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. Échantillon, simulation et fluctuation

Expérience aléatoire

.....

.....

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

.....

.....

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

. . . . . .

. . . . . .

. . . . . .

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
$s \leftarrow 0$
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
$f \leftarrow s/n$
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)
```

2. 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

. . . . . .

. . . . . .

. . . . . .

. . . . . .

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 ?

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

Lois des grands nombres

.....

.....

.....

.....

.....

.....

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

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

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

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

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.

4. Caisse à outils

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 ?

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

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
```

6. Évaluations

Devoir en temps libre n° 15 : Échantillonnage

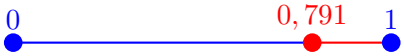
Il est rappelé que la qualité de la rédaction, la clarté et la précision des raisonnements entreront pour une part importante dans l'apprécitaion des copies. Le barème est donné à titre indicatif. Le sujet sera rendu avec la copie.

Exercice n°1 : ***Affaire Partida***

En Novembre 1976 dans le comté de Hidalgo, Rodrigo Partida était condamné à huit ans de prison pour cambriolage et tentative de viol.
Il attaqua ce jugement affirmant que la désignation des jurés de ce comté était discriminatoire pour les américains d'origine mexicaine : 79,1% de la population du comté était d'origine mexicaine mais, sur les 870 personnes convoquées pour être jurés les 11 années précédentes, seules 339 d'entre elles étaient d'origine mexicaine.

1. Simulation de la désignation d'un juré
   On étudie une fonction du tableur qui choisit un juré en tenant compte de ses origines.
   a) Quel nombre de jurés d'origine mexicaine peut-on espérer en choisissant au hasard 870 personnes dans la population de ce comté ?
   b) Avec un tableur, la fonction `ALEA()` génère un nombre aléatoire dans $[0; 1[$.
   Que renvoie `SI(Alea()<p;1;0)` ?

c) En prenant $p = 0,791$ expliquer comment cette formule permet de simuler la désignation d'un juré de ce comté en respectant les fréquences. *On pourra s'aider du schéma ci-dessous :*



2. Programmation

On procède à une simulation de 100 séries de désignation de jury.

a) Compléter la feuille de calcul à l'aide des instructions.
   - Saisir en cellule `A1` la formule `SI(ALEA()<0,791;1;0)`
   - Copier sur la plage `A2:A870`
   - Saisir en `A871` la formule `SOMME(A1:A870)/870`
   - Sélectionner la plage `A1:A871`
   - Copier sur la plage `B1:CV871`

b) Que représentent les nombres de la plage de cellules `A1:A870` ?

c) Que représente le nombre affiché dans la cellule `A871` ?

d) Que représentent les valeurs extrêmes obtenues dans la plage de cellules `A871:CV871` ?

e) Représenter avec un nuage de points la série de données de la plage `A871:CV871`.

f) A-t-on obligatoirement 688 jurés d'origine mexicaine ?
Calculer le nombre maximal de jurés d'origine mexicaine dans un jury, obtenu lors de la simulation.

3. Intervalle de fluctuation

Il s'agit d'interpréter les résultats de cette simulation.

a) Donner l'intervalle de fluctuation correspondant à la simulation antérieure.
Celui-ci confirme-t-il les observations précédentes ?

b) Dans les simulations faites sur tableur, obtient-on un nombre de jurés mexicains égal à celui de l'affaire Partida ?

c) Comment expliquer cette situation ?