Loi de Newcomb - Benford, loi uniforme et loi log-normale

Université de Grenoble - Préparation à l'agrégation

Christophe Leuridan

1 Loi de Newcomb - Benford

Quand on travaille avec des réels pouvant avoir des ordres de grandeur très variés, on utilise l'écriture scientifique : tout réel strictement positif s admet une unique décomposition sous la forme $s = t \times 10^n$ avec $t \in [1, 10[$ et $n \in \mathbf{Z}$. Le premier chiffre significatif de s est la partie entière de t, notée $\mathrm{Ent}(t)$ et appartient à $\{1; 2; 3; 4; 5; 6; 7; 8; 9\}$.

En 1881, Simon Newcomb publie un article présentant un étrange phénomène : le premier volume des tables logarithmiques est plus utilisé que le deuxième qui l'est plus que le troisième et ainsi de suite. Un savant calcul l'amène à conclure que la probabilité que le premier chiffre significatif d'un nombre, « pris dans un ensemble quelconque », soit d, est égale à $\log(1+1/d)$ où log est la fonction logarithme en base dix. Cet article passe totalement inaperçu. Par contre, 57 ans plus tard, un article de Franck Benford, motivé par la même observation et aboutissant à la même loi de probabilité, étayé d'exemples éclectiques, attire l'attention. La loi est baptisée loi de Benford. On a trouvé depuis de très nombreux exemples de données se conformant à cette loi de probabilité. [1]

Suggestion de développement : observer les premiers chiffres significatifs des populations des 26 communes de l'agglomération grenobloise (données en dernière page). Peut-on accepter l'hypothèse : ces valeurs sont des réalisations indépendantes d'une variable aléatoire suivant la loi de Newcomb-Benford ?

2 Lien avec la loi uniforme

Une première explication à la loi de Benford est la suivante. Si $S = T\ 10^N$ est une variable aléatoire strictement positive écrite en notation scientifique et si λ est un réel strictement positif, alors la variable aléatoire $S' = \lambda S$ a pour écriture scientifique $S' = T'\ 10^{N'}$ avec $N' = N + \mathrm{Ent}(\log\lambda + \log T)$ et $\log T' = \mathrm{Fra}(\log\lambda + \log T))$, en notant $\mathrm{Fra}(x) = x - \mathrm{Ent}(x)$ la partie fractionnaire d'un nombre x.

La multiplication de S par une constante $\lambda > 0$ peut par exemple correspondre à un changement d'unité (conversion de dollars en euros, de miles en kilomètres,...). Cherchons s'il existe une « loi universelle » pour la variable aléatoire T, c'est-à-dire une loi qui soit préservée par les transformations $S \mapsto S' = \lambda S$. On voudrait donc que pour tout $\lambda > 0$, $\mathrm{Fra}(\log\lambda + \log T))$ ait même loi que $\log T$. La proposition ci-dessous montre que pour cela, il faut et il suffit que $\log T$ suive la loi uniforme sur $[0, 1[$. Cela entraîne que $\mathrm{Ent}(T)$ suit la loi de Newcomb - Benford.

[1] Ce paragraphe est extrait de l'introduction de l'article « Le premier chiffre significatif fait sa loi » de *La Recherche*, janvier 1999, pp. 72–75

Proposition 1 *(Caractérisation de la loi uniforme sur $[0,1[$). Soit R une variable aléatoire à valeurs dans $[0,1[$.*

1. *La loi de R est complètement déterminée par les valeurs $\mathbf{E}[e^{i2\pi mR}]$ pour $m \in \mathbf{Z}$;*
2. *S'il existe un irrationnel α tel que $\mathrm{Fra}(\alpha + R)$ ait même loi que R, alors R suit la loi uniforme sur $[0,1[$.*
3. *Réciproquement, si R suit la loi uniforme sur $[0,1[$, alors $\mathrm{Fra}(\alpha + R)$ a même loi que R pour tout réel α.*

Démonstration. Dans le cas particulier où R possède une densité, le premier point découle de l'injectivité de la transformation qui à un élément de $L^1([0,1])$ associe ses coefficients de Fourier. Pour le cas général, on utilise le fait que toute fonction continue et 1-périodique de $\mathbf{R}$ dans $\mathbf{C}$ est limite uniforme de polynômes trigonométriques et que si $0 < a < b < 1$, l'indicatrice de $]a,b]$ peut s'écrire comme limite ponctuelle sur $[0,1[$ de fonctions continues, 1-périodiques, à valeurs dans $[0,1]$.

