Statistiques inférentielles – 2nde

A) Échantillonnage

1) Notion d'échantillons

Exemples : On propose plusieurs situations d'échantillonnage

1) Sur l'ensemble des cartes à puce produites par une entreprise en une semaine, on en prélève 200. On dit que cet ensemble de 200 cartes à puce constitue un **échantillon de taille 200** de la population de toutes les cartes à puce produites en une semaine.

2) On s'intéresse aux intentions de vote lors d'une élection. On sonde 1000 personnes en leur demandant leur intention de vote. L'ensemble de ces 1000 personnes constitue un **échantillon de taille 1000** de la population totale des électeurs.

3) On lance une pièce de monnaie 50 fois de suite (successivement) et on note les résultats obtenus. L'ensemble de ces 50 lancers constitue un **échantillon de taille 50**.

Définition : Un **échantillon de taille *n*** est constitué des résultats de *n* répétitions indépendantes de la même expérience sur l'ensemble des personnes ou objets sur lesquels porte l'étude statistique (la population).

Un **échantillon** issu d'une population est donc l'ensemble de quelques éléments de cette population.

2) Simulations d'expériences aléatoires

Exemple : On considère l'expérience aléatoire qui consiste à lancer un dé à 6 faces. Le programme Python suivant permet de simuler cette expérience.

```
from random import*

def dé():
    r=randint(1,6)
    return(r)
```

Note : La fonction **randint** renvoie un nombre aléatoire entier de 1 à 6.

```
>>> dé()
1
```

On exécute le programme et on obtient l'affichage ci-contre.
Cela signifie que le logiciel a simulé un lancer de dé et on a obtenu un « 1 ».

La règle du jeu veut que si le résultat est « 1 » ou « 6 », on gagne.
Dans le cas contraire, on perd. On répète *n* fois de suite cette expérience à deux issues (gagner ou perdre) consistant à lancer le dé.

On modifie et complète le programme Python afin de simuler *n* lancers de dé.
Le programme affiche le nombre de fois que l'on gagne.
La variable **n** désigne le nombre de lancers.
La variable **s** permet de compter le nombre de fois que l'on gagne : le dé s'arrête sur «1 » ou sur « 6 »

```
from random import*

def dé(n):
    s=0
    for k in range(n):
        r=randint(1,6)
        if r==1 or r==6:
            s=s+1
    return(s)
```

```
>>> dé(10)
3
```

On exécute le programme et on obtient l'affichage ci-contre.
Cela signifie que sur 10 lancers, on a gagné 3 fois.

3) Loi des grands nombres

Modifions le programme afin d'afficher en sortie la fréquence de jeux gagnés sur un échantillon de *n* lancers de dé.

Il suffit de remplacer dans la dernière ligne **return(s)** (l'effectif) par **return(s/n)** (la fréquence).

```
from random import*

def dé(n):
    s=0
    for k in range(n):
        r=randint(1,6)
        if r==1 or r==6:
            s=s+1
    return(s/n)
```

```
>>> dé(10)
0.2
>>> dé(100)
0.32
>>> dé(1000)
0.328
>>> dé(5000)
0.3372
>>> dé(100000)
0.33353
```

On exécute le programme pour des valeurs de *n* de plus en plus grandes.
Ci-contre les résultats obtenus à l'aide du logiciel.

On constate que, plus *n* devient grand, plus les fréquences observées semblent se rapprocher d'une valeur théorique égale à $\frac{1}{3}$

En effet, la probabilité de gagner (obtenir un « 1 » ou un « 6 ») est égale à

$$p=\frac{2}{6}=\frac{1}{3}$$

Théorème (***Loi des grands nombres***) : Lorsque *n* devient grand, sauf exception, la fréquence observée est proche de la probabilité théorique

B) Estimations

1) Estimation ponctuelle

On se propose maintenant de répéter N fois la simulation de l'expérience aléatoire précédente. Dans chaque cas, pour *n* suffisamment grand, la fréquence observée *f* devrait être proche de la probabilité théorique $p=\frac{1}{3}$

