



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

résultat quasi miraculeux confirme que C/n est théoriquement la plus équitable attribution de poids aux entiers. La densité $D_e(A) = \sum \{1/k^e, k \text{ dans } A\} / \sum \{1/k^e, k \text{ entier}\}$ est toujours définie si $e > 1$, car maintenant le dénominateur est fini (car la série de terme $1/k^e$ est de somme finie quand $e > 1$).

Comme $1/k^e$ est proche de $1/k$ si e est proche de 1, pour contourner la divergence de la série $\sum 1/k$, on définit la densité $D_z(A)$ de l'ensemble A d'entiers comme la limite, si elle existe, de $D_e(A)$ quand e tend vers 1. On la nomme parfois « densité zêta » (à cause du dénominateur qui est la fonction zêta de Riemann), d'où le « z ». Les ensembles d'entiers A dont la densité D_z existe sont les mêmes que ceux pour lesquels la densité logarithmique (voir l'encadré 3) existe, et alors les deux densités coïncident. Ce résultat remarquable est publié dans la thèse de Persi Diaconis (voir la bibliographie).

Les deux façons envisageables d'étendre la notion de densité naturelle en se fondant sur l'idée d'attribuer (autant que possible) une probabilité $1/n$ à l'entier n fonctionnent et conduisent de manière inattendue à la même notion. Le sens de ces théorèmes mathématiques est que, moyennant un petit passage par des limites pour contourner la divergence de $\sum 1/n$, il est possible de formuler une définition de « choix aléatoire équitable » entre entiers. Ainsi, la loi de Zipf est dans une position privilégiée par rapport à toutes les autres lois envisageables sur les entiers.

Benford retrouvé

Un dernier théorème s'ajoute aux arguments pratiques et mathématiques. La loi de Benford apparaît quand on examine le premier chiffre d'une série de données : la probabilité pour qu'un nombre entier pris dans un ensemble assez grand (par exemple dans le tableau des longueurs de tous les fleuves mesurées en kilomètres) commence par le chiffre i ($i = 1, 2, \dots, 9$) est $\log_{10}(1 + 1/i)$ et il y a donc plus de 30 % de chances (car $\log_{10}(2) = 0,30103$) que le premier chiffre d'un tel nombre soit 1 (voir l'article de janvier 2007 dans cette rubrique).

Une question s'est posée depuis longtemps : la liste des entiers vérifie-t-

elle la loi de Benford ? Autrement dit, le quotient (Nombre d'entiers $< n$ commençant par le chiffre i)/ n a-t-il pour limite $\log_{10}(1 + 1/i)$? Il se trouve que non : ce quotient n'a pas de limite quand n tend vers l'infini, mais oscille sans cesse, car les nombres commençant par i n'ont pas de densité naturelle. Maintenant que nous savons qu'on peut étendre la densité naturelle, reformulons la question : l'ensemble des entiers commençant par i ($i = 1, 2, \dots, 9$) a-t-il une densité logarithmique (ou, ce qui revient au même, une densité D_z) ? Comme par miracle, la réponse est oui, et la densité trouvée est $\log_{10}(1 + 1/i)$. Les entiers, traités équitablement, suivent la loi de Benford et il n'est donc pas étonnant que cette loi se rencontre si fréquemment.

À défaut de mesure de probabilité uniforme sur les entiers, la loi de Zipf joue le rôle d'une mesure naturelle. On le voit en observant la prévalence de cette loi pour des données concrètes (telles les fréquences d'utilisation des mots dans un texte). On le voit quand on s'aperçoit que la loi de Zipf est liée à la complexité de Kolmogorov des entiers. On le voit aussi quand on recherche les densités qui, pour les ensembles d'entiers, jouent le rôle de probabilités uniformes.

La preuve mathématique qu'avec ces densités en C/n , les entiers eux-mêmes vérifient la loi de Benford (et c'est vrai aussi des nombres premiers), renforce la conviction que ces densités tirées de la loi de Zipf sont les bonnes façons de choisir au hasard des nombres entiers. Le fait que la loi de Benford redonne en un certain sens la loi de Zipf est un argument supplémentaire, car il existe des explications directes de la loi de Benford, telle celle proposée par Nicolas Gauvrit et l'auteur de cette rubrique, et qu'elles sont donc aussi des explications de la loi de Zipf.

Le monde mathématique est déconcertant : l'infini dénombrable, le plus simple de tous, semble interdire qu'on en pioche les éléments au hasard équitablement, alors que le continu de l'intervalle $[0, 1]$, plus gros et plus compliqué que l'infini dénombrable, l'autorise. Heureusement, la loi de Zipf, à sa façon, joue ce rôle de probabilité uniforme sur les entiers.

L'AUTEUR



J.-P. DELAHAYE est professeur à l'Université de Lille et chercheur au Laboratoire d'informatique fondamentale de Lille (LIFL).

BIBLIOGRAPHIE

N. Gauvrit, J.-P. Delahaye et H. Zenil, *Sloane's Gap : Do mathematical and social factors explain the distribution of numbers in the OEIS ?*, *Journal of Humanistic Mathematics*, à paraître, 2012.

N. Gauvrit et J.-P. Delahaye, *Pourquoi la loi de Benford n'est pas mystérieuse*, *Mathématiques et Sciences Humaines*, vol. 182-2, pp. 7-15, 2008 (<http://msh.revues.org/10363>).

M. Newman, *Power laws, Pareto distributions and Zipf's law*, *Contemporary Physics*, vol. 46, pp. 323-351, 2005.

G. Tenenbaum, *Introduction to Analytic and Probabilistic Number Theory*, Cambridge University Press, 1995.

P. Diaconis, *Weak and Strong Averages in Probability and Theory of Numbers*, thèse de l'Université Harvard, 1974, http://www-stat.stanford.edu/ffcgates/PERSI/Persi_PhDdiss.pdf (c'est le document à lire sur les densités étendant la densité naturelle).

G. Zipf, *Human Behaviour and the Principle of Least Effort*, Addison-Wesley Publishing, 1949.