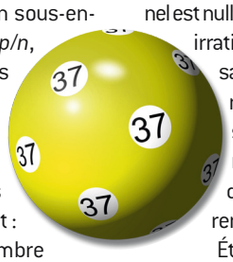
LOGIQUE & CALCUL

Les entiers ne naissent pas égaux

Il est impossible de définir une loi de probabilité uniforme sur l'ensemble des nombres entiers. Ce fait est étroitement lié à la loi de Zipf, une loi statistique dont les manifestations sont innombrables.

Jean-Paul DELAHAYE

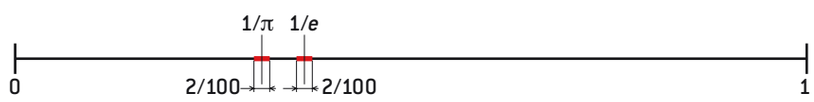
Dans un dé à six faces, le « 1 », comme chacune des autres faces, a la probabilité 1/6 de « sortir » et en lançant le dé on a une probabilité de 3/6 = 1/2 d'obtenir un nombre pair. Sur un ensemble ayant n éléments, une probabilité est équilibrée (on dit aussi « uniforme ») quand chaque élément, comme une face d'un dé, a la même probabilité. La probabilité d'un sous-ensemble de p éléments est alors p/n, le quotient du nombre de « cas favorables » sur le nombre total de cas. Cette probabilité p/n est l'outil de base pour comprendre la roulette des casinos et les tirages du loto. La question est : que se passe-t-il quand le nombre total d'éléments est infini ?



Pour l'ensemble infini des nombres réels compris entre 0 et 1, il est facile de définir une probabilité uniforme. On obtient la mesure de Lebesgue découverte par Henri Lebesgue (1875-1941) au début du XX^{e} siècle. Cette mesure sur $[0, 1]$ est fondée sur la longueur des intervalles $[a, b]$ de $[0, 1]$ auxquels on attribue la probabilité $b-a$. Elle donne un sens à l'opération « tirer un nombre réel équitablement entre 0 et 1 ». Elle indique par exemple que la probabilité pour qu'un tel nombre se trouve à moins d'un centième de $1/\pi$ ou de $1/e$ est exactement 1/25, car la zone favorable est composée des deux morceaux disjoints $[1/\pi - 1/100, 1/\pi + 1/100]$ et $[1/e - 1/100, 1/e + 1/100]$ dont la longueur totale est 4/100 = 1/25 *(voir le schéma en bas de la page)*.

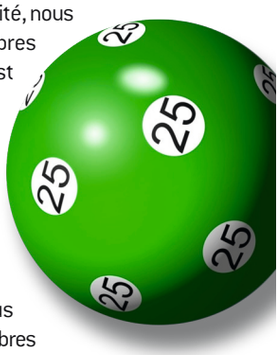
La mesure de Lebesgue montre aussi que les nombres rationnels sont négligeables : la probabilité de tirer au hasard un nombre rationnel est nulle et la probabilité de tirer un nombre irrationnel est égale à un. En effet, on sait enfermer les rationnels, qui sont numérotables par des entiers (ils sont dénombrables), dans une réunion d'intervalles de longueurs $c/2^n$, de longueur cumulée c que l'on peut rendre aussi petite que possible.

Étrangement, bien que l'infini dénombrable (celui des nombres entiers $1, 2, ..., n, ...$) précède et soit plus simple que l'infini des réels de l'intervalle $[0, 1]$ (dénommé « continu »), il est impossible d'y définir une probabilité uniforme qui jouerait, pour les entiers, le rôle que la mesure de Lebesgue joue pour $[0, 1]$. C'est très ennuyeux, car cela interdit de donner un sens mathématique au choix aléatoire équitable d'un entier. « Tirer un nombre réel au hasard entre 0 et 1 sans en favoriser aucun » est mathématiquement possible et la théorie moderne des probabilités déduit de la solution de Lebesgue des résultats concernant la physique, l'économie et toutes sortes de problèmes concrets. « Tirer un entier au hasard, sans en favoriser aucun » n'a, en revanche, mathématiquement pas de sens pour quiconque veut rester dans le cadre de la théorie des probabilités fixé en 1933 par Andreï Kolmogorov. Nous allons voir cependant qu'il existe de subtiles et merveilleuses solutions de remplacement, nommées densités, mises au point pour les besoins de l'arithmétique et qui sont importantes pour comprendre les distributions statistiques concrètes.



