L'étonnante loi de Benford

Le « 1 » apparaît en premier bien plus souvent que les autres chiffres !
Ce phénomène est imparfaitement expliqué, mais il permet de dépister les tricheurs.

Je vous propose le pari suivant : ouvrons le journal, choisissons une page au hasard et notons le premier nombre que nous rencontrons ; si le premier chiffre significatif de ce nombre est supérieur à 3, je vous donnerai 100 euros, sinon c'est vous qui me donnerez 100 euros. La proposition vous est , semble-t-il, nettement favorable : il n'y a en effet que trois chiffres qui me font gagner (1, 2, 3), alors qu'il y en a six pour vous (4, 5, 6, 7, 8, 9) ; le 0 ne compte pas, car il ne peut pas être un premier chiffre significatif. Vous pensez donc gagner environ deux fois sur trois. Serais-je idiot de vous proposer un tel pari ?

Eh bien non : si vous acceptez, je gagnerai dans plus de 60 pour cent des cas. Aussi étonnant que cela paraisse, le premier chiffre significatif d'un nombre rencontré dans un article de journal n'a pas autant de chances d'être un 1, un 2, un 3, ..., ou un 9 (la probabilité serait alors 1/9, ou 11,11 pour cent). La loi de Benford indique que, dans un contexte général comme celui d'un article de journal, les probabilités p de rencontrer les différents chiffres comme premier chiffre significatif sont, exprimées en pourcentage : $p(1) = 30{,}1$; $p(2) = 17{,}6$; $p(3) = 12{,}5$; $p(4) = 9{,}7$; $p(5) = 7{,}9$; $p(6) = 6{,}7$; $p(7) = 5{,}8$; $p(8) = 5{,}1$; $p(9) = 4{,}6$. Puisque 30,1 + 17,6 + 2,5 = 60,2, je gagnerai mon pari dans 60,2 pour cent des cas.

Quelle est donc cette loi bizarre du premier chiffre significatif, si contraire à l'intuition ?

Découverte deux fois

Universellement connue aujourd'hui sous le nom de loi de Benford, cette loi a été découverte en 1881 par l'astronome Simon Newcomb, qui avait observé une détérioration des tables de logarithmes bien plus importante pour les pages des nombres commençant par 1 que pour les pages des nombres commençant par 9. L'article qu'il publia sur le sujet dans le *Journal of Mathematics* passa inaperçu ; aussi, lorsque 57 ans plus tard, le physicien Frank Benford remarqua à son tour l'usure inégale des pages des tables de logarithmes, il crut être le premier à formuler cette loi qui, aujourd'hui porte indûment son nom.

Les articles de Newcomb et de Benford ne se contentent pas de constater, ils proposent chacun la même formule mathématique donnant la probabilité des chiffres. La probabilité que le chiffre c soit le premier chiffre significatif d'un nombre est, d'après eux, $\log_{10}(1 + 1/c)$ où $\log_{10}$ désigne le logarithme décimal : $\log_{10}(10)=1$; $\log_{10}(1)=0$; $\log_{10}(ab) = \log_{10}(a) + \log_{10}(b)$; $\log_{10}(a/b) = \log_{10}(a) - \log_{10}(b)$. Remarquons que leur proposition convient parfaitement pour une probabilité puisque la somme des probabilités est égale à 1 (un des chiffres doit apparaître au premier rang) :

$\log_{10}(1 + 1/1) + \log_{10}(1 + 1/2) + ... + \log_{10}(1 + 1/9) =$
$\log_{10}(2) + \log_{10}(3/2) + \log_{10}(4/3) + ... + \log_{10}(10/9) =$
$\log_{10}[2 \times (3/2) \times (4/3) \times ... \times (10/9)] = \log_{10}(10) = 1$

Avant d'examiner en quel sens cette loi est une loi scientifique, et comment nous pouvons l'expliquer, voyons sa généralisation.

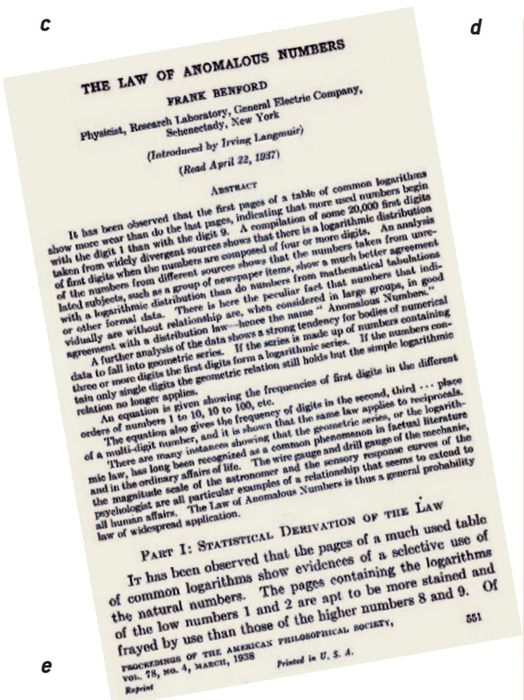
La mantisse est en rapport avec l'écriture d'un nombre en notation scientifique. La mantisse d'un nombre R est le nombre a, compris entre 1 et 10, tel que $R = a \times 10^n$ (ainsi la mantisse de 5 000 000 est 5, la mantisse de 1 024 est 1,024,