Les deux autres points se déduisent facilement du premier, le troisième pouvant aussi être montré directement à l'aide des fonctions de répartition.

Suggestions de développement :

1. Compléter la démonstration de la proposition 1.
2. Pourquoi est-il important que α soit irrationnel dans le point 2 ?
3. Est-il possible que λS ait même loi que S ?
4. Quelle est la loi de T si $\log T$ suive la loi uniforme sur $[0,1[$? Pourquoi $\mathrm{Ent}(T)$ suit-elle la loi de Newcomb - Benford ?
5. On note $s_k = t_k 10^{n_k}$ pour $1 \leq k \leq 26$ les populations des 26 communes de l'agglomération grenobloise. Peut-on accepter l'hypothèse : les valeurs $\log t_k$ sont des réalisations indépendantes d'une variable aléatoire de loi uniforme sur $[0,1[$?

3 Lois unimodales et partie fractionnaire

On appelle loi unimodale sur $\mathbf{R}$ toute loi à densité dont la densité admet un unique maximum en un point x_0, est croissante sur $]-\infty, x_0]$ et décroissante sur $[x_0, +\infty[$. La valeur du maximum de la densité donne une indication sur l'étalement de la loi : si le maximum de la densité est petit, la loi est nécessairement étalée.

Le lemme ci-dessous explique pourquoi la partie fractionnaire d'une variable aléatoire X de loi unimodale très étalée suit une loi proche de la loi uniforme sur $[0,1[$.

Théorème 2 *Soit X une variable aléatoire réelle admettant une densité f.*

1. *La variable aléatoire $R = \mathrm{Fra}(X)$ a une densité f_R sur $[0,1[$ donnée par*
$$f_R(r) = \sum_{n \in \mathbf{Z}} f(r+n).$$
2. *Si la loi de X est unimodale alors pour tout $r \in [0,1[$, $|f_R(r) - 1| \leq ||f||_\infty$.*
3. *Si f est de classe $\mathcal{C}^1$ et si f' est intégrable, alors*
$$\int_0^1 (f_R(r) - 1)^2 \, dr \leq \frac{||f'||_1^2}{12}.$$

Démonstration. Montrons les différents points.

1. On vérifie aisément que pour tout borélien $B \subset [0,1[$,

$$P[R \in B] = \int_B \Big(\sum_{n\in\mathbf{Z}} f(r+n))$$

Le cas particulier où $B \subset [0,1[$ montre que la série de fonctions $\sum f(\cdot + n)$ converge normalement dans $L^1([0,1[)$ et presque partout sur $[0,1[$.

2. Pour tout $r \in \mathbf{R}$, notons

$$S(r) = \sum_{n\in\mathbf{Z}} f(r+n)$$

Alors S est une fonction 1-périodique qui coïncide avec f_R sur $[0,1[$. En notant x_0 le point où la densité f atteint son maximum, il suffit donc de démontrer que pour $r \in [x_0, x_0{+}1]$,

$$1 - f(x_0) \leq S(r) \leq 1 + f(x_0).$$

Par croissance de f sur $]-\infty, r-1]$ et par décroissance de f sur $[r, +\infty[$, on a alors

$$\sum_{n\leq -2} f(r+n) \ \leq \ \int_{-\infty}^{r-1} f(x)dx \ \leq \ \sum_{n\leq -1} f(r+n),$$

$$\sum_{n\geq 1} f(r+n) \ \leq \ \int_{r}^{+\infty} f(x)dx \ \leq \ \sum_{n\geq 0} f(r+n),$$

d'où par addition

$$S(r) - f(r-1) - f(r) \ \leq \ 1 - \int_{r-1}^{r} f(x)dx \ \leq \ S(r).$$

On termine en remarquant que

$$(x_0 - (r-1))f(r-1) + (r - x_0)f(r) \ \leq \ \int_{r-1}^{r} f(x) \ \leq \ f(x_0)$$

et

$$(r - x_0)f(r-1) + (x_0 - (r-1))f(r) \ \leq \ f(x_0),$$

par croissance de f sur $[r-1, x_0]$ et décroissance de f sur $[x_0, r]$.

