UNE EXPLICATION POUR LA LOI DE BENFORD

La loi de Benford, qui porte sur le premier chiffre significatif des nombres, a perdu de son mystère. Parallèlement, elle a été généralisée et, ainsi, a gagné en efficacité pour détecter des données frauduleuses.

L'AUTEUR

JEAN-PAUL DELAHAYE
professeur émérite à l'université de Lille et chercheur au Centre de recherche en informatique, signal et automatique de Lille (Cristal)

La loi de Benford ou «loi du premier chiffre significatif» n'en finit pas de troubler, d'intéresser et de susciter des travaux et des applications. Plus de 130 articles scientifiques ont été publiés sur le sujet ces cinq dernières années (*voir http://www.benfordonline.net*).

Certains la trouvent mystérieuse, alors que d'autres croient en comprendre la nature et en proposent des explications. Elle a été généralisée et utilisée pour effectuer des tests et repérer des fraudes. Nous présenterons quelques-unes des idées récentes sur ce sujet étrange et passionnant.

LA LOI DU PREMIER CHIFFRE

Les nombres que l'on rencontre pour mesurer la population des villes, les distances entre étoiles ou les prix apparaissant sur les produits d'un grand supermarché montrent une propriété surprenante. Dans ces séries, la proportion de nombres dont le premier chiffre significatif est 1 est supérieure à la proportion de nombres dont le premier chiffre significatif est 2, elle-même supérieure à la proportion de nombres dont le premier chiffre significatif est 3, et ainsi de suite.

La loi du premier chiffre significatif indique précisément que dans un contexte général, et sans raisons particulières opposées, les probabilités de rencontrer les différents chiffres en tête des nombres sont respectivement:

$p(1)=30{,}1\%$, $p(2)=17{,}6\%$, $p(3)=12{,}5\%$, $p(4)=9{,}7\%$, $p(5)=7{,}9\%$, $p(6)=6{,}7\%$, $p(7)=5{,}8\%$, $p(8)=5{,}1\%$, $p(9)=4{,}6\%$.

Bien qu'aujourd'hui désignée sous le nom de loi de Benford, cette loi a été formulée la première fois par l'astronome canadien Simon Newcomb en 1881. Il avait remarqué que les premières pages des tables de logarithmes étaient plus abîmées que les suivantes. Son article fut ignoré jusqu'à ce que, 57 ans plus tard, le physicien américain Frank Benford remarquât lui aussi l'usure inégale des pages des tables numériques. Les articles de Newcomb et de Benford proposent la même formule: la probabilité de rencontrer le chiffre c comme premier chiffre d'un nombre est d'après eux $\log_{10}(c+1)-\log_{10}(c)$, où le logarithme utilisé est le logarithme décimal qui, rappelons-le, vérifie:

$\log_{10}(1)=0$;
$\log_{10}(10^n)=n$ pour n entier;
$\log_{10}(ab)=\log_{10}(a)+\log_{10}(b)$;
$\log_{10}(a/b)=\log_{10}(a)-\log_{10}(b)$.

Toutes les séries statistiques ne vérifient pas cette loi de Benford. La taille d'un humain adulte mesurée en centimètres commence, à de rares exceptions près, par 1 et ne la vérifie donc pas. De même, les plaques d'immatriculation des véhicules automobiles ont des numéros qui, dans chaque pays, sont le plus souvent bien répartis: autant de numéros commençant par 1, que par 2, etc. Pour que la loi se manifeste, il semble nécessaire que les nombres de la série envisagée prennent des valeurs variant sur plusieurs ordres de grandeur (c'est le cas des tailles des villes), et qu'ils soient assez régulièrement espacés (*voir l'encadré 1 pour d'autres détails*).

Pour certaines suites numériques, la loi de Benford n'est pas conjecturée, mais prouvée.

Jean-Paul Delahaye a récemment publié: **Le fabuleux nombre π** (Belin, 2018).

