



1. Probabilités des entiers dans l'encyclopédie de Neil Sloane

Il est impossible d'associer une probabilité uniforme aux entiers (probabilité qui donnerait à chaque entier la même probabilité). En effet, la probabilité d'un sous-ensemble des entiers est égale à la somme des probabilités de chaque élément : si la probabilité d'un entier est non nulle, la probabilité de l'ensemble des entiers est infinie, ce qui ne convient pas pour une probabilité ; si elle est nulle, alors la probabilité de l'ensemble des entiers est nulle aussi, ce qui ne convient pas mieux. Pour traiter des probabilités d'ensembles d'entiers, il faut attribuer des probabilités différentes aux entiers, de la meilleure façon possible. L'étude des entiers présents dans l'encyclopédie de Sloane est un guide pour trouver cette meilleure attribution.

L'encyclopédie des suites numériques de Neil Sloane contient un

grand nombre de fois l'entier 1, un peu moins souvent l'entier 2, etc. La courbe, notons-la $Sloane(n)$, donnant le nombre d'occurrences de n en fonction de n ressemble à un nuage de points globalement décroissants. Classés par nombre d'occurrences décroissantes, on a une loi de type $C/n^{1,3}$, où C est une constante. Le nuage de Sloane est lié à la complexité de Kolmogorov et cela justifie que l'on retrouve une loi de Zipf. La courbe $Sloane(n)$ est une version approchée de la courbe $C/2^{K(n)}$, où $K(n)$ désigne la complexité de Kolmogorov de l'entier n , qui vaut typiquement $\log_2(n)$: quand n est « complexe », $K(n)$ est grand (n est impossible à définir simplement), $1/2^{K(n)}$ est petit et donc le point correspondant de $Sloane(n)$ est bas. C'est pourquoi aussi $1024 = 2^{10}$ est situé au-dessus des autres points de même ordre de grandeur : le fait

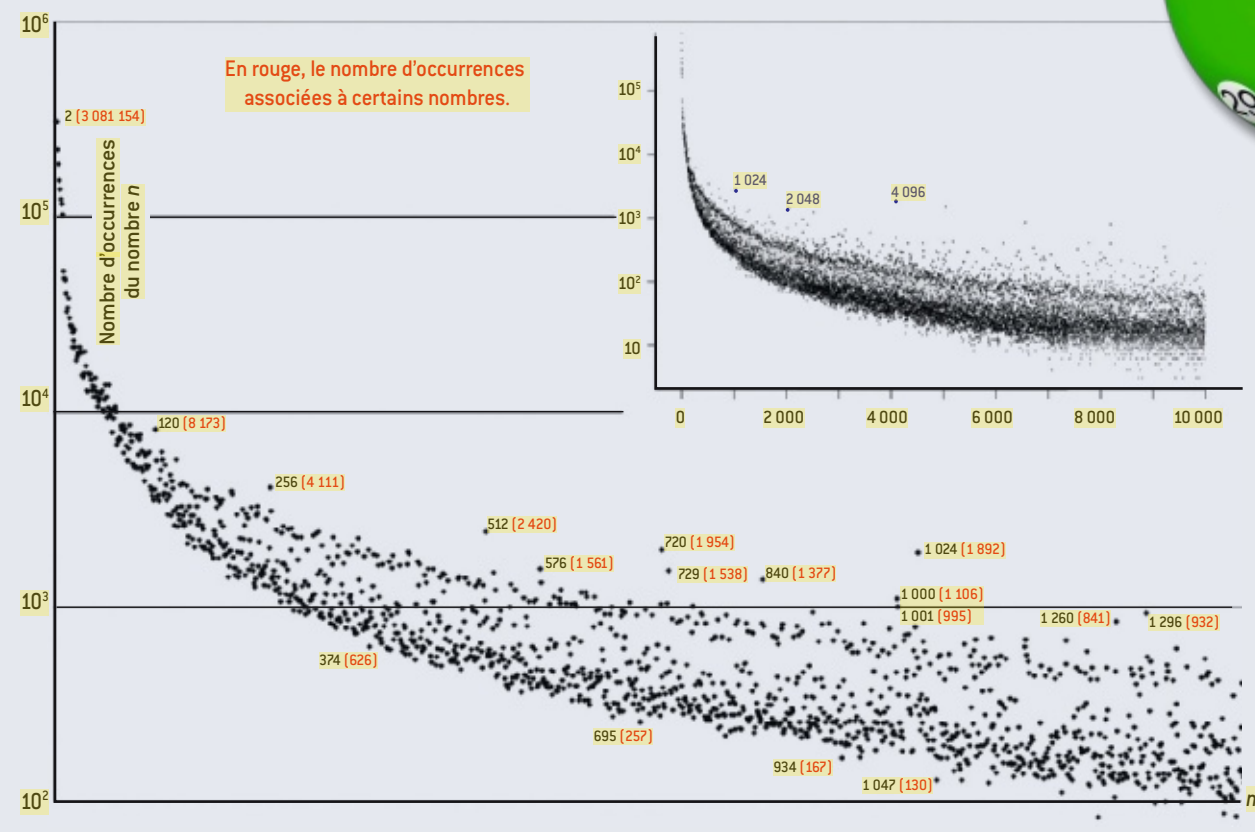
d'être une puissance de 2 a pour conséquence que sa complexité de Kolmogorov est faible et donc que son nombre d'occurrences dans la base (proportionnel à $1/2^{K(n)}$) est grand.

