre indicatif. Le sujet sera rendu avec la copie.

Exercice n°1 : ***Le mauvais lot ?***

Une marque d'électroménager a un taux de retour au service après-vente (SAV) de 7,3 %.

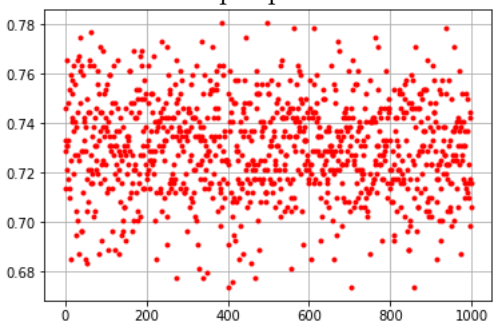
1. Expliquer comment simuler le fait qu'un appareil fabriqué par cette marque passe par le SAV ou non. On peut utiliser une fonction **Alea()** en langage naturel, générant un nombre réel entre 0 et 1 ou la commande **random()** en langage Python.
2. Recopier et compléter la fonction Python ci-dessous afin qu'elle simule un échantillon de 529 appareils de cette marque et renvoie la fréquence de ceux qui retournent au SAV.

```
def sav():
        effectif = ...
    for j in range(1,530):
        if random.random() <= ... :
            effectif = effectif + 1
    return effectif / ...
```

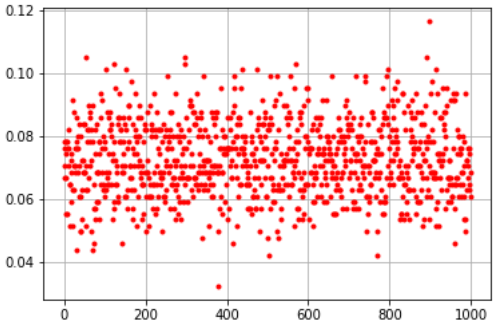
3. En lançant 1 000 fois cette fonction, on simule 1 000 échantillons de 529 appareils.

a) Un des deux graphiques proposés ci-dessous, donne la fréquence des appareils passant par le SAV dans chacun des 1 000 échantillons, lequel ? Pourquoi ?

Graphique 1



Graphique 2



b) En utilisant le bon graphique, donner un encadrement de la fréquence attendue des appareils passant par le SAV dans un échantillon de taille 529.

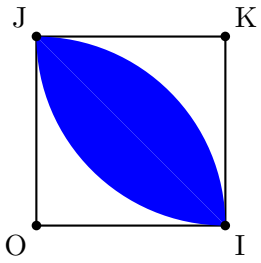
4. Une grande surface a vendu 529 appareils de cette marque et a dû gérer 69 retour vers le SAV.

a) Expliquer pourquoi la direction de cette grande surface pense que ce lot d'appareils a un taux de retour vers le SAV élevé.

b) En aurait-il-été de même si seulement 46 appareils avaient dû repasser par le SAV ?

Exercice n°2 : ***Dans l'oeil ?***

Dans un jeu video, l'ordinateur choisit au hasard un point M à l'intérieur du carré OIKJ ci-contre de côté 1. Si le point M se trouve à l'intérieur de l'oeil (zone colorée), le joueur gagne des points de vie supplémentaires.
Cet oeil est délimité par deux arcs de cercle, l'un de centre O et l'autre de centre K, chacun de rayon 1.
On souhaite savoir si un joueur a plus d'une chance sur deux de gagner des points de vie supplémentaires de cette façon.



1. On munit le plan d'un repère orthonormé (O ; I ; J).

a) Justifier qu'un point $M(x\ ;\ y)$ est à l'intérieur de l'oeil central si, et seulement si, $OM^2 \leqslant 1$ et $KM^2 \leqslant 1$.

b) On considère les fonctions OM2 et KM2 ci-contre. Que permettent-elles de faire lorsque $(x\ ;\ y)$ sont les coordonnées d'un point M ?

```
def OM2(x,y):
  return x**2+y**2

def KM2(x,y):
  return (x-1)**2+(y-1)**2
```

2. Soit un entier naturel n non nul. On souhaite simuler n fois l'expérience aléatoire et calculer la fréquence des cas où le joueur gagne des points de vie supplémentaires. On propose pour cela la fonction ci-contre. Expliquer la démarche.

```
def simul(n):
  NbSucces=0
  for i in range(n):
    x=random()
    y=random()
    if OM2(x,y)<=1 and KM2(x,y)<=1:
      NbSucces=NbSucces+1
  return NbSucces/n
```

3. On a exécuté 200 fois la fonction **simul(50)** et on a représenté graphiquement ci-contre les fréquences observées, ainsi que la droite d'équation $y = 0,5$.

   Peut-on penser qu'un joueur a plus d'une chance sur deux de gagner des points de vie supplémentaires ? Pourquoi ?