LES SÉRIES NUMÉRIQUES DE BENFORD ET SA LOI

1

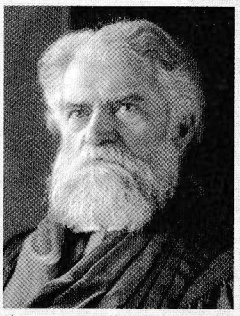
Simon Newcomb, véritable inventeur de la loi dite aujourd'hui de Benford, qu'il a formulée en 1881, était un astronome, mathématicien, économiste et statisticien né au Canada. Il a été professeur de mathématiques et d'astronomie à l'université Johns-Hopkins et fut éditeur de l'*American Journal of Mathematics*. En plus de l'anglais, il parlait le français, l'allemand, l'italien et le suédois. Il était aussi alpiniste, a publié plusieurs ouvrages de vulgarisation scientifique et un roman de science-fiction.

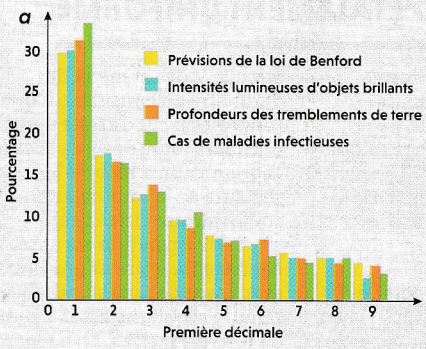
Dans son article de 1938, Franck Benford prend 20 séries de nombres (longueurs de rivière, surfaces géographiques, séries mathématiques, tables de mortalité, etc.) et calcule pour chacune la proportion de nombres commençant par 1, par 2, etc. Toutes les séries suivent à peu près la loi qu'il énonce et qui affirme que dans une série de nombres, on trouvera le plus souvent une proportion de nombres commençant par c égale à $\log_{10}(c+1) - \log_{10}(c)$. Une multitude d'autres confirmations approchées de la loi ont depuis été apportées *(graphique a)*.

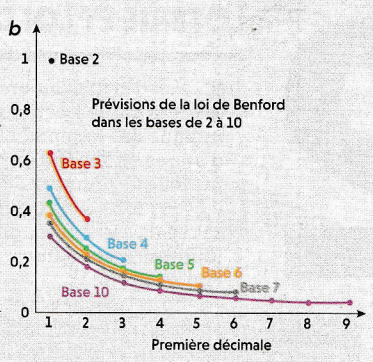
Il existe des versions de la loi de Benford pour chaque base de numération *(graphique b)*, et elles aussi ont été confrontées avec un bon succès aux séries numériques réelles. La loi de Benford n'est donc pas liée à l'usage de la base 10 par les humains. Notons que la loi reste vérifiée (sauf exception) quand on change les unités de mesure, par exemple quand on passe des kilomètres aux miles pour mesurer des distances.

La loi de Benford s'exprime sous une forme continue (dont découle la version en termes de premier chiffre significatif) : la probabilité pour que la partie fractionnaire du logarithme décimal des valeurs d'une série soit comprise entre *a* et *b* avec $a < b$ est égale à $(b - a)$, c'est-à-dire que la probabilité est uniforme sur l'intervalle [0, 1] *(voir le texte principal pour plus de détails sur ce point)*.

On a remarqué que la loi n'est pas systématiquement vraie, mais que lorsque les données sont bien étalées et ne présentent pas d'irrégularité manifeste, elle est presque toujours approximativement vraie. Quand on multiplie les données, il y a bien convergence vers des probabilités qui favorisent le 1 comme premier chiffre par rapport au 2, le 2 par rapport au 3, etc., mais les probabilités limites observées ne sont que rarement celles annoncées par la loi. Selon une grande expérience portant sur 230 ensembles de données, seuls 13 % des ensembles se conforment de très près à la loi de Benford. Aujourd'hui, on dispose d'une explication claire de cette constatation.


Simon Newcomb (1835-1909)





C'est le cas pour les puissances de 2 (2, 4, 8, 16, 32, ...). On a démontré qu'à l'infini, la proportion des puissances de 2 qui commencent par le chiffre 1 est exactement $\log_{10}(2) - \log_{10}(1)$, que la proportion des puissances de 2 qui commencent par 2 est $\log_{10}(3) - \log_{10}(2)$, etc.

