Association Mathématique du Québec

Article

La loi de Newcomb-Benford ou la loi du premier chiffre significatif

Vincent Genest, Université de Montréal,
Christian Genest, Université McGill

Résumé

La loi de Newcomb–Benford stipule que dans un jeu de données couvrant plusieurs ordres de grandeur, la proportion d'observations dont le premier chiffre significatif est d devrait approcher $\log_{10}(1 + 1/d)$. Les auteurs proposent une justification probabiliste de ce phénomène, qu'ils illustrent dans divers cas. Après avoir rappelé certaines propriétés de cette loi de probabilité, notamment sa caractérisation par invariance sous transformation linéaire de l'échelle de mesure, ils expliquent comment vérifier si elle s'ajuste bien à des données au moyen d'un test du khi-deux. Tous leurs résultats s'appuient sur des notions élémentaires de calcul différentiel et intégral.

1 Introduction

Selon l'édition 2010 du *CIA World Factbook* [4], le monde compte actuellement 194 états, dont la taille varie de 824 (le Vatican) à plus d'un milliard de personnes (la République populaire de Chine). Avec ses 34 millions d'habitants, le Canada figure au 36$^{\text{e}}$ rang des pays les plus peuplés de la planète.

L'effectif de population du Vatican commence par un 8, celui du Canada par un 3 et celui de la Chine par un 1. Si on procède ainsi pour l'ensemble des pays du monde et que l'on compile les résultats, on obtient le tableau 1. Manifestement, les états dont l'effectif commence par un 1 sont beaucoup plus nombreux que ceux dont l'effectif commence par un 3 ou par un 8... Devrait-on s'en étonner ?

Premier chiffre	1	2	3	4	5	6	7	8	9
Nombre de pays	49	36	23	24	13	14	13	10	12
Fréquence observée	0,253	0,186	0,119	0,123	0,067	0,072	0,067	0,051	0,062

Tableau 1. Répartition du premier chiffre significatif des effectifs de population de 194 pays du monde selon l'édition 2010 du *CIA World Factbook*.

Si on répète l'exercice avec d'autres ensembles de données, on constate que le phénomène se reproduit et que, dans bien des cas, la fréquence d'apparition des 1 est supérieure à celle des 2, elle-même

supérieure à celle des 3, et ainsi de suite. Cette observation a d'abord été rapportée dans la littérature scientifique par Simon Newcomb, astronome, économiste et mathématicien de renom.

Dans un article paru en 1881, Newcomb [12] avait en effet remarqué que les premières pages des tables logarithmiques, auxquelles on se référait constamment à l'époque pour toutes sortes de calculs, étaient plus usées que les dernières. Il avait alors été conduit à postuler que lorsqu'une variable aléatoire X strictement positive couvre plusieurs ordres de grandeur (ses valeurs s'étalant entre 10^{-5} et 10^5, par exemple), la partie fractionnaire de son logarithme (en base 10) est à peu près équidistribuée. Comme on le verra au §2, il s'ensuit qu'étant donné un échantillon de valeurs de X, la proportion d'observations dont le premier chiffre significatif est $d \in \{1, \ldots, 9\}$ devrait alors être proche de

$$Q_1(d) = \log_{10}(1 + 1/d).$$

Les valeurs numériques de $Q_1(1), \ldots, Q_1(9)$ sont données au tableau 2. De prime abord, la correspondance entre ce tableau et le précédent paraît très bonne, quoiqu'imparfaite. De fait, comme un test du khi-deux permet de le confirmer (voir §6 pour de plus amples détails), les écarts entre les deux séries ne sont pas statistiquement significatifs ($p \approx 0{,}77$).

Premier chiffre	1	2	3	4	5	6	7	8	9
Fréquence espérée	0,301	0,176	0,125	0,097	0,079	0,067	0,058	0,051	0,046

TABLEAU 2. Fréquence relative espérée du premier chiffre significatif selon la loi de Newcomb–Benford.

Les manifestations de la loi formulée par Newcomb [12] abondent dans la nature. Le physicien américain Frank Benford a largement contribué à populariser le phénomène par la publication, en 1938, d'un article [3] dans lequel il en donnait de nombreux exemples tirés de sources aussi variées que la Ligue américaine de baseball, les adresses de savants et une pléiade de constantes physico-chimiques telles les poids moléculaires, les chaleurs spécifiques et des données sur les corps noirs.

De nos jours, l'illustration la plus probante de la loi de Newcomb–Benford est sans doute fournie par « l'inverseur de Plouffe », qui contient à lui seul plus de 215 millions de constantes mathématiques dont environ 30,1% commencent par un 1, 17,6% commencent par un 2, etc. Cette banque de données, élaborée par le mathématicien canadien Simon Plouffe, est disponible en ligne à `http ://pi.lacim.uqam.ca/fra/`.

Le but du présent travail est de décrire en termes simples l'origine et les propriétés de la loi de Newcomb–Benford. On distingue d'abord au §2 deux versions de cette loi qui sont parfois confondues dans la pléthore d'écrits sur le sujet. Au §3, on démontre que la loi de Newcomb–Benford émerge naturellement comme cas limite d'une suite (X_n) de variables aléatoires dont l'étalement croît (dans un sens à préciser) à mesure que $n \to \infty$. On explique ensuite au §4 pourquoi cette loi de probabilité est caractérisée par son invariance sous transformation linéaire de l'échelle de mesure. Deux variantes de la loi de Newcomb–Benford sont énoncées au §5 et on montre au §6 comment il est possible de vérifier si cette loi s'ajuste bien à des données à l'aide d'un test du khi-deux. Enfin, on donne au §7 une courte biographie de Newcomb et de Benford.

À l'exception peut-être de la proposition 1 et de son corollaire, les résultats présentés ici ne sont pas nouveaux. Sans être forcément originales, leurs démonstrations s'appuient sur des notions élémentaires de probabilité et de calcul différentiel et intégral qui les rendent plus abordables que celles de Hill [8, 9], entre autres. L'article a été conçu pour être accessible à un public de niveau collégial.

2 Formalisme mathématique

Soit X un aléa strictement positif. Toute valeur x de cette variable aléatoire peut être exprimée de façon unique sous la forme

$$x = s \times 10^k,$$

où $k \in \mathbb{Z}$ est un entier relatif et s est un nombre réel compris dans l'intervalle $[1, 10)$. Le scalaire s est souvent appelé le *significande* et sa partie entière est le premier chiffre significatif de x. On dira en outre que $\log_{10}(s) \in [0, 1)$ est la *mantisse* de x, bien que cette acception ne soit pas universelle. Par exemple si $x = 1/13\ 983\ 816 \approx 7{,}15 \times 10^{-8}$ (la probabilité de gagner le gros lot au Lotto 6/49), le significande de x est $s \approx 7{,}15$ et la mantisse est $\log_{10}(s) \approx 0{,}85$.

Étant donné un entier $d \in \{1, \ldots, 10\}$, définissons l'événement

$$E(d) : \text{« le premier chiffre significatif de } X \text{ est strictement inférieur à } d \text{ ».}$$

Puisque le premier chiffre significatif de $X = S \times 10^K$ est le même que celui de son significande, cet événement peut aussi s'exprimer sous la forme

$$E(d) : \text{« le premier chiffre significatif de } S \text{ est strictement inférieur à } d \text{ »,}$$

de sorte que $E(d)$ est réalisé si et seulement si $1 \leqslant S < d$ ou encore $0 \leqslant \log_{10}(S) < \log_{10}(d)$.