1. Les données recueillies par Benford et publiées en 1938 sont d'une grande variété : toutes indiquent que le premier chiffre significatif de données statistiques (longueur des rivières, des routes, des altitudes des montagnes, de la population des villes, des aires des circonscriptions religieuses avant la Révolution française) est 1 avec une fréquence de 30 pour cent au lieu de 11 pour cent si les chiffres de 1 à 9 apparaissaient avec la même fréquence. De même le 2 est surepresenté, mais moins, et plus souvent premier chiffre significatif que « 3 ». Enfin, le 9 est le plus rare. On retrouve cette répartition bizarre détaillée sur l'histogramme *(e)* en géographie, en mathématiques pour le premier chiffre de certaines suites, dans les résultats sportifs, etc. (tableau *d*, traduit de Benford). Les proportions coïncident bien avec la série découverte par Benford (dernière ligne du tableau *d* sous l'intitulé « prédite ») : la probabilité pour qu'un nombre pris dans une grande série de données commence par le chiffre c (c = 1, 2, ..., 9) est $\log_{10}(1 + 1/c)$. En haut, Frank Benford *(a)* à l'époque de la publication de son article, en mars 1938, et les célèbres cartes de Vidal-Lablache *(b)* où le phénomène est patent quand on mesure les différentes caractéristiques géographiques. Frank Benford a dénommé dans son article *(c)* cette observation la loi des nombres anormaux : « Il a été observé que les premières pages des tables de logarithmes (utilisées avant les ordinateurs pour les calculs précis) sont plus usées que les dernières pages, ce qui indique que les nombres les plus utilisés commencent plus fréquemment par 1 que par 9. Une compilation de plus de 20 000 premiers chiffres des nombres provenant de sources très différentes montre une distribution logarithmique de ces chiffres quand les nombres ont plus de quatre chiffres. »

a

b

c



THE LAW OF ANOMALOUS NUMBERS

FRANK BENFORD

Physicist, Research Laboratory, General Electric Company, Schenectady, New York

(Introduced by Irving Langmuir)

(Read April 22, 1937)

ABSTRACT

It has been observed that the first pages of a table of common logarithms show more wear than do the last pages, indicating that more used numbers begin with the digit 1 than with the digit 9. A compilation of some 20,000 first digits taken from widely divergent sources shows that there is a logarithmic distribution of first digits when the numbers are composed of four or more digits. An analysis of the numbers from different sources shows that the numbers taken from unrelated subjects, such as a group of newspaper items, show a much better agreement with a logarithmic distribution than do numbers from mathematical tabulations or other formal data. There is here the peculiar fact that numbers that individually are without relationship are, when considered in large groups, in good agreement with a distribution law—hence the name "Anomalous Numbers."

A further analysis of the data shows a strong tendency for bodies of numerical data to fall into geometric series. If the series is made up of numbers containing three or more digits the first digits form a logarithmic series. If the numbers contain only single digits the geometric relation still holds but the simple logarithmic relation no longer applies.

An equation is given showing the frequencies of first digits in the different orders of numbers 1 to 10, 10 to 100, etc.

The equation also gives the frequency of digits in the second, third ··· place of a multi-digit number, and it is shown that the same law applies to reciprocals.

There are many instances showing that the geometric series, or the logarithmic law, has long been recognized as a common phenomenon in factual literature and in the ordinary affairs of life. The wire gauge and drill gauge of the mechanic, the magnitude scale of the astronomer and the sensory response curves of the psychologist are all particular examples of a relationship that seems to extend to all human affairs. The Law of Anomalous Numbers is thus a general probability law of widespread application.

PART I: STATISTICAL DERIVATION OF THE LAW

IT has been observed that the pages of a much used table of common logarithms show evidences of a selective use of the natural numbers. The pages containing the logarithms of the low numbers 1 and 2 are apt to be more stained and frayed by use than those of the higher numbers 8 and 9. Of

551

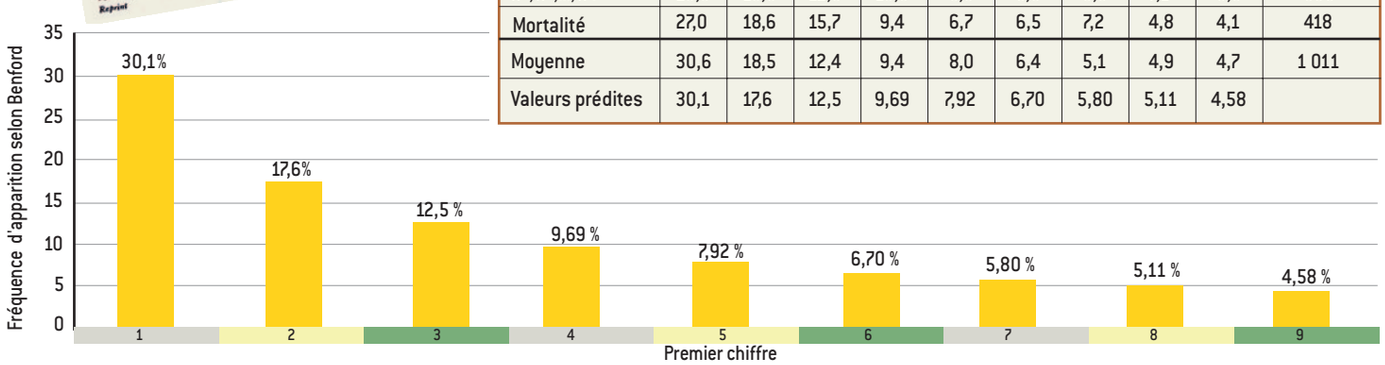
PROCEEDINGS OF THE AMERICAN PHILOSOPHICAL SOCIETY, VOL. 78, NO. 4, MARCH, 1938 Printed in U. S. A.