L'ÉTALEMENT UNIFORME ATTENDU

Peut-on formuler une explication simple de ce que nous observons et prouvons pour certaines suites numériques? Il semble que oui. Celle que nous allons présenter maintenant convainc la plupart de ceux qui font l'effort de la comprendre. Elle pourra sembler un peu compliquée, mais c'est la justification intuitive la plus générale connue, et ce qu'elle suggère est utile, comme nous allons le voir.

L'explication se fonde sur une première loi qui est rarement énoncée, car sans doute jugée trop simple, et que nous appellerons «loi d'étalement uniforme de la partie fractionnaire». Avant de la formuler, fixons des notations qui aideront à s'exprimer efficacement. Soit r un nombre réel (par exemple $r = 2{,}71828...$). Sa partie entière est le plus grand entier inférieur ou égal à r, que nous noterons $[r]$. Sa partie fractionnaire est $r - [r]$ et sera notée $\{r\}$: ainsi, $[2{,}71828] = 2$ et $\{2{,}71828\} = 0{,}71828$. En langage simple, la partie entière est ce qui se trouve devant la virgule, la partie fractionnaire ce qui se trouve derrière.

LA LOI D'ÉTALEMENT DE LA PARTIE FRACTIONNAIRE

Si l'on choisit des nombres réels r au hasard dans un intervalle large de plusieurs unités (par exemple entre 0 et 20), et que la loi qui indique >

> la probabilité de tomber sur une des valeurs possibles est assez régulière et étalée, alors la partie fractionnaire des nombres r sera, à peu de chose près, uniformément répartie dans l'intervalle des nombres entre 0 et 1.

Considérons par exemple la moyenne générale des notes (sur 20) des élèves d'une école. On trouve des nombres du type: 10,54; 12,43; 7,23; 11,97; 12,41; 13,80; 16,55... dont les parties fractionnaires sont: 0,54; 0,43; 0,23; 0,97; 0,41; 0,80; 0,55...

Les notes ne seront pas nécessairement uniformément étalées sur tout l'intervalle allant de 0 à 20, et il est même probable qu'il y aura un grand nombre de notes autour de 10 ou 11. En revanche, ce qu'indique la loi d'étalement uniforme est que la partie fractionnaire de ces notes sera, elle, bien étalée entre 0 et 1. En particulier, il y aura à peu près autant de parties fractionnaires commençant par 0 (après la virgule) que de parties fractionnaires commençant par 1, que de parties fractionnaires commençant par 2, etc. Avec sans doute de petites variations autour de 1/10, la proportion de notes sera la même pour chacune des 10 catégories.

Plus généralement, si l'on se donne deux nombres a et b compris entre 0 et 1 avec $a<b$, la proportion de notes dont la partie fractionnaire est comprise entre a et b vaut $b-a$, la longueur de l'intervalle $[a, b]$. Dans notre exemple, la proportion de notes d'élèves dont la partie fractionnaire est comprise entre 0,25 et 0,40 vaut environ 15%.

L'explication de cette loi est que, sauf cas particuliers, les parties fractionnaires des nombres ne seront pas concentrées sur la même zone de l'intervalle [0, 1]. S'il y a plus de nombres entre 12,3 et 12,4, cela ne sera pas vrai (sauf exception) entre 15,3 et 15,4. Ainsi, les irrégularités possibles de densité sur les 20 intervalles possibles entre deux entiers consécutifs se compenseront plus ou moins, ce qui uniformisera la série des parties fractionnaires, que l'on peut voir comme une sorte de moyenne de ce qui se passe sur chacun des 20 intervalles entre deux entiers. Une interprétation graphique de cette idée est

