

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

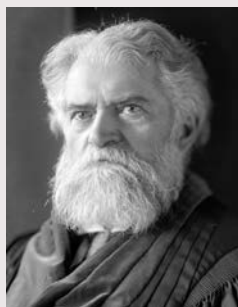
trouvera le plus souvent une proportion de nombres commençant par c égale à $\log_{10}(c+1) - \log_{10}(c)$. Une multitude d'autres confirmations approchées de la loi ont depuis été apportées (graphique a).

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.

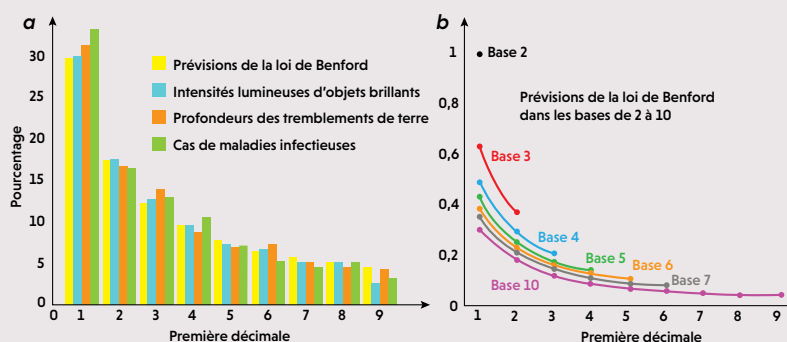
La loi de Benford s'exprime sous une forme continue (dont découle la version en termes de premier chiffre significatif) : la probabilité pour que la partie fractionnaire du logarithme décimal des valeurs d'une série

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

> 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

LOTÉRIE 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.



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 écrit : « [...] 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.

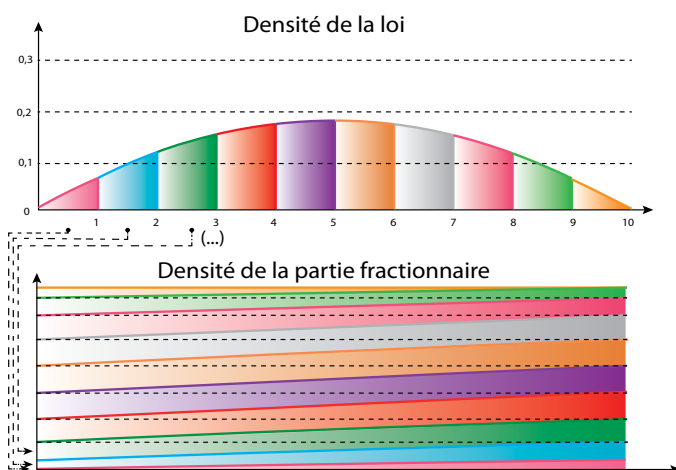
On reprend l'énoncé précédent et on l'applique non pas aux nombres r de la série considérée, mais à leur logarithme décimal, $\log_{10}(r)$. Si l'on choisit des nombres réels r au hasard sur une large plage couvrant plusieurs ordres de grandeur (par exemple entre 1 et 10^{20}), et que la loi qui indique 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 >