Chp 10 : Echantillonnage

Ce domaine des mathématiques est à la croisée des statistiques et des probabilités. L'échantillonnage est l'étude « d'une partie contenue dans un tout ». Un exemple omniprésent dans les médias est la réalisation de sondages.

I) Echantillonnage

1. Notion d'échantillon

On considère une expérience aléatoire à deux issues. Par convention :

- l'une des issues est appelée « succès », de probabilité p ;
- l'autre est appelée « échec », de probabilité $1-p$.

Déf : Soit n un entier naturel non nul.

- On répète n fois l'expérience aléatoire de façon indépendante (c'est-à-dire que les résultats obtenus au fur et à mesure n'ont pas d'influence sur les résultats suivants). La liste des n résultats obtenus est un **échantillon** de taille n.
- La fréquence de succès observée f sur un échantillon de taille n comportant k succès est donnée par la formule : $f=\frac{k}{n}$.

Exple : Dans une urne comportant des boules rouges et des boules bleues, on tire au hasard une boule, on note sa couleur puis on la remet dans l'urne. Les deux issues possibles sont Rouge (R) et Bleu (B).
En répétant 50 fois cette expérience, on obtient une liste de 50 lettres R ou B qui constitue un échantillon de taille 50.
Imaginons que lors de cette répétition, on ait tiré 19 fois une boule rouge. Si on choisit Rouge comme succès, la fréquence de succès observée dans cet échantillon est alors $f=\frac{19}{50}=0,38$.

Rq : - Si, dans l'exemple précédent, on avait effectué des tirages sans remise, la répétition de l'expérience n'aurait pas été effectuée de façon indépendante (avoir obtenu une boule rouge lors du 1^er^ tirage aurait influencé le résultat du second tirage car dans ce cas, il y aurait eu « moins de chance » de tirer une boule rouge). Cependant, lorsqu'on réalise un sondage par exemple, les personnes interrogées sont choisies au hasard et ne peuvent pas être interrogées plusieurs fois. Un échantillon de taille n de la population correspond alors au « tirage » sans remise de n individus dans cette population. Néanmoins, si le sondage concerne une population de grande taille, on admet que l'on peut assimiler le procédé à un tirage avec remise.
- Voir l'annexe pour comprendre comment simuler une expérience aléatoire à deux issues.

2. Fluctuation d'échantillonnage

On reprend l'exemple de l'urne. On sait que dans l'urne, il y a au total 10 boules dont 4 sont rouges. La probabilité de tirer une boule rouge est donc $p=0,4$.

Le tableau suivant donne la fréquence de succès observée (c'est-à-dire la fréquence de boules rouges) sur dix échantillons différents de taille 50.

Numéro de l'échantillon	1	2	3	4	5	6	7	8	9	10
Nombre de boules rouges	18	20	15	21	19	18	23	18	17	25
Fréquence de succès f	0,36	0,40	0,30	0,42	0,38	0,36	0,46	0,36	0,34	0,5

On remarque que la fréquence de succès observée varie d'un échantillon à l'autre et n'est en général pas égale à p. On dit que f fluctue autour de p ; il s'agit de la **fluctuation d'échantillonnage.**

Rq : Si on ne réalise qu'une seule fois l'expérience aléatoire et que l'on obtient donc un échantillon de taille 1, la fréquence observée de succès est égale à 0 ou à 1. Intuitivement, on sent que plus la taille de l'échantillon est grande, plus il y a de chances que f et p soient proches.

II) Estimation

1. Loi des grands nombres

Propriété : Lorsque la taille n de l'échantillon devient grande, sauf exception, la fréquence de succès f observée devient plus proche de p.

Rq : « Sauf exception » signifie que dans la très grande majorité des cas, on observe expérimentalement que la fréquence f devient proche de p. Il peut néanmoins arriver d'obtenir un échantillon constitué de n succès, ou inversement de n échecs : la fréquence f de succès est alors éloignée de p . Mais ce cas de figure est peu fréquent.

2. Estimer une probabilité ou une proportion

Principe de l'estimation : Soit p un nombre inconnu compris entre 0 et 1.

Considérons ces deux situations :

① On cherche à estimer la probabilité p d'un évènement A d'une expérience aléatoire ;

② On cherche à estimer la proportion p d'individus d'une population présentant un caractère donné.

Pour obtenir une estimation de p , on constitue un échantillon de taille n , avec n grand.

Dans la 1ère situation, cela consiste à répéter n fois l'expérience aléatoire de façon indépendante. On choisit pour succès la réalisation de l'évènement A et on calcule alors la fréquence de succès observée f dans l'échantillon.

Dans la 2ème situation, cela consiste à choisir au hasard n individus de la population. On calcule alors la fréquence f d'individus présentant le caractère étudié dans l'échantillon.

D'après la loi forte des grands nombres, pour n grand, f et p sont proches et on peut donc dire que f constitue une **estimation** de p .

Exple : Avant un référendum, le gouvernement d'un état souhaite estimer la proportion de votants qui répondront favorablement à la question posée. Pour cela, un institut organise un sondage auprès de 1200 individus de cette population. 636 personnes répondent favorablement. La fréquence observée est donc $f = \dfrac{636}{1200}$, soit $f = 0,53$.

Ainsi, une estimation de la proportion de votants favorables est $0,53$.

Rq : L'estimation trouvée dépend de l'échantillon considéré, donc il y a plusieurs estimations possibles de p .

On utilise d'ailleurs généralement plusieurs échantillons de même taille n pour réaliser une bonne estimation de p .

Un résultat mathématique permet d'affirmer que sur un grand nombre d'échantillons sur lesquels on a calculé la fréquence observée f , l'écart entre p et f (et donc l'erreur commise en approchant p par f) est inférieur ou égal à $\dfrac{1}{\sqrt{n}}$ dans au moins 95% des échantillons. Ainsi, pour n grand, il y a plus de 95% de chances que f soit une bonne estimation de p .

ANNEXE

Comment simuler informatiquement une expérience aléatoire à deux issues x_1 et x_2, de probabilité respectives p et $1-p$?

La méthode consiste à générer un nombre réel aléatoire compris entre 0 et 1.

On modélise que : ▪ le résultat de l'expérience est x_1 si le nombre aléatoire obtenu est inférieur ou égal à p .

▪ le résultat de l'expérience est x_2 sinon.

Rq : Cela repose sur le fait que lorsqu'on génère aléatoirement un nombre compris entre 0 et 1, la probabilité que le nombre obtenu soit inférieur ou égal à p (c'est-à-dire compris entre 0 et p) est égale à : $\dfrac{\text{longueur de l'intervalle } [0;p]}{\text{longueur de l'intervalle } [0;1]} = \dfrac{p-0}{1-0} = p$.

Pour mettre en œuvre cette propriété, il faut disposer de commandes permettant de générer un nombre réel aléatoire entre 0 et 1 (exclus) dans différents langages :

- Avec le langage naturel : **nombre aléatoire entre 0 et 1**
- Avec la calculatrice : **NbrAléat** (pour le TI) et **Ran#** (pour les Casio)
- Avec le tableur : **ALEA()**
- Avec Python : **random()** à condition d'avoir écrit **from random import*** en début de programme.