2 LOTERIE ET LOI D'ÉTALEMENT UNIFORME

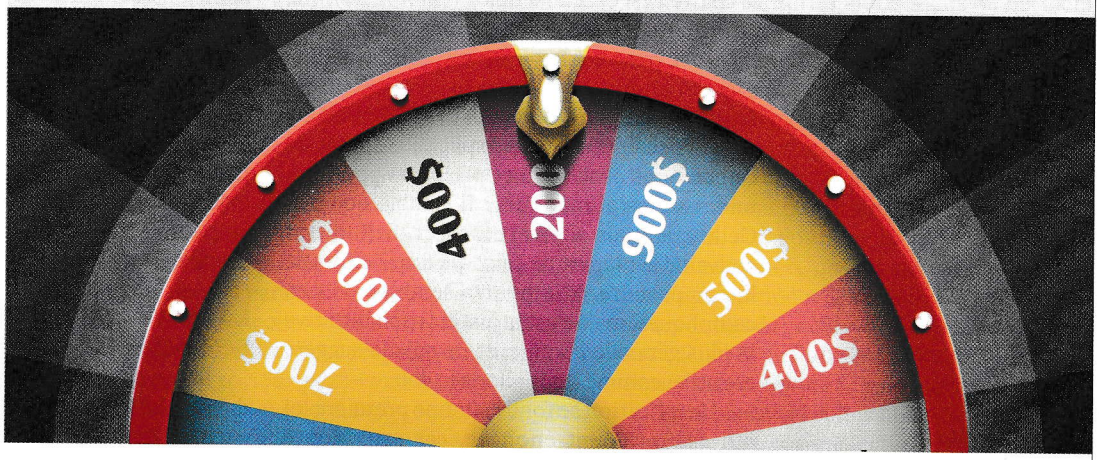
La loi d'étalement uniforme de la partie fractionnaire indique que les nombres d'une série dont on enlève ce qui précède la virgule (2,3456 devient 0,3456 ; 132,4388 devient 0,4388) se répartissent à peu près uniformément dans l'intervalle [0, 1].

Tous les joueurs de loterie acceptent (inconsciemment) la loi d'étalement uniforme de la partie fractionnaire. En effet, lancer assez fort la roue d'une loterie de un mètre de périmètre et regarder où se place la marque indiquant la case gagnante revient au même que lancer fort une boule sur une rainure circulaire de un mètre de périmètre en considérant que l'endroit où elle s'arrête après plusieurs tours désigne la case gagnante (on a recopié les secteurs de la roue de loterie sur la rainure circulaire).

Cette opération est aussi équivalente à lancer une boule avec assez de puissance sur une rainure rectiligne assez longue où tous les mètres, on a recopié les dessins de la rainure circulaire désignant les cases gagnantes.

Cette opération est encore équivalente à choisir un nombre réel assez grand, à prendre sa partie fractionnaire et à considérer un dessin des secteurs gagnants entre 0 et 1 reproduisant celui utilisé périodiquement sur la rainure rectiligne.

Si vous pensez que la loterie est équitable, c'est-à-dire que la probabilité que la marque s'arrête sur un secteur du disque qui tourne est proportionnelle à l'angle de ce secteur, alors vous pensez aussi que la partie fractionnaire de nombres au hasard assez grands est, à peu de chose près, uniformément distribuée entre 0 et 1. Vous croyez donc à la loi d'étalement uniforme de la partie fractionnaire et donc aussi à la loi de Benford qui en découle.



© MicroOne/Shutterstock.com

proposée dans l'encadré 3, ainsi qu'une analogie avec la croyance en l'équité d'une loterie dans l'encadré 2.

À côté de l'idée intuitive et un peu vague, il y a bien évidemment des formulations précises de conditions mathématiques qui assurent que l'écart à l'uniformité est faible, ou même qu'il est nul; les lecteurs intéressés pourront les consulter dans les articles de la bibliographie.

L'origine de cette loi dont on va comprendre l'importance centrale est difficile à retracer. En 1912, Henri Poincaré a exprimé un principe équivalent sans énoncer de théorème précis. Il expliquait à propos de la troisième décimale des nombres trouvés dans une table de logarithmes qu'on observera «que les dix chiffres 0, 1, 2, 3, ..., 9 sont également répartis sur cette liste et par conséquent la probabilité pour que cette troisième décimale soit paire est égale à 1/2». Pour lui, cet énoncé d'uniformité était évident; il écrivit: «[...] un instinct invincible porte à le penser».