Reprint

d

Premier chiffre	1	2	3	4	5	6	7	8	9	Échantillon
Rivières	31,0	16,4	10,7	11,3	7,2	8,6	5,5	4,2	5,1	335
Populations	33,9	20,4	14,2	8,1	7,2	6,2	4,1	3,7	2,2	3 259
Constantes	41,3	14,4	4,8	8,6	10,6	5,8	1,0	2,9	10,6	104
Journaux	30,0	16,4	12,0	10,0	8,0	6,0	6,0	5,0	5,0	100
Chaleurs spéc.	31,0	18,4	16,2	14,6	10,6	4,1	3,2	4,8	4,1	1 389
Pression	31,0	18,3	12,8	9,8	8,3	6,4	5,7	4,4	4,7	703
Puissance perdue	30,0	18,4	11,9	10,8	8,1	7,0	5,1	5,1	3,6	690
Poids moléculaires	26,7	25,2	15,4	10,8	6,7	5,1	4,1	2,8	3,2	1 800
Drainage	27,1	23,9	13,8	12,6	8,2	5,0	5,0	2,5	1,9	159
Poids atomiques	47,2	18,7	5,5	4,4	6,6	4,4	3,3	4,4	5,5	91
$n-1, \sqrt{n}$	25,7	20,3	9,7	6,8	6,6	6,8	7,2	8,0	8,9	5 000
Architecture	36,8	14,8	14,3	7,5	8,3	8,4	7,0	7,3	5,6	560
Reader's Digest	33,4	18,5	12,4	7,5	7,1	6,5	5,5	4,9	4,2	308
Coûts	32,4	18,8	10,1	10,1	9,8	5,5	4,7	5,5	3,1	741
Voltage X	27,9	17,5	14,4	9,0	8,1	7,4	5,1	5,8	4,8	707
Baseball	32,7	17,6	12,6	9,8	7,4	6,4	4,9	5,6	3,0	1 458
Rayonnement	31,0	17,3	14,1	8,7	6,6	7,0	5,2	4,7	5,4	1 165
Adresses	28,9	19,2	12,6	8,8	8,5	6,4	5,6	5,0	5,0	342
$n^1, n^2, ..., n!$	25,3	16,0	12,0	10,0	8,5	8,8	6,8	7,1	5,5	900
Mortalité	27,0	18,6	15,7	9,4	6,7	6,5	7,2	4,8	4,1	418
Moyenne	30,6	18,5	12,4	9,4	8,0	6,4	5,1	4,9	4,7	1 011
Valeurs prédites	30,1	17,6	12,5	9,69	7,92	6,70	5,80	5,11	4,58	

e



la mantisse de 0,25 est 2,5). Alors la probabilité, selon la loi de Benford, que la mantisse d'un nombre réel soit dans l'intervalle [*a, b*] est ($\log_{10}(b) - \log_{10}(a)$).

Cette généralisation redonne la version initiale de la loi de Benford pour $a = c$ et $b = c + 1$, car : $\log_{10}(c + 1) - \log_{10}(c) = \log_{10}[(c+1)/c] = \log_{10}(1+1/c)$, la probabilité que le premier chiffre soit *c*. Il résulte aussi de cette généralisation que lorsqu'une série de nombres satisfait la loi de Benford, la probabilité de trouver un second chiffre significatif égal à *y* (par exemple 5) dépend du premier chiffre *x*. Ainsi, la probabilité que le deuxième chiffre significatif soit 5 quand le premier vaut 1 est : (log(1,6) – log(1,5)) / (log(2) – log(1)) = 9,31 pour cent. La probabilité que le second chiffre significatif soit 5 quand le premier est 8 est (log(8,6) – log(8,5))/ (log(9) – log(8)) = 9,93 pour cent.

La généralisation permet d'étudier les autres chiffres et le résultat se conforme cette fois à l'intuition : si une série de nombres satisfait la loi de Benford, alors plus un chiffre est loin après le premier chiffre, plus la probabilité qu'il soit un 1, un 2, ... ou un 9 est proche de 1/10 (cette fois le 0 est possible). Pour le second chiffre significatif, les probabilités *p* sont respectivement : *p*(0)=11,96 ; *p*(1)=11,39 ; *p*(2)=10,88 ; *p*(3)=10,43 ; *p*(4)=10,03 ; *p*(5)=9,67 ; *p*(6)=9,34 ; *p*(7)=9,03 ; *p*(8)=8,76 ; *p*(9)=8,50. Pour le quatrième chiffre, toutes les probabilités sont comprises entre 9,98 pour cent et 10,02 pour cent.

La loi est-elle toujours valable ?