Probabilités non uniformes

Pour déterminer la probabilité d'un ensemble de nombres entiers positifs en oubliant temporairement l'exigence d'uniformité, nous prenons une série (a_n) de nombres positifs ou nuls dont la somme est égale à 1, par exemple $a_n = 1/2^n$ qui vérifie bien $1/2 + 1/4 + ... + 1/2^n + ... = 1$. Une fois (a_n) choisi, la probabilité d'un ensemble d'entiers A est la somme des termes a_n dont l'indice n est un élément de A.



Avec le choix $a_n = 1/2^n$, nous trouvons que l'ensemble I des nombres impairs $\{1, 3, ..., 2n+1, ...\}$ a pour probabilité $1/2 + 1/8 + 1/32 + ... + 1/2^{2n+1} + ... = 2/3$ et que l'ensemble P des nombres pairs $\{2, 4, ..., 2n, ...\}$ a pour probabilité 1/3. Ce n'est guère satisfaisant : cela ne semble

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

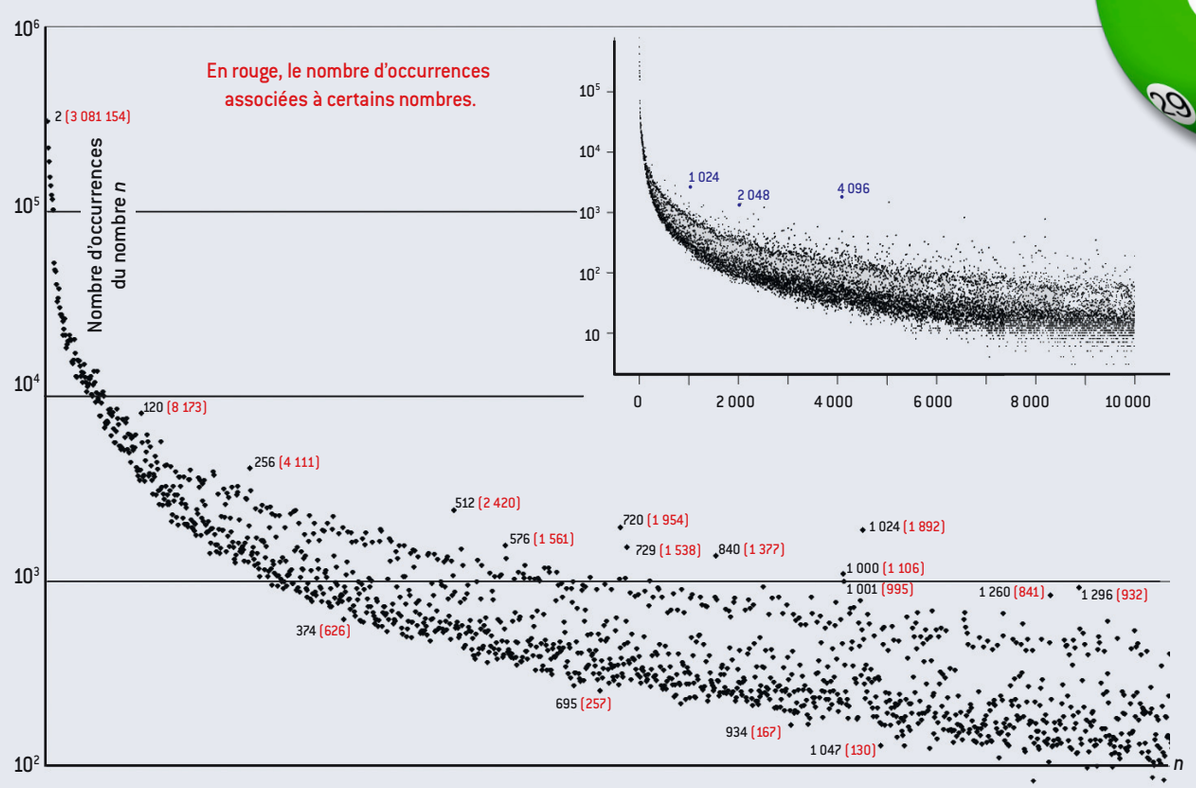
Il est impossible d'associer une probabilité uniforme aux entiers (probabilité qui donnerait à chaque entier la même probabilité). En effet, la probabilité d'un sous-ensemble des entiers est égale à la somme des probabilités de chaque élément : si la probabilité d'un entier est non nulle, la probabilité 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 ce 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.