L'idée de l'uniformisation des parties fractionnaires a été reprise par le mathématicien d'origine croate William Feller dans son célèbre traité de théorie des probabilités de 1966, mais comme on l'a noté depuis, en voulant formuler un énoncé mathématique, il s'est trompé. Sans connaître ni le texte de Poincaré ni celui de Feller, et cette fois avec un énoncé et une démonstration exacts, Nicolas Gauvrit et moi avons retrouvé l'idée en 2008, en même temps qu'un autre énoncé précis différent et correct était proposé indépendamment par les chercheurs allemands Lutz Dümbgen et Christoph Leuenberger. L'impossibilité de classer les résultats mathématiques pour les retrouver facilement comme on classe les mots d'un dictionnaire a pour conséquence que, souvent, des résultats qui ne se rattachent pas clairement à un domaine précis des mathématiques sont découverts indépendamment plusieurs fois.

ÉVIDENCE DE LA LOI DE BENFORD

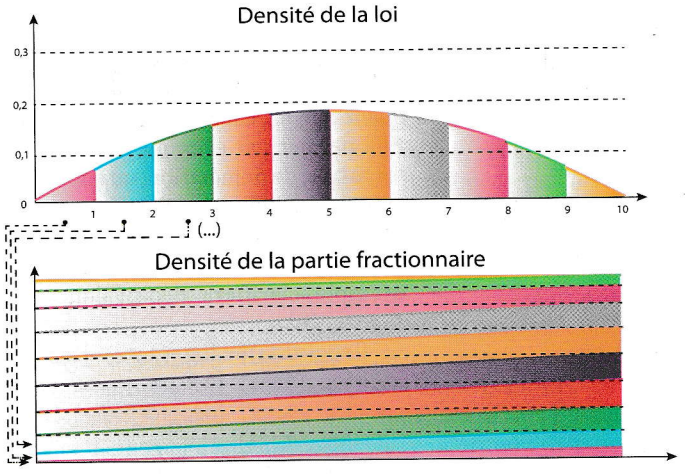
Grâce à cette loi d'étalement à la fois intuitive et formalisable, on va obtenir une justification naturelle et simple de la loi de Benford. L'idée consiste simplement à appliquer un logarithme décimal à la loi précédente... et à réfléchir un peu.

On reprend l'énoncé précédent et on l'applique non pas aux nombres r de la série considérée, mais à leur logarithme décimal, $\log_{10}(r)$. Si l'on choisit des nombres réels r au hasard sur une large plage couvrant plusieurs ordres de grandeur (par exemple entre 1 et 10^{20}), et que la loi qui indique la probabilité de tomber sur une des valeurs possibles est assez régulière et étalée, alors les parties fractionnaires des logarithmes décimaux des nombres, c'est-à-dire les $\{\log_{10}(r)\}$, seront, à peu de chose près, uniformément réparties entre 0 et 1.

3

LA LOI D'ÉTALEMENT DE LA PARTIE FRACTIONNAIRE

La partie fractionnaire $\{r\}$ d'un nombre r est ce qui est derrière la virgule. Par exemple, $\{52{,}98734\} = 0{,}98734$. La loi d'étalement uniforme de la partie fractionnaire permet de déduire la loi de Benford. En voici l'énoncé : si l'on choisit au hasard des nombres réels sur une plage large de plusieurs unités (par exemple entre 1 et 10), et que la loi donnant la probabilité de tomber sur une des valeurs possibles est assez régulière, alors la partie fractionnaire des nombres qu'on trouvera sera, à peu de chose près, uniformément répartie entre 0 et 1. La figure ci-dessous, due à Nicolas Gauvrit, illustre l'idée intuitive de la loi d'étalement uniforme de la partie fractionnaire. Dans le graphique du bas, les tranches de la densité de départ sont superposées après réduction proportionnelle de taille, pour former la densité de la partie fractionnaire. Les pentes des tranches se compensent à peu près. Dans le cas d'une fonction linéaire sur chaque intervalle $[n, n+1]$, la compensation se fait parfaitement, et donc la partie fractionnaire est dans un tel cas parfaitement uniforme.



