Étude statistique de suites

Guilhem Repetto

L'équirépartition est une notion à première vue très intuitive, mais qui conduit à des résultats variés dans différents domaines des mathématiques, comme ceux des systèmes dynamiques et des méthodes d'intégration de fonctions.

On étudie quelques propriétés des suites équiréparties, puis le critère de Weyl, qui permet de déterminer si certaines suites sont équiréparties. On étudie aussi un théorème vrai *dans le cas général*. Le point de vue purement mathématique est complété par une approche expérimentale, où l'on se propose de vérifier la loi de Benford, liée à l'équirépartition, dans des données issues du commerce de détail et des arbres de France métropolitaine.

Table des matières

1 L'équirépartition modulo 1

Notations, définitions :

- ▷ La partie fractionnaire du réel x est notée $\{x\}$ et est définie par $\{x\} = x - \lfloor x \rfloor$.
- ▷ $\mathbb{1}_E$ est la fonction indicatrice d'un ensemble E.
- ▷ Une suite (u_n) est *équirépartie modulo 1* si pour tous les réels $(a,b) \in [0;1]^2$ avec $a < b$, la limite suivante est vérifiée :

$$\lim_{N\to+\infty} \frac{1}{N+1} \sum_{n=0}^{N} \mathbb{1}_{[a;b[}(\{u_n\}) = \int_0^1 \mathbb{1}_{[a;b[}(x)\,\mathrm{d}x = b - a$$

- ▷ *Loi de Benford discrète* : une variable X suit une loi de Benford discrète si $\forall d \in [\![1;9]\!], \mathbb{P}(X = d) = \log\left(1 + \frac{1}{d}\right)$.
- ▷ *Loi de Benford continue* : une variable X suit une loi de Benford continue si $\forall a, b \in [1;10[$ où $a < b$, $\mathbb{P}(X \in [a,b[) = \log b - \log a$.
- ▷ Une suite $(u_n)_n$ *suit la loi de Benford continue* si et seulement si la suite $(\log u_n)_n$ est équirépartie modulo 1.

A Premières propriétés

Les trois propriétés suivantes sont des conséquences directes de la définition :

Propriété 1.1 : les décalages ne changent pas l'équirépartition

Soit (u_n) équirépartie modulo 1, et $\lambda \in \mathbb{R}$. Alors $(u_n + \lambda)_{n\in\mathbb{N}}$ est équirépartie modulo 1.

Preuve de la propriété 1.1 : Soient $0 \leq a < b \leq 1$. On a

$$\begin{aligned} \{x_n + \lambda\} \in [a,b] &\iff \exists Z \in \mathbb{Z}, x_n + \lambda + Z \in [a,b] \\ &\iff \{x_n\} \in \begin{cases} [a-\lambda, b-\lambda] \text{ si } \lambda \leq a \\ [0, b-\lambda] \cup [a-\lambda+1, 1] \text{ sinon} \end{cases} \end{aligned}$$

et le résultat découle de l'équirépartition de $(x_n)_n$. □

Propriété 1.2 : convergence de la différence

Soit (u_n) équirépartie modulo 1, et (v_n) une suite telle que

$$\lim_{n\to+\infty} u_n - v_n = \lambda \in \mathbb{R}$$

Alors (v_n) est équirépartie modulo 1.

Preuve de la propriété 1.2 : D'après la propriété 1, il suffit de considérer $\lambda = 0$. Notons $\varepsilon_n = x_n - y_n$, et considérons $]a;b[\in [0;1]$ et η tel que $0 < \eta < \min\left(a, 1-b, \frac{b-a}{2}\right)$. Il existe alors $N \in \mathbb{N}$ tel que $(n \geq N \implies -\eta \leq \varepsilon_n \leq \eta)$. Soit donc $n \geq N$. On a

$$\begin{aligned} & a + \eta \leq \{x_n\} \leq b - \eta \\ \implies \quad & a \leq \{y_n\} \leq b \qquad (i) \\ \implies \quad & a - \eta \leq \{x_n\} \leq b + \eta \qquad (ii) \end{aligned}$$

Ainsi :

$$\begin{aligned} b-a-2\eta &= \lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N} \mathbb{1}_{[a+\eta,b-\eta]}(\{x_n\}) \\ &\overset{(i)}{\leq} \lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N} \mathbb{1}_{[a,b]}(\{y_n\}) \\ &\overset{(ii)}{\leq} \lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N} \mathbb{1}_{[a-\eta,b+\eta]}(\{x_n\}) \\ &= b-a+2\eta \end{aligned}$$

ce qui termine la preuve. □

Propriété 1.3 : racines carrées

Si $\alpha \neq 0$, alors la suite $(\alpha\sqrt{n})_n$ est équirépartie modulo 1.