3. Les coefficients de Fourier de S sont donnés par

$$c_m(S) = \int_0^1 S(r)e^{-i2\pi mr}\, dr = \int_{\mathbf{R}} f(t)e^{-i2\pi mt}\, dt$$

pour $m \in \mathbf{Z}$. En particulier $c_0(S) = 1$.

Comme f et f' sont intégrables sur $\mathbf{R}$, f possède des limites finies en $+\infty$ et $-\infty$, nécessairement nulles. Pour $m \neq 0$, on a donc

$$c_m(S) = \frac{1}{i2\pi m} \int_{\mathbf{R}} f'(t)e^{-i2\pi mt}\, dt,$$

d'où $|c_m(S)| \leq ||f'||/(2\pi m)$. A l'aide de la formule de Parseval, on en déduit que

$$\int_0^1 (S(r)-1)^2 = \sum_{m\neq 0} |c_m(S)|^2 \leq \frac{||f'||_1^2}{12}.$$

Ce résultat explique pourquoi l'on rencontre souvent la loi de Newcomb - Benford pour des données réparties sur des ordres de grandeurs très différents : il suffit que leurs logarithmes en base 10 puissent être vus comme des réalisations d'une variable aléatoire réelle de loi unimodale et très étalée.

Suggestions de développement :

1. Compléter la démonstration de la proposition 2.
2. Que donne la majoration du point 3 lorsque f est unimodale ?
3. Montrer que si X admet une densité f bornée, alors sa fonction quantile $F_X^{\leftarrow}$ vérifie $F_X^{\leftarrow}(v) - F_X^{\leftarrow}(u) \geq (v-u)/||f||_\infty$ pour $0 < u < v < 1$. En déduire que

$$\mathrm{Var}(X) \geq \frac{1}{12||f||_\infty^2}$$

Cette inégalité peut-elle être une égalité ?
Indication : on pourra montrer que si U et V sont des variables aléatoires indépendantes de loi uniforme sur $]0,1[$,

$$\mathbf{E}\left[\left(F_X^{\leftarrow}(V) - F_X^{\leftarrow}(U)\right)^2\right] = 2\ \mathrm{Var}(X).$$

4. Regarder la qualité de l'encadrement de la densité de R dans le cas où X suit la loi exponentielle de paramètre 1 ou la loi gaussienne de variance 1. L'estimation est-elle fine ?

4 Loi log-normale

On se propose ici d'expliquer pourquoi les valeurs boursières obéissent à la loi de Newcomb-Benford.

Notons s le cours d'une action un jour donné (pris comme origine des temps) et S_n son cours n jours plus tard. Supposons que l'action évolue suivant des taux aléatoires indépendants et identiquement distribués. Autrement dit, $S_n = S_{n-1} \times \eta_n$ où $\eta_1, \eta_2, ...$ sont des variables aléatoires strictement positives i.i.d.. Si les variables aléatoires $\xi_n = \ln \eta_n$ sont de carré intégrable, on peut approcher la loi de $\log S_n$ par une loi normale, autrement dit la loi de S_n par une loi log-normale.

Quand $n \to +\infty$, la loi de $\log S_n$ est donc proche d'une loi normale de variance élevée. Sous réserve que $\log S_n$ possède une densité unimodale proche de la densité de la loi gaussienne de même espérance et de même variance, on peut appliquer le résultat de la partie précédente : la loi de $\mathrm{Ent}(\log S_n)$ est proche de la loi uniforme sur $[0,1[$ et la loi du premier chiffre significatif de S_n est proche de la loi de Newcomb-Benford.