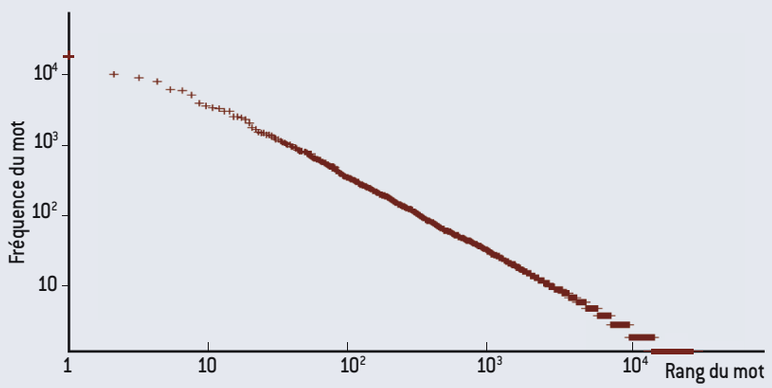
À cause des phénomènes culturels d'entraînements et de mode, les mathématiciens concentrent leur attention sur les nombres les plus simples, ce qui pousse les points correspondants de la courbe vers le haut, créant une zone évidée dans le graphe : le fossé.



2. La loi de Zipf

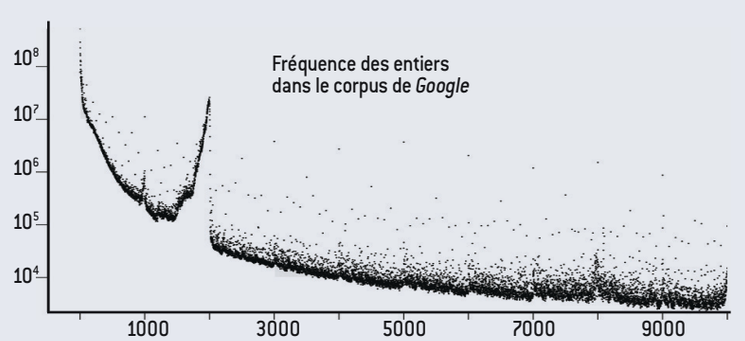
Si l'on classe les mots d'un texte par fréquence décroissante, on constate le plus souvent que la fréquence du mot en position n est C/n, où C est une constante. Cette loi, généralisée sous la forme C/n^e (où e est proche de 1), porte le nom de loi de Zipf.

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 $f_1, f_2, ..., f_n, ...$, 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 $an + b$ ($n = 1, 2, \ldots$) 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 nécessairement, 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 : $D_{log}(A) = \lim_{n \to \infty} (\sum\{1/k\,;\, k \text{ dans A}, k \leq n\} / \sum\{1/k\,;\, k \leq n\})$.

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 $1 + 1/2 + \ldots + 1/k$ 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.

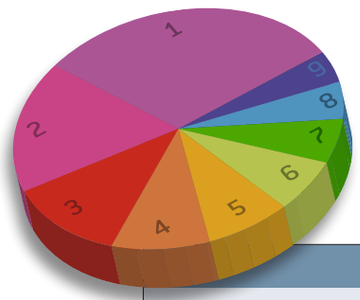
On a généralisé la loi de Zipf par une loi de type $f_n = C/n^e$, où e est un exposant proche de 1. On la nomme aussi loi de puissance ou loi de Pareto (même quand e n'est pas proche de 1). Nous ne considérerons que les situations où des objets sont classés par ordre de fréquence décroissante et où la 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 $i = 1, 2, ..., 9$) 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 ($i = 1, ..., 9$).

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 $Proba\,(n) \approx 1/2^{K(n)}$, 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 ($i = 1, 2, ..., 9$) 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 ($i = 1, 2, ..., 9$) 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.