

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



4. De Benford à Zipf, et inversement

La loi de Benford indique qu'en prenant des nombres au hasard en grande quantité (par exemple des données géographiques), les nombres commenceront plus fréquemment par 1 que par 2, par 2 que par 3, etc. La probabilité $p(i)$ associée au chiffre i (pour $i = 1, 2, \dots, 9$) est $\log_{10}(1 + 1/i)$:

$p(1) = 30,1\%$, $p(2) = 17,6\%$, $p(3) = 12,5\%$, $p(4) = 9,7\%$, $p(5) = 7,9\%$, $p(6) = 6,7\%$, $p(7) = 5,8\%$, $p(8) = 5,1\%$, $p(9) = 4,6\%$.

La loi de Zipf, selon laquelle la distribution en C/n est une distribution naturelle pour des entiers, est liée d'une façon étonnante à la loi de Benford. D'une part, « Zipf »

donne « Benford » : la densité logarithmique, qui est fondée sur l'attribution du poids $1/n$ à l'entier n , est une traduction mathématique rigoureuse de la loi de Zipf généralisant la densité naturelle ; or elle attribue, comme la loi de Benford, la densité $\log_{10}(1 + 1/i)$ aux entiers commençant par i ($i = 1, \dots, 9$).

D'autre part, « Benford » donne « Zipf » : la loi de Benford pour une base de numération d quelconque indique que l'entier i est le premier chiffre d'un entier n avec une probabilité $\log_d(1 + 1/i)$, ce qui, lorsque i est assez grand, est en gros proportionnel à C/i , avec $C = 1/\log(d)$, ce qu'indique de son côté la loi de Zipf.

cette distribution que nous allons présenter établit un lien entre probabilité naturelle sur les entiers, complexité de Kolmogorov et loi de Zipf, ce que l'on peut voir comme un fondement théorique à la loi de Zipf.

La complexité de Kolmogorov $K(n)$ d'un entier est la taille du plus petit programme qui engendre n (dans un langage de programmation fixé). Les entiers ayant une définition simple (par exemple $2^{1000000}$) ont une faible complexité de Kolmogorov et ont en conséquence beaucoup de définitions assez simples. La théorie indique aussi que $K(n)$ est lié à la probabilité qu'un programme tiré au hasard produise n , par la relation $\text{Proba}[n] \approx 1/2^{K(n)}$, et donc à la probabilité que les suites mathématiques collectées dans l'encyclopédie de N. Sloane contiennent cet entier un grand nombre de fois. Cette dernière remarque suppose qu'on assimile ces suites numériques à des programmes, ce qui n'a rien d'absurde. Autrement dit, le nuage de Sloane est une version mathématique, concrète et humaine de la courbe représentant la fonction $1/2^{K(n)}$ déterminée par la complexité des entiers. Or la théorie indique que cette courbe suit une loi de Zipf, car $K(n)$ vaut typiquement $\log_2(n)$ et que $1/2^{\log_2(n)} = 1/n$. La complexité de Kolmogorov indiquait par avance que l'on a une loi de Zipf pour le nuage de Sloane, ce que l'étude statistique du nuage a confirmé.

Tout cela est remarquable et nous aide à comprendre la loi de Zipf. Même si d'autres mécanismes engendrent une loi de Zipf, celui lié à la théorie de la complexité est particulièrement frappant, puisqu'il indique un lien avec une distribution de probabilité naturelle fondée sur la complexité des entiers, lien

validé par l'étude du nuage de Sloane. Cette analyse est encore renforcée par l'étude d'un autre nuage étonnant, indépendant du premier, et qui est cette fois tiré d'une base de données de cinq millions de livres numérisés par la Société Google et exploitée statistiquement par une équipe de chercheurs de l'Université Harvard réunie autour de Jean-Baptiste Michel. L'article de cette rubrique dans *Pour la Science* d'août 2011 présentait cette mise à disposition et la science naissante qu'on prétend en déduire, dénommée « culturomique ».

Le nuage des cinq millions de livres

Jim Fowler, mathématicien à l'Université de l'Ohio, aux États-Unis, a calculé et représenté l'équivalent du nuage de Sloane pour les cinq millions de livres de la base de données de Google. Il a compté le nombre d'occurrences de chacun des nombres entiers de 1 à 10 000 et a dessiné les 10 000 points correspondants. La courbe obtenue (voir l'encadré 2) présente le même aspect général que le nuage de Sloane. Cependant, plusieurs différences sautent aux yeux... et s'expliquent facilement.

On observe notamment une sorte de sursaut violent de la courbe avant et du côté de l'entier 2 000. Cette perturbation correspond évidemment aux nombres désignant des années proches, car ces nombres sont cités fréquemment dans les textes des livres. Il s'agit d'un effet temporel et la même courbe tracée en 2100 montrera un sursaut équivalent du côté de 2 100. Un autre effet, cette fois « décimal », est aussi très net : les nombres ronds comme 1 000, 5 000, 8 000

ou même 7 500, 8 500, etc., sont plus cités que leurs voisins et se retrouvent donc au-dessus de la courbe générale. Notre usage du système décimal et la pratique des arrondis faussent la fréquence d'utilisation des nombres entiers en favorisant de manière prévisible certains d'entre eux.

L'étude détaillée du nuage montre d'autres étrangetés pas toujours aussi faciles à expliquer, mais qui proviennent de conventions, de codes ou de règles d'usage favorisant certains entiers. Malgré ces effets perturbateurs divers, la courbe (éventuellement redessinée pour classer les nombres par fréquence décroissante) se conforme très bien à une loi de Zipf d'exposant e proche de 1, comme celle du nuage de Sloane.

La comparaison des deux courbes est amusante et riche d'enseignement : là où l'intérêt mathématique favorisait les puissances de 2 et les nombres avec de nombreux diviseurs, c'est maintenant les conventions liées à la numération décimale ou à la mention des années du calendrier qui créent les plus fortes déviations. Aussi bien le nuage « mathématique » de Sloane que le nuage « culturel » des cinq millions de livres exhibent un tracé conforme à une loi en C/n^e . Ces traitements sur des données massives confirment que les répartitions de probabilité en C/n^e (e proche de 1) tiennent un rôle spécifique et doivent être retenues par défaut comme probabilités naturelles sur les entiers. Deux arguments de nature mathématique vont s'ajouter à celui de la complexité de Kolmogorov et justifier notre affirmation.

Un premier argument porte sur la densité naturelle (voir l'encadré 3). Un second