



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 $2^{1\,000\,000}$) 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