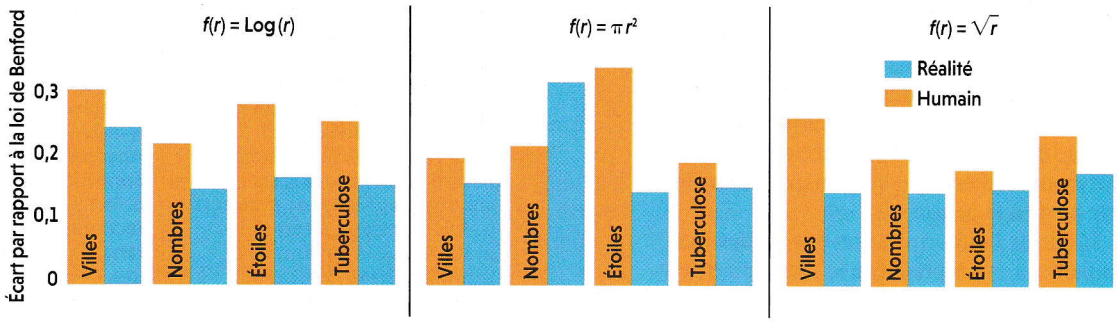
On ne le voit pas d'emblée, mais ce qu'on vient d'énoncer est la loi de Benford (ou plus exactement une loi plus puissante parfois dénommée «loi de Benford continue»). En effet, affirmer que c est le premier chiffre significatif du nombre r équivaut à énoncer que $\log_{10}(c) \leq \{\log_{10}(r)\} < \log_{10}(c+1)$, ce que l'on justifiera un peu plus loin.

Les parties fractionnaires des images par $\log_{10}$ des nombres r dont le premier chiffre significatif est c occupent donc dans l'intervalle [0, 1] un intervalle de longueur $\log_{10}(c+1) - \log_{10}(c)$, ce qui signifie, si l'on admet l'uniforme répartition, que leur >

4 LA LOI DE BENFORD GÉNÉRALE SOUMISE À DES TESTS

La loi de Benford générale affirme que la partie fractionnaire des nombres $f(r)$ obtenus à partir d'une série de nombres r sera uniformément étalée sur l'intervalle [0, 1] et cela pour toute fonction continue croissante f. Comme la loi de Benford classique, fondée sur $f(r) = \log_{10}(r)$, on peut utiliser la loi générale pour repérer des données truquées.

Afin de s'assurer que la méthode est bonne, on a demandé à des sujets humains de créer de fausses données et on a mesuré la qualité de ce qu'ils produisaient en utilisant la loi de Benford générale pour trois fonctions différentes : $f(r) = \text{Log}(r)$, $f(r) = \pi r^2$ et $f(r) = \sqrt{r}$. On a utilisé des données provenant de quatre sources différentes qui ont conduit à 12 comparaisons entre la réalité et ce que produit un humain simulant la réalité *(voir le protocole détaillé dans le texte principal)*. Dans 11 cas sur 12, la simulation par l'humain est moins conforme à la loi de Benford que les données réelles. Il est donc possible, en effectuant des tests fondés sur la loi de Benford, de détecter des données artificielles produites par des humains.



> proportion est $\log_{10}(c+1) - \log_{10}(c)$. C'est exactement ce qu'exprime la loi de Benford formulée au sujet du premier chiffre significatif en base décimale !

La loi de Benford continue est effectivement un peu plus puissante que la loi n'évoquant que les chiffres significatifs en base 10. Elle permet par exemple de retrouver, dans toute base de numération, un énoncé évoquant le premier chiffre significatif.

La justification de l'équivalence utilisée plus haut se fait rigoureusement, mais un exemple clarifiera l'équivalence. Prenons r=7234. On a :

$$\{\log_{10}(7234)\} = \log_{10}(7234) - [\log_{10}(7234)]$$
$$= \log_{10}(7234) - [3{,}8593\ldots] = \log_{10}(7234) - 3$$
$$= \log_{10}(7234) - \log_{10}(10^3) = \log_{10}(7234/1000)$$
$$= \log_{10}(7{,}234).$$

Puisque $\log_{10}$ est une fonction croissante, on a : $\log_{10}(7) < \log_{10}(7{,}234) < \log_{10}(8)$.

Soustraire $[\log_{10}(r)]$ à $\log_{10}(r)$ ramène la valeur dans le $\log_{10}$ entre 1 et 10 et donc l'encadrement entre $\log_{10}(c)$ et $\log_{10}(c+1)$ repère le premier chiffre significatif. D'où l'équivalence indiquée.

