

R

ENDEZ-VOUS

P.74 Logique & calcul
 P.80 Art & science
 P.82 Idées de physique
 P.86 Science & fiction
 P.92 Chroniques de l'évolution
 P.96 Science & gastronomie
 P.98 À picorer

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

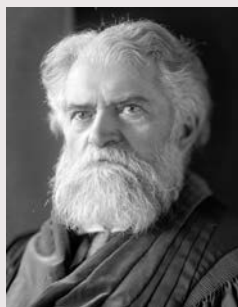
trouvera le plus souvent une proportion de nombres commençant par c égale à $\log_{10}(c+1) - \log_{10}(c)$. Une multitude d'autres confirmations approchées de la loi ont depuis été apportées (graphique a).

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.

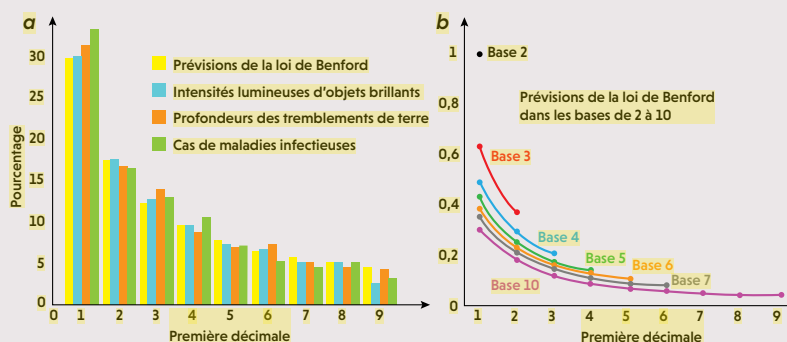
La loi de Benford s'exprime sous une forme continue (dont découle la version en termes de premier chiffre significatif) : la probabilité pour que la partie fractionnaire du logarithme décimal des valeurs d'une série

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 $r=2,71828\dots$). 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