Or $\log_{10}(X) = \log_{10}(S) + K$ pour un certain entier $K \in \mathbb{Z}$ dont la valeur n'a aucune influence sur le premier chiffre significatif. L'événement $E(d)$ se produit donc si et seulement si

$$\log_{10}(X) \in \bigcup_{k \in \mathbb{Z}} [k, k + \log_{10}(d)).$$

De façon plus générale, si on pose

$$\mathcal{M}(\epsilon) = \bigcup_{k \in \mathbb{Z}} [k, k + \epsilon)$$

pour tout $\epsilon \in [0, 1]$, la probabilité

$$P(\epsilon) = \Pr\{\log_{10}(X) \in \mathcal{M}(\epsilon)\} \tag{1}$$

est aussi celle de l'événement $0 \leqslant \log_{10}(S) < \epsilon$.

Définition 1 *On dit qu'une variable aléatoire X obéit à la loi faible de Newcomb–Benford si et seulement si $P(\epsilon) = \epsilon$ pour tout $\epsilon \in \{\log_{10}(1), \log_{10}(2), \ldots, \log_{10}(10)\}$.*

Définition 2 *On dit qu'une variable aléatoire X obéit à la loi forte de Newcomb–Benford si et seulement si $P(\epsilon) = \epsilon$ pour tout $\epsilon \in [0, 1]$.*

La loi de Newcomb–Benford a d'abord été formulée dans sa version forte par Newcomb [12]. La version faible, qui en découle, a été énoncée par Benford [3]. Toutefois, celui-ci ne semble pas avoir été au fait des travaux de Newcomb ou des observations déjà rapportées sur le phénomène par d'illustres prédécesseurs dont Poincaré [16]. La forme faible de la loi de Newcomb–Benford est néanmoins suffisante pour déterminer la probabilité que l'entier $d \in \{1, \ldots, 9\}$ soit le premier chiffre significatif de X. Pour calculer explicitement cette probabilité, il suffit de remarquer que l'événement « obtenir d comme premier chiffre significatif » s'écrit $E(d+1) \setminus E(d)$. Ainsi, on trouve

$$\begin{aligned} \Pr\{E(d+1)\} - \Pr\{E(d)\} &= \log_{10}(d+1) - \log_{10}(d) \\ &= \log_{10}(1 + 1/d), \end{aligned}$$

ce qui est bien le résultat annoncé dans l'introduction.

3 Justification de la loi de Newcomb–Benford

C'est une chose d'affirmer que le premier chiffre significatif vaut d avec probabilité $\log_{10}(1 + 1/d)$; c'en est une autre d'expliquer *pourquoi* il en est ainsi… ou plutôt *quand* on peut s'attendre à ce qu'il en soit ainsi. Car la loi de Newcomb–Benford n'est pas toujours vérifiée. À l'évidence, un échantillon de tailles d'adultes exprimées en mètres n'obéit pas à la loi de Newcomb–Benford puisque presque toutes les mesures commencent par un 1 !

Pour comprendre la genèse de la loi de Newcomb–Benford, imaginons un grand ensemble d'observations d'une certaine variable X s'étalant entre 10^{-7} et 10^{7} (par exemple) et traçons l'histogramme des valeurs de $\log_{10}(X)$. Le résultat pourrait ressembler à la figure 1, sur laquelle l'abscisse a aussi été partiellement graduée en fonction de la variable X elle-même pour illustrer les ordres de grandeur.

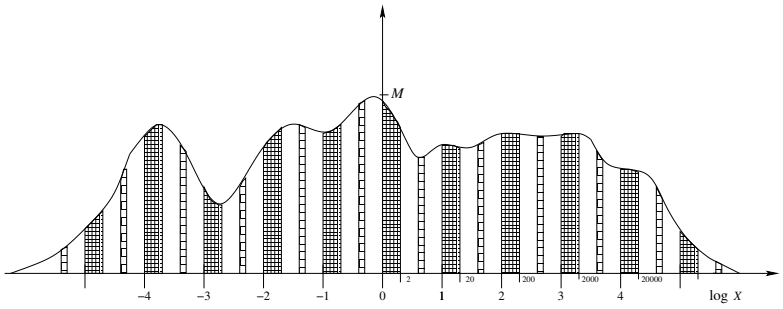


FIG. 1 – Distribution de $\log_{10}(X)$ pour une variable fictive X couvrant plusieurs ordres de grandeur. L'abscisse est partiellement graduée en fonction de X. Les plages quadrillées et échelonnées représentent respectivement la proportion des valeurs de X commençant par un 1 et un 4.

Sur la figure 1, on a quadrillé l'aire sous la courbe pour les valeurs de $\log_{10}(X)$ comprises entre k et $k+\log_{10}(2)$ pour un entier $k \in \mathbb{Z}$ quelconque ; ceci correspond donc aux valeurs de X dont la mantisse $\log_{10}(S)$ se situe entre $\log_{10}(1) = 0$ et $\log_{10}(2) = 0{,}301$. De même, les surfaces échelonnées sont celles pour lesquelles S varie entre 4 et 5. Ainsi, l'aire des zones quadrillées représente la proportion de valeurs de X commençant par un 1 et celle des zones échelonnées reflète la proportion des données commençant par un 4.

De toute évidence, la surface des zones quadrillées est plus grande que celle des zones échelonnées. Autrement dit, la distribution représentée à la figure 1 compte beaucoup plus de nombres commençant par un 1 que de nombres commençant par un 4. De fait, c'est aussi le cas pour chacun des ordres de grandeur pris séparément.

Le graphique suggère en outre que la variable $\log_{10}(S)$ est distribuée à peu près uniformément sur l'ensemble du domaine de X, bien que ce ne soit pas forcément le cas pour chacun des ordres de grandeur pris séparément.

Le résultat suivant, inspiré des travaux de Gauvrit et Delahaye [7], formalise l'argument ci-dessus et permet d'expliquer pourquoi la loi de Newcomb–Benford se vérifie dans de nombreux cas concrets.

Proposition 1 *Soit X un aléa strictement positif de densité continue. Supposons que la densité f de $\log_{10}(X)$ soit majorée par une constante $M > 0$ et qu'il existe des entiers $K \in \mathbb{N}$ et $L \in \mathbb{Z}$ tels que f soit croissante sur $(-\infty, L]$ et décroissante sur $[L + 2K, \infty)$. Pour tout $\epsilon \in [0, 1]$, on a alors $|P(\epsilon) - \epsilon| \leqslant 2(K + 1)M$.*

La démonstration du résultat, fournie à l'annexe A, repose sur le fait que l'on peut écrire

$$P(\epsilon) = \int_{\mathcal{M}(\epsilon)} f(y)\,\mathrm{d}y = \sum_{k\in\mathbb{Z}} \int_k^{k+\epsilon} f(y)\,\mathrm{d}y.$$

Pour saisir en quoi la proposition 1 permet de justifier la loi de Newcomb–Benford, considérons une suite (X_n) d'aléas strictement positifs dont les fonctions de répartition sont continûment dérivables. Soit f_n la densité de $\log_{10}(X_n)$, dont on suppose qu'elle est majorée par une constante $M_n > 0$. Admettons de plus que $f_n(x)$ soit strictement décroissante en $|x|$ lorsque $x \notin [L_n, L_n + 2K_n]$, où $K_n \in \mathbb{N}$ et $L_n \in \mathbb{Z}$ pour tout $n \in \mathbb{N}$. Étant donné $\epsilon \in [0, 1]$, posons en outre

$$P_n(\epsilon) = \Pr\{\log_{10}(X_n) \in \mathcal{M}(\epsilon)\}.$$

Le majorant M_n peut être interprété comme un indice d'étalement de $\log_{10}(X_n)$. En effet, si un aléa générique Y de variance finie admet une densité continue majorée par une constante $M > 0$, alors

$$\mathrm{var}(Y) \geqslant \frac{1}{12M^2}\,. \tag{2}$$

Ce résultat, rapporté entre autres par Fréchet [6] sans démonstration, est établi à l'annexe B. Or dans le cas présent, si la suite (K_n) est bornée et si $M_n \to 0$ lorsque $n \to \infty$, il découle de la proposition 1 que $P_n(\epsilon) \to \epsilon$ pour tout $\epsilon \in [0, 1]$. En d'autres termes, plus la loi de $\log_{10}(X_n)$ est étalée dans ce sens particulier, plus X_n se conformera de près à la loi forte de Newcomb–Benford.