ON VEUT DES PREUVES !

Comme précédemment, le principe énoncé, c'est-à-dire la loi de Benford continue, est un peu vague, et il faudrait préciser quand elle ne s'applique pas, à l'aide de résultats mathématiques démontrés.

De tels énoncés mathématiques existent ; ils sont un peu compliqués à formuler *(voir l'article de Michel Valadier cité dans la bibliographie)*. Ils consistent à formuler des hypothèses exprimant l'idée d'étalement et de régularité, et, en fonction de la précision avec laquelle les hypothèses sont imposées, ils affirment que la série de nombres vérifie la loi de Benford avec une précision que le théorème explicite.

Il faut être un peu prudent avec la formulation non formelle de la loi de Benford ; elle n'est que la traduction d'une intuition « invincible » comme l'écrit Poincaré, mais qui reste une intuition. Tout dépend de ce qu'on nomme « large plage de plusieurs ordres de grandeur » et « loi étalée et régulière ». Cependant, l'énoncé informel montre pourquoi la loi est souvent vérifiée, au moins approximativement.

La loi informelle explique aussi une autre propriété de la loi de Benford remarquée depuis longtemps : en augmentant la taille d'une série de nombres, on ne tend pas toujours vers les valeurs annoncées des probabilités $\log_{10}(c+1) - \log_{10}(c)$. L'explication est claire : si les nombres suivent une loi précise, la compensation entre les intervalles quand on passe aux parties fractionnaires des logarithmes ne se fera qu'exceptionnellement, et donc en multipliant les données, on ne converge pas vers une loi parfaitement uniforme sur [0, 1], mais vers une loi approximativement uniforme.

Comprendre est le meilleur moyen d'avancer et c'est ici le cas. L'identification de l'origine de la loi de Benford suggère une méthode simple de la généraliser : remplacer la fonction

$\log_{10}$ par une autre du même type, c'est-à-dire croissante et continue.

Si f est une fonction continue croissante, si l'on choisit des nombres réels r au hasard sur une large plage, et si la loi qui indique la probabilité de tomber sur une des valeurs possibles est assez régulière et étalée, alors la partie fractionnaire des $f(r)$ sera, à peu de choses près, uniformément répartie entre 0 et 1.

Là encore, des théorèmes sont possibles. Malheureusement, pour les fonctions f que l'on peut envisager (par exemple $f(x)=\sqrt{x}$, $f(x)=x^2$, $f(x)=\exp(x)$) il n'y a pas de traduction simple de la loi générale en termes de premier chiffre significatif. Ces généralisations ne sont donc pas aussi spectaculaires que la loi obtenue avec $f(x)=\log_{10}(x)$. Elles sont cependant utiles, car elles procurent des outils de détection de fraude.

UTILISER LA LOI DE BENFORD GÉNÉRALE POUR DÉTECTER DES FRAUDES

La loi de Benford a maintes fois été utilisée pour détecter des fraudes. Un livre récent est même consacré à ce sujet (Mark Nigrini, *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*, Wiley, 2012).

Le principe est simple: s'ils s'étalent régulièrement sur plusieurs ordres de grandeur, les nombres apparaissant dans des comptes ou des statistiques doivent, sauf raisons particulières, vérifier la loi de Benford. Si ce sont des nombres inventés, leur auteur risque d'avoir voulu en créer à peu près autant qui commencent par 1 que par 2, 3, etc. N'étant pas averti de la propriété qu'énonce la loi de Benford, le faussaire ne la respectera pas. Si les données suspectes ne concordent pas avec la loi de Benford, on conclura que les données ont été falsifiées.

Cela s'est produit à propos de l'expertise de données financières et fiscales, cela s'est produit en science où des tests ont repéré des données d'expériences frauduleuses. La loi de Benford a aussi été utilisée pour distinguer des images numériques artificielles d'images numériques naturelles ou pour identifier quelles étaient les images d'une série qui contenaient des données cachées (stéganographie).