Preuve de la propriété 1.3 : Soient $0 \leq a < b < 1$ et $N \geq 1$. Notons $I_N = \{n \in [\![1;N]\!] \mid a \leq \{\alpha\sqrt{n}\} < b\}$. Soit $n \in [\![1;N]\!]$. On a $n \in I_N$ si et seulement s'il existe $k \in \mathbb{N}$ vérifiant

$$(a+k)^2 \leq \alpha^2 n \leq (b+k)^2$$

En notant $K_N = \max\{k \in \mathbb{N} \mid (a+k)^2 \leq \alpha^2 N\}$, on déduit l'expression

$$I_N = \bigcup_{0\leq k\leq K_N} \left[\left(\frac{a+k}{\alpha}\right)^2 ; \left(\frac{b+k}{\alpha}\right)^2\right] \cap \mathbb{N} \quad \text{(union disjointe)}$$

Remarquons que $K_N = \left\lfloor \alpha\sqrt{N} - a \right\rfloor \sim \alpha\sqrt{N}$. De plus, si u et v sont deux réels quelconques, on observe que

$$v-u-1 \leq \operatorname{Card}\{[u;v] \cap \mathbb{Z}\} \leq v-u+1$$

ce qui permet de donner un encadrement de $S_N(a,b) = \operatorname{Card} I_N$:

$$\sum_{k=0}^{K_N} ((b-a)(b-a+2k)-1) \leq S_N \leq \sum_{k=0}^{K_N} ((b-a)(b-a+2k)+1)$$

et un simple calcul d'équivalent des deux sommes montre que

$$S_N(a,b) \sim N(b-a)$$

Cela montre en particulier que $(\sqrt{n})_n$ est équirépartie modulo 1. □

B Le critère de Weyl

Le critère de Weyl donne une condition nécessaire et suffisante d'équirépartition, et permet de traiter de nouveaux cas de suites.

Propriété 1.4 :

La suite (u_n) est équirépartie modulo 1 si et seulement si

$$\forall f \in \mathcal{C}^0([0;1],\mathbb{R}), \quad \lim_{N\to+\infty} \frac{1}{N+1}\sum_{n=0}^{N} f(\{u_n\}) = \int_0^1 f(x)\,\mathrm{d}x$$

Preuve de la propriété 1.4 : On suppose, sans perte de généralité, que (u_n) est à valeurs dans $[0,1[$ pour plus de clarté.

$\boxed{\Rightarrow}$ Soit f continue. Soit $\varepsilon > 0$. On construit une fonction en escalier f_ε telle que $\|f - f_\varepsilon\|_\infty \leq \varepsilon$. On a alors

$$\begin{aligned}\frac{1}{N+1}\sum_{n=0}^{N} f(u_n) - \int_0^1 f(x)\,\mathrm{d}x =& \frac{1}{N+1}\sum_{n=0}^{N}\left(f(u_n) - f_\varepsilon(u_n)\right)\\ &+ \frac{1}{N+1}\sum_{n=0}^{N} f_\varepsilon(u_n) - \int_0^1 f_\varepsilon\\ &+ \int_0^1 (f_\varepsilon - f)\end{aligned}$$

d'où

$$\left|\frac{1}{N+1}\sum_{n=0}^{N} f(u_n) - \int_0^1 f(x)\,\mathrm{d}x\right| \leq \varepsilon + \left(\frac{1}{N+1}\sum_{n=0}^{N} f_\varepsilon(u_n) - \int_0^1 f_\varepsilon(x)\,\mathrm{d}x\right) + \varepsilon$$

En notant $f_\varepsilon = \sum_{i=0}^{d} f(a_i)\mathbb{1}_{[a_i;a_{i+1}[}$, où $a_0, \ldots, a_{d+1}$ est une subdivision adaptée à f_ε, comme (u_n) est équirépartie, on constate que

$$\begin{aligned}\frac{1}{N+1}\sum_{n=0}^{N} f_\varepsilon(u_n) &= \sum_{i=0}^{d}\frac{1}{N+1}\sum_{n=0}^{N} f(a_i)\mathbb{1}_{[a_i;a_{i+1}[}(u_n)\\ &\xrightarrow[N\to+\infty]{} \sum_{i=0}^{d} f(a_i)(a_{i+1}-a_i)\\ &= \int_0^1 f_\varepsilon(x)\,\mathrm{d}x\end{aligned}$$

Ainsi, pour N assez grand, on a $\left|\frac{1}{N+1}\sum_{n=0}^{N} f(u_n) - \int_0^1 f(x)\,\mathrm{d}x\right| \leq 3\varepsilon$, ce qui était attendu.