Un énoncé formel de ce résultat est donné ci-dessous pour mémoire.

Corollaire 1 *Soit (X_n) une suite d'aléas strictement positifs de densités continues. Supposons que pour tout $n \in \mathbb{N}$, la densité f_n de $\log_{10}(X_n)$ soit majorée par une constante $M_n > 0$ et qu'il existe des entiers $K_n \in \mathbb{N}$ et $L_n \in \mathbb{Z}$ tels que f_n soit croissante sur $(-\infty, L_n]$ et décroissante sur $[L_n+2K_n, \infty)$. Supposons en outre que*

$$\lim_{n\to\infty} M_n = 0 \quad \textit{et} \quad \limsup_{n\to\infty} K_n < \infty.$$

Alors pour tout $\epsilon \in [0,1]$,

$$\lim_{n\to\infty} P_n(\epsilon) = \epsilon.$$

Nous allons maintenant donner quelques illustrations élémentaires de la proposition 1. À notre connaissance, le principal intérêt de ces exemples réside dans leur commodité de calcul.

Exemple 1 Soit (Z_n) une suite d'aléas mutuellement indépendants tels que pour tout $n \in \mathbb{N}$, la variable $\log_{10}(Z_n)$ a une distribution gaussienne (ou normale) d'espérance $\mu \in \mathbb{R}$ et de variance $\sigma^2 \in (0, \infty)$; de façon équivalente, on dit que Z_n obéit à une loi log-normale. La variable

$$Y_n = \log_{10}(Z_1 \times \cdots \times Z_n) = \sum_{i=1}^{n} \log_{10}(Z_i) \tag{3}$$

est alors gaussienne, d'espérance $n\mu$ et de variance $n\sigma^2$. Par conséquent, sa densité est définie en tout $y \in \mathbb{R}$ par

$$f_n(y) = \left(\frac{1}{2\pi n\sigma^2}\right)^{1/2} \exp\left\{-\frac{1}{2}\,\frac{(y-n\mu)^2}{n\sigma^2}\right\}.$$

Cette densité satisfait les conditions de la proposition 1. En effet, puisqu'elle atteint son unique maximum en $n\mu$ (appelé le mode), on peut prendre par exemple $L_n = \lfloor n\mu \rfloor$ et $K_n = 1$, où de manière générale $\lfloor t \rfloor$ dénote la partie entière de $t \in \mathbb{R}$. De plus, f_n est majorée par $M_n = f_n(n\mu) = 1/\sqrt{2\pi n\sigma^2}$. Puisque $M_n \to 0$ rapidement quand $n \to \infty$, la distribution du premier chiffre significatif de $X_n = Z_1 \times \cdots \times Z_n$ suivra donc d'assez près la loi de Newcomb–Benford, même pour n petit.

De fait, il n'y a rien de particulier au choix de la loi gaussienne dans l'exemple 1, sinon qu'il permet le calcul explicite de la loi de Y_n et du majorant M_n de sa densité. Supposons de façon plus générale que les variables $\log_{10}(Z_1), \log_{10}(Z_2), \ldots$ soient mutuellement indépendantes et de même loi. Si cette loi est de variance finie, il découle alors de la version classique du Théorème central limite que la distribution de $Y_n = \log_{10}(Z_1) + \cdots + \log_{10}(Z_n)$ est asymptotiquement normale. En invoquant ce résultat, on devrait donc arriver à démontrer que la variable $X_n = Z_1 \times \cdots \times Z_n$ se conforme de mieux en mieux à la loi de Newcomb–Benford, à mesure que $n \to \infty$.

Un argument à cet effet déborderait largement le cadre du présent article et ne sera donc pas présenté. Remarquons toutefois que l'existence des moments de $\log_{10}(Z_n)$ n'est pas une condition préalable à l'obtention de la loi de Newcomb–Benford. Supposons par exemple que pour tout $n \in \mathbb{N}$,

$\log_{10}(Z_n)$ obéisse plutôt à une loi de Cauchy–Lorentz (ou de Breit–Wigner) standard, dont la densité est donnée en tout $y \in \mathbb{R}$ par

$$f(y) = \frac{1}{\pi(1+y^2)}\,.$$

Cette loi de probabilité ne possède ni espérance ni variance. On peut néanmoins vérifier (grâce à la transformée de Fourier) que la densité de la variable Y_n définie en (3) est alors donnée en tout $y \in \mathbb{R}$ par

$$f_n(y) = \frac{n}{\pi(n^2+y^2)}\,.$$

Cette densité étant décroissante en $|y|$ pour tout $y \in \mathbb{R}$, on est à nouveau dans les conditions d'application de la proposition 1 ; en effet, il suffit de prendre $L_n = 0$ et $K_n = 1$ (entre autres choix) et de noter que $M_n = f_n(0) = 1/(n\pi) \to 0$ quand $n \to \infty$.

L'exemple suivant reprend en substance un résultat de Adhikari et Sarkar [1].

Exemple 2 Soit (Z_n) une suite de variables aléatoires mutuellement indépendantes équidistribuées sur l'intervalle $(0,1)$. La variable $-\ln(Z_n)$ obéit alors à une loi exponentielle standard (c'est-à-dire d'espérance 1) puisque pour tout $t > 0$, on a

$$\Pr\{-\ln(Z_n) > t\} = \Pr(Z_n \leqslant e^{-t}) = e^{-t}.$$

Il s'ensuit que $T_n = -\ln(Z_1 \times \cdots \times Z_n)$, qui est une somme de variables aléatoires exponentielles indépendantes, obéit à une loi Gamma de paramètres $\alpha = n$ et $\beta = 1$, encore appelée loi de Erlang de paramètre n (il s'agit à nouveau d'un résultat classique que l'on peut aisément démontrer en examinant la transformée de Laplace de T_n). La densité de T_n est ainsi donnée en tout $t > 0$ par

$$g_n(t) = \frac{1}{(n-1)!}\, t^{n-1} e^{-t}.$$

On se convainc sans peine que son maximum se produit en $t = n-1$.

