

3. Densités naturelle et logarithmique

Si l'on nomme densité naturelle, pour un ensemble A d'entiers, la limite (à condition qu'elle existe), quand n tend vers l'infini, du quotient [nombre d'éléments de A inférieurs à n]/ n , on trouve que la densité naturelle des nombres pairs est $1/2$, comme celle des nombres impairs. Plus généralement, la densité naturelle des termes d'une suite arithmétique $an + b$ ($n = 1, 2, \dots$) existe toujours et vaut $1/a$. Malheureusement, les ensembles pour lesquels la densité naturelle est définie sont rares. Par exemple, l'ensemble des entiers qui, écrits en base 2, ont un nombre pair de chiffres n'a pas de densité naturelle. De plus, la réunion de deux ensembles ayant des densités naturelles n'en a pas né-

cessairement, ce qui empêche en particulier de considérer cette densité comme une probabilité.

Une découverte provient de l'étude de la densité logarithmique définie pour un ensemble A d'entiers par : $D_{\log}(A) = \lim_{n \rightarrow \infty} (\sum \{1/k; k \text{ dans } A, k \leq n\} / \sum \{1/k; k \leq n\})$.

En d'autres termes, on retient $1/k$ pour chaque entier k présent dans A et inférieur à n ; on additionne ces $1/k$ et l'on normalise en divisant par la somme des $1/k$ pour k inférieur à n ; cela donne un poids d_n à l'ensemble A , dont on prend la limite (si elle existe) quand n tend vers l'infini pour tenir compte de tous les éléments de A quand l'ensemble est infini.

On a démontré que si l'ensemble d'entiers A possède une densité naturelle, alors A possède aussi une densité logarithmique et elles sont identiques. Autrement dit, en attribuant le poids $1/n$ à l'entier n , et en faisant tendre n vers l'infini après normalisation, on étend la notion de densité naturelle sans en perdre les bonnes propriétés. C'est utile car, avec cette extension, des ensembles plus nombreux ont maintenant une densité, c'est-à-dire une sorte de probabilité, et on a préservé l'idée que les nombres pairs et les nombres impairs avaient une densité $1/2$.

Notons que l'on ne peut pas poser : $D_{\log}(A) = \sum \{1/k; k \text{ dans } A\} / \sum \{1/k; k \text{ entier}\}$, car la série harmonique $1 + 1/2 +$

$\dots + 1/k$ tend vers l'infini quand k tend vers l'infini.

Dans un cas concret, s'il n'y a qu'un nombre fini d'objets à prendre en compte, par exemple le nombre des mots dans un texte (fini), on peut choisir d'attribuer un poids C/n à chaque entier n et fixer C pour que le total des probabilités soit égal à 1, cela sans qu'il faille faire tendre n vers l'infini.

Dans le cas infini, la même idée est presque utilisable et c'est ce que propose la densité logarithmique. Autrement dit, affecter le poids C/n à chaque entier est un choix naturel, car il permet de définir une notion de densité qui étend la densité naturelle à chaque fois qu'elle est définie.

à avoir remarqué la nature hyperbolique de la fréquence d'usage des mots fut le sténographe français J.-B. Estoup. »

Passons sur cette navrante iniquité et indiquons que, depuis, on a vu ou cru voir des lois de Zipf dans un nombre considérable de situations concrètes. La loi semblerait concerner aussi bien la taille des villes classées par population décroissante, le nombre de citations que reçoit un article scientifique, le nombre d'articles qu'écrit un chercheur durant sa carrière, le nombre d'espèces par genre dans la classification des êtres vivants, le nombre de visites que reçoit une page Internet, l'expression des gènes, la musique, les cratères lunaires, les tremblements de terre, les taches solaires, etc.

On a généralisé la loi de Zipf par une loi de type $f_n = C/n^e$, où e est un exposant proche de 1. On la nomme aussi loi de puissance ou loi de Pareto (même quand e n'est pas proche de 1). Nous ne considérerons que les situations où des objets sont classés par ordre de fréquence décroissante et où la fréquence trouvée pour l'objet en position n est donc assimilable à une probabilité associée au nombre entier n . La régularité avec laquelle apparaissent des distributions de probabilités de type C/n^e (e proche de 1) pour les entiers

suggère qu'à la place de l'impossible probabilité uniforme sur les entiers, nous devrions, quand rien ne s'y oppose, considérer que les entiers ont une probabilité naturelle de la forme C/n^e (e proche de 1).

Récemment, grâce à la puissance des systèmes informatiques, deux nouveaux types de données, que nous nommerons « nuage de Sloane » et « nuage des cinq millions de livres », ont été constitués et ont appuyé l'idée qu'il y a quelque chose de particulier concernant les distributions de probabilité en C/n^e pour e proche de 1. Il semble de plus en plus certain que les distributions de probabilité sur les entiers de type loi de Zipf jouent un rôle spécifique et nullement fortuit.

Le nuage de Sloane

Le nuage de Sloane est obtenu en comptant le nombre d'occurrences de l'entier n dans l'encyclopédie des suites numériques de Neil Sloane (<http://oeis.org/>). Chacune des 200 000 suites de l'encyclopédie, que N. Sloane réunit depuis 1965 avec l'aide de la communauté mathématique, est stockée (on n'y garde qu'un nombre limité de termes, environ 150 caractères pour chaque suite). Seules les suites présentant un intérêt mathématique

sont retenues. La base de suites – qui est aussi une base de nombres entiers – associe une probabilité à chaque nombre entier, déduite de son nombre d'apparitions dans la base : le nombre d'occurrences d'un entier dans la base est une mesure de son importance mathématique et peut-être de sa probabilité objective d'apparition, si une telle notion a un sens. Le nombre d'occurrences de l'entier n dans la base, notons-le $Sloane(n)$, ne suit pas une courbe décroissante régulière. Cela est dû à ce que les mathématiciens concentrent leur attention sur certains entiers plus que sur d'autres. Les puissances de 2 (2, 4, 8, 16, ...) sont par exemple nettement favorisées, de même que les nombres premiers, ou les nombres ayant beaucoup de facteurs.

La courbe $Sloane(n)$, que nous nommons nuage de Sloane (voir l'encadré 2), doit être vue comme une représentation de l'intérêt mathématique relatif des nombres entiers. L'étude de ce nuage, suggérée par Philippe Guglielmetti, a été menée dans un article paru en 2011. Sous sa forme de nuage, ou redessinée pour classer les entiers par nombres d'occurrences décroissantes, on a affaire à une courbe proche de $C/n^{1.3}$, une loi de Zipf d'exposant 1,3. C'est tout à fait remarquable et la justification théorique de