Chapitre 13 : Echantillonnage

Faut-il croire aux sondages ? : https://video-streaming.orange.fr/actu-politique/l-oeil-du-20h-faut-il-croire-les-sondages-CNT000001elHNQ.html

I) Définition

Définition : En statistiques, **l'échantillonnage** consiste à étudier une population à partir d'un échantillon (un extrait de cette population).

Rappel : Une population est un ensemble de "trucs" : les voitures françaises, les lycéens, ...

L'idée étant d'étendre les caractéristiques observées sur l'échantillon à l'ensemble de la population ou de déterminer la crédibilité de certaines hypothèses que l'on cherche à tester.

Exemples d'échantillons :

Lancer 100 fois un dé et noter le résultat obtenu.
Interroger 100 personnes et noter leur candidat politique préféré.
Dans une usine de médicaments, prélever 100 boîtes et tester leur conformité.

Définition :

Un **échantillon de taille *n*** est une liste de n résultats obtenus par n répétitions indépendantes d'une même expérience aléatoire.

Epreuve de Bernoulli :

Une épreuve de Bernoulli est une expérience qui n'a que deux issues possibles.
L'une des deux issues sera appelé "succés" et on notera p sa probabilité. L'autre issue sera appelé échec et sa probabilité sera alors de 1-p.

Exemple : obtenir pile ou face, gagner ou perdre, conforme ou non conforme,...

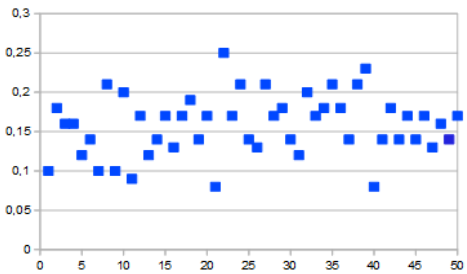
Définition :

Dans un n - échantillon, si on note k le nombre de succés, on peut calculer la **fréquence de succés** par $f = \frac{k}{n}$.

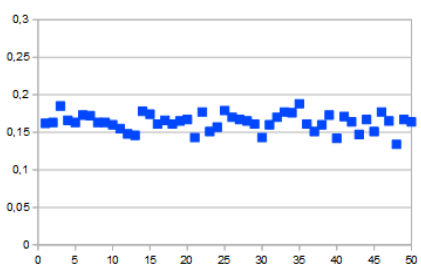
II) Loi des grands nombres

On lance un dé à 6 faces. Naturellement, vous savez que nous avons une chance sur 6 d'obtenir un "un".
Cela ne signifie en aucun cas que si nous lançons six fois le dé, nous sommes sûr d'avoir un "un".

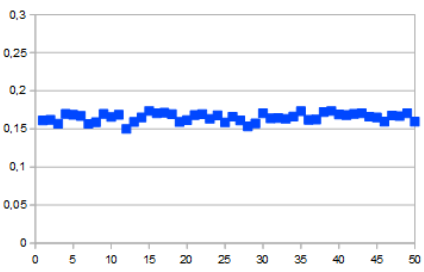
Lançons maintenant n fois ce même dé, nous obtenons un n - échantillon dont nous pouvons calculer la fréquence de succés.
Reproduisons ceci 50 fois, chaque point représente un échantillon



50 échantillons de 100 dés



50 échantillons de 1000 dés



50 échantillons de 5000 dés

On observe que les fréquences semblent fluctuer autour d'une valeur proche de 0,16, c'est ce que l'on appelle la **fluctuation d'échantillonnage**. Et plus la taille de l'échantillon est grande (plus on lance de dés donc), moins cette fluctuation est visible.

En remarquant que $\frac{1}{6} \approx 0{,}16$, on peut conjecturer que plus l'échantillon est grand, plus la fréquence observée se rapproche de la probabilité théorique.

Loi des grands nombres :

Lorsque la taille *n* de l'échantillon devient assez grande, sauf exception, les fréquences de succès *f* observées sont moins dispersées et sont assez proches de la probabilité *p*.

Que veut dire le sauf exception ?
Cela veut dire que cela marche quasiment tout le temps néanmoins il faut être conscient qu'il est possible d'obtenir 1000 "un" en lançant 1000 dés, la fréquence de succés est alors de 1, mais la probabilité que cette situation se produise est tellement faible que l'on considérera que ce n'est pas une situation normale.

III) Estimations

Considérons maintenant l'expérience suivante : Je demande à 100 personnes, prises au hasard, si elles croient aux ovnis. Considérons que "oui" est le succés. Et je répète cette expérience 50 fois. J'obtiens ces résultats :

50 échantillons de taille 100

Puisque nous avons remarqué que les fréquences de succès fluctuaient autour de la probabilité (de la proportion), nous aurions envie d'annoncé qu'environ 50% des français croient aux ovnis.

Principe de l'estimation :

On constitue des échantillons de taille *n* "assez grand".
Les fréquences observées donnent des valeurs approchées de la probabilité (de la proportion).

Erreur :

L'erreur commise est $|f - p|$. C'est l'écart entre la valeur mesurée et la valeur théorique.

Propriété :

Sauf exception, $|f - p| \leqslant \frac{1}{\sqrt{n}}$

Voilà pourquoi *n* doit être assez grand, sinon cela ne sert à rien. En pratique, on considère que n = 1000 donne une marge d'erreur acceptable. *(les fameux + ou - 3 points des sondages).*

IV) Simulation

Nous avons vu comment simuler le lancé de plusieurs pièces sur un tableur.
On peut également le faire sur Python.

Il faut commencer par aller chercher les fonctions aléatoires : `from random import *`

La fonction randint (n,p) renvoie un entier aléatoire entre n et p compris :

```
from random import *

de=randint(1,6)
print(de)
```

donc ce programme simule le lancé d'un dé et stocke le résultat dans une variable "de".

Avec une boucle, on simule plusieurs dés :

```
from random import *

for i in range(6):
    de=randint(1,6)
    print(de)
```

La fonction rand() renvoie un réel compris entre 0 et 1.

```
p=random()
print(p)
```

Simuler un échantillon :

Imaginons que 27% de la population croient aux extra-terrestres. Je voudrais créer un 200 échantillons (c'est-à-dire simuler informatiquement une liste de 200 personnes et compter combien croient aux extra-terrestres).

Pour chaque personne, nous allons générer un réel aléatoire entre 0 et 1 ; s'il est plus petit que 27% (=0,27) , cette personne sera comptée comme croyant aux extra-terrestres et s'il est plus grand que 0,27, elle sera comptée comme ne croyant pas aux extra-terrestres.

```
from random import *

ovni=0
for i in range(200):
    a=random()
    if a<=0.27:
        ovni=ovni+1

print(ovni)
```