14

Echantillonnage

Au programme

- √ Simuler un échantillon
- √ Observer la loi des grands nombres à l'aide de simulations
- √ Estimer une probabilité ou une proportion.

Table des matières

I Etude d'un exemple

On a lancé 24 fois un dé bien équilibré et on a obtenu les résultats suivants :

Nombre obtenu	1	2	3	4	5	6	Total
Nombre d'apparitions	4	3	6	2	5	4	24
Fréquence d'apparition	$\frac{4}{24}=0,166\dots$	$\frac{3}{24}=0,125$	$\frac{6}{24}=0,25$	$\frac{2}{24}0,083\dots$	$\frac{5}{24}$	$\frac{4}{24}=0,166\dots$	1

Les nombres écrits en troisième ligne du tableau ne sont bien sûr pas des probabilités (les six probabilités de chacun des six événements sont égales à $\frac{1}{6}$), mais ce sont des fréquences. **Une probabilité n'est pas une fréquence**. Une probabilité est un nombre théorique abstrait qui essaie de « prédire l'avenir », ce qui n'est pas le cas d'une fréquence qui « mesure du concret ».

On recommence l'expérience qui consiste à lancer plusieurs fois un dé bien équilibré. Mais cette fois-ci, on a lancé $240\,000$ fois ce dé (cette expérience peut être simulée par un ordinateur) et on a obtenu les résultats suivants :

Nombre obtenu	1	2	3	4	5	6	Total
Nombre d'apparitions	39 764	40 021	40 753	39 801	39 583	40 078	240 000
Fréquence d'apparition	$0,165\,6\dots$	$0,166\,7\dots$	$0,169\,8\dots$	$0,165\,8\dots$	$0,164\,9\dots$	$0,166\,9\dots$	1

Les fréquences d'apparition sont beaucoup plus proches les unes des autres, toutes égales à un peu plus que $0,16$. Ces fréquences sont en fait toutes proches de la probabilité théorique qui est $p=\frac{1}{6}=0,166\,666\,6\dots$

Un théorème de mathématiques (hors de portée d'un(e) élève de seconde) dit que ce qui vient de se passer « n'est pas un hasard » car quand on effectue un grand nombre de fois la même expérience, les fréquences observées sont proches de la probabilité théorique (et si on pouvait faire une infinité d'expérience, la fréquence observée serait la probabilité théorique). Ce théorème est connu sous le nom de **loi des grands nombres**. On va un peu expliquer ce théorème dans les paragraphes suivants.

II Echantillonnage

A Définition d'un échantillon

On s'intéresse à une expérience aléatoire à deux issues (par exemple, obtenir le 1 ou ne pas obtenir le 1 après un lancer de dé bien équilibré).

L'une des issues est conventionnellement appelée « **succès** » et a une probabilité p (par exemple, $p=\frac{1}{6}$ si le succès est « obtenir le 1 » après un lancer de dé bien équilibré).

L'autre issue est conventionnellement appelée « **échec** » et a une probabilité $1-p$ (par exemple, $1-p=\frac{5}{6}$ si l'échec est « ne pas obtenir le 1 » après un lancer de dé bien équilibré).

Cette expérience peut être répétée, **de manière indépendante** ou pas. Par exemple, dans une urne contenant un certain nombre de boules blanches et un certain nombre de boules noires, on tire une boule et on obtient une blanche. Si on recommence cette expérience en remettant d'abord la boule blanche, la probabilité d'obtenir une boule blanche au deuxième tirage est la même qu'au premier tirage. Dans ce cas, on a effectué des tirages **de manière indépendante**. Si par contre, on ne remet pas la boule blanche tirée en premier, la probabilité d'obtenir une boule blanche a changé car les proportions de boules blanches et noires ne sont plus les mêmes. Dans ce cas, les deux tirages ne se font pas de manière indépendante.

Définition 1

Soit n un entier naturel non nul.

On effectue n fois une même expérience aléatoire, **de manière indépendante**, c'est-à-dire de manière à ce que les différents résultats obtenus à chaque tentative n'aient aucune influence sur les résultats suivants.