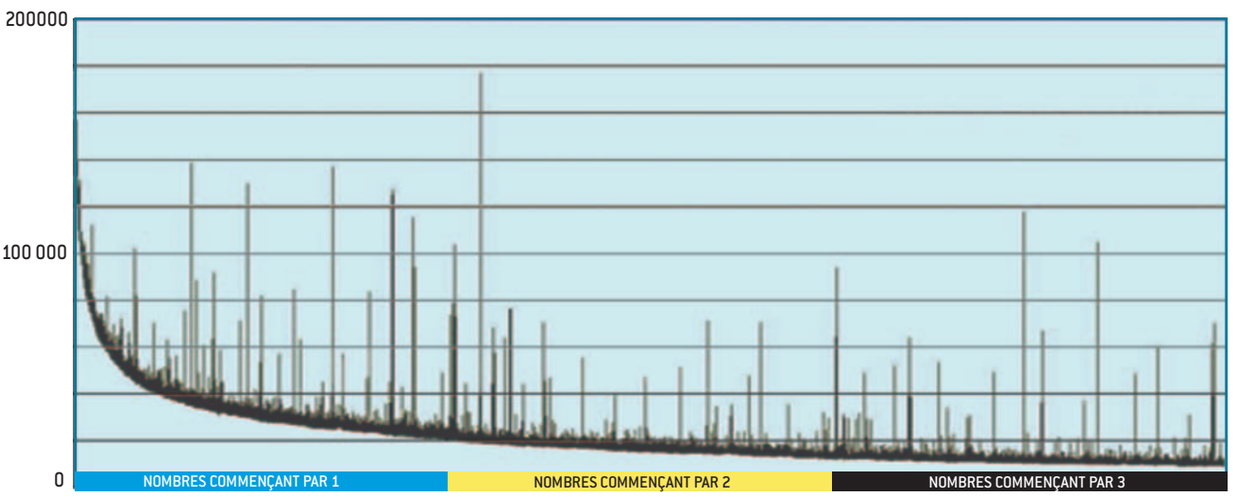
Immanquablement, face à cette prétendue « loi », nous nous posons de nombreuses questions : est-elle vraiment toujours vérifiée ? En quel sens s'agit-il d'une loi ? D'où provient-elle et que signifie-t-elle ?

Bien sûr, la réponse à la première question est : non, la loi n'est pas systématiquement vérifiée. Ainsi, les numéros de téléphone que l'on trouve dans un annuaire ne la vérifient pas, car les opérateurs distribuent les numéros en fonction de critères particuliers et systématiques (les numéros commencent d'ailleurs souvent par 0). Les tailles des êtres humains adultes mesurées en mètres commencent dans plus de 95 pour cent des cas par un 1 et donc la « loi de Benford » n'est pas satisfaite pour des données statistiques sur les tailles. Les prix de vente d'un modèle particulier de voiture neuve varient peu d'un concessionnaire à l'autre et donc ne vérifient pas la loi : à chaque fois qu'une donnée statistique est contrainte de rester dans un intervalle étroit, la loi de Benford n'est pas satisfaite.

Benford, pour tester son idée, avait collecté 20 229 observations diverses incluant des données géographiques, des résultats sportifs, des valeurs de constantes physiques, des prix, des séries mathématiques, etc. Il avait constaté, d'une part, que quand on regroupe toutes les données, la loi était satisfaite avec une bonne précision, comme on le voit sur le tableau de la figure 1. D'autre part, de nombreuses séries la satisfaisaient « à peu près » : elles montraient au moins une nette décroissance de la fréquence des chiffres en fonction de leur rang. Ces constatations ont été souvent refaites.

Cependant, des études récentes fondées sur des collectes de nombres bien plus volumineuses et un examen statistique plus fin tempèrent l'enthousiasme. P. Scott et M. Fasli, de l'Université de l'Essex, ont examiné 230 séries statistiques incluant au total près de 500 000 données. Ils remarquent que le phénomène de décroissance des probabilités d'un chiffre au suivant est quasi général, mais qu'il ne suit de près la loi de Benford que dans 29 séries. La loi des logarithmes est une approximation, souvent grossière, de ce qui est observé.

Les lois déterminant réellement la probabilité d'apparition du premier chiffre significatif sont, dans le monde réel, des variantes de la loi pure fixée par les logarithmes que Newcomb et Benford ont formulée. Le cas des chiffres provenant de données financières est caractéristique : les chiffres de ce domaine s'écartent assez nettement de la loi de Benford. Reste que même si l'ajustement n'est pas parfait et ne le devient pas



2. La table des constantes numériques de Simon Plouffe contient un milliard de constantes (occupant 45 gigaoctets de mémoire !) dont bien sûr π, *e*, $\sqrt{2}$, $\sqrt{3}$, sin(1), logπ, exp($\sqrt{2}$), etc., mais aussi des sommes de séries et des valeurs de fonctions classiques évaluées en divers points, etc., Les statistiques sur le premier chiffre des constantes présentes dans la table le 6 novembre 2006 sont : 38,046 % des entrées commencent par 1 ; 15,433 % commencent par 2 ; 10,871 % commencent par 3 ; 8,195 % commencent par 4 ; 6,913 % commencent par 5 ; 6,066 % commencent par 6 ; 5,225 % commencent par 7 ; 4,777 % commencent par 8 et 4,475 % des entrées commencent par 9. Ce qui donne à peu près les statistiques de la loi de Benford, mais avec un écart à la loi non négligeable compte tenu du très grand nombre de constantes considérées. La courbe tracée montre des pics qui sont dus à ce que S. Plouffe a étudié certains nombres avec une attention particulière.

nécessairement quand on augmente le nombre des données examinées, la loi de Benford pour de nombreuses séries numériques réelles est pertinente en première approximation et bien meilleure que l'attribution d'une probabilité de 1/9 à chaque chiffre que nous souffle l'intuition. Pourquoi cela ?