Un problème apparaît. À force de parler de la loi de Benford, les truqueurs de données pourraient en être informés. Ils s'appliqueraient à produire des données inventées qui la respectent et passeraient ainsi à travers les outils de repérage statistique fondés sur la loi de Benford. La loi générale permet de contrer ce risque: en utilisant les variantes avec diverses fonctions f, on repérera les données falsifiées.

Pour tester cette méthode, une série d'expériences a été réalisée par un groupe de chercheurs réunis autour de Nicolas Gauvrit. Nous décrivons ici l'une des expériences menées. Les productions pseudoaléatoires humaines ont été examinées dans quatre contextes où l'on constate la loi de Benford sur les chiffres significatifs.

Un groupe de 169 adultes, recrutés *via* les réseaux sociaux ou par courrier électronique, a participé à cette expérience. Leurs âges variaient de 13 à 73 ans. Les participants ont été répartis au hasard en quatre groupes pour s'occuper de données:

(a) sur la population des 5000 villes américaines les plus peuplées,

(b) sur des constantes mathématiques provenant des tables de Simon Plouffe,

(c) sur les distances en années-lumière entre la Terre et les étoiles visibles les plus proches,

(d) sur les nombres de cas de tuberculose par pays pour l'année 2012.

Dans chaque groupe, les participants ont été informés que l'on avait sélectionné au hasard une série de 30 nombres dans les données réelles, et qu'ils devaient tenter de produire ce qu'ils pensaient être une série analogue et plausible de 30 nombres. Des séries de 30 nombres provenant des bases de données réelles ont aussi été constituées.

Pour chaque ensemble de 30 valeurs de r (fabriquées ou réelles), on a examiné les distributions des parties fractionnaires des $f(r)$, avec $f(r)=\mathrm{Log}(r)$, $f(r)=\pi r^2$ et $f(r)=\sqrt{r}$. L'écart par rapport à la loi de Benford générale a été mesuré par une méthode classique (la statistique de Kolmogorov-Smirnov). Comme prévu, les données fabriquées par les humains se conforment moins bien à la loi de Benford que les données réelles (*voir l'encadré 4*).

Sur les 12 comparaisons (3 fonctions, 4 types de données), une seule exception a été notée provenant de la comparaison des données de la table de constantes de Simon Plouffe testées avec la loi de Benford pour la fonction $f(r)=\pi r^2$: dans ce cas, les humains sont en moyenne plus conformes à ce que prévoit la loi de Benford que les données réelles!

La conclusion est donc claire: les humains se font repérer par la loi de Benford générale quand ils tentent de fabriquer de fausses données.

D'autres résultats de l'étude menée suggèrent aussi que, selon le type de données dont on veut contrôler l'authenticité, certains choix de la fonction remplaçant $\log_{10}$ dans la loi de Benford générale sont préférables à d'autres. Comprendre pourquoi sera important pour créer des outils de détection auxquels plus aucun fraudeur ne pourra échapper. Mais, c'est certain, grâce à la loi de Benford générale, la panoplie des chasseurs de tricheurs numériques s'est enrichie d'une nouvelle arme puissante. ■

BIBLIOGRAPHIE

N. Gauvrit *et al.*, **Generalized Benford's law as a lie detector**, *Advances in Cognitive Psychology*, vol. 13(2), pp. 121-127, 2017.

M. Valadier, **The Benford phenomenon for random variables. Discussion of Feller's way**, prépublication arXiv:1203.2518, 2012.

N. Gauvrit et J.-P. Delahaye, **Scatter and regularity imply Benford's Law... and more**, dans Hector Zenil (éd.), *Randomness through computation : Some answers, more questions*, World Scientific, ch. 4, pp. 53-69, 2011.

N. Gauvrit et J.-P. Delahaye, **La loi de Benford générale**, *Math. Sci. Hum. Math. Soc. Sci.*, vol. 186, pp. 5-15, 2009.

N. Gauvrit et J.-P. Delahaye, **Pourquoi la loi de Benford n'est pas mystérieuse**, *Math. Sci. Hum. Math. Soc. Sci.*, vol. 182, pp. 7-15, 2008.

J.-P. Delahaye, **L'étonnante loi de Benford**, *Pour la Science* n° 351, pp. 90-95, janvier 2007.