

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.

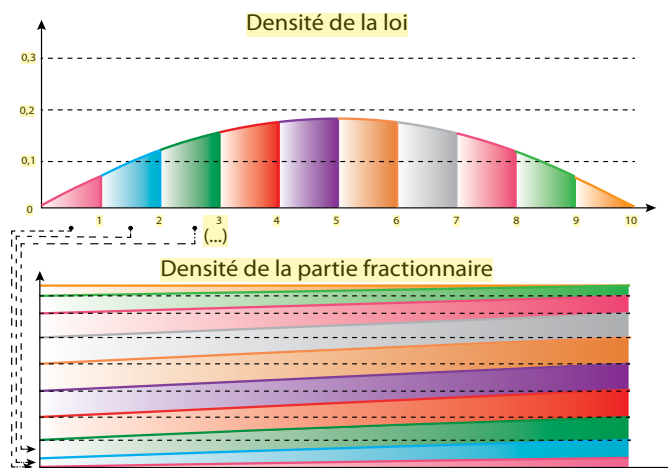
On reprend l'énoncé précédent et on l'applique non pas aux nombres r de la série considérée, mais à leur logarithme décimal, $\log_{10}(r)$. Si l'on choisit des nombres réels r au hasard sur une large plage couvrant plusieurs ordres de grandeur (par exemple entre 1 et 10^{20}), et que la loi qui indique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

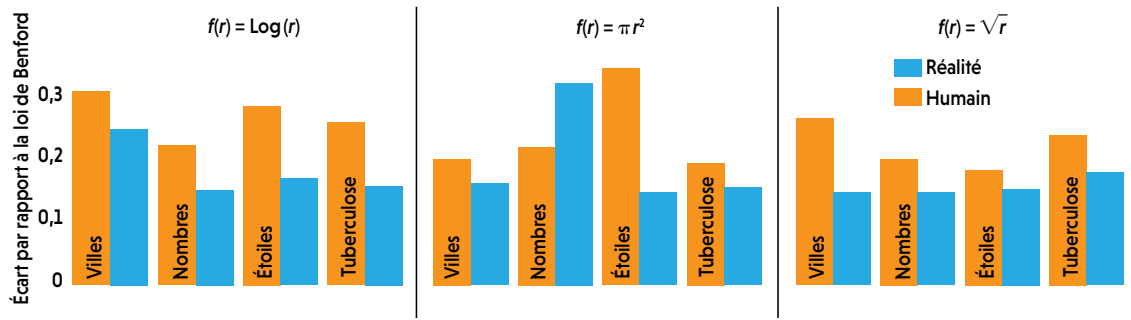
LA LOI DE BENFORD GÉNÉRALE SOUMISE À DES TESTS

4

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 :

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

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