Suites mathématiques et Benford

Considérons d'abord les suites purement mathématiques pour lesquelles un ensemble de résultats, certains récents, éclaire la règle et ses exceptions.

En 1968, le mathématicien russe Vladimir Arnold, associé à André Avez, démontra que la suite numérique (2^n) satisfait la loi de Benford à l'infini, dans le sens suivant : la proportion des éléments de la suite considérée jusqu'à n, dont le premier chiffre est 1, tend vers $\log_{10}(2)$ quand n tend vers l'infini, conformément à la loi de Benford et il en va de même pour les autres chiffres c pour lesquels la limite est bien le $\log_{10}(1 + 1/c)$ attendu.

Ce résultat a depuis été généralisé. On a d'abord montré que la suite (r^n), r étant un nombre réel positif, satisfait la loi de Benford à l'infini pourvu que $\log_{10}(r)$ ne soit pas un nombre rationnel (c'est-à-dire un rapport de deux entiers). Les suites (3^n), (4^n), (5^n) satisfont donc la loi de Benford, alors que (10^n) et $(\sqrt{10})^n$ ne la satisfont pas, ce qu'il est facile de constater.

En 2005, Paul Jolissaint de l'Institut de mathématiques de Neuchâtel a démontré un résultat général concernant les suites définies par une relation de récurrence du type $x(n) = a_1x(n-1) + a_2x(n-2) + ... + a_px(n-p)$. Ce résultat indique que la suite de Fibonacci 1, 1, 2, 3, 5, 8, ... (définie par $x(0) = x(1) = 1$ et $x(n) = x(n-1) + x(n-2)$) satisfait la loi de Benford à l'infini. Le résultat s'applique aussi à la suite obtenue en ne retenant qu'un élément sur deux de ces suites ou en ne retenant que les éléments dont le numéro est un carré parfait, et plus généralement à toutes les suites extraites en ne retenant que les termes dont le numéro est donné par un polynôme (fixé) à coefficients entiers ne prenant que des valeurs positives différentes. Il est remarquable que la démonstration de ce beau théorème s'appuie sur un résultat de 1916 dû au mathématicien Hermann Weyl (1885-1955).

D'autres démonstrations récentes établissent que les suites $n!$ et n^n satisfont aussi parfaitement la loi de Benford à l'infini. Les coefficients du binôme de Newton (qu'on trouve dans le triangle de Pascal) satisfont eux aussi la loi de Benford. Nous savons encore que si (x_n) satisfait la loi de Benford, il en va de même de (cx_n), c constante positive quelconque, ou même de (x_n^p), p étant un entier quelconque non nul.

En revanche, il a été démontré que, pour b positif et p_n le n-ième nombre premier, aucune des suites : $(b, 2b, 3b,..., nb, ...)$, $(1, 2^b, 3^b,..., n^b, ...)$, $(\log_b 1, \log_b 2, \log_b 3,..., \log_b n,...)$, $(1, 2, 3, 5,..., p_n,...)$, $(\log_b 2, \log_b 3, \log_b 5,..., \log_b p_n,...)$ ne vérifie la loi de Benford parce que les fréquences associées aux chiffres ne convergent pas.

La table de constantes mathématiques que Simon Plouffe collecte depuis de nombreuses années et qui contient aujourd'hui plus d'un milliard de nombres est intéressante, car elle satisfait en gros la loi de Benford. S. Plouffe utilise d'ailleurs cette loi pour détecter des erreurs qui pourraient s'y glisser. Les exceptions en certains points à la loi s'expliquent par la façon dont il a engendré sa collection en s'intéressant à des nombres particuliers.

3. Cesàro explique Benford!

Les numéros des maisons dans les rues obéissent à la loi de Benford à la condition de faire des moyennes successives inventées par le mathématicien Ernesto Cesàro (1859-1906). On désigne par $f_1(n)$ la fréquence du chiffre 1 comme premier chiffre dans la suite des n premiers nombres :
1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21...

De même, $f_2(n)$ est la fréquence du chiffre 2 comme premier chiffre dans la suite des n premiers nombres :
1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21...,
et ainsi de suite pour $f_3(n)$, $f_4(n)$, $f_5(n)$..., $f_9(n)$.