$\boxed{\Leftarrow}$ Soient $0 \leq a < b \leq 1$, et $\varepsilon > 0$. On construit sans peine deux fonctions affines par morceaux f_ε^- et f_ε^+ telles que

$$\begin{aligned}f_\varepsilon^- \leq \mathbb{1}_{[a,b[} \leq f_\varepsilon^+\\ \int_0^1 \left(f_\varepsilon^+ - f_\varepsilon^-\right) \leq \varepsilon\end{aligned}$$

On a alors

$$\frac{1}{N+1}\sum_{n=0}^{N} f_\varepsilon^-(u_n) \leq \frac{1}{N+1}\sum_{n=0}^{N}\mathbb{1}_{[a,b[}(u_n) \leq \frac{1}{N+1}\sum_{n=0}^{N} f_\varepsilon^+(u_n)$$

En considérant les limites des termes d'encadrement, on a à partir d'un certain rang :

$$\int_0^1 f_\varepsilon^-(x)\,\mathrm{d}x - \varepsilon \leq \frac{1}{N+1}\sum_{n=0}^{N}\mathbb{1}_{[a,b[}(u_n) \leq \int_0^1 f_\varepsilon^+(x)\,\mathrm{d}x + \varepsilon$$

et par construction de f_ε^- et f_ε^+ :

$$b - a - 2\varepsilon \leq \frac{1}{N+1}\sum_{n=0}^{N}\mathbb{1}_{[a,b[}(u_n) \leq b - a + 2\varepsilon$$

Donc (u_n) est équirépartie modulo 1. □

Théorème 1.1 : critère de Weyl

La suite (u_n) est équirépartie modulo 1 si et seulement si

$$\forall k \in \mathbb{N}^*, \quad \lim_{N\to+\infty} \sum_{n=0}^{N} \exp(2i\pi k u_n) = 0$$

Preuve du théorème de Weyl (1.1) :

$\boxed{\Rightarrow}$ D'après la propriété précédente.

$\boxed{\Leftarrow}$ Soit $\varepsilon > 0$. Le théorème polynomial de Weierstrass assure l'existence d'un polynôme trigonométrique Ψ, combinaison linéaire finie des fonctions $(x \mapsto \exp(2ik\pi x))_{k\in\mathbb{Z}}$, tel que

$$\|f - \Psi\|_{\infty,[0;1]} \leq \varepsilon$$

On procède donc à la majoration suivante, valable à partir d'un certain rang :

$$\begin{aligned}\left|\int_0^1 f(x)\,\mathrm{d}x - \frac{1}{N+1}\sum_{n=0}^{N} f(u_n)\right| &\leq \left|\int_0^1 f(x) - \Psi(x)\,\mathrm{d}x\right| \\ &+ \left|\int_0^1 \Psi(x)\,\mathrm{d}x - \frac{1}{N+1}\sum_{n=0}^{N}\Psi(u_n)\right| \\ &+ \left|\frac{1}{N+1}\sum_{n=0}^{N} f(u_n) - \Psi(u_n)\right| \\ &\leq 3\varepsilon\end{aligned}$$

grâce à l'hypothèse initiale ainsi que $\int_0^1 \Psi(x)\,\mathrm{d}x = 0$. □

Un corollaire est le suivant, qui traite le cas des suites arithmétiques :

Théorème 1.2 : Bohl-Sierpinski-Weyl

La suite $(x\,n)_n$ est équirépartie modulo 1 si et seulement si $x \in \mathbb{R} \setminus \mathbb{Q}$.

Preuve du théorème de Bohl-Sierpinski-Weyl (1.2) : Si $x \in \mathbb{R} \setminus \mathbb{Q}$, alors $\forall k \in \mathbb{N}^*, \exp(2i\pi kx) \neq 1$. Cela entraîne :

$$\frac{1}{N}\left|\sum_{n=0}^{N-1}\exp(2i\pi knx)\right| = \frac{|\exp(2i\pi kNx) - 1|}{N\,|\exp(2i\pi kx) - 1|} \leq \frac{2}{N\,|\exp(2i\pi kx) - 1|} \xrightarrow[N\to\infty]{} 0$$

Sinon, en écrivant $x = \frac{p}{q}$ où $(p, q) \in \mathbb{Z} \times \mathbb{N}^*$, on observe que la suite $(\exp(2i\pi qnx))_n$ est constante égale à 1, ce qui ne satisfait pas le critère de Weyl. □

C Résultats vrais dans le cas général

Définition : Un sous-ensemble E de $\mathbb{R}$ est *négligeable* lorsque pour tout $\varepsilon > 0$, il existe une famille d'intervalles $(I_n)_n$ telle que $E \in \cup_n I_n$ et $\sum_n \ell(I_n) < \varepsilon$, où $\ell(I_j)$ est la longueur de l'intervalle I_j. Une propriété vérifiée sur le complémentaire d'un ensemble négligeable est *vraie presque partout*. Intuitivement, si on choisit un réel "au hasard", alors la propriété en question est vérifiée pour ce réel.

Je reproduis ici un résultat simple, trouvé dans [1] :

Théorème 1.3 : Généralisation du théorème de Bohl-Sierpinski-Weyl

Soit $(u_n)_n$ une suite d'entiers injective. Alors $(xu_n)_n$ est équirépartie modulo 1 pour presque tout réel x.