Si on s'intéresse maintenant à la variable $Y_n = \log_{10}(Z_1 \times \cdots \times Z_n) = -T_n/\ln(10)$, laquelle est toujours négative, il est clair que pour tout $y < 0$, on a

$$\Pr(Y_n \leqslant y) = \Pr(T_n > -y\ln(10)\} = 1 - \int_0^{-y\ln(10)} g_n(t)\ \mathrm{d}t.$$

En dérivant par rapport à y les deux membres de l'équation, on trouve que la densité de Y_n est

$$f_n(y) = \ln(10) \times g_n\{-y\ln(10)\}$$

en tout $y < 0$. Cette densité répond aux conditions de la proposition 1. En effet, puisqu'elle atteint son maximum en $-(n-1)/\ln(10)$, on peut prendre (par exemple) $L_n = \lfloor -(n-1)/\ln(10) \rfloor$ et $K_n = 1$. De plus,

$$M_n = \ln(10) \times g_n(n-1) = \frac{\ln(10)}{(n-1)!}\,(n-1)^{n-1} e^{1-n}$$

majore f_n. Vu que $M_n \to 0$ rapidement quand $n \to \infty$, la variable $X_n = Y_1 \times \cdots \times Y_n$ épousera de très près la loi de Newcomb–Benford, même pour n petit.

Remarque 1 Un simple coup d'œil à la démonstration donnée à l'annexe A permet de constater que la borne $2(K+1)M$ énoncée dans la proposition 1 n'est pas optimale. Le résultat pourrait donc être raffiné, mais il est suffisamment général pour les fins présentes. À titre d'illustration, le tableau 3 donne la répartition du premier chiffre significatif dans un échantillon de taille 50 000 de valeurs de $X = Z_1 \times \cdots \times Z_5$ lorsque $Z_1, \ldots, Z_5$ forment un échantillon aléatoire de loi uniforme sur l'intervalle $(0, 1)$. Les fréquences espérées sous la loi de Newcomb–Benford y sont également données. L'ajustement entre les deux séries est excellent, comme en témoigne le test du khi-deux ($p \approx 0{,}88$).

Premier chiffre	1	2	3	4	5	6	7	8	9
Fréquence observée	15 127	8 919	6 107	4 736	3 927	3 284	2 926	2 608	2 366
Fréquence espérée	15 051	8 805	6 247	4 845	3 959	3 347	2 900	2 558	2 288

TABLEAU 3. Fréquences observées du premier chiffre significatif dans un échantillon aléatoire de taille 50 000 du produit $Z_1 \times \cdots \times Z_5$ de cinq nombres choisis aléatoirement dans l'intervalle $(0, 1)$, ainsi que les fréquences espérées correspondantes sous la loi de Newcomb–Benford.

4 Caractérisation par invariance

On appelle transformation linéaire d'échelle toute application de la forme $X \mapsto cX$ pour un certain $c > 0$. Songeons par exemple à la conversion d'un montant X d'une devise à une autre. Puisque

$$\log_{10}(cX) = \log_{10}(c) + \log_{10}(X),$$

la loi de $\log_{10}(cX)$ est la même que celle de $\log_{10}(X)$, à une translation près. Si la densité de $\log_{10}(X)$ satisfait les conditions de la proposition 1 pour certains choix de $K \in \mathbb{N}$ et $L \in \mathbb{Z}$, la densité de $\log_{10}(cX)$ y répond aussi si on remplace K par $K+1$ et L par $L + \lfloor \log_{10}(c) \rfloor$.

La proposition suivante montre que cette propriété d'invariance est non seulement satisfaite par la loi de Newcomb–Benford, mais qu'elle la caractérise.

Proposition 2 *Une variable aléatoire X strictement positive obéit à la version forte de la loi de Newcomb–Benford si et seulement si, pour tout $c > 0$ et tout $\epsilon \in [0, 1]$, on a*

$$P_c(\epsilon) \equiv \Pr\{\log_{10}(cX) \in \mathcal{M}(\epsilon)\} = P(\epsilon), \tag{4}$$

où $P(\epsilon)$ est défini en (1).

La démonstration du résultat s'appuie sur le fait que pour tout $\delta \in [0, 1]$, on a

$$\bigcup_{k \in \mathbb{Z}} [k - \delta, k) = \mathbb{R} \setminus \mathcal{M}(1 - \delta)$$

et donc

$$\Pr\left\{\log_{10}(X) \in \bigcup_{k \in \mathbb{Z}} [k - \delta, k)\right\} = 1 - P(1 - \delta).$$

Démonstration Fixons $c \in (0,\infty)$ et $\epsilon \in [0,1]$. Si $\delta \in [0,1)$ dénote la mantisse de c, alors

$$\log_{10}(cX) \in \mathcal{M}(\epsilon) \quad \Leftrightarrow \quad \log_{10}(X) \in \bigcup_{k\in\mathbb{Z}} [k-\delta, k-\delta+\epsilon).$$

Deux cas peuvent alors se produire. Si $\delta \in [0,\epsilon)$, on peut écrire

$$\bigcup_{k\in\mathbb{Z}} [k-\delta, k-\delta+\epsilon) \quad = \quad \bigcup_{k\in\mathbb{Z}} [k, k+\epsilon-\delta) \quad \cup \quad \bigcup_{k\in\mathbb{Z}} [k-\delta, k)$$

comme la réunion de deux ensembles disjoints, de sorte que

$$P_c(\epsilon) = P(\epsilon-\delta) + 1 - P(1-\delta). \tag{5}$$

Si $\delta \in (\epsilon, 1)$, alors

$$\bigcup_{k\in\mathbb{Z}} [k-\delta, k-\delta+\epsilon) \quad = \quad \bigcup_{k\in\mathbb{Z}} [k, k+\epsilon+1-\delta) \quad \setminus \quad \bigcup_{k\in\mathbb{Z}} [k, k+1-\delta)$$

et il s'ensuit que

$$P_c(\epsilon) = P(\epsilon+1-\delta) - P(1-\delta). \tag{6}$$

Si X obéit à la version forte de la loi de Newcomb–Benford, on a $P(\epsilon) = \epsilon$ pour tout $\epsilon \in [0,1]$. On déduit donc des équations (5) et (6) que $P_c(\epsilon) = \epsilon$ quels que soient $c \in (0,\infty)$ et $\epsilon \in [0,1]$.

Pour démontrer la réciproque, il suffit d'observer que lorsque l'équation (4) est vérifiée, la relation (6) entraîne que

$$P(x) + P(y) = P(x+y)$$

pour tous $x, y \in [0,1]$ tels que $x+y \leqslant 1$. On reconnaît ici l'équation fonctionnelle de Cauchy. Puisque P est une fonction de probabilité, elle est évidemment bornée ; un résultat de Kestelman [11] permet alors de conclure que la seule solution possible est l'identité, à savoir $P(\epsilon) = \epsilon$ pour tout $\epsilon \in [0,1]$.

On peut facilement se convaincre de la justesse du résultat de Kestelman dans le cas particulier où la fonction P est continue sur $[0,1]$ et dérivable sur $(0,1)$. En effet, on a

$$\frac{P(x+z) - P(x)}{z} = \frac{P(y+z) - P(y)}{z}$$

pour tous $x, y \in (0,1)$ et tout $z \in (0,1)$ suffisamment petit. En prenant la limite lorsque $z \to 0$, on déduit que $P'(x) = P'(y)$ pour tous $x, y \in (0,1)$, ce qui entraîne que la fonction P' est constante sur l'intervalle $(0,1)$. Puisque $P(0) = 0$ et $P(1) = 1$, il s'ensuit que $P(x) = x$ pour tout $x \in [0,1]$. □

Remarque 2 Il existe d'autres caractérisations de la loi de Newcomb–Benford. En particulier, Hill [8] a démontré qu'elle est la seule qui soit invariante par changement de base. Janvresse et de la Rue [10] ont également montré que cette loi émerge naturellement dans le cas où les observations sont issues d'un mélange de lois uniformes. Une connaissance du calcul différentiel et intégral suffit pour comprendre le résultat de Janvresse et de la Rue ; les travaux de Hill font cependant appel à des notions beaucoup plus avancées.