On effectue alors la moyenne à la Cesàro des $f_1(n)$:
$s_1(n) = [f_1(1)+f_1(2)+f_1(3)+f_1(4)+f_1(5)+f_1(6)+f_1(7)+...+f_1(n)]/n.$

Et de même :
$s_2(n) = [f_2(1)+f_2(2)+f_2(3)+f_2(4)+f_2(5)+f_2(6)+f_2(7)+...+f_2(n)]/n.$
Puis on réitère pour les moyennes de Cesàro $t_1(n)$, $t_2(n)$:
$t_1(n) = [s_1(1)+s_1(2)+s_1(3)+s_1(4)+s_1(5)+s_1(6)+s_1(7)+...+s_1(n)]/n,$
et :
$t_2(n) = [s_2(1)+s_2(2)+s_2(3)+s_2(4)+s_2(5)+s_2(6)+s_2(7)+...+s_2(n)]/n.$

La valeur de $f_1(n)$ varie entre 1/9 et 5/9 et la valeur de la suite $f_9(n)$ entre 1/81 et 1/9. En continuant ainsi, on obtient des valeurs qui oscillent de moins en moins, et B. Flehinger a démontré qu'en poursuivant ces calculs de moyennes l'intervalle de variation de ces sommes se réduit, à l'infini, à la valeur attendue, soit $\log_{10}(2)$ pour celle associée au 1.

Cette convergence d'une suite à la Cesàro est une idée intéressante dans la mesure où elle fait converger des suites qui étaient divergentes. L'exemple souvent cité est la suite 01010101... qui converge vers 1/2 au sens de Cesàro.

Cesàro à la rescousse

Reste que les résultats qu'on obtient dans le monde parfait des mathématiques ne justifient pas ce que Newcomb et Benford ont vu pour les séries de nombres provenant du monde réel. Aussi de nombreuses études tentent d'expliquer rationnellement l'étrange phénomène. Quatre types d'explications sont proposés.

Pour les numéros des maisons qu'on trouve dans des adresses collectées dans un annuaire, et qui vérifient assez bien la loi de Benford, une idée simple vient à l'esprit. Si une rue possède 50 numéros, alors plus d'un cinquième des numéros commencent par un 1 (à cause de 10, 11, 12,..., 19) ; si elle en possède 20 ou 200, plus de la moitié des numéros commencent par un 1. Il est donc parfaitement normal que dans une rue dont la longueur est inconnue, on trouve en moyenne plus souvent des numéros commençant par 1 que par 9 (et plus généralement par le chiffre c que par $c + 1$).

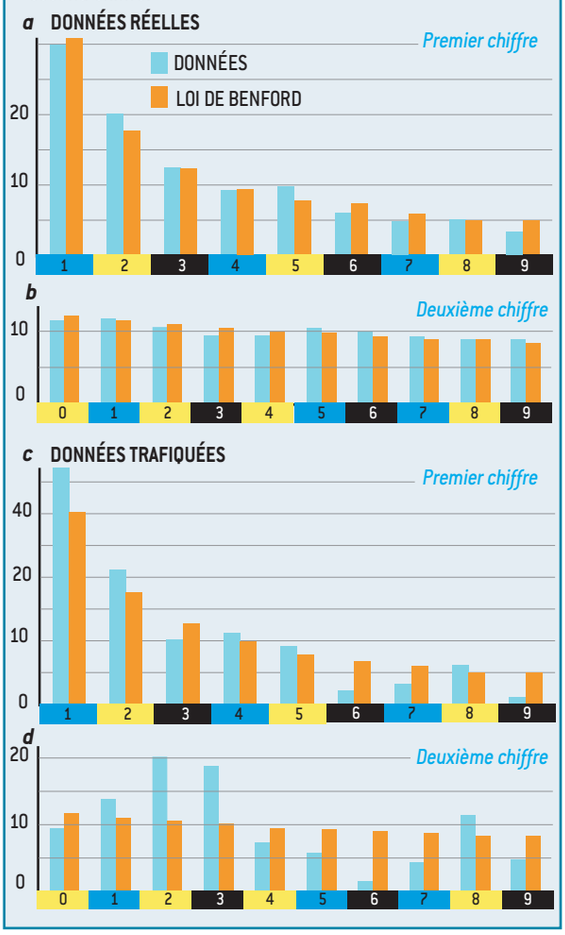
Si on note $f_1(n)$ la fréquence des entiers commençant par 1 parmi les entiers compris entre 1 et n (de même on notera $f_2(n)$ la fréquence pour le 2, $f_3(n)$ la fréquence pour le 3, etc.), on s'aperçoit que la suite $f_1(n)$ n'est pas convergente et qu'elle oscille indéfiniment entre 1/9 et 5/9. L'oscillation est de plus en plus lente, mais son amplitude reste de 4/9. La suite $f_9(n)$ oscille, elle, entre 1/81 et 1/9. Comment s'en sortir ?

4. La détection des fraudes

L'analyse des chiffres *(digital analysis),* discipline récente, s'assure de la cohérence interne et de la vraisemblance de grandes quantités de données numériques. Elle explore systématiquement les chiffres des séries étudiées pour y repérer des anomalies de fréquence signifiant souvent que les données ont été manipulées, falsifiées ou inventées. La loi de Benford est un outil de cette discipline (voir : http://www.aicpa.org/pubs/jofa/may1999/nigrini.htm)

Ce type d'analyse a été utilisé avec succès par les services fiscaux américains pour repérer des fraudeurs. Récemment une étude statistique minutieuse a été menée par A. Saville sur les données fournies par 34 entreprises dont 17 étaient connues pour avoir manipulé leurs comptes. Une pure analyse des chiffres, en se limitant aux séries dont il était présumé qu'elles devaient vérifier la loi de Benford (ce n'est pas le cas de toutes les séries) a conduit à identifier les 17 entreprises classées suspectes. Le test n'est cependant pas parfait, puisque quatre autres entreprises *a priori* non suspectes ont aussi été désignées par les tests. Voir *Adrian Saville, Using Benford's Law to Detect Data Error and Fraud, South African Journal of Economic and Management Sciences,* 2006, pp. 341-354.