Preuve du théorème 1.3 : On considère d'abord le cas $x \in [0;1[$.
Pour $k, N \in \mathbb{N}^*$ fixés, on note

$$S(N,x) = \frac{1}{N}\sum_{n=0}^{N-1} \exp(2ik\pi u_n x)$$

En élevant au carré, puis en intégrant sur $[0;1[$:

$$|S(N,x)|^2 = \frac{1}{N^2}\sum_{m=0}^{N-1}\sum_{n=0}^{N-1} \exp(2ik\pi(u_m - u_n)x)$$

$$\int_0^1 |S(N,x)|^2 \,\mathrm{d}x = \frac{1}{N^2}\sum_{m=0}^{N-1}\sum_{n=0}^{N-1}\int_0^1 \exp(2ik\pi(u_m-u_n)x)\,\mathrm{d}x$$

$$= \frac{1}{N} \text{ (seuls les termes où } m = n \text{ contribuent)}$$

Comme cette dernière expression est le terme général d'une série divergente, on se ramène à une série convergente en ne considérant que les $S(N^2,x)$, et en sommant :

$$\sum_{N=1}^{+\infty}\int_0^1 \left|S(N^2,x)\right|^2 \mathrm{d}x = \frac{\pi^2}{6} < \infty$$

Par le théorème de convergence monotone :

$$\int_0^1 \sum_{N=1}^{+\infty} \left|S(N^2,x)\right|^2 \mathrm{d}x < \infty$$

ce qui montre que $\sum_{N=1}^{+\infty}\left|S(N^2,x)\right|^2 \mathrm{d}x < \infty$ pour presque tout x dans $[0;1[$. Par conséquent, pour ces mêmes x, on a $\lim_{N\to\infty}\left|S(N^2,x)\right|^2 = 0$.

Pour nous ramener à N, notons m l'unique entier tel que $m^2 \le N < (m+1)^2$. On a alors

$$|S(N,x)| \le \left|\frac{1}{N}\sum_{n=0}^{m^2+2m}\exp(2ik\pi u_n x)\right| \le \left|S(m^2,x)\right| + \frac{2m}{N} \le \left|S(m^2,x)\right| + \frac{2}{\sqrt{N}}$$

Cette dernière inégalité montre finalement

$$\exists \lim_{N\to\infty} S(N,x) = 0 \text{ pour presque tout } x \in [0;1[$$

Le critère de Weyl permet de conclure que la suite étudiée est équirépartie pour presque tout $x \in [0;1[$.

En remarquant que pour tout $m \in \mathbb{Z}$, on a $\{u_n(m+x)\} = \{u_n x\}$, on obtient que l'ensemble $\{x \in [m; m+1[\mid S(N,x) \not\to 0\}$ est un translaté de $\{x \in [0;1[\mid S(N,x) \not\to 0\}$, donc négligeable. L'ensemble des x pour lesquels la suite n'est pas équirépartie est donc bien négligeable en tant que réunion dénombrable d'ensembles négligeables. □

Le théorème de Koksma, dont la preuve est assez technique, concerne les suites géométriques :

Théorème 1.4 : Koksma (admis)

Pour presque tout réel $x > 1$, la suite géométrique $(x^n)_n$ est équirépartie modulo 1.

Cependant, on dispose d'une classe infinie de contre-exemples à ce théorème :

Définition : un *nombre de Pisot-Vijayaraghavan* est un entier algébrique tel que toutes les autres racines de son polynôme minimal sont de module strictement inférieur à 1.

Propriété 1.5 : contre-exemple des nombres de Pisot

Si θ est un nombre de Pisot, alors $(\theta^n)_n$ n'est pas équirépartie modulo 1.

Preuve de la propriété 1.5 : on note $\Pi \in \mathbb{Z}_r[X]$ le polynôme minimal de θ, et $\alpha_1, \ldots, \alpha_{r-1}$ les autres racines complexes de Π, comptées avec multiplicité. Ce polynôme est le polynôme caractéristique de sa matrice compagnon $C_\Pi \in \mathcal{M}_r(\mathbb{Z})$, qui est trigonalisable (Π est scindé sur $\mathbb{C}$) en une matrice dont la diagonale est $\theta, \alpha_1 \ldots, \alpha_{r-1}$. Ainsi, comme la trace est un invariant de similitude, on a $\forall n \in \mathbb{N}, \theta^n + \alpha_1^n + \cdots + \alpha_{r-1}^n = \mathrm{Tr}(C_\Pi^n)$. Or, $\mathrm{Tr}(C_\Pi^n) \in \mathbb{Z}$ car C_Π^n reste dans $\mathcal{M}_r(\mathbb{Z})$. Le résultat découle du fait que $\forall i \in [\![1; r-1]\!], |\alpha_i| < 1$, ce qui force $\{\theta^n\}$ à se rapprocher de 0 ou de 1, empêchant ainsi son équirépartition. □

Par exemple, la suite $(\phi^n)_n$, où ϕ est le nombre d'or, n'est pas équirépartie modulo 1.

