L'étonnante loi de Benford

Paul Jolissaint, HEP-BEJUNE et Université de Neuchâtel, pajolissaint@sunrise.ch

1 Introduction

La répartition des chiffres $1, \ldots, 9$ en tant que premiers chiffres significatifs (par exemple 2 est le premier chiffre significatif de **2**345.6, 7 est celui de 0.0**7**8) dans un grand nombre d'ensembles de données numériques est contre-intuitive. Découverte une première fois par S. Newcomb [9] à la fin du XIX^{e} siècle en constatant que les premières pages des tables de logarithmes étaient plus usées que les suivantes, l'observation est passée inaperçue, puis a été redécouverte indépendamment en 1938 par F. Benford. Ce dernier a rassemblé 20'229 données numériques provenant de sources diverses et a appelé cette règle la **loi des nombres anormaux** [1]. Le nom de la loi a été attribué à Benford car, contrairement à son prédécesseur malheureux, son article a été abondamment lu. De très bonnes introductions "grand public" en français sont présentées dans [3] et [6], et une présentation mathématiquement rigoureuse et complète est contenue dans la monographie récente [2].

Nous allons motiver notre sujet par la description d'un exemple typique de fraude. En 1993, Wayne J. Nelson, un employé du Trésor de l'état d'Arizona, est reconnu coupable d'avoir détourné près de 2 millions de dollars en versant à des personnes fictives 23 chèques dont voici la liste des dates et des montants en dollars correspondants :

(a) Le 9 octobre 1992 : 1′927.48 et 27′902.31.

(b) Le 14 octobre 1992 : 86′241.90, 72′117.46, 81′321.75, 81′321.75 et 97′473.96.

(c) Le 19 octobre 1992 : 93′249.11, 89′658.17, 87′776.89, 92′105.83, 79′949.16, 87′602.93, 96′879.27, 91′806.47, 84′991.67, 90′831.83, 93′766.67, 88′338.72, 94′639.49, 83′709.28, 96′412.21, 88′432.86 et 71′552.16.

Voici quelques indices de fraude :
- Le fraudeur a commencé par de petites valeurs, puis les montants et leur nombre ont augmenté.
- Tous les montants étaient inférieurs à $100'000 : des montants supérieurs auraient sans doute fait l'objet de vérifications par un supérieur hiérarchique.
- Les chiffres significatifs sont trop grands : plus de 90 % admettent 7,8 ou 9 comme premier chiffre. De plus, chacune des paires de premiers chiffres 87, 88, 93 et 96 a été utilisée deux fois dans les 23 montants.

Or, la dernière observation est en contradiction avec la *loi de Benford*. Avant d'en donner la définition précise dans le paragraphe suivant, signalons que, lorsque l'on relève au hasard un nombre assez grand (disons au moins 200) de données numériques de l'un des ensembles suivants
- valeurs boursières
- listes de prix d'articles divers [1]
- valeurs numériques variées tirées de journaux
- grandeurs géographiques (populations de villes, superficies des lacs d'un certain continent, débit de rivières, etc.)
- dans une certaine mesure : les numéros des maisons dans les rues (aux USA, par exemple)
- ...

on constate que le premier chiffre significatif de ces valeurs est plus souvent 1 que 2, qui est plus fréquent que 3, etc., le chiffre le moins fréquent étant 9.

Et cette constatation reste vraie même si l'on change les unités !

1. Mon collègue Didier Müller du Lycée cantonal de Porrentruy m'a communiqué dernièrement l'anecdote suivante : en guise de préliminaire à la présentation de la loi de Benford dans l'une de ses classes, il avait demandé à ses élèves de relever les prix d'articles variés dans des supermarchés ou dans des magazines. Or, les valeurs présentées par un des élèves ne satisfaisaient pas la loi : l'élève a dû avouer qu'il n'avait pas réalisé le travail demandé et avait inventé les valeurs présentées !

2 Définitions

Donnons tout d'abord une première définition élémentaire de la loi de Benford.

Définition 2.1 *Un ensemble de valeurs numériques suit la* ***loi de Benford*** *si, pour chaque chiffre* $d \in \{1, \ldots, 9\}$ *la proportion de valeurs qui commencent par* d *est*

$$\log\left(\frac{d+1}{d}\right),$$

où $\log(x)$ *désigne le logarithme en base* 10 *du nombre* $x > 0$.