Les chercheurs ont également examiné, (voir ci-dessous), le premier chiffre (*a* et *c*) et le deuxième chiffre (*b* et *d*) du coefficient de corrélation de données réelles publiées dans un journal américain de sociologie (http://ideas.repec.org/e/pdi71.html), et de données trafiquées. La tricherie est bien visible sur le second chiffre.



Pour adoucir ces oscillations, l'idée est de faire la moyenne des $f_1(k)$ pour k variant de 1 jusqu'à n : $[f_1(1)+f_1(2)+..+ f_1(n)]/n$, c'est ce que l'on nomme la moyenne de Cesàro $s_1(n)$ de la suite $f_1(n)$. On obtient alors une nouvelle suite qui ne converge toujours pas, mais qui varie cette fois dans un intervalle plus étroit. En recommençant ce procédé de moyenne, on obtient des suites qui varient dans des intervalles de plus en plus étroits et B. Flehinger a démontré en 1966 que l'intervalle qu'on obtient en poursuivant ces calculs de moyennes de moyennes s'approche, à l'infini, de la valeur attendue, soit $\log_{10}(1 + 1/1) = \log_{10}(2)$.

Pour les autres chiffres, le constat est analogue et donc, dans le sens très particulier des moyennes itérées de Cesàro, on peut dire que la fréquence des nombres commençant par c dans l'ensemble de tous les entiers est $\log_{10}(1 + 1/c)$.

D'autres méthodes de ce type permettent d'attribuer une « mesure » (une fréquence) à l'ensemble des nombres entiers commençant par 1 (ou 2, ..., ou 9). Toutes celles connues et possédant des propriétés raisonnables (par exemple que l'ensemble des nombres pairs ait pour mesure 1/2) conduisent au même résultat et accordent à l'ensemble des nombres entiers commençant par c la mesure $\log_{10}(1 + 1/c)$ conforme à ce que dicte la loi de Benford. On peut interpréter ces résultats comme la démonstration que la suite des entiers 1, 2, 3, ..., n, ... vérifie une forme « faible » (à la Cesàro) de la loi de Benford, et que c'est pourquoi les numéros de rues satisfont à peu près la loi de Benford.

Lorsqu'on utilise ces mesures, on constate que le quotient de la densité des nombres premiers commençant par c, sur la densité de tous les nombres premiers est exactement $\log_{10}(1 + 1/c)$. Les nombres premiers qui au sens de la fréquence ne suivent pas la loi de Benford (car il n'y a pas convergence de la fréquence) la vérifient donc au sens de ces mesures.

Aussi merveilleux que soient ces résultats, ils ne suffisent pas à expliquer tous les cas concrets où les données statistiques sont conformes à la loi de Benford. Il faut trouver d'autres explications.

Invariances et mélanges

Les statisticiens se sont longtemps étonnés : la loi de Benford est satisfaite pour les longueurs des fleuves quand on mesure celles-ci en kilomètres, mais aussi quand on les mesure en miles ou avec n'importe quelle unité de longueur. Cette invariance par multiplication des données par une constante a été étudiée et Roger Pinkham a démontré en 1961 que la seule loi sur les mantisses invariante par multiplication (la probabilité que la mantisse de x soit comprise entre a et b est égale à la probabilité p_r que c fois la mantisse de x soit comprise entre a et b : $p_r(a < \text{mantisse}(x) < b) = p_r(a < c \times \text{mantisse}(x) < b)$) est la loi de Benford.

En clair, s'il existe une loi pour les mantisses et que, comme on s'y attend, cette loi ne dépend pas des unités de mesure utilisées pour collecter les données, alors cette loi ne peut être que celle de Benford. Un résultat analogue concernant le changement de base de numération a été démontré par T. Hill en 1995. Ces deux résultats, associés à ceux sur les mesures cités plus haut, établissent de manière

concordante que la seule loi envisageable pour le premier chiffre significatif est la loi de Benford et non pas la loi équitable du 1/9 que nous soufflait notre trompeuse intuition.

Même si la loi de Benford est la seule possible pour les mantisses et le premier chiffre significatif, cela ne démontre pas que pour des données réelles, c'est elle que l'on rencontrera. Une troisième catégorie de résultats nous approche un peu plus de ce but.

Un théorème de T. Hill de 1996 montre qu'un bon mélange de lois quelconques – ne vérifiant pas individuellement la loi de Benford – donne des nombres vérifiant globalement la loi de Benford. Sous une forme un peu plus précise le résultat est le suivant.

Si nous choisissons au hasard une distribution de probabilités (dans un ensemble varié de distributions ne vérifiant pas individuellement la loi de Benford), et si nous prenons un nombre selon cette distribution et que nous recommencions un grand nombre de fois cette double opération de choix, alors si le processus général de ces choix est indépendant de la base (ou des unités de mesure), la série produite vérifiera la loi de Benford.

Ce théorème expliquerait pourquoi quand on prend tous les nombres trouvés dans un journal – ils proviennent de lois variées et sans liens précis qui toutes ne vérifient pas la loi de Benford – alors on obtient un ensemble de nombres se conformant assez bien à la loi de Benford.