On veut calculer la proportion des cas pour lesquels l'écart entre *f* et *p* est inférieur ou égale à $\frac{1}{\sqrt{n}}$

Après avoir importé le module **math**, nécessaire pour utiliser la fonction **abs** (valeur absolue), on complète le programme précédent avec la fonction **estim**.

```
from math import*

def estim(N,n):
    c=0
    for k in range(N):
        f=dé(n)
        if abs(f-1/3)<=1/sqrt(n):
            c=c+1
    return(c/N)
```

abs(f-1/3) est l'écart entre *f* et $\frac{1}{3}$

$\sqrt{n}$ se note **sqrt(n)**.

```
>>> estim(10,10000)
0.9
>>> estim(50,10000)
0.96
>>> estim(100,10000)
0.96
>>> estim(100,10000)
0.94
```

On exécute le programme pour différentes valeurs de N en choisissant *n* suffisamment grand, soit *n* = 10000.

On trouve des valeurs proches de 0,95 ce qui signifie que dans 95 % des cas, l'écart entre la fréquence observée *f* et la probabilité *p* est inférieur ou égale à 0,01. En effet : $\frac{1}{\sqrt{n}}=\frac{1}{\sqrt{10000}}=0{,}01$

Principe de l'estimation : Pour *n* assez grand, *f* donne une bonne estimation de *p* dans environ 95 % des cas.

```
from random import*
from math import*
def dé(n):
s=0
for k in range(n):
r=randint(1,6)
if r==1 or r==6:
s=s+1
return(s/n)
def estim(N,n):
c=0
for k in range(N):
f=dé(n)
if abs(f-1/3)<=1/sqrt(n):
c=c+1
return(c/N)
```

2) Les intervalles de confiance

Définition : On appelle intervalle $[a\,;b]$ l'ensemble de tous les nombres réels compris entre *a* et *b* (inclus) ; ainsi $x\in[a\,;b]$ revient à $a\leqslant x\leqslant b$

Avec les conclusions de la partie précédente on peut ainsi définir la notion d'« intervalle de confiance »

Définition : Soit p la fréquence théorique d'un processus aléatoire ; l'estimation de la fréquence observée sur un échantillon de taille n parmi une population de taille N est : $I_C=\left[p-\frac{1}{\sqrt{n}}\,;\,p+\frac{1}{\sqrt{n}}\right]$ à 95 % de fiabilité

Exemple : On lance un dé à 4 faces (tétraédrique) ; on étudie la fréquence d'obtenir le « 4 » sur un échantillon de plusieurs simulations de taille n
On note f la fréquence observée sur chaque simulation

- pour $n=20$ on a $I_C=\left[\frac{1}{4}-\frac{1}{\sqrt{20}}\,;\frac{1}{4}+\frac{1}{\sqrt{20}}\right]=[0{,}026\,;0{,}474]$
- pour $n=50$ on a $I_C=\left[\frac{1}{4}-\frac{1}{\sqrt{50}}\,;\frac{1}{4}+\frac{1}{\sqrt{50}}\right]=[0{,}109\,;0{,}391]$
- pour $n=100$ on a $I_C=\left[\frac{1}{4}-\frac{1}{\sqrt{100}}\,;\frac{1}{4}+\frac{1}{\sqrt{100}}\right]=[0{,}15\,;0{,}35]$
- pour $n=500$ on a $I_C=\left[\frac{1}{4}-\frac{1}{\sqrt{500}}\,;\frac{1}{4}+\frac{1}{\sqrt{500}}\right]=[0{,}205\,;0{,}295]$
- pour $n=1000$ on a $I_C=\left[\frac{1}{4}-\frac{1}{\sqrt{1000}}\,;\frac{1}{4}+\frac{1}{\sqrt{1000}}\right]=[0{,}218\,;0{,}282]$

Ainsi, on peut déduire que plus la taille de l'échantillon est grand, plus l'intervalle de confiance est précis

Exercices :