Explicitement, cela donne :

Chiffre	Fréquence théorique
1	$\log(2/1) \cong 0.301$
2	$\log(3/2) \cong 0.176$
3	$\log(4/3) \cong 0.125$
4	$\log(5/4) \cong 0.097$
5	$\log(6/5) \cong 0.079$
6	$\log(7/6) \cong 0.067$
7	$\log(8/7) \cong 0.058$
8	$\log(9/8) \cong 0.051$
9	$\log(10/9) \cong 0.046$

Avant de définir la loi générale, introduisons deux notations. Soit $x > 0$; son ***significande*** est l'unique nombre $S(x) \in [1, 10)$ tel que

$$x = S(x) \cdot 10^k$$

pour un certain entier k (nécessairement unique). Si $x > 0$, on note $D_1(x)$ le **premier chiffre significatif** de x ; c'est la partie entière de $S(x)$ et $D_1(x) \in \{1, \ldots, 9\}$.

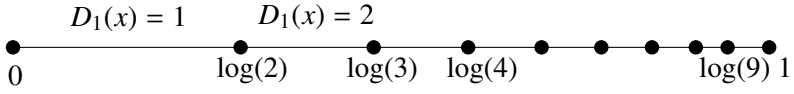
Nous pouvons maintenant interpréter le fait que, pour un certain $x > 0$, on ait $D_1(x) = d$. Cela signifie que $d \cdot 10^k \leq x < (d+1) \cdot 10^k$ pour un certain entier k. Autrement dit,

$$d \leq S(x) < d+1 \quad \text{ou encore} \quad \log(S(x)) \in [\log(d), \log(d+1)).$$

Or, $\log\left(\frac{d+1}{d}\right) = \log(d+1) - \log(d)$ est la **longueur** de l'intervalle

$$[\log(d), \log(d+1)).$$

Ainsi, dire qu'un ensemble de valeurs positives $\{x_1, \ldots, x_n, \ldots\}$ satisfait la loi de Benford au sens de la définition 2.1 revient à dire que, pour chacun des chiffres $d \in \{1, \ldots, 9\}$, la proportion des valeurs k telles que $D_1(x_k) = d$ est égale à la longueur de l'intervalle $[\log(d), \log(d+1))$, c'est-à-dire à $\log(d+1) - \log(d) = \log\left(\frac{d+1}{d}\right)$. L'ensemble des valeurs $\{\log(S(x_1)), \ldots, \log(S(x_n)), \ldots\}$ est donc **uniformément réparti dans l'intervalle** $[0, 1]$.



Cela mène à la définition générale suivante pour les suites de nombres positifs $(x_n)_{n \geq 1} = (x_1, x_2, \ldots,)$:

Définition 2.2 *Une suite* (x_n) *satisfait la* ***loi de Benford*** *si, pour tout* $t \in [0, 1]$,

$$\lim_{N \to \infty} \frac{\#\{1 \leq n \leq N : \log(S(x_n)) < t\}}{N} = t.$$

Remarquons que, pour tout $x > 0$, $\log(S(x)) = \langle \log(x) \rangle$, où $\langle \cdot \rangle$ désigne la partie fractionnaire (appelée aussi **mantisse** du logarithme). La loi générale implique bien la première puisque, pour tout digit $1 \leq d \leq 9$, la proportion des valeurs x_k telles que $d \leq D_1(x_k) < d+1$ est égale à la différence

$$\lim_{N\to\infty} \frac{\#\{1 \leq n \leq N : \log(S(x_n)) < \log(d+1)\}}{N} - \lim_{N\to\infty} \frac{\#\{1 \leq n \leq N : \log(S(x_n)) < \log(d)\}}{N} = \log(d+1) - \log(d).$$

La preuve rigoureuse que certaines suites satisfont la loi de Benford repose sur le critère suivant dû à H. Weyl [10] : tout d'abord, disons qu'une suite $(a_n) \subset \mathbb{R}$ est **uniformément distribuée modulo 1** (abrégé u.d. mod1) si, pour tout $t \in [0,1]$,

$$\lim_{N\to\infty} \frac{\#\{1 \leq n \leq N : \langle a_n \rangle < t\}}{N} = t.$$