Pourtant, même si ce dernier résultat explique que des données mélangées satisfont la loi de Benford, il semble nécessaire aussi que des processus probabilistes simples engendrent des données numériques vérifiant la loi de Benford. On imagine alors des algorithmes élémentaires donnant des séries aléatoires conformes à la loi de Benford et on argumente que le monde réel agit comme ces algorithmes. Par exemple, si l'on choisit au hasard uniformément des nombres x entre deux entiers a et b et qu'on calcule c^x (c positif), alors la série de nombres ainsi engendrés vérifie la loi de Benford.

D'autres résultats du même genre (multiplications de nombres aléatoires entre eux, exponentielle de loi normale, etc.) décrivent des mécanismes probabilistes qui produisent des séries numériques satisfaisant la loi de Benford, au moins approximativement.

Reste cependant une insatisfaction : les processus proposés ne recouvrent pas la généralité des cas dont les données numériques vérifient la loi de Benford. On ne voit pas ainsi d'explication à ce que la table des constantes mathématiques de S. Plouffe se conforme à la loi de Benford.

Fraudes et interrogations

Même imparfaitement analysée, la loi logarithmique sur les chiffres est utilisée. Des statisticiens convaincus que, sauf cas exceptionnels, des données doivent la vérifier l'utilisent pour repérer des comptabilités truquées ou le manque de sérieux de séries statistiques économiques suspectées d'être bidonnées. Le test de la loi de Benford serait même un outil courant dans la chasse aux fraudes fiscales.

Une étude de psychologie expérimentale menée par Andrea Dickmann, de l'Institut fédéral suisse de technologie à Zurich, a validé cette méthode. Des sujets à qui on demande de créer des données fantaisistes les produisent sans respecter la loi de Benford (ou très imparfaitement) et produisent donc des données faciles à distinguer des séries réelles. Quelqu'un d'informé peut simuler des données et tromper un test fondé sur le premier chiffre significatif ; en revanche, A. Dickmann montre que si l'on prend en compte le second chiffre significatif, la fraude devient si manifeste qu'aucun sujet humain ne semble en mesure d'échapper à ce repérage statistique.

La loi de Benford apparaît si importante à certains, qu'ils ont suggéré d'en tenir compte pour la conception des ordinateurs afin que ceux-ci exploitent le fait qu'ils rencontreront plus souvent des nombres commençant par 1 que par 9 (et plus généralement c que $c + 1$). Un tel ajustement de la structure des ordinateurs permettrait de mieux gérer leurs mémoires et leurs calculs qu'on ne le fait aujourd'hui, puisque nos machines sont conçues en supposant implicitement que chaque chiffre possède la probabilité 1/9 d'apparaître comme premier chiffre significatif.

La grande question qui reste posée par la loi de Benford est finalement : est-ce nous humains ou le monde physique dans lequel nous vivons, ou une raison plus générale encore qui explique les nombreuses séries numériques s'ajustant à la loi logarithmique de Benford ?

Le fait que la loi ne soit pas spécifique de la base 10, les constantes mathématiques de S. Plouffe, les théorèmes généraux d'invariance, les algorithmes probabilistes de génération et les démonstrations que des catégories importantes de suites mathématiques se conforment à la loi de Benford, tout cela nous a fait progresser. Il semble à peu près certain aujourd'hui que l'explication ultime n'est pas liée à notre façon d'examiner le monde et de le mettre en chiffres, ni à une propriété singulière de notre univers physique (et que d'autres univers physiques possibles pourraient ne pas posséder), mais résulte de principes généraux encore incomplètement identifiés et que seules les mathématiques conduiront à formuler entièrement... peut-être.

Jean-Paul DELAHAYE est professeur d'informatique à l'Univ. de Lille.

Auteur & Bibliographie

C. CALDWELL, *Does Benford's law apply to prime numbers ?,* 2006 : http://primes.utm.edu/notes/faq/BenfordsLaw.html

P. JOLISSAINT, *Loi de Benford, relations de récurrence et suites équidistribuées*, 2005 : w3.jura.ch/ijsla/Benford.pdf

P. N. POSCH, *A survey on sequences and distribution functions satisfying the first-digit-law*, 2004 : http://www.posch.net/

E. JANVRESSE et T. DE LA RUE, *La loi de Benford,* in *Quadrature* n°48, pp. 5-9, avril-juin, 2003.

P. D. SCOTT et M. FASLI. *Benford's law: an empirical investigation and a novel explanation* 2001 : http://citeseer.ist.psu.edu/709593.html

T. HILL, *A statistical derivation of the significant-digit law,* in *Statistical Science,* 10(4) pp. 354-363, 1995.

T. HILL, *Base-invariance implies benford's law,* in *Proc. American Mathematical Society*, 123(2) pp. 887-895, 1995.

V. ARNOLD et A. AVEZ, *Ergodic problems of classical mechanics,* Benjamin, 1968.

F. BENFORD, *The law of anomalous numbers*, in *Proc. American Philosophical Society*, 78(4) pp. 551-772, 1938.

S. NEWCOMB, *Note on the frequency of the use of the digits in natural numbers*, in *Amer. Jour. Math.,* 4 pp. 39-40, 1881.