5 Variations sur le thème

Deux variantes de la loi de Newcomb–Benford sont présentées ici. En se servant de la proposition 1, on déduit d'abord au §5.1 la distribution du $\ell^{\text{ième}}$ chiffre significatif. On présente ensuite au §5.2 une généralisation de la loi de Newcomb–Benford au cas de nombres exprimés en d'autres bases.

5.1 Loi du $\ell^{\text{ième}}$ chiffre significatif

Les conséquences de la loi forte de Newcomb–Benford s'étendent au delà de la loi du premier chiffre significatif. Considérons par exemple l'événement

« les ℓ premiers chiffres significatifs de X sont $d_1, \ldots, d_\ell$, dans l'ordre ».

Celui-ci se produit si et seulement si

$$\sum_{i=1}^{\ell} d_i \times 10^{-i+1} \leqslant S < 10^{-\ell+1} + \sum_{i=1}^{\ell} d_i \times 10^{-i+1}, \tag{7}$$

où S est le significande de X. Ainsi, les quatre premiers chiffres de X sont 1763 si $1{,}763 \leqslant S < 1{,}764$, événement qui s'exprime sous la forme (7) avec $\ell = 4$, $d_1 = 1$, $d_2 = 7$, $d_3 = 6$ et $d_4 = 3$. Sous la loi forte de Newcomb–Benford, la probabilité de cet événement est donnée par

$$\log_{10}\left(10^{-\ell+1} + \sum_{i=1}^{\ell} d_i \times 10^{-i+1}\right) - \log_{10}\left(\sum_{i=1}^{\ell} d_i \times 10^{-i+1}\right),$$

ce qui s'écrit aussi

$$\log_{10}\left(1 + 10^{-\ell+1} \Big/ \sum_{i=1}^{\ell} d_i \times 10^{-i+1}\right).$$

On trouve donc que

$$\Pr\{\text{les } \ell \text{ premiers chiffres significatifs de } X \text{ sont } d_1, \ldots, d_\ell\} = \log_{10}\left(1 + \frac{1}{\text{“}d_1 \cdots d_\ell\text{”}}\right), \tag{8}$$

où “$d_1 \cdots d_\ell$” est l'entier représenté par la concaténation des chiffres $d_1, \ldots, d_\ell$. Ainsi, la probabilité que le développement décimal de X commence par 1763 est $\log_{10}(1 + 1/1763) \approx 2{,}46 \times 10^{-4}$.

La formule (8) permet de calculer aisément la probabilité de tout événement d'intérêt.

Exemple 3 Soit X une variable aléatoire strictement positive obéissant à la loi forte de Newcomb–Benford. Étant donné un entier $d \in \{0, \ldots, 9\}$, la probabilité que le $\ell^{\text{ième}}$ chiffre significatif de X soit égal à d est donnée (pour $\ell \geqslant 2$) par

$$Q_\ell(d) = \sum_{d_1=1}^{9} \sum_{d_2=0}^{9} \cdots \sum_{d_{\ell-1}=0}^{9} \log_{10}\left(1 + 1 \Big/ \sum_{i=1}^{\ell-1} d_i \times 10^{\ell-i} + d\right).$$

En particulier, la formule se réduit à l'expression suivante dans le cas $\ell = 2$:

$$Q_2(d) = \sum_{d_1=1}^{9} \log_{10}\left(1 + \frac{1}{10\,d_1 + d}\right).$$

Les valeurs numériques de $Q_2(d)$ pour $d \in \{0, \ldots, 9\}$ sont rapportées au tableau 4. On constate que la fréquence relative du second chiffre significatif prédite par la loi de Newcomb–Benford est presque uniforme. De fait, on peut montrer que si $\ell \to \infty$, $Q_\ell(d)$ tend vers $1/10$ pour tout $d \in \{0, \ldots, 9\}$.

Second chiffre	0	1	2	3	4	5	6	7	8	9
Fréquence espérée	0,120	0,114	0,109	0,104	0,100	0,097	0,093	0,090	0,088	0,085

TABLEAU 4. Fréquence relative espérée du second chiffre significatif selon la loi de Newcomb–Benford.

Remarque 3 Il est amusant de noter que sous la loi de Newcomb–Benford, les chiffres successifs du développement décimal d'un nombre aléatoire ne sont pas stochastiquement indépendants. D'après le tableau 4, par exemple, la probabilité d'obtenir un 9 comme second chiffre significatif est d'environ 8,5% alors que si on sait déjà que le premier chiffre significatif est un 6, la formule (8) permet de conclure que la probabilité *conditionnelle* d'avoir un 9 comme deuxième chiffre significatif vaut

$$\frac{\log_{10}(1 + 1/69)}{\log_{10}(1 + 1/6)} \approx 9{,}3\%.$$

5.2 Loi de Newcomb–Benford en base b

Le lecteur attentif aura remarqué que dans les développements entourant les propositions 1 et 2, le fait que le logarithme de X soit exprimé en base 10 ne joue aucun rôle particulier. De fait, les résultats précédents restent valables *mutatis mutandis* pour tout choix de base $B = b + 1$, où $b \in \mathbb{N}$.

Pour comprendre pourquoi il en est ainsi, commençons par observer que tout nombre x peut s'exprimer en base B sous la forme

$$x = s \times B^k,$$

où $k \in \mathbb{Z}$ est un entier relatif et $s \in [1, B)$ est le significande (en base B). Par suite, l'événement

$$E(d) = \text{« le premier chiffre significatif de } X \text{ est strictement inférieur à } d \text{ »}$$

se produit si et seulement si

$$\log_B(X) \in \mathcal{M}(\log_B(d)),$$

où $\mathcal{M}(\epsilon)$ conserve le même sens qu'auparavant. Les démonstrations des propositions 1 et 2 peuvent donc être reprises intégralement et il s'ensuit entre autres que si X obéit à la loi forte de Newcomb–Benford, la fréquence relative d'apparition de l'entier $d \in \{1, \ldots, b\}$ comme premier chiffre significatif de son développement en base B est alors donnée par

$$\log_B(1 + 1/d) = \frac{\log_{10}(1 + 1/d)}{\log_{10}(B)} = \frac{\ln(1 + 1/d)}{\ln(B)}.$$

Ainsi, la probabilité que le premier chiffre significatif soit d diminue à mesure que B augmente, et vice versa. Le cas limite se produit quand $B = 2$, c'est-à-dire quand les nombres sont exprimés sous forme binaire. Dans ce cas, le premier chiffre significatif est forcément $d = 1$ et de fait, $\log_2(2) = 1$.

6 Applications pratiques

Après être longtemps restée dans l'ombre, la loi de Newcomb–Benford suscite actuellement un engouement considérable. Le site `http ://www.benfordonline.net/`, qui s'est donné pour mandat de répertorier les écrits sur le sujet, fait état de plus de trente articles par an depuis le début des années 2000, alors qu'il n'en recense qu'une soixantaine jusqu'en 1975 ; voir l'article de Raimi [17] pour une synthèse critique des travaux parus à cette date.