Théorème 2.3 *(Théorème de Weyl) La suite (a_n) est u.d.* mod1 *si et seulement si, pour tout $k \in \mathbb{Z}$, $k \neq 0$, on a*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} e^{2k\pi i a_n} = 0.$$

Idée de la preuve. Dire que $(a_n)_{n\geq 1}$ est u.d. mod1 revient à dire que la loi de probabilité de la suite des parties fractionnaires $(\langle a_n \rangle)_{n\geq 1}$ est la loi uniforme sur $[0,1]$. Or, la fonction caractéristique (transformée de Fourier) de la loi uniforme est la distribution de Dirac δ_0 sur $\mathbb{Z}$:

$$\delta_0(k) = \begin{cases} 1 & \text{si } k = 0 \\ 0 & \text{si } k \neq 0. \end{cases}$$

D'un autre côté, la fonction caractéristique de la distribution

$$F_N(t) = \frac{\#\{1 \leq n \leq N : \langle a_n \rangle < t\}}{N}$$

est la fonction $k \mapsto \frac{1}{N}\sum_{n=1}^{N} e^{2k\pi i \langle a_n \rangle} = \frac{1}{N}\sum_{n=1}^{N} e^{2k\pi i a_n}$. Le théorème de continuité de Lévy-Cramér permet de conclure. □

Exemple Soit $a > 0$ un nombre irrationnel. Alors la suite $(na)_{n\geq 1}$ est u.d. mod1. En effet, fixons $k \neq 0$. On a

$$\begin{aligned} \frac{1}{N}\sum_{n=1}^{N} e^{2k\pi i n a} &= \frac{1}{N}\sum_{n=1}^{N} (e^{2k\pi i a})^n \\ &= e^{2k\pi i a} \frac{1 - e^{2k\pi i N a}}{N(1 - e^{2k\pi i a})} \to_{N\to\infty} 0. \end{aligned}$$

Par suite, si $r > 0$ est un nombre réel, la suite géométrique $(r^n)_{n\geq 1}$ satisfait la loi de Benford si (et seulement si) $\log(r)$ est irrationnel puisque $\log(r^n) = n\log(r)$ pour tout n.

Remarque Cet exemple permet de comprendre pourquoi un grand nombre d'ensembles de valeurs numériques satisfont (approximativement) la loi de Benford : si l'on classe un tel ensemble de valeurs par ordre croissant et si la suite ainsi obtenue est approximativement une suite géométrique de raison $r > 0$ telle que $\log(r)$ est irrationnel, alors la suite satisfait approximativement la loi de Benford.

3 Exemples et applications

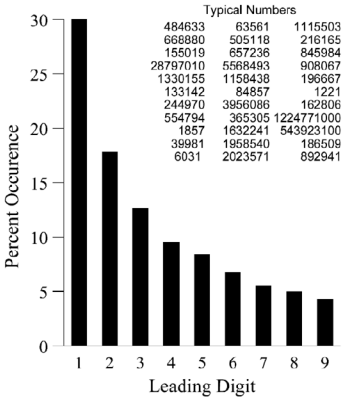
Donnons tout d'abord quelques exemples d'ensembles de valeurs qui ne satisfont *pas* la loi de Benford :

1. L'ensemble des numéros de téléphone d'une région donnée ; des ensembles de numéros de cartes de crédit, ou des ensembles de codes de sécurité ou de contrôle (code ISBN), numéros AVS.
2. Les ensembles qui suivent une loi normale : les valeurs sont plus ou moins concentrées autour de la moyenne. Par exemple, l'ensemble des tailles des adultes d'une région donnée (quelle que soit l'unité choisie !).

3. Tout ensemble de nombres (pseudo-)aléatoires : un tel ensemble satisfait une répartition uniforme.

Les ensembles suivants en revanche satisfont au moins approximativement la loi de Benford.

1. De nombreuses études montrent que les comptabilités des grandes entreprises et leurs revenus imposables suivent la loi de Benford ; cela a permis à M. Nigrini de proposer des tests qui utilisent la loi de Benford comme indice de fraude.