1) On lance 1 pièce ; déterminer l'intervalle de confiance de la fréquence du « pile » pour un échantillon de taille 150
2) On lance 1 dé à 6 faces ; déterminer l'intervalle de confiance de la fréquence du « 4 » pour un échantillon de taille 250
3) On lance 1 dé à 8 faces ; déterminer l'intervalle de confiance de la fréquence du « 7 » pour un échantillon de taille 500

C) Tests d'hypothèses

1) Étude de plusieurs pièces

On souhaite vérifier si une pièce de monnaie est bien équilibrée (non truquée)
Pour cela, on effectue plusieurs simulations de n lancers

On se propose de tester 5 pièces différentes :

- pièce A : avec 20 lancers on obtient 12 « pile »
- pièce B : avec 50 lancers on obtient 33 « pile »
- pièce C : avec 80 lancers on obtient 28 « pile »
- pièce D : avec 120 lancers on obtient 68 « pile »
- pièce E : avec 150 lancers on obtient 62 « pile »

Quelles sont les pièces « truquées » et les pièces « non truquées » ?

2) Prise de décision

Propriétés : Soit une fréquence théorique p et une fréquence observée f sur un échantillon de taille n , on note l'hypothèse $H_0 : « f \simeq p »$ à 95 % de fiabilité

- si $f \in \left[p-\frac{1}{\sqrt{n}} ; p+\frac{1}{\sqrt{n}} \right]$ alors on peut estimer que H_0 est vraie avec une fiabilité de 95 %
- si $f \notin \left[p-\frac{1}{\sqrt{n}} ; p+\frac{1}{\sqrt{n}} \right]$ alors on peut estimer que H_0 est fausse avec une risque de 5 %

Exemples : On applique ce test d'hypothèse aux 5 pièces précédentes

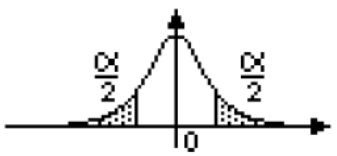
- pièce A : avec 20 lancers on obtient 12 « pile »
  on a $f=\frac{12}{20}=0,6$ et $I_C=\left[\frac{1}{2}-\frac{1}{\sqrt{20}} ; \frac{1}{2}+\frac{1}{\sqrt{20}}\right]=[0,276 ; 0,723]$
  donc $f \in I_C$ donc la pièce A n'est pas truquée

- pièce B : avec 50 lancers on obtient 33 « pile »
  on a $f=\frac{33}{50} \simeq 0,67$ et $I_C=\left[\frac{1}{2}-\frac{1}{\sqrt{50}} ; \frac{1}{2}+\frac{1}{\sqrt{50}}\right]=[0,359 ; 0,641]$
  donc $f \notin I_C$ donc la pièce B est truquée

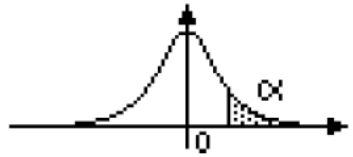
- pièce C : avec 80 lancers on obtient 28 « pile »
  on a $f=\frac{28}{80}=0,35$ et $I_C=\left[\frac{1}{2}-\frac{1}{\sqrt{80}} ; \frac{1}{2}+\frac{1}{\sqrt{80}}\right]=[0,388 ; 0,612]$
  donc $f \notin I_C$ donc la pièce C est truquée

- pièce D : avec 120 lancers on obtient 68 « pile »
  on a $f=\frac{68}{120} \simeq 0,57$ et $I_C=\left[\frac{1}{2}-\frac{1}{\sqrt{80}} ; \frac{1}{2}+\frac{1}{\sqrt{80}}\right]=[0,409 ; 0,591]$
  donc $f \in I_C$ donc la pièce D n'est pas truquée

- pièce E : avec 150 lancers on obtient 62 « pile »
  on a $f=\frac{62}{150} \simeq 0,413$ et $I_C=\left[\frac{1}{2}-\frac{1}{\sqrt{150}} ; \frac{1}{2}+\frac{1}{\sqrt{150}}\right]=[0,418 ; 0,582]$
  donc $f \notin I_C$ donc la pièce E est truquée



Test bilatéral



Test unilatéral