Même si elle a fait couler beaucoup d'encre, la loi de Newcomb–Benford a généralement été perçue jusqu'à récemment comme une simple curiosité mathématique. Le comptable américain Mark Nigrini semble avoir été le premier à lui conférer un sens pratique. Dans sa thèse [13], il a proposé d'utiliser la loi de Newcomb–Benford pour la détection de fraudes financières dans le cadre de vérifications comptables. Selon lui, les fréquences des premiers chiffres significatifs dans les états financiers d'une entreprise devraient se conformer à la loi de Newcomb–Benford. Des écarts indus pourraient donc suggérer d'éventuelles malversations. Par ses nombreux écrits, et plus particulièrement [14], Nigrini a fortement contribué à propager cette idée, qui a depuis été reprise par différents auteurs (voir par exemple [5, 18]) ; des logiciels commerciaux ont même été développés à cette fin.

Sans se prononcer sur l'à-propos d'une telle démarche, expliquons brièvement comment il est possible de tester qu'un échantillon aléatoire $V_1, \ldots, V_m$ d'une variable V se conforme (ou non) à la loi de Newcomb–Benford. Appelons $v_1, \ldots, v_q$ les valeurs possibles de V. Il pourrait s'agir par exemple des entiers $1, \ldots, 9$ ou des entiers $10, \ldots, 99$ selon que l'on s'intéresse au premier ou aux deux premiers chiffres significatifs. Les données peuvent alors être colligées et présentées comme suit :

Valeur de V	v_1	v_2	$\cdots$	v_q	Total
Fréquence relative observée	h_1	h_2	$\cdots$	h_q	1

Supposons maintenant que l'on veuille vérifier si ces fréquences sont cohérentes avec les probabilités stipulées par une loi $\mathcal{L}$ (celle de Newcomb–Benford, dans notre cas de figure), notées comme suit :

Valeur de V	v_1	v_2	$\cdots$	v_q	Total
Probabilité sous la loi $\mathcal{L}$	ℓ_1	ℓ_2	$\cdots$	ℓ_q	1

Si la loi $\mathcal{L}$ s'avère un bon modèle pour les observations, on s'attend à ce que dans un échantillon aléatoire de taille m, les fréquences relatives observées $h_1, \ldots, h_q$ des différentes valeurs possibles de V soient relativement proches des valeurs théoriques, $\ell_1, \ldots, \ell_q$. Le test d'adéquation du khi-deux consiste à mesurer la « distance » entre ces deux séries de fréquences au moyen de la statistique

$$W_m = m \sum_{i=1}^{q} \frac{(h_i - \ell_i)^2}{\ell_i} \,.$$

Si les fréquences théoriques sont les bonnes et si la taille m de l'échantillon est suffisamment grande, la statistique W_m se comporte alors (voir par exemple Allard [2]) comme une variable du khi-deux à $\nu = q-1$ degrés de liberté. Autrement dit, si l'expérience de prélever un échantillon aléatoire de taille m de la loi de Newcomb–Benford était répétée un très grand nombre de fois de façon indépendante et dans les mêmes conditions, « l'histogramme idéalisé » des valeurs prises par W_m serait donné, en tout $w \in (0, \infty)$, par

$$\chi^2_\nu(w) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}\, w^{\nu/2-1} e^{-w/2}.$$

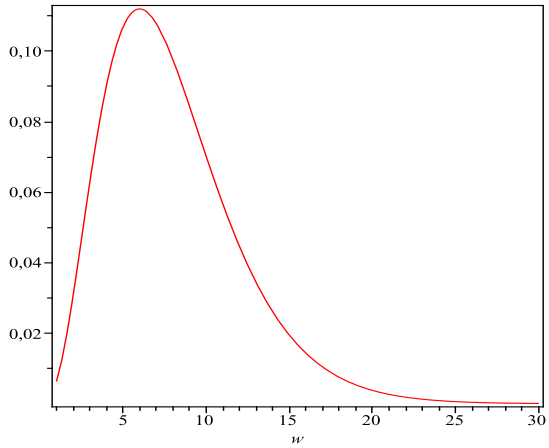


FIG. 2 – Densité de la loi du khi-deux à huit degrés de liberté.

La figure 2 illustre la densité de la loi du khi-deux à $\nu = 8$ degrés de liberté, qui représente la distribution des valeurs de la statistique W_m lorsque la loi de Newcomb–Benford est valable pour les fréquences observées du premier chiffre significatif. Si la valeur observée w_m de W_m se situe trop loin dans la queue de la distribution (disons autour de 25 ou 30, par exemple), ceci aura pour effet de jeter du discrédit sur l'hypothèse à l'effet que les données sont conformes au modèle.

La pratique statistique consiste donc à rejeter la loi de Newcomb–Benford comme modèle lorsque le quantile correspondant à w_m est trop grand. En d'autres termes, le modèle est discrédité quand la probabilité d'observer une valeur supérieure à w_m est trop petite. Cette probabilité, appelée le *seuil observé* du test, est donnée par

$$p = \Pr(W_m \geqslant w_m) = \int_{w_m}^{\infty} \frac{1}{2^{\nu/2}\Gamma(\nu/2)}\, w^{\nu/2-1} e^{-w/2}\, \mathrm{d}w.$$

À titre d'exemple, on pourrait s'entendre *a priori* pour rejeter l'hypothèse nulle si le seuil est inférieur à 5%, voire à 1% ou à 0,1%.

Exemple 4 Supposons que l'on veuille tester si la loi faible de Newcomb–Benford, dont les probabilités sont données au tableau 2, s'ajuste bien aux fréquences relatives rapportées au tableau 1. Il suffit alors de calculer

$$w_m = 194 \left\{ \frac{(0{,}253 - 0{,}301)^2}{0{,}301} + \cdots + \frac{(0{,}062 - 0{,}046)^2}{0{,}046} \right\} \approx 4{,}84.$$

Puisque $\nu = 8$, le seuil observé du test est $p \approx 0{,}77$. On n'est donc pas du tout en mesure de réfuter l'hypothèse nulle. En d'autres termes, ces données sont cohérentes avec la loi de Newcomb–Benford.

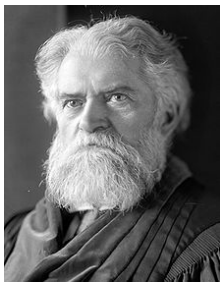
Le test du khi-deux est facile d'emploi et puissant. Si on s'en sert pour éprouver l'hypothèse

$$H_0 : \text{« les observations proviennent de la loi } \mathcal{L} \text{ »}$$

et qu'en réalité les données proviennent d'une loi $\mathcal{L}^* \neq \mathcal{L}$, l'hypothèse nulle sera éventuellement rejetée à n'importe quel seuil prédéterminé, pourvu que l'on dispose d'un échantillon de taille m suffisamment grande. De fait, la probabilité de rejeter H_0 tend vers 1 quand $m \to \infty$, même lorsque $\mathcal{L}^*$ diffère très peu de $\mathcal{L}$. Si l'échantillon est grand, on pourrait donc être amené à rejeter la loi $\mathcal{L}$ dans des cas où l'approximation qu'elle fournit paraît pourtant satisfaisante à toutes fins utiles.

À titre d'exemple, Nigrini et Miller [15] montrent que la loi de Newcomb–Benford est un excellent prédicteur de la répartition des deux premiers chiffres significatifs dans un échantillon de $m = 457\,440$ données hydrologiques. Ils trouvent néanmoins que $w_m = 122{,}595$, ce qui correspond au 99$^{\text{e}}$ centile de la loi du khi-deux à $\nu = 89$ degrés de liberté. Le seuil observé est donc relativement faible ($p \approx 0{,}01$), mais vu la taille de l'échantillon, il n'y a pas lieu de s'alarmer. Ici comme dans bien d'autres situations, la loi *exacte* du phénomène n'est peut-être pas celle de Newcomb–Benford, mais il n'empêche qu'elle propose une approximation simple, utile, facile à motiver, voire élégante.