La liste des n résultats obtenus s'appelle un **échantillon de taille** n.

Définition 2

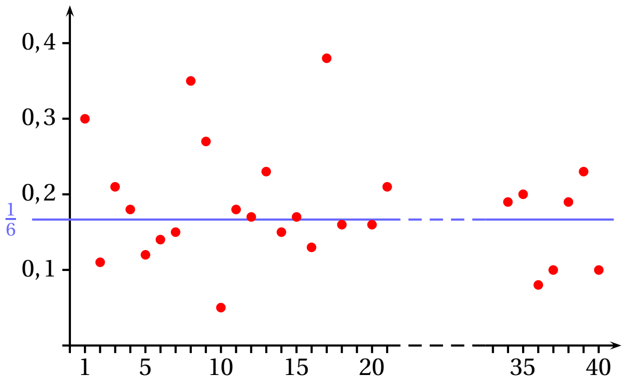
Soit k un entier naturel tel que $0 \leqslant k \leqslant n$.

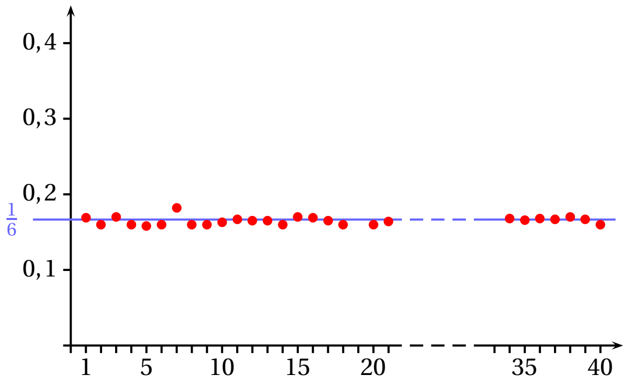
Si on a obtenu k succès sur un échantillon de taille n, la **fréquence observée de succès** est $f = \dfrac{k}{n}$.

B Fluctuation d'échantillonnage

Dans l'exemple du premier paragraphe, on a effectué une fois l'expérience qui consistait à lancer 24 fois un dé à six faces et aussi on a effectué une fois l'expérience qui consistait à lancer $240\,000$ fois un dé à six faces. De manière plus générale, on peut effectuer N fois l'expérience qui consiste à lancer n fois un dé à six faces bien équilibré. On a ainsi à disposition N échantillons de taille n. On peut s'intéresser à la fréquence d'apparition du 1 dans chacun des N échantillons.

Voici deux graphiques, l'un avec $\mathrm{N} = 40$ et $n = 24$ et un avec $\mathrm{N} = 40$ et $n = 240\,000$ (obtenus par simulation sur ordinateur).
En abscisse, on a le numéro de l'échantillon (qui est donc un numéro entre 1 et $\mathrm{N} = 40$) et en ordonnée, on a la fréquence observée d'apparition du 1 dans l'échantillon correspondant.





Les fréquences observées sur un échantillon de taille n évoluent ou encore **fluctuent** autour de la probabilité théorique $p = \dfrac{1}{6}$. Ce phénomène s'appelle la **fluctuation d'échantillonnage**.

On constate aussi que cette fluctuation semble diminuer quand n augmente et que cette fluctuation semble petite quand n est grand.

C La loi des grands nombres

On peut démontrer à plus haut niveau le théorème qui suit, appelé loi des grands nombres (en toute rigueur, loi faible des grands nombres et il y a aussi une loi forte des grands nombres). L'énoncé ci-dessous est très imprécis et flou. Ce n'est pas le vrai énoncé de la loi des grands nombres. Mais il n'est pas possible de l'améliorer dans une classe de seconde.

Theoreme 1

Lorsque la taille n de l'échantillon est grande, sauf exception, la fréquence observée est proche de la probabilité p.