2 Conséquence observable de l'équirépartition : la loi de Benford

A Existence d'une loi du premier chiffre dans la nature

Sous certaines conditions, la variable aléatoire donnant le premier chiffre significatif d'un nombre choisi dans une série de valeurs "suffisamment étalées en ordre de grandeur" ne semble pas suivre la loi uniforme sur $[\![1; 9]\!]$, mais plutôt la loi de Benford.

Ainsi, d'après le théorème 1.2, la suite $(k^n)_n$, où k n'est pas une puissance de 10, suit la loi de Benford, car dans ce cas $\log k$ est irrationnel.

Les auteurs de [5] précisent cependant que cette loi n'est pas toujours vérifiée sur des données issues de la nature ou produites par les activités humaines. Par exemple, les séries de nombres pseudo-aléatoires produits par des humains ne la vérifient pas. Ils montrent que si le logarithme d'une variable continue X est "suffisamment étalé", alors X suit la loi de Benford. Leur résultat repose sur le théorème suivant :

Théorème 2.1 : convergence en loi vers la loi uniforme (admis)

Soit X une variable réelle de densité f, où

(i) f atteint un maximum M en un point a

(ii) f est croissante sur $]-\infty; a]$, décroissante sur $[a; +\infty[$

Alors,

$$\forall x \in [0; 1[, \quad |P(\{X\} < x) - x| < 2M$$

En outre, si (X_n) est une suite de telles variables telles que $M_n \to 0$, alors $\{X_n\}$ converge en loi vers la loi uniforme sur $[0; 1[$.

Pour obtenir que $\{\log X_n\}$ converge en loi vers une loi uniforme, il suffit d'adapter le théorème :

Propriété 2.1 : convergence en loi du logarithme

Soit X une variable réelle strictement positive de densité f, où

(i) $x \mapsto xf(x)$ atteint un maximum M en un point a

(ii) $x \mapsto xf(x)$ est croissante sur $]0; a]$, décroissante sur $[a; +\infty[$

Alors,

$$\forall x \in [0; 1[, \quad |P(\{\log X\} < x) - x| < 2M \ln 10$$

Ainsi, si (X_n) est une suite de telles variables telles que $M_n \to 0$, alors $\{\log X_n\}$ converge en loi vers la loi uniforme sur $[0; 1[$. On conclut alors grâce à la définition que X_n converge en loi vers la loi de Benford.

Les auteurs de [5] font remarquer que, sous certaines conditions de régularité de f, on peut de la même façon obtenir la convergence en loi d'un grand nombre de fonctions de X. La loi du logarithme est la plus facile à remarquer, car elle concerne le premier chiffre. Quelques exemples sont réunis dans le tableau suivant :

φ	Si $\cdots$ vérifie (i) et (ii), alors $\varphi(X_n)$ converge vers la loi uniforme
$x \mapsto \log x$	$x \mapsto xf(x)$
$x \mapsto \sqrt{x}$	$x \mapsto \sqrt{x}f(x)$
$x \mapsto x^2$	$x \mapsto x^{-1}f(x)$
$x \mapsto x^n, n \in \mathbb{N}^*$	$x \mapsto x^{-n+1}f(x)$
$x \mapsto e^x$	$x \mapsto e^{-x}f(x)$

B Échantillon de prix relevés dans un supermarché

Pour tenter de vérifier la loi de Benford sur des données réelles, j'ai choisi d'utiliser les prix d'un rayon de supermarché, qui sont disponibles sur une page web du magasin. Ces prix ne sont pas disponibles sous forme de base de données, mais sont contenus dans le texte de la page. J'explique dans cette partie les solutions que j'ai utilisées pour extraire ces prix.

B.1 Langage des prix et expression régulière

On note $C = \{␣, 0, 1, \dots, 9\}$, et Σ l'ensemble des caractères ASCII. Le langage des prix $L \subsetneq \Sigma$ est l'ensemble des mots sur Σ qui représentent un prix valide en euros. Une expression régulière $\mathcal{E}$ décrivant L est

$$\mathcal{E} = C^+ \cdot (\text{“,”} + \text{“.”}) \cdot C^+ \cdot (\varepsilon + \text{“␣”}) \cdot \text{“€”}$$

Certains prix des pages web utilisées font aussi figurer les prix au kilo ou au litre, ce qui ne correspond pas au prix d'un article réel. On définit alors un nouveau langage L_0 afin d'éliminer ce type de prix. Une expression régulière $\mathcal{E}_0$ décrivant L_0 est

$$\mathcal{E}_0 = \mathcal{E} \cdot \text{“␣”} \cdot (\Sigma \setminus \{\text{“/”}\})$$

Si $u' \in L_0$, il existe une unique décomposition $u' = uw$ avec $u \in L$, où u est un prix valide que nous souhaitons prendre en compte. On peut alors convertir u en flottant, en appliquant d'abord la fonction φ suivante qui supprime les caractères indésirables, puis la fonction de conversion `float`.

$$\varphi(a) = \begin{cases} \varepsilon \text{ si } a \in \{␣, €\} \\ \text{. si } a \in \{\text{“,”}, \text{“.”}\} \\ a \text{ sinon} \end{cases} \quad \text{puis} \quad \varphi(a_1 \dots a_n) = \varphi(a_1) \dots \varphi(a_n)$$

On calcule enfin le premier chiffre par divisions successives par 10.