Mais on ne peut pas vérifier une loi statistique sur une seule donnée ! Qu'en est-il si l'on suit les cours $S_n^{(1)}, \dots, S_n^{(d)}$ de d actions valant initialement $s_1, \dots, s_d$?

On suppose alors que

$$(\frac{S_n^{(1)}}{S_{n-1}^{(1)}}, \dots, \frac{S_n^{(d)}}{S_{n-1}^{(d)}}) = (\eta_n^{(1)}, \dots, \eta_n^{(d)})$$

où les variables aléatoires $(\eta_n^{(1)}, \dots, \eta_n^{(d)})_{n\geq 1}$ à valeurs dans $(\mathbf{R}_+^*)^d$ sont i.i.d., mais on ne suppose pas que les composantes $\eta_n^{(1)}, \dots, \eta_n^{(d)}$ soient i.i.d. pour n fixé.

Notons ϕ la fonction caractéristique de $(\xi_n^{(1)}, \ldots, \xi_n^{(d)}) = (\log \eta_n^{(1)}, \ldots, \log \eta_n^{(d)})$ et

$$(R_n^{(1)}, \ldots, R_n^{(d)}) = (\mathrm{Fra}(\log S_n^{(1)}), \ldots, \mathrm{Fra}(\log S_n^{(d)})).$$

Alors quels que soient les entiers relatifs $k_1, \ldots, k_d$,

$$\mathbf{E}[\exp(i2\pi(k_1 R_n^{(1)} + \cdots + k_d R_n^{(d)}))] \quad = \quad \mathbf{E}[\exp(i2\pi(k_1 \log S_n^{(1)} + \cdots + k_d \log S_n^{(d)}))]$$

Par indépendance et équidistribution des variables aléatoires $(\xi_k^{(1)}, \ldots, \xi_k^{(d)})$, on a donc

$$\mathbf{E}[\exp(i2\pi(k_1 R_n^{(1)} + \cdots + k_d R_n^{(d)}))] \quad = \quad e^{i2\pi(k_1 \log s_1 + \cdots + k_d \log s_d)} \phi(2\pi k_1, \ldots, 2\pi k_d)^n$$

Mais si $(k_1, \ldots, k_d) \neq (0, \ldots, 0)$, la variable aléatoire $k_1\eta_1 + \cdots + k_d\eta_d$ possède une densité si bien que

$$|\phi(2\pi k_1, \ldots, 2\pi k_d)| = \mathbf{E}[\exp(i2\pi(k_1\eta_1 + \cdots + k_d\eta_d))] < 1,$$

d'où

$$\mathbf{E}[\exp(i2\pi(k_1 R_n^{(1)} + \cdots + k_d R_n^{(d)}))] \to 0.$$

Cela montre que la loi de $(R_n^{(1)}, \ldots, R_n^{(d)})$ tend vers la loi uniforme sur $[0, 1[$.

Suggestions de développement :

1. Justifier et préciser l'approximation de la loi de $\log S_n$ par une loi log-normale.
2. Justifier les calculs montrant la convergence en loi de $(R_n^{(1)}, \ldots, R_n^{(d)})$. On pourra se limiter au cas où $d = 1$ ou au cas où les $\eta_n^{(1)}, \ldots, \eta_n^{(d)}$ sont i.i.d..
3. Discuter les hypothèses d'indépendance et d'équidistribution. Pourquoi est-il peu réaliste de supposer l'indépendance entre les différentes actions ?

5 Données numériques

Populations des 26 communes de la comunauté d'agglomération de Grenoble (valeurs au recensement de 1999).

Claix	7 610
Corenc	3 949
Domène	6 444
Echirolles	33 169
Eybens	9 546
Fontaine	23 586
Gières	6 165
Grenoble	156 203
La Tronche	6 672
Le Fontanil	2 474
Le Gua	2 848
Meylan	19 044
Murianette	619
Noyarey	2 104
Poisat	2 116
Pont de Claix	11 612
Sassenage	9 964
Seyssins	6 937
Seyssinet	13 207
St Égrève	15 691
St Martin d'H	35 927
St Martin le V	5 233
St Paul de V	1 878
Varces, A et R	6 383
Veurey Voroize	1 346
Vif	8 198