L'énoncé qui précède est l'énoncé du programme officiel. C'est une version vulgarisée de la loi des grands nombres. Les mots « grande », « proche », « sauf exception » sont très imprécis et n'ont pas de réelle signification. Mais encore une fois, en classe de seconde, il n'est pas possible d'être plus précis. Explicitons un peu plus le morceau de phrase « sauf exception ». Quand on lance $240\,000$ fois un dé à six faces bien équilibré, il est toujours possible de n'obtenir que des 1 par exemple ce qui fournirait une fréquence $f = 1$, très éloignée de la probabilité $p = \dfrac{1}{6}$. Mais la probabilité que ceci se produise est très faible et dans la grande majorité des cas, si on lance $240\,000$ fois un dé à six faces bien équilibré, la fréquence d'apparition du 1 sera très proche de $p = \dfrac{1}{6}$.

III Estimation

A Estimation d'une probabilité

On considère une certaine expérience aléatoire puis A un certain événement associé à cette expérience aléatoire. On veut estimer la probabilité p de cet événement, p étant inconnue.

Le principe de l'estimation est le suivant :

- On choisit un entier n assez grand (taille de l'échantillon) puis un entier naturel non nul N (nombre d'échantillons).
- on effectue une première fois n simulations de l'expérience (par ordinateur ou par soi même si on a du temps devant soi), le succès étant « l'événement A est réalisé » et on calcule la fréquence observée de succès f_1.
- on recommence jusqu'à obtenir N échantillons de taille n et on obtient des fréquences observées de succès $f_1, f_2, \ldots, f_{\mathrm{N}}$.
- La plupart des fréquences observées sont proches de la probabilité à déterminer, ce qui permet d'en donner une **estimation**.

Par exemple, on lance 10 fois une pièce de monnaie bien équilibrée et on veut une estimation de la probabilité de l'événement « on obtient 5 fois Pile ».

On fait effectue 10 fois ($\mathrm{N} = 10$) l'expérience qui consiste à effectuer $10\,000$ fois ($n = 10\,000$) dix lancers d'une pièce de monnaie bien équilibrée. On obtient ainsi 10 échantillons de taille 500. On obtient des fréquences observées f_1, $f_2, \ldots, f_{10}$.

n^{o} de l'échantillon	1	2	3	4	5	6	7	8	9	10
Fréquence	$0,237$	$0,245$	$0,247$	$0,241$	$0,269$	$0,248$	$0,247$	$0,244$	$0,247$	$0,248$

Une bonne estimation de p semble être $p = 0,24$.

On peut montrer à plus haut niveau que la probabilité exacte est

$$p = \frac{10 \times 9 \times 8 \times 7 \times 6}{5 \times 4 \times 3 \times 2} \times \frac{1}{2^{10}} = \frac{3 \times 7 \times 3}{2^8} = \frac{63}{256} = 0,246\ldots$$

B Estimation d'une proportion

On veut cette fois-ci estimer dans une population de grande taille (par exemple, la population française ou aussi le parc automobile américain) la proportion d'individus possédant un certain caractère (par exemple, le nombre de français votant pour un certain(e) candidat(e) ou le nombre de voitures du parc américain ayant déjà eu un accident).

Le principe est le même que pour estimer une probabilité à la nuance près que les échantillons sont prélevés directement dans la population (en effectuant des sondages). Les fréquences observées sur différents échantillons fournissent une bonne estimation de la proportion.

C Précision de l'estimation

Quand on cherche à estimer une probabilité ou une proportion p à partir de l'observation d'une fréquence f dans un échantillon de taille n, il est intuitivement clair que l'estimation sera d'autant plus précise que n sera choisi grand.

On peut démontrer à plus haut niveau que dans une grande majorité des cas (environ 95%), on a

$$f - \frac{1}{\sqrt{n}} \leqslant p \leqslant f + \frac{1}{\sqrt{n}}$$

ou encore $|p - f| \leqslant \dfrac{1}{\sqrt{n}}$. $|p - f|$ est l'erreur commise quand on approche p par f. Elle est inférieure à $\dfrac{1}{\sqrt{n}}$.

Dans l'exemple du paragraphe « estimation d'une probabilité », nous avions choisi des échantillons de taille $n = 10\,000$. Puisque

$$\frac{1}{\sqrt{10\,000}} = 0,01$$

la plupart des fréquences observées ne s'écartaient pas à plus de $0,01$ de la valeur exacte de la probabilité théorique p.