B.2 Premier algorithme d'extraction des prix

Soit une chaîne de caractères $c_0 c_1 \dots c_N$. On utilise l'algorithme suivant :

Algorithme 2.1 : extraction naïve des prix

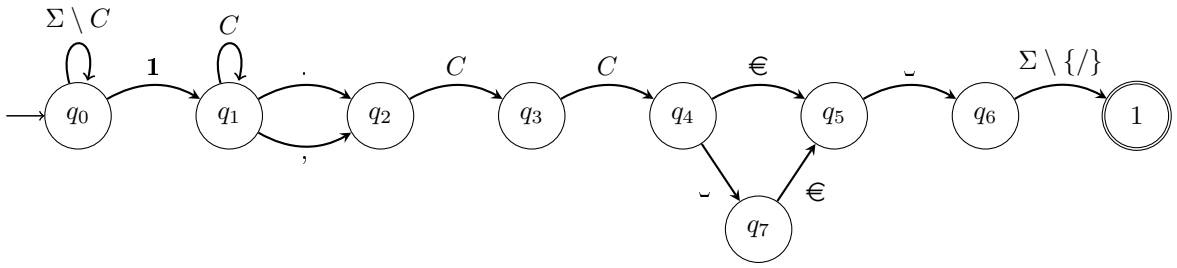
- Créer un tableau t d'entiers à neuf cases initialisées à 0, noter $p = 0$
- Tant que $p < N$, parcourir la chaîne de caractères à partir de c_p jusqu'au premier indice i tel que $c_i = €$
- Considérer le plus grand mot $u = c_{i-k} c_{i-k+1} \dots c_i c_{i+1} \dots c_{i+l}$ où $\forall j \in [\![i-k; i+l]\!] \setminus \{i\}, c_j \in C$.
-
  - Si $c_{i+l+1} = /$, ne rien faire (cas d'un prix au litre ou au kilo)
  - Sinon : calculer le premier chiffre de $(\texttt{float} \circ \varphi)(u)$, noté d, et ajouter 1 à $t.[d]$
- Définir $p = i + 1$
- Fin Tant que
- Renvoyer t

La complexité spatiale est donc en $\mathcal{O}(1)$, et la complexité temporelle en $\mathcal{O}(N + \ln P)$, où P est le plus grand prix présent dans le fichier, donc en $\mathcal{O}(N)$ dans notre cas car $P \leq 10^4$ et $N \gg P$. Un inconvénient est le mouvement de va-et-vient dans le texte lorsque l'algorithme détecte le symbole €.

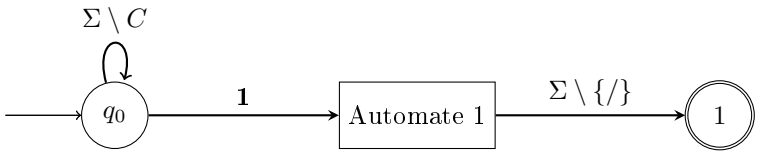
B.3 Second algorithme, et idées d'amélioration

Comme on peut se contenter de stocker le premier chiffre des prix, on peut utiliser un automate déterministe, afin de n'effectuer qu'une seule lecture à sens unique du fichier issu de la page web.

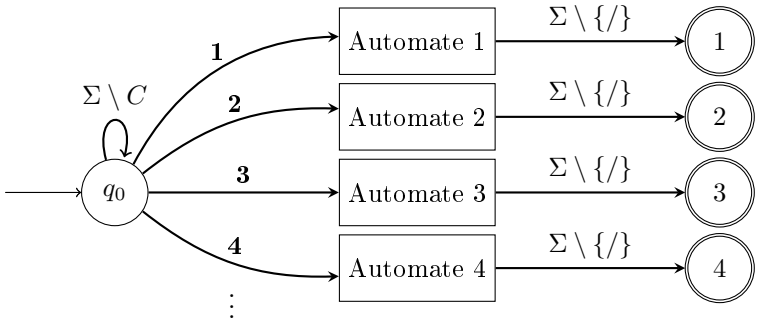
Un automate déterministe reconnaissant les prix commençant par 1 est le suivant :