D'après S. W. Smith, 2007

2. Les populations des villes américaines : lors du recensement de juillet 2009 par exemple, 19'509 villes ont été recensées, de *Abbeville (Alabama)* à *Yoder (Wyoming)*, et on constate que la loi de Benford est relativement bien satisfaite :

Chiffre	**Fréquence observée**	**Fréquence théorique**
1	0.294	0.301
2	0.181	0.176
3	0.120	0.125
4	0.094	0.097
5	0.079	0.079
6	0.070	0.067
7	0.060	0.058
8	0.053	0.051
9	0.046	0.046

3. (P. Diaconis [5]) La suite $(n!)$ satisfait la loi de Benford mais pas la suite des nombres premiers $p_1 = 2, p_2 = 3, \ldots$.
4. (P. Jolissaint [7], [8]) Soit $r > 0$ un nombre réel tel que $\log(r)$ soit irrationnel, et soit $P(n)$ un polynôme de degré positif, à coefficients entiers et tel que $P(n) \to +\infty$ lorsque $n \to +\infty$. Alors la sous-suite $(r^{P(n)})_{n\geq 1}$ satisfait encore la loi de Benford. Ainsi, par exemple les suites (2^{n^2}) ou (2^{n^3-n}) satisfont la loi de Benford.
5. Plus généralement, c'est le cas de très nombreuses suites qui satisfont une relation de récurrence linéaire telle que la suite de Fibonacci (F_n) par exemple ainsi que chaque sous-suite $(F_{P(n)})$ où P désigne un polynôme comme ci-dessus.
6. (H. Deligny et P. Jolissaint [4]) Soit ℓ un entier positif tel que $\log(\ell) \notin \mathbb{Q}$; alors la sous-suite (p_{ℓ^n}) de la suite des nombres premiers satisfait la loi de Benford.
7. (Berger et Hill [2]) Soit I un intervalle compact, soit $f : I \to \mathbb{R}$ une fonction analytique et soit $x^* \in I$ un zéro de f. Soit $(x_n)_{n\geq 0}$ une suite obtenue par la formule de Newton :

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

Alors pour presque toute valeur initiale x_0 proche de x^*, les suites $(|x_n - x^*|)$ et $(|x_{n+1} - x_n|)$ satisfont la loi de Benford.

8. Le nombre d'articles publiés par année sur la loi de Benford depuis 1938 (*sic*).

Mentionnons pour terminer trois applications plus ou moins récentes.

(1) Détection de fraudes (erreurs ou falsifications de données) dans les comptabilités et les audits : Mark Nigrini a amassé dès le début des années 1990 un grand nombre de preuves empiriques qui justifient l'usage de la loi de Benford comme indicateur de fraude. Si la fraude est délibérée, les données suivent rarement la loi de Benford.

(2) Traitement d'images (Pérez-González et al.) : l'application à la transformée en cosinus discrète d'images aide à déterminer si elles contiennent des messages cachés (stéganographie). La méthode n'est pas plus fiable que d'autres, mais plus simple.

(3) En référence avec l'exemple 2 ci-dessus, si un modèle mathématique est conçu pour décrire l'évolution de populations ou d'un ensemble de données qui satisfont la loi de Benford, une condition de pertinence du modèle est qu'il devra lui aussi satisfaire cette loi.

Références

[1] F. Benford. The law of anomalous numbers. *Proc. Amer. Philosiphical Soc.*, 78 :551–572, 1938.

[2] A. Berger and T. P. Hill. *An Introduction to Benford's Law*. Princeton University Press, Princeton and Oxford, 2015.

[3] J.-P. Delahaye. L'étonnante loi de Benford. *Pour La Science*, 351 :90–95, 2007.

[4] H. Deligny and P. Jolissaint. Relations de récurrence linéaires, primitivité et loi de Benford. *Elem. Math.*, 68 :9–21, 2013.

[5] P. Diaconis. The distribution of leading digits and uniform distribution mod 1. *Ann. Prob.*, 5 :72–81, 1977.

[6] T. Hill. Le premier chiffre significatif fait sa loi. *La Recherche*, 316 :72–75, 1999.

[7] P. Jolissaint. Loi de Benford, relations de récurrence et suites équidistribuées. *Elem. Math.*, 60 :10–18, 2005.

[8] P. Jolissaint. Loi de Benford, relations de récurrence et suites équidistribuées II. *Elem. Math.*, 64 :21–36, 2009.

[9] S. Newcomb. Note on the frequency of use of the different digits in natural numbers. *Amer. J. Math.*, 4 :39–40, 1881.

[10] H. Weyl. Ueber die Gleichverteilung von Zahlen mod 1. *Math. Ann.*, 77 :313–352, 1916.