Cette représentation concrète de la complexité de Kolmogorov par le nuage de Sloane est légèrement biaisée par un effet culturel et social qu'on voit clairement sous la forme d'une zone claire, le fossé de Sloane, séparant le nuage en une partie inférieure et une partie supérieure.

Ce fossé étudié récemment (voir la bibliographie) est dû à ce que la communauté mathématique qui contribue à l'encyclopédie s'intéresse aux nombres entiers ayant une faible complexité de Kolmogorov (ce sont les nombres simples à définir et ayant plusieurs définitions), mais qu'elle s'y intéresse

d'une manière qui augmente la fréquence des nombres simples.

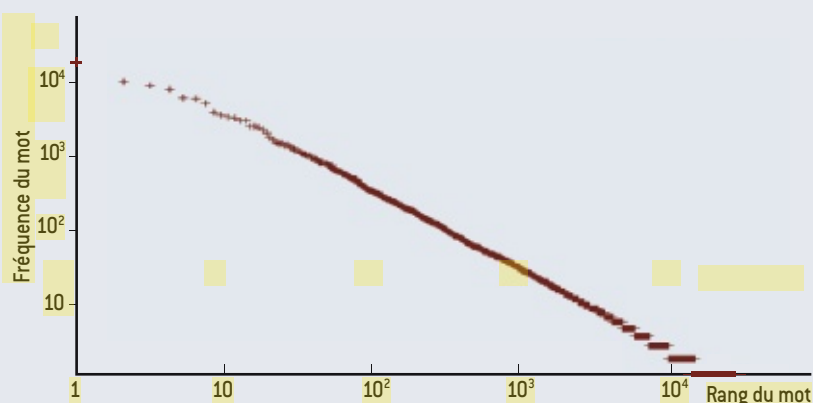
À cause des phénomènes culturels d'entraînements et de mode, les mathématiciens concentrent leur attention sur les nombres les plus simples, ce qui pousse les points correspondants de la courbe vers le haut, créant une zone évidée dans le graphe : le fossé.



2. La loi de Zipf

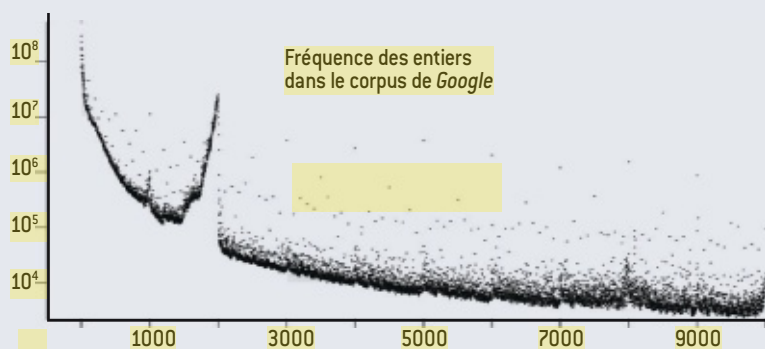
Si l'on classe les mots d'un texte par fréquence décroissante, on constate le plus souvent que la fréquence du mot en position n est C/n , où C est une constante. Cette loi, généralisée sous la forme C/n^e (où e est proche de 1), porte le nom de loi de Zipf.

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)