Les transitions sont étiquetées soit par un ensemble de lettres, soit par une seule. Pour chaque état, si la lettre lue ne correspond à aucune transition, on retourne à l'état initial q_0 par une transition non-représentée (par souci de lisibilité). On symbolise alors cet automate par



ce qui permet, en construisant des automates similaires pour chaque premier chiffre, de construire un nouvel automate dont les états finaux indiquent le premier chiffre du prix reconnu :



Voici un algorithme qui ne lit qu'une seule fois le fichier des prix :

Algorithme 2.2 : extraction des prix

- Créer un tableau t d'entiers à neuf cases initialisées à 0, mettre l'état de l'automate à q_0
- Tant qu'il reste des caractères non-lus, lire le caractère suivant et calculer le nouvel état
  - Si l'état est un état final i, ajouter 1 à $t.[i]$
  - Sinon, ne rien faire
- Fin Tant que
- Renvoyer t

La complexité spatiale est donc en $\mathcal{O}(1)$, et la complexité temporelle en $\mathcal{O}(N)$. De plus, cet algorithme semble plus simple que le premier, et moins susceptible de provoquer des erreurs une fois l'automate correctement défini. On pourrait appliquer des méthodes complémentaires pour trouver un automate plus simple, et ainsi optimiser l'utilisation de mémoire. L'algorithme de Knuth–Morris–Pratt permet par exemple de rechercher un mot donné dans un texte, pour un coût temporel $\mathcal{O}(m+N)$ avec m la taille du mot et N celle du texte, et un coût spatial de $\mathcal{O}(m)$.

C Arbres des forêts de France métropolitaine

Pour tenter de vérifier la loi de Benford dans la nature, je dispose des données mesurées sur 350 641 arbres répartis sur tout le territoire français [6] J'ai accès à la circonférence de leur tronc, leur hauteur, ainsi que

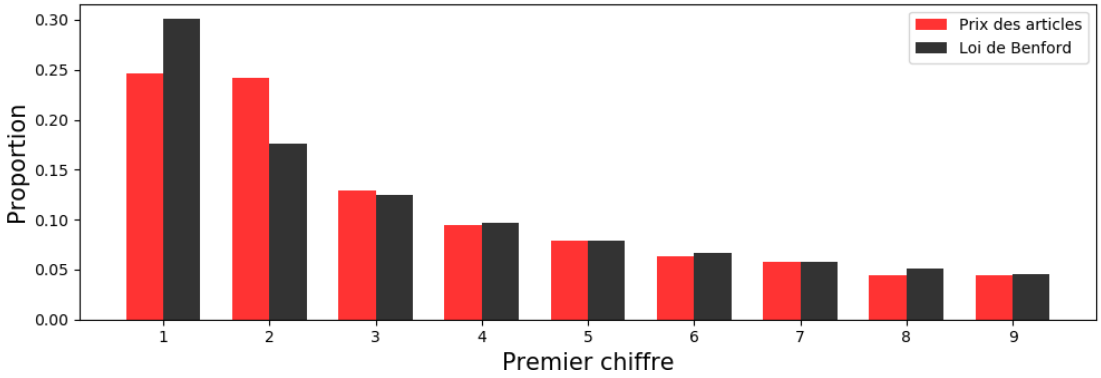


FIGURE 1 – Loi de Benford sur 12296 articles du rayon "bébé" de Carrefour 04/11/21

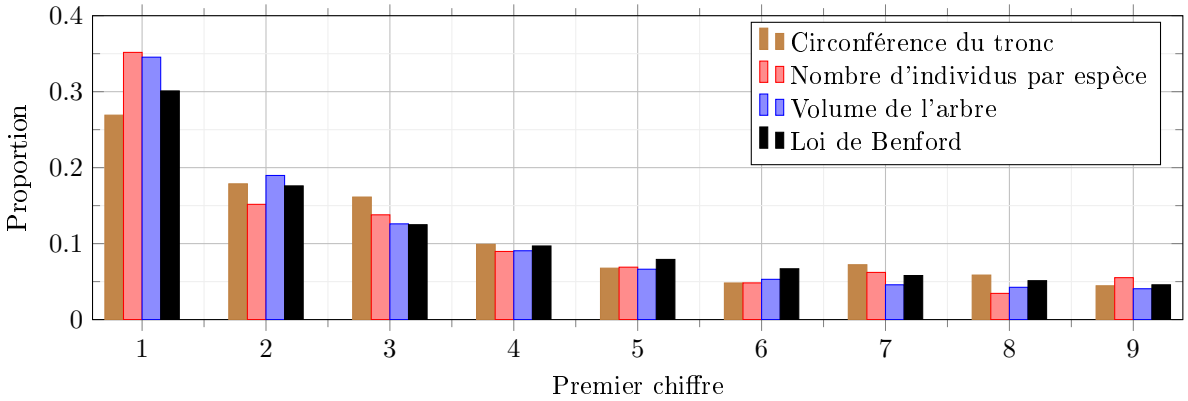


FIGURE 2 – La loi de Benford semble vérifiée sur les caractéristiques de 350641 arbres français

le volume estimé de chaque arbre. Les unités sont le mètre et le mètre cube. Je prends aussi en compte le nombre d'individus de chacune des cent quarante-neuf espèces. Pour chaque plage de données et chaque chiffre $c \in [\![1;9]\!]$, je compte le nombre d'arbres dont le premier chiffre de la donnée considérée est c.