7 Notes historiques



Il est intéressant de noter que la loi de Newcomb–Benford fut d'abord formulée par un Canadien. En effet, Simon Newcomb[1] est né à Wallace Bridge (Nouvelle-Écosse) en 1835. Il manifesta dans sa jeunesse une grande précocité en mathématiques et un intérêt marqué pour l'astronomie. Autodidacte, polyglotte et polymathe, il fut d'abord calculateur au bureau de l'*American Ephemeris and Nautical Almanac* à Cambridge (Massachusetts). Diplômé de *Harvard University* en 1858, il resta à l'emploi du bureau jusqu'en 1861 puis devint professeur de mathématiques au *United States Naval Observatory*. Il occupa ce poste jusqu'à sa retraite, en 1897, et mourut à Washington en 1909.

[1]Photo par Harris & Ewing, source : Wikimedia Commons, *Prints and Photographs Division* de la Bibliothèque du Congrès des États-Unis sous le numéro d'identification hec.16238

Au cours de sa carrière, Newcomb eut une énorme influence sur les activités de l'observatoire. Sa plus grande contribution fut sans doute la restructuration de la théorie et des moyens de calcul des tables lunaires et planétaires de l'almanac. Auteur de plus de 500 publications scientifiques dans divers domaines, dont les mathématiques, la mécanique céleste et l'économie, il rédigea aussi une autobiographie et un roman de science fiction intitulé *His Wisdom the Defender*. Il occupa en outre plusieurs postes d'influence dans des organismes scientifiques. Il fut notamment président de l'*American Mathematical Society* (1897–1898) et président fondateur de l'*American Astronomical Society* (1899–1905). Un cratère lunaire et l'astéroïde Newcombia honorent sa mémoire.

Frank Albert Benford, quant à lui, est né à Johnstown (Pennsylvanie) en 1883. Physicien diplômé de *University of Michigan* en 1910, il a fait carrière en recherche pendant 38 ans chez *General Electric*, à Schenectady (état de New-York). Auteur d'une centaine d'articles scientifiques en optique et en mathématiques, il est également titulaire d'une vingtaine de brevets. En plus de ses travaux sur la loi qui porte son nom, Benford s'est démarqué par la conception d'un instrument permettant de mesurer l'indice de réfraction du verre. Il est décédé en 1948.

Annexe A : Démonstration de la proposition 1

Soit X un aléa strictement positif de densité continue. Supposons que la densité f de $Y = \log_{10}(X)$ soit majorée par une constante $M > 0$ et admettons qu'il existe des entiers $K \in \mathbb{N}$ et $L \in \mathbb{Z}$ tels que f soit croissante sur $(-\infty, L]$ et décroissante sur $[L + 2K, \infty)$. L'objectif de cette annexe est de démontrer que pour tout $\epsilon \in [0, 1]$, on a $|P(\epsilon) - \epsilon| \leqslant 2(K + 1)M$.

L'énoncé étant trivial lorsque $\epsilon = 0$, fixons $\epsilon \in (0, 1]$. Pour tout entier $k < L$, la croissance de f sur l'intervalle $(k, k + 1)$ entraîne que

$$\begin{aligned}\frac{1-\epsilon}{\epsilon}\int_k^{k+\epsilon} f(y)\,\mathrm{d}y &\leqslant \frac{1-\epsilon}{\epsilon}\int_k^{k+\epsilon} f(k+\epsilon)\,\mathrm{d}y \\ &= (1-\epsilon)f(k+\epsilon) = \int_{k+\epsilon}^{k+1} f(k+\epsilon)\,\mathrm{d}y \leqslant \int_{k+\epsilon}^{k+1} f(y)\,\mathrm{d}y.\end{aligned}$$

En remplaçant le terme d'extrême gauche de l'inéquation ci-dessus par

$$\frac{1}{\epsilon}\int_k^{k+\epsilon} f(y)\,\mathrm{d}y - \int_k^{k+\epsilon} f(y)\,\mathrm{d}y,$$

on déduit que

$$\frac{1}{\epsilon}\int_k^{k+\epsilon} f(y)\,\mathrm{d}y \leqslant \int_k^{k+\epsilon} f(y)\,\mathrm{d}y + \int_{k+\epsilon}^{k+1} f(y)\,\mathrm{d}y = \int_k^{k+1} f(y)\,\mathrm{d}y. \tag{9}$$

De même, si $k > L + 2K$, la décroissance de f sur l'intervalle $(k + \epsilon - 1, k + \epsilon)$ permet d'écrire

$$\begin{aligned}\frac{1-\epsilon}{\epsilon}\int_k^{k+\epsilon} f(y)\,\mathrm{d}y &\leqslant \frac{1-\epsilon}{\epsilon}\int_k^{k+\epsilon} f(k)\,\mathrm{d}y \\ &= (1-\epsilon)f(k) = \int_{k-1+\epsilon}^{k} f(k)\,\mathrm{d}y \leqslant \int_{k-1+\epsilon}^{k} f(y)\,\mathrm{d}y,\end{aligned}$$

d'où l'on tire

$$\frac{1}{\epsilon}\int_k^{k+\epsilon} f(y)\,\mathrm{d}y \leqslant \int_{k-1+\epsilon}^{k+\epsilon} f(y)\,\mathrm{d}y \leqslant \int_{k-1}^{k} f(y)\,\mathrm{d}y, \tag{10}$$

où la dernière inégalité est justifiée par le fait que f est décroissante sur l'intervalle $(k-1,k)$.

Par ailleurs, si $k \in \{L,\ldots,L+2K\}$, on a

$$\frac{1}{\epsilon}\int_k^{k+\epsilon} f(y)\,\mathrm{d}y \leqslant M. \tag{11}$$

En combinant les inégalités (9), (10) et (11), on trouve

$$\frac{P(\epsilon)}{\epsilon} = \frac{1}{\epsilon}\sum_{k\in\mathbb{Z}}\int_k^{k+\epsilon} f(y)\,\mathrm{d}y \leqslant (2K+1)M + \int_{\mathbb{R}\setminus[L,L+2K]} f(y)\,\mathrm{d}y \leqslant (2K+1)M+1,$$

puisque f est une densité. Il s'ensuit que

$$P(\epsilon) \leqslant \epsilon + (2K+1)M\epsilon \leqslant \epsilon + (2K+1)M$$

et donc que $P(\epsilon)-\epsilon \leqslant (2K+1)M \leqslant 2(K+1)M$.

