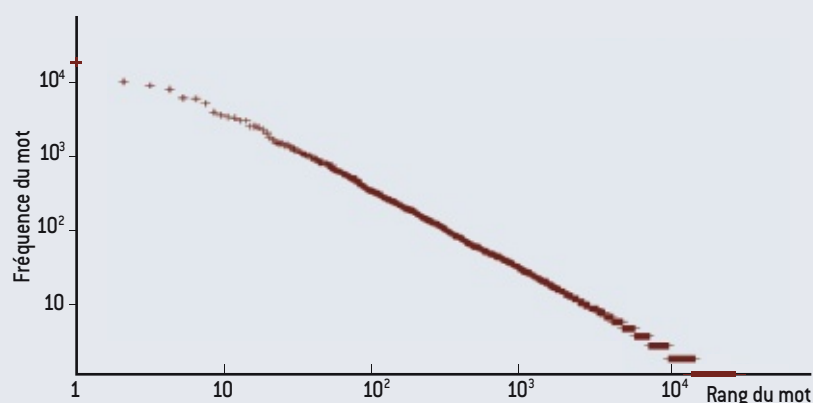


2. La loi de Zipf

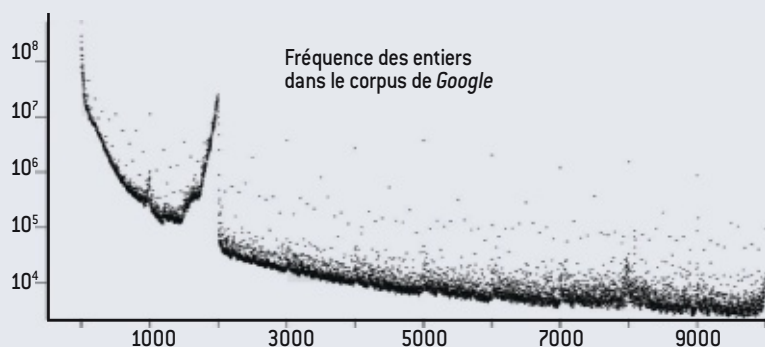
Si l'on classe les mots d'un texte par fréquence décroissante, on constate le plus souvent que la fréquence du mot en position n est C/n , où C est une constante. Cette loi, généralisée sous la forme C/n^e (où e est proche de 1), porte le nom de loi de Zipf.

Le travail de Zipf concernait le roman *Ulysse* de James Joyce. On retrouve la loi de Zipf dans de nombreux domaines, bien au-delà de la linguistique. Une série d'arguments présentés dans le texte de l'article conduisent à la voir comme une forme de loi, aussi équitable que possible, sur les entiers.



Analogue au nuage de Sloane de la page précédente, mais totalement indépendante, la courbe ci-dessous a été obtenue par Jim Fowler en utilisant les données de cinq millions de livres numérisés par la Société Google. Comme pour l'encyclopédie de Sloane, on a compté le nombre d'occurrences des entiers écrits en chiffres dans l'énorme base de textes de plus de 100 milliards de mots, constituée par le projet Culturomique autour de Jean-Baptiste Michel, de l'Université Harvard.

L'allure générale du « nuage des cinq millions de livres » est semblable à celle du nuage de Sloane. On a affaire à une loi de Zipf. Les biais sociaux produisent des effets de nature différente. Les nombres qui sont mentionnés en excès le sont non plus pour des raisons d'intérêt mathématique reconnu, mais pour des raisons contextuelles ou liées à l'usage du système décimal, qui favorise par exemple les nombres ronds (10, 50, 100, 1 000, 1 500, 5 000, etc.). Entre 1 et 100, un graphique à plus grande échelle ferait apparaître que le nombre 12 et ses multiples sont favorisés, ce qui est lié au rôle particulier du nombre 12 dans notre culture.



pas équitable, puisqu'il y a autant de nombres pairs que de nombres impairs et que ce sont deux ensembles d'entiers de même structure. Attribuer le poids $1/2^n$ à n n'est sans doute pas une bonne idée, mais il y a bien d'autres choix possibles pour $\{a_n\}$. N'y a-t-il pas, parmi toutes les séries $\{a_n\}$ utilisables pour « peser » les entiers et les ensembles d'entiers, certaines séries plus naturelles ou meilleures que d'autres ? Si oui, comment les trouver ? L'exigence que les nombres pairs et impairs aient chacun le poids $1/2$ servira de critère.

George Zipf compte les mots

La recherche de ce substitut à l'impossible probabilité uniforme sur les entiers semble avoir une solution. Nous allons voir qu'elle est liée à la théorie de la complexité de Kolmogorov, qui est la tentative mathématique la plus générale pour définir la notion de simplicité. Mais avant d'explorer les aspects théoriques et peut-être spéculatifs de cette solution, laissons-nous guider par des données concrètes.

En 1949, George Zipf examine la fréquence d'usage des mots dans le roman *Ulysse* de James Joyce. L'œuvre comporte 260 430 mots, dont 29 899 mots différents. Le mot le plus courant, *the*, apparaît environ deux fois plus que le deuxième (*of*), trois fois plus que le troisième, ..., 100 fois plus que le centième, etc. Dit autrement : en classant les mots par fréquence décroissante, la fréquence f_n du mot classé en position n est environ C/n , où C est une constante. Cette remarque qui s'étend à d'autres textes et d'autres langues est à l'origine du nom « loi de Zipf » donné à ce type de relations. Si des objets, pris dans un ensemble comportant éventuellement des répétitions et classés par fréquence décroissante $f_1, f_2, \dots, f_n, \dots$, vérifient approximativement une relation du type $f_n = C/n$, on dit qu'ils suivent une loi de Zipf.

L'histoire est souvent injuste. En effet, dès 1916, le Français Jean-Baptiste Estoup [1886-1950] avait fait une observation équivalente en étudiant des textes français. Zipf le savait puisqu'il indique dans son livre de 1949 : « La première personne (à ma connaissance) »

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 : $D_{\log}(A) =$

$$\lim_{n \rightarrow \infty} \left(\frac{\sum \{1/k ; k \text{ dans } A, k \leq n\}}{\sum \{1/k ; k \leq n\}} \right).$$

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