La figure 2 reprend les paramètres pour lesquels la loi semble vérifiée, et la figure 3 représente le cas de la hauteur, qui ne la vérifie pas.

Une raison semble être l'étendue en ordre de grandeur des plages de données, car :

- les volumes vont de $1,52 \times 10^{-4}\,m^3$ à $23,44\,m^3$, soit **cinq** ordres de grandeur
- les populations vont de 1 individu, comme pour le paulownia ou le saule faux daphné, à 30 041 pour le chêne pédonculé, soit un total de **cinq** ordres de grandeur
- les circonférences vont de $0.23\,m$ à $8.69\,m$, soit **deux** ordres de grandeur
- les hauteurs vont de $1,3\,m$ à $49,4\,m$, ce qui représente **un** ordre de grandeur

3 Conclusion

La vérification de la loi de Benford dans la nature semble dépendre du nombre d'ordres de grandeur dans les données utilisées. Elle a été découverte en 1881 par l'astronome américain Simon Newcomb qui avait remarqué que certaines pages des tables de logarithmes, correspondant aux nombres commençant par le chiffre 1, étaient plus usées que d'autres, suggérant que certains nombres issus de mesures étaient plus fréquents que d'autres. Il est frappant de constater qu'elle se vérifie aujourd'hui dans des domaines variés, grâce aux quantités très importantes de données disponibles, dont Newcomb n'aurait pas rêvé. De plus, le phénomène s'explique bien par l'équirépartition, notion simple à définir. La loi de Benford perd de son mystère lorsqu'on la replace dans la grande famille de lois concernant les variables aléatoires continues. Si l'on impose certaines conditions à la fonction de densité, on peut créer autant de "lois de Benford" que l'on souhaite.

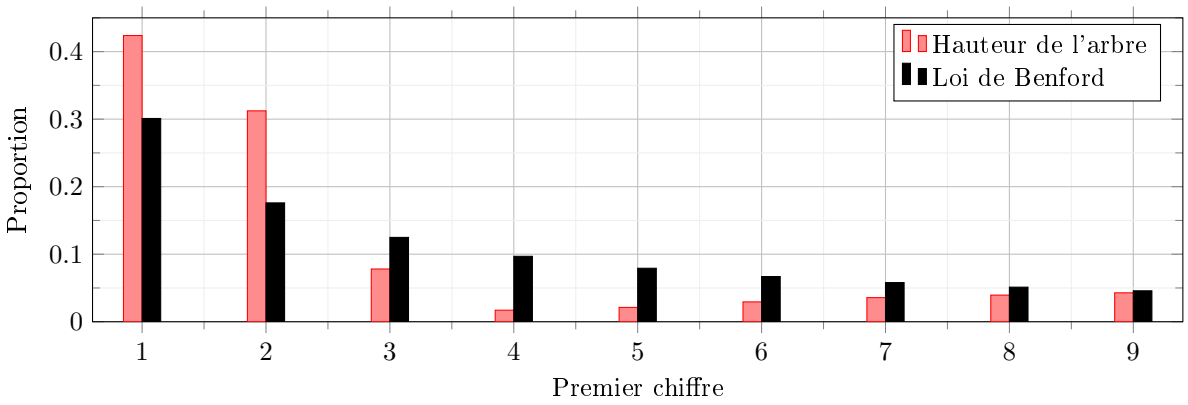


FIGURE 3 – La loi de Benford ne semble pas vérifiée sur la hauteur des arbres

Références

[1] *Uniform Distribution of Sequences (p.1 - 39)*, L. Kuipers, H. Niederreiter, Wiley-interscience, 1974.

[2] *A First Course in Dynamics with a Panorama of Recent Developments (p.96 - 112)*, Boris Hasselblatt, Anatole Katok, Cambridge University Press, 2003.

[3] *Thèmes d'analyse (p.139 - 165)*, Jean-Marie Exbrayat, Michel Alessandri, Masson, 1997.

[4] *La suite des puissances de 3/2*, F.D. et M.M.F., La Recherche, 2001.

[5] *Pourquoi la loi de Benford n'est pas mystérieuse*, Nicolas Gauvrit, Jean-Paul Delahaye, Mathematics and social sciences, n°182, 2008.

[6] *Institut national de l'information géographique et forestière*, `inventaire-forestier.ign.fr/dataifn/DonneesBrutes/requetage`, consulté le 10 janvier 2022.