La démonstration de l'inégalité $\epsilon - P(\epsilon) \leqslant 2(K+1)M$ est analogue. Pour tout entier $k < L$, la croissance de f sur l'intervalle $(k-1+\epsilon, k+\epsilon)$ fait en sorte que

$$\frac{1-\epsilon}{\epsilon}\int_k^{k+\epsilon} f(y)\,\mathrm{d}y \geqslant (1-\epsilon)f(k) \geqslant \int_{k-1+\epsilon}^{k} f(y)\,\mathrm{d}y.$$

Par conséquent,

$$\frac{1}{\epsilon}\int_k^{k+\epsilon} f(y)\,\mathrm{d}y \geqslant \int_{k-1+\epsilon}^{k+\epsilon} f(y)\,\mathrm{d}y \geqslant \int_{k-1}^{k} f(y)\,\mathrm{d}y. \tag{12}$$

De façon semblable si $k > L+2K$, la décroissance de f sur l'intervalle $(k-1,k)$ conduit à

$$\frac{1-\epsilon}{\epsilon}\int_k^{k+\epsilon} f(y)\,\mathrm{d}y \geqslant (1-\epsilon)f(k+\epsilon) \geqslant \int_{k+\epsilon}^{k+1} f(y)\,\mathrm{d}y,$$

d'où il découle que

$$\frac{1}{\epsilon}\int_k^{k+\epsilon} f(y)\,\mathrm{d}y \geqslant \int_{k}^{k+1} f(y)\,\mathrm{d}y. \tag{13}$$

En se servant des inégalités (12) et (13), on conclut que

$$\begin{aligned}\frac{P(\epsilon)}{\epsilon} &\geqslant \sum_{k=-\infty}^{L-1}\int_{k-1}^{k} f(y)\,\mathrm{d}y + \sum_{k=L+2K+1}^{\infty}\int_k^{k+1} f(y)\,\mathrm{d}y \\ &= 1 - \int_{L-1}^{L+2K+1} f(y)\,\mathrm{d}y \geqslant 1-2(K+1)M\end{aligned}$$

pour toute valeur possible de $\epsilon \in (0,1]$. Ceci achève la démonstration. □

Annexe B : Démonstration de la borne de Fréchet

Pour démontrer la validité de l'inégalité (2), on peut supposer sans perte de généralité que la variable Y est d'espérance nulle. Soit f la densité de cette variable, supposée continue et majorée par une constante $M > 0$. On pose $I_M = [-1/(2M), 1/(2M)]$ et $J_M = \mathbb{R} \setminus I_M$. On a alors

$$\frac{1}{12M^2} = \int_{I_M} My^2 \,\mathrm{d}y = \int_{I_M} y^2 f(y) \,\mathrm{d}y + \int_{I_M} y^2 \{M - f(y)\} \,\mathrm{d}y.$$

On va montrer ci-dessous que

$$\int_{I_M} y^2 \{M - f(y)\} \,\mathrm{d}y \leqslant \int_{J_M} y^2 f(y) \,\mathrm{d}y, \tag{14}$$

ce qui permettra de déduire que

$$\frac{1}{12M^2} \leqslant \int_{I_M} y^2 f(y) \,\mathrm{d}y + \int_{J_M} y^2 f(y) \,\mathrm{d}y = \int_{\mathbb{R}} y^2 f(y) \,\mathrm{d}y.$$

Puisque le terme de droite n'est autre que $\mathrm{E}(Y^2) = \mathrm{var}(Y)$, on pourra alors conclure.

Partant du fait que $y^2 \leqslant 1/(4M^2)$ pour tout $y \in I_M$, on a d'abord

$$\int_{I_M} y^2 \{M - f(y)\} \,\mathrm{d}y \leqslant \frac{1}{4M^2} \int_{I_M} \{M - f(y)\} \,\mathrm{d}y, \tag{15}$$

où l'on s'est servi de l'hypothèse à l'effet que $M - f(y) \geqslant 0$ en tout $y \in \mathbb{R}$. Or, la définition de I_M et le fait que f est une densité entraînent que

$$\int_{I_M} \{M - f(y)\} \,\mathrm{d}y = 1 - \int_{I_M} f(y) \,\mathrm{d}y = \int_{J_M} f(y) \,\mathrm{d}y.$$

En exploitant cette identité et l'inégalité $y^2 \geqslant 1/(4M^2)$ valable pour tout $y \in J_M$, on déduit que

$$\frac{1}{4M^2} \int_{I_M} \{M - f(y)\} \,\mathrm{d}y = \frac{1}{4M^2} \int_{J_M} f(y) \,\mathrm{d}y \leqslant \int_{J_M} y^2 f(y) \,\mathrm{d}y. \tag{16}$$

En conjuguant les relations (15) et (16), on obtient l'inégalité (14), ce qui complète l'argument. □

Remerciements

Vincent Genest a bénéficié d'une bourse de recherche Alexander-Graham-Bell du Conseil de recherches en sciences naturelles et en génie du Canada, qu'il remercie de son appui. Christian Genest est également reconnaissant envers cet organisme et le Fonds québécois de la recherche sur la nature et les technologies, qui soutiennent ses travaux par le biais de leurs programmes de subventions.

Références

[1] Adhikari, A.K. & Sarkar, B.P. *Distribution of most significant digit in certain functions whose arguments are random variables*. Sankhyā : The Indian Journal of Statistics, Série B, Vol. 30 (1968), no 1/2, pp. 47–58.

[2] Allard, Jacques. *Concepts fondamentaux de la statistique*. Addison-Wesley, Montréal, 1992.

[3] Benford, Frank. *The law of anomalous numbers*. Proceedings of the American Philosophical Society, Vol. 78 (1938), no 4, pp. 551–572.

[4] *CIA World Factbook*, édition 2010, mise à jour toutes les deux semaines, disponible sur internet à https ://www.cia.gov/library/publications/the-world-factbook/index.html.

[5] Fewster, Rachel M. *A simple explanation of Benford's law*. The American Statistician, Vol. 63 (2009), no 1, pp. 26–32.

[6] Fréchet, Maurice. *Sur une limitation très générale de la dispersion de la médiane*. Journal de la Société de statistique de Paris, Vol. 81 (1940), no 1, pp. 67–77.

[7] Gauvrit, Nicolas et Delahaye, Jean-Paul. *Pourquoi la loi de Benford n'est pas mystérieuse*. Mathématiques et sciences humaines, Vol. 46 (2008), no 2, pp. 7–15.

[8] Hill, Theodore P. *Base-invariance implies Benford's law*. Proceedings of the American Mathematical Society, Vol. 123 (1995), no 3, pp. 887–895.

[9] Hill, Theodore P. *A statistical derivation of the significant-digit law*. Statistical Science, Vol. 10 (1995), no 4, pp. 354–363.

[10] Janvresse, Élise et de la Rue, Thierry. *From uniform distributions to Benford's law*. Journal of Applied Probability, Vol. 41 (2004), no 4, pp. 1203–1210.

[11] Kestelman, Hyman. *On the functional equation* $f(x+y) = f(x)+f(y)$. Polska Akademia Nauk. Fundamenta Mathematicae, Vol. 34 (1947), pp. 144–147.

[12] Newcomb, Simon. *Note on the frequency of use of the different digits in natural numbers*. American Journal of Mathematics, Vol. 4 (1881), no 1, pp. 39–40.

[13] Nigrini, Mark J. *The Detection of Income Tax Evasion Through an Analysis of Digital Frequencies*. Thèse, Université de Cincinnati, Ohio, 1992.

[14] Nigrini, Mark J. *A taxpayer compliance application of Benfordís law*. Journal of the American Taxation Association, Vol. 18 (1996), no 1, pp. 72–91.

[15] Nigrini, Mark J. & Miller, Steven J. *Benford's law applied to hydrology data—Results and relevance to other geophysical data*. Mathematical Geology, Vol. 39 (2007), no 5, pp. 469–490.

[16] Poincaré, Henri. Répartition des décimales dans une table numérique. Dans *Calcul des probabilités*, Gauthier-Villars, Paris, 1912, pp. 313–320.

[17] Raimi, Ralph A. *The first digit problem*. The American Mathematical Monthly, Vol. 83 (1976), no 7, pp. 521–538.

[18] Rousseau, Christiane. *Apprendre à frauder ou à détecter les fraudes ?* Accromath, Vol. 5 (2010), no 2, pp. 2–7.